Explainable Spectral Super-Resolution Based on a Single RGB Image

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

In

Electronics and Computer Engineering

at Massey University, Manawatu, New Zealand

YUAN CHANG

2023

Abstract

Hyperspectral imaging offers fine spectral measurements of target surfaces, finding utility in various fields. However, traditional hyperspectral systems grapple with high-cost issues. On the other hand, conventional RGB cameras, which provide relatively coarse measurements of surface spectra, are widely accessible. Consequently, the recovery of spectral information from RGB images has emerged as a popular approach for low-cost hyperspectral imaging, a venture also known as single-image spectral super-resolution. Yet, existing methods, mostly rooted in deep convolutional neural networks, tend to suffer from limited interpretability.

In our research, we propose an explainable method for single-image spectral super-resolution. This method relies on the RGBPQR colour space, a low-dimensional spectral data model representing the spectrum. Leveraging the RGBPQR spectral model, we can transform the spectral reconstruction task into a regression problem. To tackle the metamerism issue, we analysed existing spectral super-resolution networks and discovered that these networks often depend on local textural information as context to mitigate metamerism. Informed by this insight, we utilized features extracted from multiscale local binary patterns as contextual information to design our explainable method.

Furthermore, in this study, we discussed the error measurements and loss functions employed in this research area and proposed a new error measurement that can represent performance more accurately. We also endeavoured to put forward a method for quantitatively measuring the ability to resolve metamerism, a critical problem in spectral super-resolution. Through our research, we offered a simple, low-dimensional, and explainable spectral super-resolution solution.

Key words: Hyperspectral imaging, spectral super-resolution.

Acknowledgements

First and foremost, I would like to extend my deepest gratitude to my supervisor, Donald Bailey. Your steadfast support and invaluable guidance have been instrumental throughout this research. Your profound expertise and insightful feedback have proved essential in navigating the complexities of my work. I am eternally thankful for your patience, encouragement, and unwavering belief in my abilities. Without a doubt, you stand as the best teacher I have ever encountered in my entire life. I would also like to extend my sincere gratitude to my co-supervisor, Steven Le Moan. Your expertise and invaluable input have been instrumental in enhancing the depth and quality of this research. Your ideas and constructive criticisms have played a crucial role in shaping my work, and I am truly thankful for your contributions.

To my parents, thank you for your endless love and belief in me. Your sacrifices and unwavering support have been my bedrock, providing me with the strength to persevere even in the most challenging times. This achievement is as much yours as it is mine, and I am eternally grateful for your ceaseless encouragement.

I extend my gratitude to my friends who have provided valuable assistance throughout my research. Additionally, I would like to express my thanks to ChatGPT for their grammar check of this work.

Last but by no means least, my deepest thanks go to my partner, Xuan Song. Your love, patience, and unwavering support have been constant. Thank you for cheering me on during my triumphs, lifting me up during the lows, and being my steadfast companion on this journey. Your belief in me, even when I doubted myself, has been the driving force behind my perseverance and determination to complete this research.

This journey would not have been possible without each one of you, and I am immeasurably thankful for your presence in my life and academic journey.

Table of Contents

List of Figures

List of Tables

Symbols and Abbreviations

CNN	Convolutional Neural Network
DNN	Deep Neural Network
GAN	Generative Adversarial Network
K-SVD	K-singular value decomposition
LBP	Local Binary Pattern
HSI	Hyperspectral Imaging
PCA	Principal Component Analysis
RGB	Red, Green, Blue
MAE	Mean Relative Error
MRAE	Mean Relative Absolute Error
NTIRE	New Trends in Image Restoration and Enhancement
ReLu	Rectified Linear unit
RMSE	Root Mean Square Error
RSNR	Peak Signal-to-Noise Ratio
SAM	Spectral Angle Mapper
SSR	Spectral super-resolution
A	Additional information
2D	2 Dimensional
3D	3 Dimensional
C	Camera function
E	The residual between the original spectra and the up-sampled spectra
H	Entropy
I	Identity matrix
L	Illuminance
N_c	Tristimulus vector
N_p	PQR weights (coefficients)
PQR	Residual E represented by the first three PCA weights
R	Spectral radiance

S	Spectral reflectance
T	Reconstruction (up-sampling) function
V	PQR bases
ΔE	CIE Colour Difference
λ	Wavelength
μ	Mean
P	RGB value

Chapter 1. Low-cost Hyperspectral Imaging

1.1. Hyperspectral Imaging

The reflectance of a material surface is defined as its effectiveness in reflecting radiant energy. It is the fraction of the reflected electromagnetic power over the incident power and is a function of the wavelength. The spectral reflectance represents many of the intrinsic features of the object, which are independent of both the illumination and the sensor's responses. Human eyes are sensitive to spectral reflectance in three bands, with three types of cones (long, medium, and short wavelengths, roughly corresponding to the perception of red, green, and blue hues, respectively). Conventional RGB cameras, which are designed to capture colour scenes for human visualization, also separate the visible range into three spectral bands: red, green, and blue.

However, all colour (spectral) imaging encounters a key challenge: capturing data with one spectral dimension and two spatial dimensions using a 2D sensor. This problem occurs because each pixel in commercial camera sensors can only provide intensity measurements in one channel. Camera sensors have a finite number of pixels, which need to be allocated for capturing either spectral or spatial information. Increasing spectral resolution often requires dividing the sensor into more spectral bands, which in turn reduces the number of pixels available for each band. This can result in lower spatial resolution, as fewer pixels are used to capture the spatial details of the scene. Therefore, there is always a trade-off between spectral resolution and spatial resolution. For example, RGB cameras commonly use a Bayer filter, which groups the pixels into 2×2 blocks, with one red, two green, and one blue colour filter in each block. To generate a full-colour image, a process called demosaicing is used. During this process, the missing colour channels for each pixel are estimated using the information from neighbouring pixels. Various interpolation algorithms can be used to perform this task, with different levels of complexity and quality.

However, three broadband filters cannot provide a precise measurement of spectral reflectance in many applications (Sowmya *et al.*, 2019). Precise measurement of spectral reflectance is necessary to capture the physical and chemical properties of various materials, such as human skin, plant

canopies, and art paintings, in a non-destructive manner. To accomplish this, a hyperspectral or multispectral device is typically required.

In contrast to the human eye and conventional RGB cameras, spectral imaging separates incoming light into a considerably larger number of bands. The difference between multispectral and hyperspectral imaging is the number of channels (bands) captured (Oh *et al.*, 2016). Multispectral imaging generally captures a small number of channels (less than 10) usually of quite wide bandwidth in the visible and near-infrared range, while hyperspectral imaging usually records tens to hundreds of spectral bands in the visible and near-infrared ranges (Adão *et al*., 2017). Figure 1 compares the spectral resolution of RGB, multispectral and hyperspectral cameras.

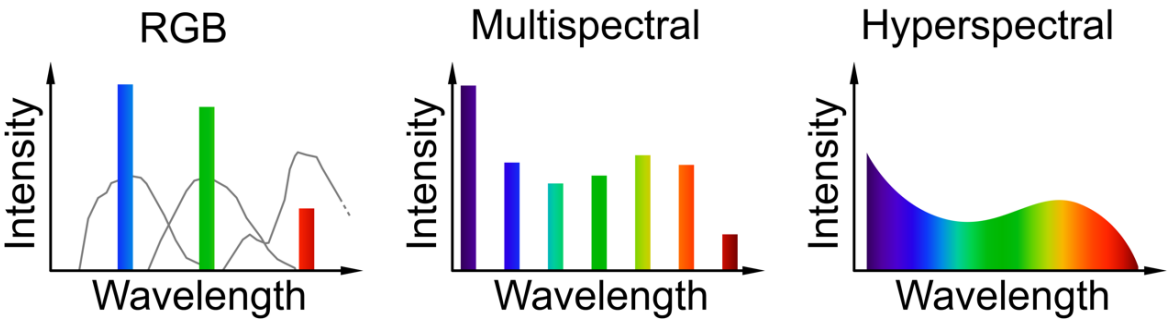


Figure 1. Sketch of spectral resolution for RGB, multispectral and hyperspectral camera.

(https://commons.wikimedia.org/wiki/File:Spectral_sampling_RGB_multispectral_hyperspectral_imaging.svg)

The goal of multispectral and hyperspectral imaging (HSI) is to record the radiance or reflectance spectrum at each pixel in a scene of interest (Valero *et al.*, 2007). The measured pixel intensities of a camera at one spatial position i_k can be modelled as the product of the spectral reflectance of the point $s(\lambda)$, the spectral power distribution of the illuminant $l(\lambda)$, and the camera spectral sensitivities $c_k(\lambda)$, where k is the channel number, integrating over the spectral range.

$$i_k = \int c_k(\lambda) l(\lambda) s(\lambda) d\lambda \qquad 1\text{-}1$$

For an RGB camera, k is in $\{1, 2, 3\}$ for the red, green and blue channel. For a hyperspectral camera, with $10\ \mathrm{nm}$ spectral resolution as an example, the visible range ($400\ \mathrm{nm}$ to $700\ \mathrm{nm}$) would have 31 bands. With the additional bands, hyperspectral cameras can provide a more precise measurement of the spectrum. Generally, the collected spectral data is represented as a 3D data cube, expressed in $x \times y \times \lambda$, where $x \times y$ represents the 2D spatial coordinates and λ represents 1D spectral dimension. However, capturing such high spectral resolution with a 2D sensor makes the hyperspectral system complex, and there is always a trade-off between spectral resolution, spatial resolution and frame rate. Section 1.2 provides an overview of the current approaches utilized for hyperspectral imaging.

1.2. Direct Approach for Hyperspectral Imaging

The classic (direct) hyperspectral approach, as opposed to computational techniques, uses optical elements to measure spectral radiance in each band. The major approaches can be classified into four categories, as demonstrated in Figure 2, which shows how these methods capture spatial and spectral information. All measured results from the direct methods should be further reshaped into the 3D hyperspectral data cube. The four categories are:

- **Point Scanning**

  This technique captures the spectral reflectance of a single pixel at a time and reconstructs the entire spatial scene by scanning the point of interest in 2 dimensions through the scene. The advantage of point scanning is that it can achieve high spectral resolution. However, the disadvantage of this method is that achieving high spatial resolution requires a huge amount of time and effort during the operation (Kamruzzaman & Sun, 2016) (Lodhi *et al.*, 2019) (Lu & Fei, 2014) (Willett *et al.*, 2013).

- **Line Scanning**

  Compared to the point scanning approach, the line scanning approach measures a line of pixels (1 spatial dimension) at a time. By spreading the spectra, the spectral dimension is mapped to the second dimension of a 2D sensor. To obtain a 2D spatial measurement, the line scanning device still requires line-by-line scanning through the target scene (Lodhi *et al*., 2019) which still takes time to operate.

- **Spectral Scanning**

  Spectral scanning measures one spectral channel of an image scene at a time. The spectral resolution is obtained by shifting a series of spectral filters in front of the image sensor. Compared to the previous two approaches, spectral scanning typically offers higher spatial resolution but a lower spectral resolution (Nieves, 2020).

- **Snapshot**

  Also known as single-shot multi-point spectrometry, the snapshot sensor can capture the full spectrum of an image scene without multiple operations. It works by applying multiple prisms for image division and wavelength discrimination before the light reaches the sensor. Like a single-chip colour camera, each pixel only captures a single wavelength. Snapshot sensors aim for higher frame rates, but the spatial and spectral resolution are commonly lower than the other techniques (Hagen & Kudenov, 2013) (Adão *et al*., 2013).

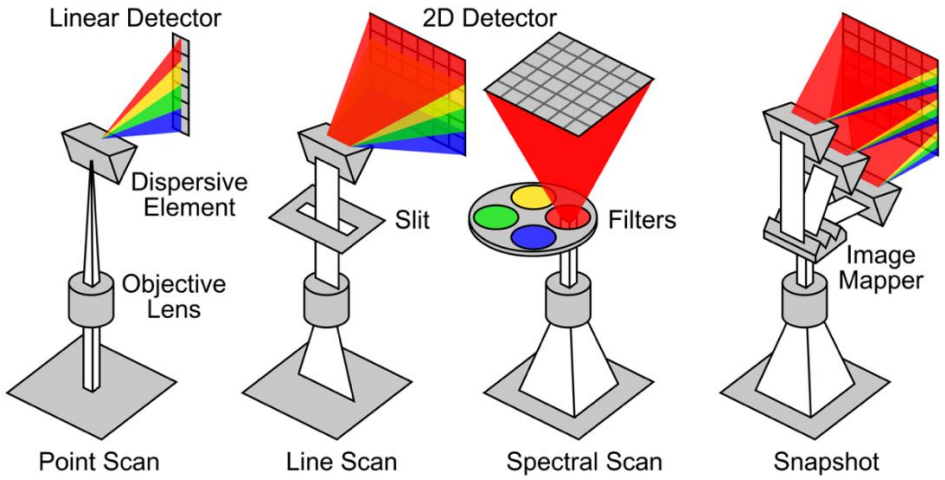


Figure 2. How different direct hyperspectral methods capture spatial and spectral information.

(https://commons.wikimedia.org/wiki/File:Multispectral_imaging_approaches.svg)

However, compared to conventional RGB imaging, all direct hyperspectral approaches are relatively complex and hard to build (Cao *et al.*, 2011) (Gao *et al.*, 2010). Due to their complexity, HSI devices typically cost tens of thousands of dollars (CBRNE Tech Index Webpage). In addition to the financial cost, the huge amount of data provided by HSI also requires high computation and memory costs.

Meanwhile, there is an increasing requirement for hyperspectral data in many applications. Hyperspectral imaging is a useful tool that enables a deep analysis of the physical properties of a scene in a non-destructive manner. HSI has been used extensively in remote sensing for agricultural (Muñoz-Huerta *et al.*, 2013) (Adão *et al.*, 2017) and environmental applications (Gardner & Sharp, 2010). As a non-destructive technique, HSI has been applied in medical science for cancer detection (Lu & Fei, 2014) (Halicek *et al.*, 2019), as well as in fields such as food science (D. G. Kim *et al.*, 2009) (Kamruzzaman & Sun, 2016), and history and cultural heritage (Akhtar *et al.*, 2014).

Given the increasing need for hyperspectral data and the high cost of existing HSI devices, the goal of achieving low-cost and easy-operational hyperspectral imaging has become a popular research topic in recent years. RGB sensors, which have red, green, and blue channels in the visible range, are common devices found in consumer cameras, webcams, and smartphones. Compared to HSI sensors, RGB sensors are cheaper and easier to operate. Therefore, obtaining low-cost hyperspectral images via RGB sensors has become a popular research topic. However, recovering the missing spectral data when using an RGB sensor requires additional sources of information. In the next section, we will introduce the approaches used to obtain hyperspectral images using an RGB sensor and explain where the additional spectral information comes from.

1.3. Using an RGB Sensor to Obtain Spectral Images

Compared to the direct approach, using an RGB sensor reduces the number of spectral measurements, offering a lower-cost solution. Given that a single RGB sensor can only provide wide and overlapping spectral measurements in the visible range, additional measurements are needed to recover the missing spectral details. Depending on the source of this extra information, RGB sensor based low-cost HSI approaches can be categorized into the following groups.

The **first** approach works by applying a sequence of spectral filters in front of the RGB sensor. Valero *et al.* (2007) presented a low-cost hyperspectral solution with a conventional RGB digital camera with up to 3 colour filters to recover the spectral data of natural scenes. However, like the spectral scanning HSI method, this requires multiple shots of the same scene, so it is limited mostly to controlled environments and static targets. Imai *et al.* (2002) discussed filter selection when measuring spectral reflectance. They compared narrowband and wideband filtering for multispectral imaging and showed that although the narrow band is theoretically more robust, in some cases good results are provided by a wideband filter set.

The **second** approach uses multiple sets of illumination to provide extra information for the RGB sensor to recover the spectral features. Chi *et al*. (2010) presented a novel active imaging approach that uses optimized wideband filtered illumination to obtain multispectral reflectance information. They used filters to model an additional light source with a known spectrum and improved the ability to recover material reflectance information. This is different from filter-based spectral imaging because the authors put the filter in front of the additional light source rather than the sensor. However, this method has its limitations. This restricts its use in an indoor environment. The ambient illumination cannot be too intense compared to the additional light source; otherwise, changes in the radiometric response to adding a light source would be too small to detect.

The **third** approach uses multiple RGB sensors to provide extra information to recover the reflectance/radiance spectrum of the target. The idea is to use the different spectral responses from different RGB sensors to increase the effective number of spectral bands captured. Oh *et al.* (2016) provided a framework for reconstructing hyperspectral images by using multiple consumer-level digital cameras. In particular, due to the differences in spectral sensitivities of the cameras, different cameras yield different RGB measurements for the same spectral signal. They introduced an algorithm that can combine and convert these different RGB measurements into a single hyperspectral image for both indoor and outdoor scenes. In this method, three cameras with known spectral responses were used. However, this approach also has its limitations, as the spectral response difference between any two consumer RGB cameras is typically quite small, which limits the amount of information that can be provided by adding each extra camera. There is also the difficulty of registering the images captured by cameras from slightly different locations.

In today's world, there is a growing availability of hyperspectral datasets, leading researchers to develop computational spectral reconstruction models that use existing spectral data to recover missing spectral information from RGB input. As machine learning and deep learning algorithms continue to advance, more reconstruction models have been proposed. This cost-effective and data-driven approach to hyperspectral solutions is commonly referred to as spectral super-resolution. The remainder of the thesis will focus on spectral super-resolution.

1.4. Spectral Super-Resolution (SSR)

Spectral super-resolution refers to a technique that aims to enhance the spectral resolution of a given image or data beyond the capabilities of the original sensor or system. In this research, we limit the discussion of single-image spectral super-resolution to methods that use algorithms or computational methods to estimate and reconstruct high-resolution spectral information in the visible range from an RGB image. This can be achieved through various approaches, such as statistical methods, machine learning, or deep learning. However, this problem is ill-posed because many possible high-resolution spectra could give rise to the same RGB (metamerism), and there is not enough information in the RGB to uniquely determine the high-resolution spectrum. In mathematics, inverting equation 1-1 with 3 equations from the RGB channels is under-determined. Therefore, before delving into the existing spectral super-resolution methods, it is necessary to explain why it is possible to recover spectral information from an RGB input and what factors limit the accuracy of the reconstruction.

1.4.1. Why Single Image Spectral Super-Resolution is Possible?

To better understand why single image spectral super-resolution is possible, it is necessary to first study the intrinsic dimensionality of natural spectral data. Figure 3 shows the spectral reflectance of a point on a flower (indicated by the red point on the left) from the NTIRE 2022 dataset (Arad *et al*., 2022). It is clear that in the visible range, the spectrum of the flower appears relatively smooth. Not only the flower but also the spectral reflectance of most natural materials are usually smooth, with a high correlation between neighbouring wavelengths (Burns, 2020). The smoothness and correlation between neighbouring bands often result in measured hyperspectral data that contain large redundancy, which enables it to be represented in a low-dimensional manner.

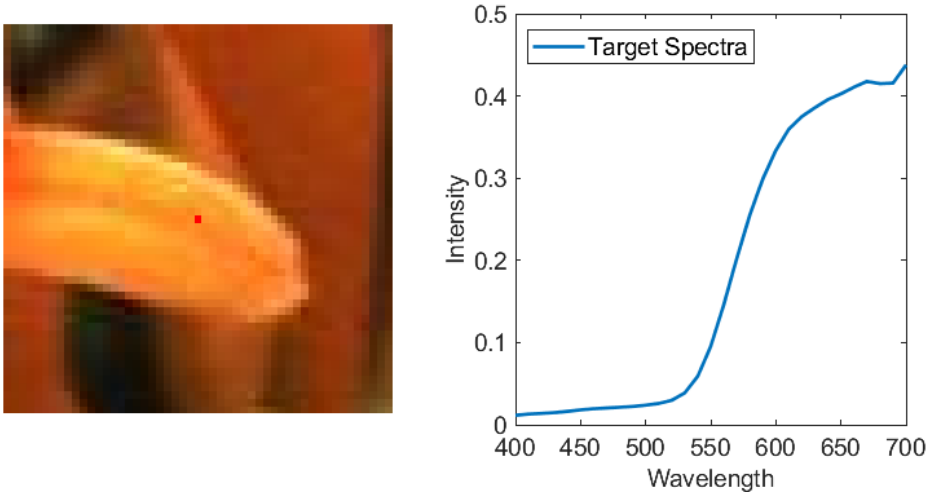


Figure 3. Spectra of flower from the NTIRE 2022 HSI dataset.

Two commonly used methods for reducing the dimensionality of hyperspectral data are feature selection and feature extraction. Feature selection aims to identify a subset of wavebands that contain the most information while minimizing redundancy. However, selecting a limited number of wavebands will unavoidably result in some information loss, no matter how carefully the wavebands are chosen.

Compared to feature selection, feature extraction is a more popular method. Principal Component Analysis (PCA) is the most famous and commonly used feature extraction method. It allows the hyperspectral data to be represented in a lower-dimensional space while preserving the most important information (Sowmya *et al.*, 2019).

PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance of the data lies on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. Consider n samples, each of p measured feature values, arranged as an $n \times p$ data matrix, X, in which the sample mean of each column has been shifted to zero. Mathematically, the transformation is defined by a set of l of p-dimensional unit basis vectors B, where $l \leq p$. These map each sample to a new vector of principal component weights (or scores) W given by:

$$W = XB \qquad 1\text{-}2$$

in such a way that the individual components of W (considered over the data set) successively inherit the maximum possible variance from X. Since $X^T X$ is a positive semidefinite matrix the transform vectors B are given from the eigenvectors of $X^T X$ corresponding to the l largest eigenvalues.

Utilizing the principal components results in the minimum squared error for any linear transformation of a given dimensionality, which explains the widespread use of PCA in various applications. PCA effectively retains the most essential information while minimizing the reconstruction error, making PCA a popular choice for tasks such as data compression, noise reduction, and feature extraction.

Cohen (1964) used PCA to determine the sufficient dimensionality required to represent spectral reflectance data. The data set consisted of 433 chips from the Munsell Book of Colour, which were measured using a spectrophotometer with a 10 nm resolution from 380 nm to 770 nm. After PCA, the first three components were found to be sufficient to model the spectral reflectance. Component one extracted 92.25% of the cumulative variance, while the first two extracted 97.72% and the first three extracted 99.68%.

However, as 433 different spectral reflectance are not sufficient to represent all types of reflectance, Maloney (1986) extended Cohen's work by including the Munsell colour samples and an additional 337 spectral reflectances. After analysing this larger set, Maloney found that using 5 to 7 components provided an accurate representation of the spectral data. However, the Munsell colour samples only contain limited types of spectra. In recent years, after analysing published natural hyperspectral databases, a dimensionality of 8-10 is commonly considered sufficient to represent spectral data (Khodr & Younes, 2011).

We applied PCA to the newest and biggest hyperspectral dataset, NTIRE 2022 (Arad *et al*., 2022), until the time of this report. We randomly selected 1000 samples from each of the 900 hyperspectral images, resulting in a total of 900,000 samples. After applying PCA, we found that the first component alone could explain 88% of the total variance, while the first three components could explain over 95% of the total variance. Figure 4 shows the error in the represented spectra measured by root mean square error (RMSE) as a function of the number of components from PCA. When more than 6 components are used, the RMSE is less than 0.004.

PCA is a widely used linear dimensionality reduction method for hyperspectral data analysis due to its speed and efficiency (Farrel & Mersereau, 2005) . However, there are cases where PCA is not suitable for feature extraction from hyperspectral images (Cheriyadat & Bruce, 2003). For example, Galliani *et al.* (2017) showed that PCA-based methods perform poorly in hyperspectral image-based remote sensing river detection since the PCA is not suitable to cover the underlying non-linear space of the spectral data. In addition to, PCA, there are several other linear and nonlinear dimensionality reduction methods, such as Locally Linear Embedding (Roweis & Saul, 2000), Kernel PCA (Vapnik, 1999), and Stochastic Neighbour Embedding (Lafon & Lee, 2006). A detailed review of these methods can be found in review papers by Khodr (Khodr & Younes, 2011) and Silva & Melo-Pinto (2021). In recent years, methods such as t-SNE and UMAP have also been used to visualize the data structure of high-dimensional data (Devassy & George, 2020) (Vermeulen *et al.*, 2021).

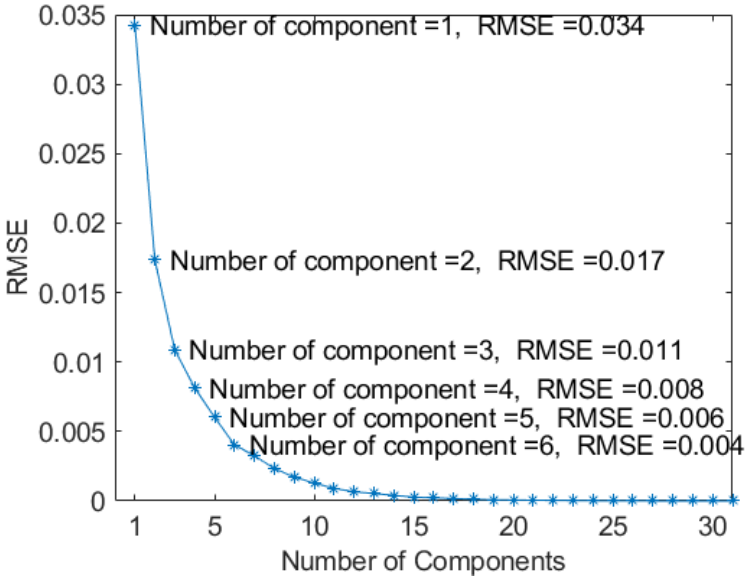


Figure 4. Error in the represented spectra decreases with the number of components from PCA.

In many applications, the large data size of hyperspectral imaging can pose computational and memory challenges during processing. High-dimensional hyperspectral data can create problems for data analysis, parameter estimation, and classification. This is known as Hughes's phenomenon (Hughes, 1968) (Pal & Foody, 2010), whereas the dimensionality of the data increases, the volume of the space increases exponentially, leading to data becoming sparse. This sparsity makes it difficult for algorithms to identify meaningful patterns, and it often results in a decrease in the performance of the learning algorithm. One approach to address this challenge is to apply data dimensionality reduction techniques to the hyperspectral data before processing.

Chakrabarti and Zickler (2011) conducted a statistical analysis of spectral data in a patch manner. The authors found that 20 PCA components could account for 99% of the total variance in 8×8 spectral patches collected from over 50 hyperspectral images. By using Gaussian mixture models, they found that 8 dimensions could accurately describe the high-frequency spatial and spectral information. Therefore, if a group of spectral reflectance values from image patches can be well represented by an 8-dimensional model, then it is reasonable to assume that the spectral reflectance of each pixel can also be represented by an 8 or less-dimensional model.

The spectral reflectance of an object often exhibits a smooth and redundant nature, allowing it to be well represented in a lower-dimensional manner. This characteristic makes the task of reconstructing spectral reflectance from RGB values less daunting and more feasible. Despite the lower-dimensional representation capturing the majority of the spectral information, a big challenge persists in estimating the missing spectral detail from RGB images. This challenge is known as metamerism, a phenomenon where different spectra produce identical RGB values under a specific

illuminant. This makes it challenging to distinguish and accurately recover the true spectral reflectance without incorporating additional or contextual information. The following section introduces the metamerism problem.

1.4.2. What limits the Spectral Super-Resolution?

Metamerism is defined as different spectral reflectance producing equal camera sensor responses under the same illuminant. This phenomenon occurs because the 3 channels of the consumer camera are insufficient to represent all the degrees of freedom needed to specify different spectra. In Figure 5, an example of metamers is shown, where the blue line in the plot corresponds to the spectra of a green man-made painting, while the red line corresponds to the spectra of green plant leaves. Although the spectra of these two surfaces have different distributions in the visible range, from a human perceptual (CIE 1964) standpoint, they are perceived as the same colour.

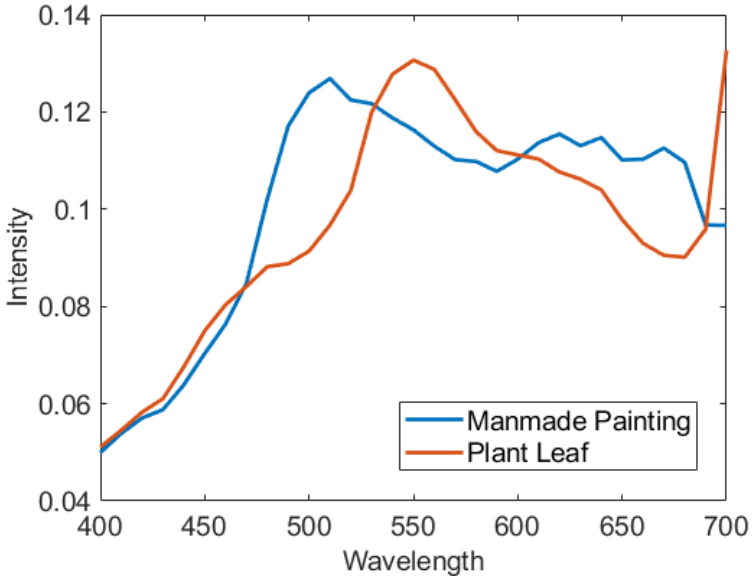


Figure 5. Example of metamers.

Foster and Amano (2006) analysed the frequency of metameric surfaces. After testing 50 hyperspectral images, the researchers announced that the frequency when the surface becomes undistinguishable under one light source is about 1% to 10%, which is sufficiently large to affect visual inferences about material identity. However, when under a second light source, the frequency of undistinguishable pairs reduces to between 0.001% and 0.1% (Kulappurath & Shamey, 2021). For spectral super-resolution, since additional illumination is usually not available, metamerism is not a problem that can be ignored.

Due to metamerism, there is no one-to-one mapping from RGB values to spectra. Instead, for a given RGB value, there may be multiple possible spectra corresponding to it. As a result, spectral super-resolution becomes a complex non-linear problem, whereas, without metamerism, it can be simply solved for example through a lookup table-based method. Previous research has shown that accurate recovery of surface spectral reflectance is usually impossible exclusively with conventional RGB cameras without additional measurements or relevant prior knowledge (Park *et al.*, 2007). Data-driven approaches to spectral super-resolution rely on this prior knowledge derived through training.

1.5. Summary

High spectral resolution can provide more detailed information about the composition and characteristics of the observed materials in a scene. This enhanced spectral information can lead to more accurate and precise classification, identification, and quantification of materials. However, acquiring high-resolution hyperspectral imagery can be expensive, as it often requires specialized sensors and platforms. Therefore, spectral super-resolution, which can help reduce costs by generating high-resolution spectral data from more readily available RGB images, has become a popular research topic.

In this thesis, we will analyse existing single-image spectral super-resolution methods, identify their limitations, and propose a novel approach that addresses these limitations. Through a comprehensive examination of current techniques, we will uncover areas of potential improvement and leverage these insights to develop an innovative method for spectral super-resolution. The rest of this thesis is organized as follows:

In Chapter 2, we provide a comprehensive review of existing methods that reconstruct spectral reflectance from RGB values, highlighting their key principles, advantages, and limitations. We also discuss the underlying concepts, such as evaluation metrics, optimization techniques, and machine learning approaches employed in these methods. At the end of Chapter 2 after reviewing existing methods detailed research goals are identified.

In Chapter 3, we initially integrate the RGBPQR spectral model into the spectral super-resolution task, offering a detailed explanation as to why this model is particularly suitable for resolving spectral reconstruction tasks.

In Chapter 4, we explore the potential of utilizing local spatial information as context for addressing the metamerism problem, by examining how incorporating spatial context can contribute to resolving spectral ambiguities that arise due to metamerism.

In Chapter 5, we analyse how spatial information is utilized by deep learning methods from two perspectives and how the spatial information can be used in a more explainable manner.

In Chapter 6, we present a novel single image spectral super-resolution model based on previous Chapters. The results of our proposed method are compared against the baseline methods. We discuss the strengths and weaknesses of our approach, along with the implications of our findings.

Finally, we conclude the thesis by summarizing our findings, discussing the contributions of our research, and outlining potential avenues for future work in the field of single-image spectral super-resolution.

Chapter 2. Recovering Spectral Information from RGB Images

With the increasing computational capabilities and more powerful machine learning and deep learning, models have been developed and shown great potential in solving regression and classification tasks. Meanwhile, there has been a dramatic increase in the number of accessible hyperspectral databases. Consequently, recovering spectral information using trained machine learning or deep learning models has become feasible. Several deep-learning research groups have concentrated on this topic and achieved remarkable reconstruction accuracy. This chapter aims to review the current state-of-the-art computational methods used to reconstruct hyperspectral images from RGB images. Before reviewing the state of the art, it is necessary to introduce the performance evaluation metrics and the existing datasets.

2.1. What is a Good Reconstruction?

It is important to be able to quantify the quality of the spectral reconstruction. This section reviews the commonly used error measurements when comparing the reconstructed and original spectral reflectance. These error measurements are also used as loss functions when training convolutional neural networks to recover spectral information.

2.1.1. Evaluation Metrics in Spectral Super-Resolution

- **Root Mean Square Error (RMSE)**

Mean square error (MSE) is a statistical measure used to quantify the differences between a set of predicted values and their corresponding actual values. It represents the power of the error signal between the reconstructed spectral reflectance and the original spectral (ground truth), and is defined as:

$$MSE = \frac{\sum_{i,c}\left(P_{gt_{i_c}} - P_{rec_{i_c}}\right)^2}{\left|P_{gt}\right|} \quad 2\text{-}1$$

where the $P_{gt_{i_c}}$ and $P_{rec_{i_c}}$ refer to the value of the i-th pixel on the c spectral channel in the ground truth and the reconstructed image respectively, and the $\left|P_{gt}\right|$ is the volume of the data cube (the number of pixels times the number of spectral channels).

The root mean square error (RMSE) represents the error in terms of amplitude rather than power:

$$RMSE = \sqrt{\frac{\sum_{i,c}\left(P_{gt_{i_c}} - P_{rec_{i_c}}\right)^2}{\left|P_{gt}\right|}} \quad 2\text{-}2$$

A related error measurement based on the RMSE is Peak Signal to Noise Ratio (PSNR):

$$PSNR = 10\log_{10}\left(\frac{peak^2}{MSE}\right) \quad 2\text{-}3$$

Peak Signal-to-Noise Ratio is defined as the ratio between the maximum possible signal value of the reconstructed spectrum (e.g., 255 for 8 bits) and the power of noise or error. Since many ratios have a wide dynamic range, PSNR is usually expressed in terms of the logarithmic decibel scale.

- **Mean Relative Absolute Error (MRAE)**

The Absolute Error is defined as the absolute difference between the original spectrum and the reconstructed spectrum.

$$MAE = \frac{\sum_{i,c}\left|P_{gt_{i_c}} - P_{rec_{i_c}}\right|}{\left|P_{gt}\right|} \quad 2\text{-}4$$

While the Relative Absolute Error (RAE) is defined by the ratio between the Absolute Error and the magnitude of the original spectrum. The Mean Relative Absolute Error (MRAE) measures the average RAE of the whole image scene. MRAE was chosen to rank the results for the NTIRE (Arad, *et.al.*, 2018; 2020; 2022) spectral recovery competitions. The MRAE is commonly used as the loss function when training deep neural networks to recover spectral information (Lin & Finlayson, 2021).

$$MRAE = \frac{\sum_{i,c} \frac{\left|P_{gt_{i_c}} - P_{rec_{i_c}}\right|}{P_{gt_{i_c}}}}{\left|P_{gt}\right|} \quad 2\text{-}5$$

When multiplied by 100% the MRAE gives the mean absolute percentage error.

- **Spectral Angle Mapper (SAM)**

Spectral Angle Mapper (SAM) determines the differences between the original spectrum and the recovered spectrum by treating them as two vectors and calculating the angle between them:

$$SAM = \cos^{-1}\left(\frac{\sum_c P_{gt_c} P_{rec_c}}{\sqrt{\sum_c {P_{gt_c}}^2}\sqrt{\sum_c {P_{rec_c}}^2}}\right) \quad 2\text{-}6$$

where the P_{gt_c} and P_{rec_c} refer to the value of the c-th spectral channel of the ground truth and the reconstructed pixel respectively. The mean of SAM from the whole image scene can be used as a general error evaluation. SAM is independent of the exposure when the image was taken, and it is commonly used in hyperspectral image classification (Murphy *et al.*, 2012) (Ayhan & Kwan, 2017).

- **CIE Delta E (ΔE)**

The CIE Delta E measures the difference between two spectra from an aspect of human perception. The reconstructed and original spectra first need to be converted into a colourimetric value with the same camera spectral response function and illumination. Then the colourimetric value needs to be converted into CIELAB colour space, which implies specifying a white point (illuminant/camera). Here we use (L_{gt}, a_{gt}, b_{gt}) and $(L_{rec}, a_{rec}, b_{rec})$ to represent the converted LAB value for the original and reconstructed spectrum. The ΔE 1964 is defined as:

$$\Delta E = \sqrt{\left(L_{gt} - L_{rec}\right)^2 + \left(a_{gt} - b_{rec}\right)^2 + \left(b_{gt} - b_{rec}\right)^2} \quad 2\text{-}7$$

There are two updates for ΔE 1964 addressed to perceptual non-uniformities in 1994 and 2000 prospectively. ΔE is measured on a scale from 0 to 100, where 0 refers to the same colour. Commonly, when ΔE less than 1 the colour difference is not noticeable from a human perceptual standpoint; while when ΔE is between 1 and 2, the colour difference could be noticed with a close observation; when ΔE is less than 10 the difference could be observed from a glance; while when ΔE is larger than 50, the colours are more opposite than similar (Lee & Powers, 2005).

2.1.2. Comparison Between Evaluation Metrics

The RMSE and MRAE both focus on the difference in intensity between the recovered spectrum and the reference spectrum at one or multiple wavelengths. They share similarities, such as not being concerned with the direction of the error and having a negative orientation, which means that lower values indicate a smaller error. However, there are also differences between them.

Because errors are squared before averaging, the RMSE is a natural loss function for regressions that minimizes least squares error. In some cases, the RMSE can be easily optimized with linear approaches due to its least square characteristic. The RMSE gives a relatively high weight to large errors, which means it is more sensitive to outliers than the MRAE. When larger errors are undesirable for an application, the RMSE would be a useful error measurement.

From an interpretive perspective, the MRAE is a better choice since it provides a straightforward average error metric. Additionally, the MRAE has an advantage over the MAE and RMSE in that it is insensitive to scaling, which can occur due to exposure. When both the original spectrum P_{gt} and the recovered spectrum P_{rec} are scaled by a factor of s, the RMSE is also scaled by s as $RMSE(sP_{gt}, sP_{rec}) = s * RMSE(P_{gt}, P_{rec})$, as is the MAE. However, the MRAE will remain the same as $MRAE(sP_{gt}, sP_{rec}) = MRAE(P_{gt}, P_{rec})$. In this case, a factor of s in RMSE might not represent a big difference in the spectrum and can disproportionately influence the error. However, when the original and recovered spectrum have different scaling, both RMSE and MRAE will be affected. The RMSE is also sensitive to outliers because it squares the error terms. This can lead to a greater emphasis on larger errors and may not be desirable if the goal is to focus on the overall error distribution rather than the extreme cases. These reasons explain why the current CNN-based spectral super-resolution methods chose MRAE as the loss function.

However, MRAE has two disadvantages. First, it produces infinite or undefined values for zero or close to zero values (Kim & Kim, 2016). This gives more weight to wavelength (channels) where the ground truth spectrum is close to 0. Second, and more importantly, when using the relative model as a loss function, the system will systematically give predictions that are lower than the ground truth (Kolassa & Martin, 2011). Let $P_{gt_{i_{c1}}}$ and $P_{gt_{i_{c2}}}$ represent the intensity of the ground truth spectrum at ith pixel at $c1$th spectral channel and $c2$th spectral channel respectively, and $P_{gt_{i_{c1}}} \gg P_{gt_{i_{c2}}}$. Here, we assume both spectral channels have the same error e. When measuring the relative error, the error in the $c2^{\text{th}}$ the channel would appear higher. As a result, the measured error is biased. When performing spectral reconstruction, using MRAE as the loss function would give high weights to low-intensity spectral channels. In spectral imaging, information is typically encoded in the spectral content or distribution across different wavelengths. Since wavelengths with higher intensity play an important role in determining the distribution of a spectrum, giving high weight to channels with low

intensity would be inadvisable in some cases. Additionally, some applications may prioritize specific wavelengths, such as the red edge near 680 nm which is used to estimate the chlorophyll content of plant leaves. In these cases, a weighted error measurement that emphasizes those wavelengths would be more appropriate. Overall, the choice of error measurement in spectral super-resolution should be based on the specific application and its priorities.

Although ΔE is based on human perception and relates to colour differences, it cannot directly quantify differences between corresponding spectra. This limitation becomes particularly pronounced when dealing with spectral samples affected by metamerism. Even if they appear similarly coloured to the human eye, the shapes of their spectra could differ significantly.

For applications that primarily focus on the shape of the spectrum, such as classification, spectra in the same shape but different scales should be normalized to the same magnitude. In such cases, the SAM, which is not affected by scaling effects, would be a better choice.

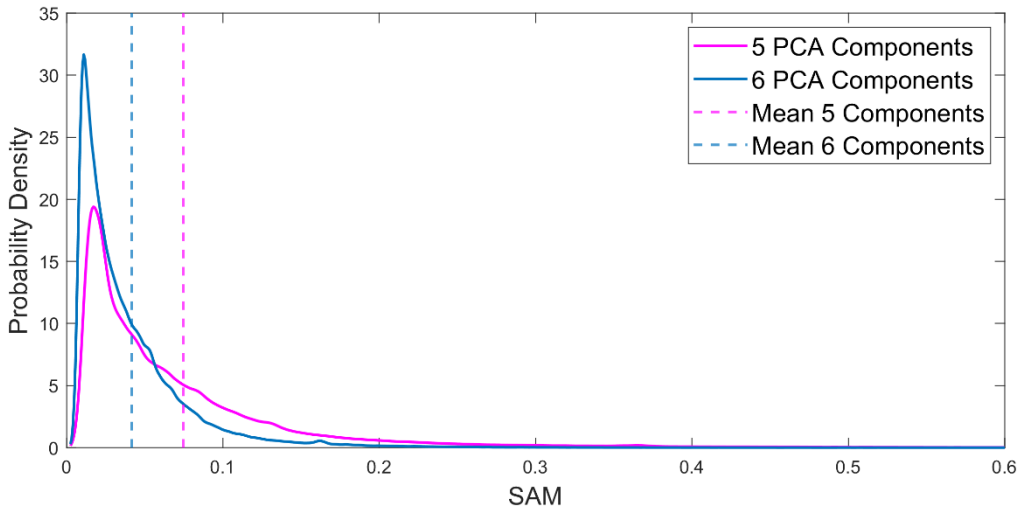
2.1.3. Using 95% Error Instead of Mean Error Measurement

The error metrics we've discussed earlier all utilize an aggregate measurement—specifically the mean—across samples to evaluate the accuracy of reconstructed spectra relative to the original spectra. Nevertheless, employing mean error as a representative gauge of spectral similarity comes with several inherent drawbacks. These include sensitivity to outliers—where even a single aberrant point can significantly skew the average error. Moreover, the mean error fails to account for skewness: if the error distribution is lopsided, the mean may not be the most representative indicator of central tendency, contrary to expectations.

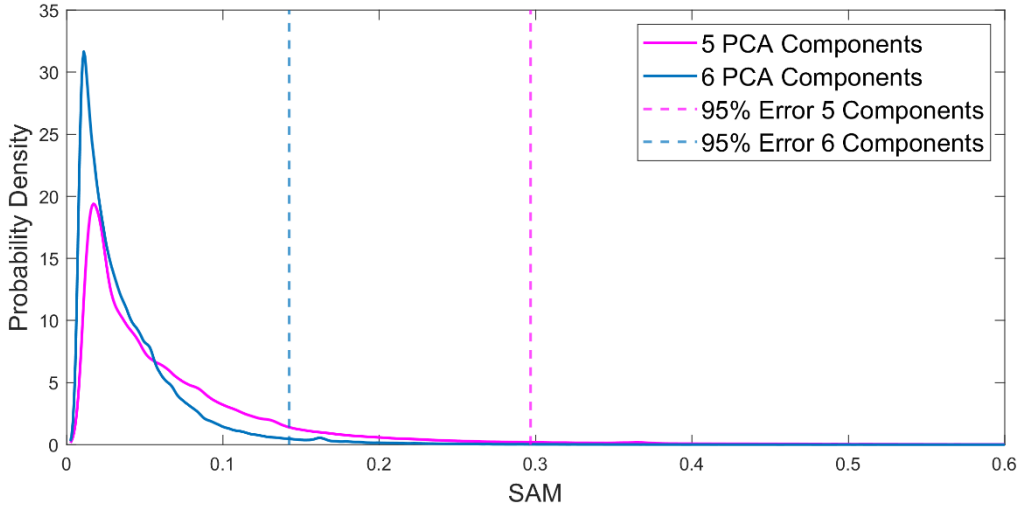
Consider Figure 6, which depicts the distribution of SAM error when using 5 (distribution 1) and 6 (distribution 2) PCA components, as introduced in the previous section, to represent spectral data. For spectral data, the reconstruction or representation error isn't likely to follow a Gaussian distribution; instead, it typically exhibits a skewed distribution with a long tail, characterized by outliers. Consequently, the average error fails to provide a comprehensive depiction of the error distribution in spectral data reconstruction or representation.

A further limitation of using average error to gauge the performance of spectral data reconstruction or representation is the difficulty of defining a significant difference. Frequently, a newly proposed method might report an improvement in reconstruction accuracy of as little as 0.001 in the MRAE. However, it's challenging to ascertain whether such a marginal improvement is genuinely significant or simply due to an enhanced reconstruction of a specific outlier.

To overcome these limitations, we propose a comprehensive analysis of the error measurement distribution when comparing reconstruction performances. Consider Figure 6, for example, the mean errors when using 5 and 6 PCA components are relatively close, offering limited insight. Yet, when we present the error distribution, it becomes evident that incorporating the 6th component increases the likelihood of a lower SAM error for a larger number of samples. In addition to analysing the error distribution, we suggest using the 95th percentile error as an overall error measurement in this study. The 95% error is defined as the threshold below which 95th percentile of samples' errors fall. Compared with the average error, the 95th percentile error is less influenced by a few outliers and also accounts for the skewness of the error distribution. Moreover, the 95th percentile error can differentiate the performance of different reconstructions more distinctly, as shown in Figure 6 (b).



(a)



(b)

Figure 6. The distribution of SAM when using 5 and 6 PCA components to represent spectral data. Dot lines in (a) show the mean of SAM while (b) show the 95th percentile (95%) error.

Hence, in this research, we opt for using the error distribution and the 95th percentile error instead of the average error when analysing the performance of reconstruction methods. It's critical for other researchers in the field of spectral reconstruction to consider the limitations we've highlighted in existing error measurements.

2.2. Hyperspectral Databases

This section provides a review of existing open-source hyperspectral databases. Table 1 summaries these databases, while the detailed information for each dataset follows the table. Early databases, such as CAVE, are relatively small. However, in recent years, competitions like NTIRE have made more hyperspectral data available. Having an overview of these datasets is crucial, as they will be involved in subsequent discussions.

Table 1. Overview of the existing open-source hyperspectral dataset.

	Number of images	Spatial resolution	Spectral resolution	Spectral range	Number of bands	Scenes
CAVE (Yasuma *et al*., 2010)	32	512×512	10 nm	400-700 nm	31	Objects, colour checker
ICVL (Arad & Ben-Shahar, 2016)	203	1392×1300 & 1392×1100	10 nm	400-700 nm	31	urban, rural, indoor, objects, plant
NTIRE 2018 (Arad *et al*., 2018)	286	1392×1300 & 1392×1100	10 nm	400-700 nm	31	urban, rural, indoor, objects, plant
NTIRE 2020 (Arad *et al*., 2020)	510	512×482	10 nm	400-700 nm	31	urban, rural, indoor, objects, plant
NTIRE 2022 (Arad *et al*., 2022)	1000	512×482	10 nm	400-700 nm	31	urban, rural, indoor, objects, plant

The **CAVE** database images were captured in indoor settings under CIE Standard Illuminant D65 lighting conditions. This dataset boasts a diverse array of subjects, encompassing both genuine and artificial fruits, various food items, beverages, and more.

ICVL database, introduced by Arad and Shahar in conjunction with their sparse coding method in 2016, is a valuable resource for hyperspectral imaging studies. It consists of 203 images captured under natural lighting conditions. Encompassing a diverse range of indoor and outdoor scenes, the ICVL database provides a broad spectrum of visual data for in-depth exploration and analysis.

NTIRE 2018 is an extension of the ICVL database with an increased number of images.

NTIRE 2020 were captured using a mobile hyperspectral camera (Specim-IQ) taken under natural lighting conditions.

NTIRE 2022 serves as an expanded version of the NTIRE 2020 database. As the most extensive hyperspectral database available, NTIRE 2022 houses a diverse array of image scenes and subjects, spanning both indoor and outdoor environments. This comprehensive collection thus provides a rich resource for the study and understanding of various hyperspectral imaging applications.

It's important to note that the datasets mentioned in our study are all hyperspectral images, rather than the spectral reflectance measurements of the observed materials. In essence, the data encapsulates the spectral radiance under illuminant, captured by a hyperspectral camera. These databases are appropriate for research topics such as low-cost hyperspectral imaging (including spectral super-resolution), but they cannot be employed as spectral reflectance data. If an application necessitates spectral-reflectance data, it becomes necessary to estimate the illuminant from the hyperspectral images. For databases like CAVE, where the illuminant is known, this process is relatively straightforward. However, for databases such as NTIRE, estimating the illuminant becomes a necessary step, which increases the complexity of the task. This illuminant estimation requires additional steps and algorithms, which may also introduce potential sources of error or uncertainty in the spectral-reflectance data.

Furthermore, the RGB values in these databases are simulated from the captured spectra using RGB camera models with various settings. They do not represent actual data obtained directly from real-world scenes. This is a crucial distinction to consider when applying or interpreting the data from these databases.

2.3. A linear Mapping from RGB to Spectrum

In the previous chapter, we discussed how spectral data can be represented accurately as a linear combination of a few basis functions. In earlier works (Marimont & Wandell, 1992) (F. H. Imai & Berns, 1999), researchers attempted to reconstruct a spectrum with linear basis functions, without using prior spectral data. Equation 2-8 represents the target spectrum as s, which can be represented as a linear combination of N basis functions b, with weights (coefficients) w.

$$s = \sum_{i=1}^{N} b_i w_i \qquad 2\text{-}8$$

However, there are two challenges for this approach: first, finding the basis functions that can accurately represent the spectral data, and second, estimating the corresponding weights/coefficients for each basis function. One simple approach is to utilize the camera function as a reference basis function and calculate the weights based on the RGB values. As discussed in Chapter 1, the RGB values of the camera can be modelled as:

$$\rho_k = \int c_k(\lambda) l(\lambda) s(\lambda) c_k(\lambda) d\lambda \qquad 2\text{-}9$$

with k in $\{1,2,3\}$ (Adão *et al*., 2017). If the spectrum is represented by a series of discrete wavelength samples, then Equation 2-9 can be represented in matrix form:

$$\mathrm{P} = CLS \qquad 2\text{-}10$$

where P contains the tristimulus vectors representing the RGB values of each sample point, C is the camera's spectral response functions, L represents the illumination and S contains the spectral reflectances of each sample. Given that low-cost HSI aims to recover the spectral radiance measured by a hyperspectral device using an RGB camera, we simplify our discussion by replacing LS with the spectral radiance (R) measured by the camera. Then 2-10 can be further written as:

$$\mathrm{P} = CR \qquad 2\text{-}11$$

With 2-11, the RGB values P could be up-sampled into spectra by using the pseudo-inverse of C:

$$\hat{R} = C^{+}\mathrm{P} \qquad 2\text{-}12$$

However, unconstrained inversion of the camera function usually results in an unsmooth up-sampled spectrum. This is because the spectral sensitivities of the red, green, and blue channels overlap in the visible range. Using the RGB values directly as weights of the inverted camera function may lead to a large different up-sampled spectrum compared to the original spectrum.

Instead of using the camera function, Glassner (1989) proposed a simple matrix inversion method that used three basis functions (constant, sine, and cosine) to derive the spectrum from an RGB triplet. Sun *et al*. (1999) used three smooth Gaussian functions as basis functions based on RGB values as weights for those functions. However, these approaches could result in reflectance spectra with negative values, which is physically impossible. Smits (1999) addressed this issue by considering the smoothness of the recovered spectrum and representing it as a linear combination of non-negative white and colour spectral bases. In addition to the basis functions corresponding to RGB colour space, Smits introduced basis functions for cyan, magenta, and yellow to smooth the

recovered spectra. However, the up-sampled spectrum is poorly correlated to the measured spectrum, except for being smooth.

Meng (2015) proposed a multiple basis function method, where the spectral basis is precomputed based on regularly sampled grid points on the XYZ colour space. Then, the bases are used as a lookup table to up-sample the spectra based on a linear combination of weights calculated by a colourimetric triplet. Burns (2020) redefined smoothness by minimizing the squared slope of the reconstructed spectrum between neighbouring spectral bands. At the same time, the authors considered the colour constancy of the recovered spectrum when designing the reconstruction functions, by adding an extra constraint that requires the recovered spectrum to be associated with the original XYZ value which is referenced to a given illuminant.

However, solely considering the smoothness and non-negativity of the spectrum is not enough to ensure that the recovered spectrum is close to the original. As more and more spectral databases became available, researchers began exploring methods for learning the basis functions from spectral data priors. The following sections will review these prior-based methods, we start the review with traditional models and then consider deep models.

2.4. Traditional Prior Based Methods

2.4.1. PCA Basis

Principal Component Analysis (PCA) is the most popular method for linear dimensionality reduction, and it has long been used to represent spectral reflectance data. Equation 2-13 shows that spectra S can be represented as the product of the weight matrix W and the PCA basis function B, in a lossless manner when the number of PCA components (bases) is equal to the number of wavelength bands.

$$R = BW. \qquad 2\text{-}13$$

Cohen (1964) first used PCA to represent colour spectral data in a low-dimensional manner. Early researchers attempted to use principal components as basis functions when reconstructing spectral information from RGB data. Fairman and Brill (2004) considered obtaining spectrum, principal-component basis, and tristimulus values from each other with a linear transform with the first three components. However, when using a limited number of principal components to represent the data, important information may reside in the variance explained by subsequent components. To solve this problem, instead of using a single set of PCA bases, Fairman and Brill learned multiple PCA

bases for different colour clusters. From the result, directed bases for each cluster could increase the reconstruction accuracy compared with using general bases for all spectra.

Ayala *et al.* (2006) learn 10 subgroups of PCA basis based on Munsell hues. For each group, the first three components are used to reconstruct spectra from RGB values. Hajipour and Shams-Neteri (2017) used a shallow neural network to classify spectral data into subgroups and used the first component as the basis function for each group. Then the spectra were represented by a linear combination of all learned bases. However, in this method, the possible metamerism within each cluster is ignored by assuming spectra with similar colours are also likely to be similar.

In addition to using multiple groups, Agahian *et al.* (2008) proposed a weighted PCA-based method, where the weights are determined to minimise the colour difference (ΔE). Otsu *et. al.* (2018) further improved the PCA-based method, the authors involve colour consistency as a constraint by ensuring the generated spectrum can be converted into the reference tristimulus colours with the same camera function.

Another challenge of using PCA bases to reconstruct spectral reflectance is determining how to correctly estimate the weights of PCA bases for given tristimulus colours. To better explain this problem, here we assume the camera function is C, and the tristimulus colours are given by:

$$\mathrm{P} = CR = CBW \qquad 2\text{-}14$$

The weight $W^{(A)}$ needs to be estimated from CB as:

$$W^{(A)} = \left(CB^{(A)}\right)^{+}\mathrm{P} \qquad 2\text{-}15$$

where A is the number of bases. Accurately estimating more weights for components means a higher reconstruction accuracy. When the number of bases equals the number of colour channels, then CB is a full-rank square matrix and can be inverted. The weights of the first three bases can be estimated as:

$$W^{(3)} = \left(CB^{(3)}\right)^{+}\mathrm{P} \qquad 2\text{-}16$$

When the number of bases is larger than the number of colour channels, the problem becomes ill-posed (the number of unknowns is larger than the number of equations). Although the weights can be estimated by the pseudo-inverse of CB, the estimate differs significantly from the optimal weights obtained by projecting the spectra onto the PCA bases. In many cases, the estimated weights are out of the range of the expected PCA variance from the prior. As a result, the recovered spectral

reflectance would have a higher error. That is the reason why previously listed methods only used the first three components as the basis function.

In considering this problem we derived a new PCA coefficient estimation method to recover spectral reflectance from RGB images (Chang *et al.*, 2021). The main idea of the proposed method is to keep each of the estimated weights within the range of the PCA distribution of the data prior while ensuring that the recovered spectral reflectance can be converted to the given RGB value. Rather than solving 2-15 by pseudo-inverse, we consider all possible solutions and select the most likely. With more than 3 bases, the system of equations is under-determined. With 4 principal components, the solutions to 2-15 fall on a line in the 4D PCA weight space. With 5 bases, the solutions to 2-15 fall on a plane on the 5D PCA weight space, and so on for more bases. It is this ambiguity in solutions that results in camera metamerism.

Let $W^{(A)}$ represent all the solutions of 2-15. Because the set of possible weights in $W^{(A)}$ represent a linear system and because the first three weights can be determined from 2-16, then the set of solutions can use a parametric representation as a linear combination of the remaining $A-3$ weights, which can be the parameterised by P. Therefore, we assume:

$$W^{(A)} = PK + W_0 \qquad 2\text{-}17$$

where K (size A by $A-3$) determines the linear relation between P and the first three weights $W^{(3)}$, and W_0 is a constant array that satisfies 2-16 when $P = 0$, this parameterisation sets:

$$K^T = \left[\left(K^{(3)}\right)^T | I\right] \qquad 2\text{-}18$$

where $K^{(3)}$ is an $A-3$ by 3 weight matrix, and I is the $A-3$ by $A-3$ identity matrix i.e., the 4^{th} and higher weight corresponds to the parameter, P. To determine W_0 we partition $B^{(3)}C$ between the first 3 and remaining bases as:

$$CB^{(A)} = \left[CB^{(3)} | CB^{(A-3)}\right] \qquad 2\text{-}19$$

and set $P = 0$ to give:

$$W_0{}^T = \left[\left(W^{(3)}\right)^T | \mathbf{0}\right] \qquad 2\text{-}20$$

where $\mathbf{0}$ is a zero matrix, and $W^{(3)}$ is the solution to 2-16 using the first 3 bases:

$$\mathrm{P} = CB^{(3)}W^{(3)} \qquad 2\text{-}21$$

When the remaining $A-3$ weights are added, to maintain colour consistency we have:

$$\mathrm{P} = CB^{(3)}(PK + W_0) \tag{2-22}$$

From 2-20, and comparing 2-21 and 2-22, it is clear that:

$$\mathbf{0} = CB^{(3)}PK \tag{2-23}$$

To satisfy 2-23 with different P, K needs to satisfy:

$$CB^{(3)}K^{(3)} = -CB^{(A-3)} \tag{2-24}$$

Therefore, the linear relation that gives the first three weights from P for all metamerism solutions is given by:

$$K^{(3)} = -\left(CB^{(3)}\right)^{-1}CB^{(A-3)} \tag{2-25}$$

To select a unique solution, we assume the data is Gaussian distributed with the variance of the ith bases σ_i^2. The highest probability (most likely) solution is therefore going to be close to the centre of the principal components. The most likely spectrum from the metamerism set can then be found by weighting each axis by the corresponding standard deviation (Mahalanobis distance):

$$P = argmin \sum_i \frac{w_i^2(P)}{\sigma_i^2} \tag{2-26}$$

Since the relation between the first three weights in $W^{(A)}$ and P is linear, this gives a quadratic function with a single well-defined solution. Then we can use the estimated weights to recover the spectral reflectance. For example, consider $A = 4$. Since, w_1, w_2 & w_3 can be linearly represented by w_4, then the possible solutions in the w_1, w_2 & w_3 space with the same RGB colour will be located on a line. What we did was select the most likely solution based on the distribution of the data prior.

The proposed method was tested using the CAVE database and compared to directly inverting 2-15. The results showed that our proposed method achieved higher accuracy. However, it's important to note that this method has its limitations. PCA assumes that the hyperspectral data are Gaussian distributed, which is not generally true. Additionally, the PCA model is not precise enough to accurately model the spectral reflectance in the training dataset.

2.4.2. Dictionary Learning Based Methods

Since spectral data can be represented with a linear combination of a few basis vectors, representing the spectral data sparsely with atoms from an overcomplete dictionary could provide better accuracy compared to using a limited number of principal components (Xing *et al.*, 2012). Arad *et. al.* (2016) first introduced a dictionary learning-based spectral super-resolution model, which uses K-SVD (K-singular value decomposition) to learn an overcomplete dictionary of the hyperspectral data. Let D_h be the computed dictionary, and h_i be the atoms:

$$D_h = \{h_1, h_2, \dots, h_n\} \quad \text{2-27}$$

For each spectrum, it could be represented sparsely by a linear combination of a few atoms from the dictionary. The spectral dictionary needs to be converted into RGB by a known camera sensitivity function.

$$D_\rho = D_h C = \{h_1 C, h_2 C, \dots, h_3 C\} \quad \text{2-28}$$

When reconstructing, the dictionary representation of an RGB pixel is computed with the Orthogonal Match Pursuit (OMP) by finding the best-matching RGB atoms within the converted dictionary which satisfy:

$$\mathrm{P} = D_\rho W \quad \text{2-29}$$

Then the spectra would be represented with the same linear combination of the corresponding spectra as:

$$R = D_h W \quad \text{2-30}$$

Arad's sparse coding method was published with the largest hyperspectral database at that time and has long been cited. Since this method only takes 'sparsity' as a constraint when doing the reconstruction, many works with better accuracy then Arad's work have been published later.

Inspired by Arad, Fu *et. al.* (2018) introduced multiple non-negative sparse dictionaries. In this work, the hyperspectral data is clustered first using K-means clustering. Then for each cluster, a spectral dictionary and the corresponding RGB dictionary are learned. The authors also considered the non-negativity of the spectral data. During reconstruction, for a given RGB input, the nearest cluster centre is found (using Euclidean distance). Then, the corresponding RGB dictionary and spectral dictionary of that cluster are used to reconstruct the spectral reflectance.

Aeschbacher et al. (2017) focused on better estimating the weights of the learned bases. This method could be understood as an update of Arad's work. It uses the same method to create hyperspectral and RGB dictionaries. Different from Arad, which converts the spectral dictionary to RGB and uses orthogonal matching pursuit to estimate the weights globally from the whole dictionary, the authors use a linear combination of local neighbour bases from the dictionary to represent the spectra, achieved by the A+ algorithm (Timofte et al., 2015) for spatial super-resolution.

The above sparse coding methods are all pixel-wise, based solely on the RGB of a pixel to recover its spectral detail, without additional information, the reconstructed results are influenced by metamerism. Li *et. al.* (2017) improved the dictionary-based method by introducing local features as additional information. The local texture was learned from 16×16 image patches by Gabor filtering (Manjunath & Ma, 1996), and then converted to feature vectors. The authors employed both the RGB value and feature vector to train an over-complex dictionary with 200 atoms. During reconstruction, the local feature vector and the RGB value acted as a constraint when selecting the bases. Since there exists a linear transform between the spectrum and RGB value, the locally linear relation learned from the spectral dictionary was maintained. Therefore, utilizing the given RGB value and local feature reduced the ambiguity of the reconstructed spectra. However, in terms of the ability to resolve metamerism, the authors assumed that high reconstruction errors were all due to metamerism rather than collecting metamerism data, which limits the reliability of the method. Additionally, the method did not consider colour consistency, leading to a potential mismatch between the recovered and original spectra in terms of colour.

Geng *et al.* (2019) developed a spatial dictionary-based method. Rather than building a pixel-wise dictionary, the authors constructed a dictionary for a 3×3 image patch centred on the target pixel. Nevertheless, this approach solely focuses on the similarity of spectra within a small patch and does not explicitly address the issue of resolving metamerism. Thus, its effectiveness in utilizing information from very local neighbours to address metamerism is yet to be determined.

Different from previous works, Nguyen *et al*. (2014) proposed a radial basis function-based method. The spectral reflectance was modelled by a linear combination of 45-50 radial basis functions, and the weight of each function was estimated from a linear least squares method. However, the result using the radial basis function doesn't show an advantage compared to other dictionary-based methods even with more dimensions.

In summation, dictionary-based methods rely on pre-existing data as a reference to facilitate the translation from RGB to the spectral domain. Early contributions, such as Arad's work, employed sparsity as the only constraint, thereby limiting the accuracy of spectral reconstruction. For more precise reconstruction, additional constraints are necessary, with efforts currently being made to incorporate elements like local linear relationships, spectral correlation, and local texture. These

added constraints have all demonstrated a reduction in reconstruction error compared to approaches that merely focus on sparsity. According to Li's report (2017), the integration of spatial information could significantly reduce the RMSE of reconstructed images from 0.12 to 0.03 in comparison to Arad's methodology. However, colour consistency has not yet been considered as a constraint in current practices. Additionally, the investigation of local spatial details, such as texture, remains relatively unexplored. Pertinent questions about the essential types of spatial information and the minimum spatial range necessary for accurate representation remain unanswered.

2.4.3. Other Traditional Methods

Manifold learning has also been used in single-image spectral super-resolution by Jia *et. al.* (2017). In their work, spectral reflectance samples from the training dataset were first projected to a low dimensional (3D) space via a non-linear dimensionally reduction method called ISOMAP. The spectral sample also needs to be converted to an RGB value. Then, the authors learned mapping from the 3D RGB space to the 3D spectral space. When doing reconstruction, their relationship can be used to recover the high-dimensional spectral data from the embedding space.

Akhtar and Mian (2018) proposed a method to recover the spectrum from RGB images of known spectral quantization by modelling natural spectra under Gaussian Processes and combining them with the RGB images. The given images were first separated into 512×512 patches. During the training process, patches from training hyperspectral images were extracted and clustered. Multiple sets of Gaussian Processes were learned from training hyperspectral image patches. Those processes were then transformed to match the spectral quantization of the RGB images. After training, the RGB image patch was extracted and matched with the RGB transformations of the hyperspectral clusters. Then the representation codes were combined with the original Gaussian Processes to construct the desired hyperspectral image.

In contrast to the dictionary-based approach, manifold learning techniques learn inherent data structures as constraints derived from the data prior. On the other hand, methods based on Gaussian Processes leverage the correlation between neighbouring wavelengths and the similarity of spatial structures as constraints. These methods thereby offer additional perspectives on the problem of spectral reconstruction.

Table 2 provides an overview of existing prior-based methods, including their training data, constraints, and performance. However, several error measurements are absent from the table. This absence is due, in some cases, to the authors using different error measurements and, in others, to the authors normalizing the data, rendering it non-comparable with the results listed. Despite being

extensively researched, prior-based methods continue to present limitations, which will be discussed in the following section.

Table 2. Overview of the existing prior-based methods with the reconstruction accuracy compassion.

Category	Method	Constraints	RGB consistency	Training data	RMSE	MRAE	SAM
Dictionary learning	Sparse Coding (Arad & Ben-Shahar, 2016)	sparsity	no	NTIRE 2018 (12-bit)	51.48	0.081	5.01
				NTIRE 2018 (8-bit)	2.63	-	-
				NTIRE 2020	0.033	0.078	6.46
				CAVE	2.74	-	-
	SR A+ (Aeschbacher *et al*., 2017)	sparsity, local Euclidean linearity	no	NTIRE 2018 (12-bit)	26.09	0.045	2.83
				NTIRE 2022 (8-bit)	0.023	0.073	4.61
	Multiple Sparse Coding (Fu *et al*., 2018)	Spectral correlation	no	CAVE	-	-	11.32
	Spatially Constrained Dictionary Learning (Geng *et al*., 2019)	Local texture	no	CAVE	-	-	-
	Locally Linear Embedded Sparse Coding (Li *et al*., 2017)	Local texture	no	NTIRE 2018 (8-bit)	-	-	-
Manifold Mapping	Manifold Mapping (Jia *et al*., 2017)	Low-dimensional manifold	no	CAVE	-	-	-
Gaussian Process	Gaussian Process (Akhtar & Mian, 2018)	Spectral physics, spatial structure similarity	no	NTIRE 2018 (8-bit)	-	-	3.68

2.4.4 Limitations of the Traditional Methods

The conventional dictionary learning approach is not without its limitations. As the complexity of the spectral database increases, the representation power of a dictionary may not be enough to capture complex nonlinear patterns in the data accurately. Consequently, the learned dictionary may not be able to learn the spectral details. Although using sub-dictionaries can improve reconstruction accuracy, this comes at the cost of increased model size and slower processing speeds. Furthermore, the quality of the learned dictionary is highly dependent on the training data. If the

training data is not representative of the dataset, the reconstruction accuracy of the learned dictionary may be poor, as observed in Arad's work.

The dictionary-based approach utilizes limited features from prior data; for instance, pixel-wise methods often neglect the potential use of spatial structure similarity as a constraint. Although there have been attempts to extract local spatial information, the efficient and effective utilization of this information remains unclear. Moreover, the colour consistency of the reconstructed spectra is frequently neglected, resulting in spectra that may exhibit different RGB values than the original ones. As such, it becomes essential to consider RGB consistency as a constraint in future works. Most importantly, traditional methods especially those that are pixel-wise are not particularly effective in addressing the issue of metamerism. Consequently, there's a pressing need to develop methods that more adequately tackle these problems.

2.5. Deep Model-Based Methods

With the development of computational ability and more hyperspectral datasets becoming available, many deep neural network-based spectral super-resolution methods have been published. Compared to dictionary learning, deep neural networks show an advantage in several aspects.

- Representation power: Deep learning models have a higher capacity to represent complex, non-linear patterns in data. This is achieved by the use of multiple layers of artificial neurons, each performing a non-linear transformation of the input data.
- End-to-end training: Deep learning models are trained end-to-end, which means that the entire model is optimized to minimize a given loss function. This enables the model to learn complex features that are specific to reduce the error in the reconstructed spectrum. Dictionary learning methods, on the other hand, typically rely on pre-defined dictionaries, which may not be optimized for reducing the reconstruction error.
- Extract local and non-local features: The convolutional deep learning models are capable of extracting both local and non-local spatial features from images. CNNs use filters to extract local features such as edges, corners, and textures. This operation allows CNNs to learn and recognize spatial hierarchies or patterns within the input image patch. To further expand the captured spatial area while simultaneously minimizing dimensionality, pooling layers are commonly integrated within the CNN architecture. While the early layers of a CNN extract local features, the deeper layers are capable of recognizing more abstract and global features. As we move further into the network, each successive layer uses the outputs from previous layers (which represent local features) as its input, allowing the network to gradually learn increasingly complex and abstract features. This hierarchical feature learning enables the model to recognize larger patterns that span across the entire image,

representing non-local or global information. These provide deep-learning approaches to the potential context to resolve metamerism when an appropriate loss function is used.

These advantages have made more researchers choose deep models when recovering spectral information from RGB. The rest of this section reviews representative deep convolutional neural network methods.

2.5.1. Overview of Deep Convolutional Neural Network-Based Methods

Figure 7 shows a classification of the existing deep methods based on the network structure provided by Zhang *et al*. (2021). In this report, we review the state-of-the-art based on this classification. Several more recent works have been added.

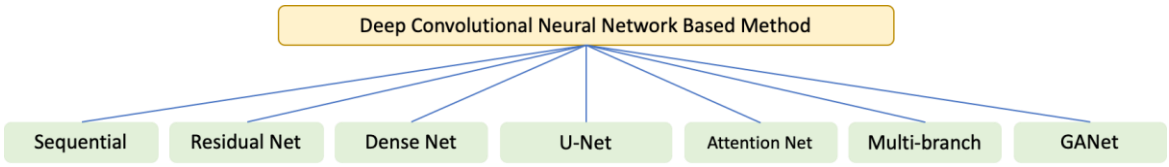


Figure 7. Classification of deep convolutional neural network-based spectral super-resolution methods by Zhang *et al*. (2021).

Table 3 provides an overview of the existing single image spectral super-resolution methods that employ CNNs. Note that this table is not exhaustive, with key methods selected from each family of techniques. The table includes pertinent information about each method such as their abbreviation, reference, network depth, number of parameters, training image patch size, the loss function employed, whether colour consistency is used as a constraint, and the dataset associated with the reconstruction accuracy. The depth of a network, signifying the number of its layers, in conjunction with the total number of parameters, serves as an indication of the network's complexity. However, due to the fact that not all networks are open-sourced and certain details such as the number of parameters is often not provided in the respective papers, some information is absent from Table 3. The dataset each network was trained and tested on is distinguished by colour. The reconstruction error is sourced from the papers, including the original paper in which the network was published and any others where the network was used as a benchmark for comparison. When a network is used for comparative purposes, it is important to bear in mind that changes in the database used and the training strategy employed may cause variances in the network's reconstruction performance from its original setting. Networks trained on the same dataset can be directly compared. However, it is not appropriate to compare reconstruction accuracies across different datasets, given that they possess different scales and formats. It's noteworthy that the original NTIRE 2018 data was in 12 bits, but some researchers have converted this data into 8 bits, a distinction made clear in the table.

Table 3. Overview of selected existing CNN-based methods with the reconstruction accuracy compassion.

Category	Networks	Depth	Number of parameters	Patch Size	Loss function	RGB consistency	Training data	RMSE	MRAE	SAM
Sequential Network	HSCNN (Xiong *et al*., 2017)	5	-	11×11	MSE	no	NTIRE 2018 (12-bit)	17.01	0.019	-
	HS2DNet (Koundinya *et al*., 2018)	5	-	64×64	MAE	no	CAVE	3.05	-	-
							NTIRE 2018 (12-bit)	21.39	0.020	-
	HS3DNet (Koundinya *et al*., 2018)						CAVE	2.86	-	-
							NTIRE 2018 (12-bit)	20.01	0.018	-
	RHSCNN (Han *et al*., 2018)	6	-	15×15	MSE	no	CAVE	4.78	-	7.37
Residual Network	HSRNET (Can & Timofte, 2018)	6	-	36×36	RMSE	no	CAVE	2.55		
	HSRNET+ (Kaya *et al.*, 2019)	21	-	_	RMSE	yes	CAVE	2.55		
							NTIRE 2018 (12-bit)	17.27	-	-
	HSCNN R (Shi *et al*., 2018)	21	-	50×50	MRAE	no	NTIRE 2018 (12-bit)	13.91	0.015	1.05
							NTIRE 2020	0.014	0.037	2.63
	EDSR (Lim *et al*., 2017)	-	2.42M	48×48	MRAE	no	NTIRE 2022	0.043	0.327	-
	MIRNET (Zamir *et al*., 2020)	6	3.75M	128×128	MRAE	no	NTIRE 2022	0.017	0.189	-
U-Net	HSUNET (Stiebel *et al.*, 2018)	5	-	32×32	ΔE	yes	NTIRE 2018 (12-bit)	15.34	0.016	-
							NTIRE 2020	0.015	0.039	2.74
	MSCNN (Yan *et al.*, 2018b)	10	-	64×64	MSE	no	NTIRE 2018 (12-bit)	19.28	0.023	1.46
							NTIRE 2020	0.024	0.072	4.91
	MXRUNNET (Banerjee & Palrecha, 2020)	56	31.35M	-	perceptual loss function (Johnson *et al.*, 2016)	no	NTIRE 2020	-	0.045	-
	HINET (Chen *et al*., 2021)	5	5.21M	256×256	MRAE	no	NTIRE 2022	0.030	0.203	-

Dense Network	HSDNET (Galliani *et al*., 2017)	6	-	64×64	Euclidean loss	no	CAVE	4.76	-	-
							NTIRE 2018 (12-bit)	20.98	0.027	1.57
							NTIRE 2020	0.025	0.085	4.34
	HSCNN D (Shi *et al*., 2018)	21	4.65M	50×50	MRAE	no	NTIRE 2018 (12-bit)	13.12	0.014	0.99
							NTIRE 2022	0.059	0.381	-
Attention Networks	AWAN (Li *et al*., 2020)	61	4.04M	64×64	MRAE, RGB consistency	yes	NTIRE 2018 (12-bit)	10.24	0.011	-
							NTIRE 2020	0.011	0.031	2.16
							NTIRE 2022	0.037	0.250	-
	HRNET (Zhao *et al*., 2020)	57	31.70M	256×256	MRAE	no	NTIRE 2018 (12-bit)	-	-	1.01
							NTIRE 2020	0.014	0.031	2.16
							NTIRE 2022	0.055	0.347	-
	RAAUENT (Li *et al*., 2020)	-	-	64×64	MRAE	no	NTIRE 2020	0.012	0.035	2.40
	RPANET (Peng *et al*., 2020)	36	-	64×64	MRAE	no	NTIRE 2020	-	0.037	-
	MST++ (Cai *et al*., 2022)		1.62M	256×256	MRAE	no	NTIRE 2022	0.025	0.164	-
	HDNET (Hu *et al*., 2022)	-	2.66M	256×256	MRAE	no	NTIRE 2022	0.032	0.205	-
Multi-branch Networks	LWRDANET (Nathan *et al*., 2020)	40	-	64×64	MRAE	no	NTIRE 2020	-	0.055	-
	PFMNET (Lei *et al*., 2020)	9	-	64×64	RMSE	no	CAVE	4.54	-	7.07
							NTIRE 2018 (8-bit)	1.03	-	0.99
	SRFNET (He *et al*., 2021)	-	-	128×128	SAM	yes	CAVE	-	-	7.62
GAN	CGAN (Alvarez-Gila *et al*., 2017)	8	-	256×256	GAN	no	NTIRE 2018 (8-bit)	1.46	-	-
	SAGAN (Liu & Zhao, 2020)	10	-	256×256	GAN	no	NTIRE 2018 (8-bit)	1.44	-	-
	TCGAN (Liu *et al*., 2022)	50	-	256×256	GAN	no	Textile	0.027	-	-

Additionally, even though the spectral data in the NTIRE 2022 dataset is an extension of the NTIRE 2020 dataset, the associated RGB values for training are different. The NTIRE 2022 RGB values incorporate a variety of exposures, which adds to the complexity of learning the mapping from RGB to spectrum. Therefore, even though the NTIRE 2020 database is similar to the NTIRE 2022, the reconstruction accuracies cannot be directly compared. In the following sections, a detailed review of the listed networks will be presented.

2.5.2. Deep Convolutional Neural Network-Based Methods

- **Sequential Networks**

A sequential network, also known as a sequential model, is a type of neural network architecture in which the layers are arranged sequentially, one after the other. This means that the output of one layer serves as the input for the next layer and so on until the final output is produced. Compared with more recent networks, they are relatively simple to implement and understand. In a sequential CNN architecture, the input image is passed through a series of convolutional layers and ends with a convolutional layer that produces the recovered image. Figure 8 shows a glimpse of a sequential network, where ***I*** represents the input image or image patch, while ***H*** represents the reconstructed hyperspectral image. The convolutional layers extract features from the input image by applying a set of learned filters. These extracted features capture different levels of abstraction and are used as context when reconstructing the spectrum from the input RGB image.

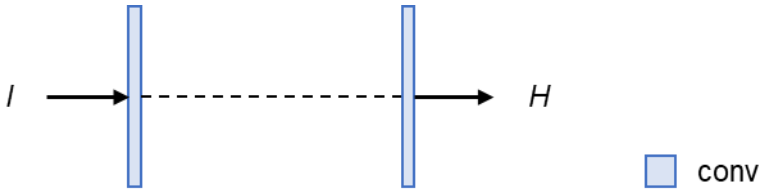


Figure 8. A glimpse of a sequential network.

Xiong *et al.* (2017) introduced the Hyperspectral Convolutional Neural Network (HSCNN), where the RGB image is first up-sampled to spectra based on Smits' work (1999). Then, the up-sampled spectra are passed through five feature mapping layers, each followed by a ReLU (Rectified Linear Unit) layer. The output of the last feature mapping layer is reconstructed to form the hyperspectral scene. This structure initially used in spatial super-resolution (J. Kim *et al.*, 2016), is adapted to learn a mapping from the roughly estimated spectrum to the ground truth. The HSCNN was trained to minimize the RMSE between the original and reconstructed spectrum. Even though the structure of the HCSNN was simple, it achieved relatively good reconstruction accuracy, which inspired many later works. The HSCNN model exhibited an interesting phenomenon, wherein its accuracy did not

improve with an increase in network depth. When the depth of the network is 5, the HSCNN achieves the best performance. This can be attributed to the absence of pooling layers in the HSCNN architecture. When the model is made deeper, it may tend to overfit the training data, leading to poor performance on testing data.

Koundinya *et al.* (2018) presented a method based on 2D and 3D CNN kernels (HS2/3DNET). In HS2/3DNET, the input RGB image is processed through several convolutional layers, each followed by an activation layer. This work is trained on 64×64 patches, with the sizes of the 2D and 3D kernels being 3×3 and $3 \times 3 \times 3$, respectively. The 3D kernel is designed to incorporate the RGB channels, as the R, G and B channels are correlated. From the result, the 3D CNN slightly improves the accuracy of the reconstruction compared to the 2D kernel. However, given the relatively simple structure of this network, it does not demonstrate any significant advantages over other networks.

Han *et al.* (2018) adapted an image spatial super-resolution network (Dong *et al*., 2015) to solve single image spectral super-resolution (RHSCNN). The network consists of three convolutional layers, with the first two followed by a ReLU layer. However, discrepancies remain between the recovered spectrum and the original spectrum, particularly in terms of high-frequency spectral content. To address these residuals, the author incorporated an additional reconstruction network with the same structure as the previous network. This network was trained on 15×15 image patches by minimizing the Euclidean distance between the recovered and reference patches. Unlike the residual network discussed later, the RHSCNN simply increases network depth by adding repetitive structures. The reported reconstruction accuracy indicates that this approach doesn't provide any distinct advantages.

However, due to the relatively simplistic architecture inherent to sequential networks, the models reviewed face challenges in learning complex data structures and mitigating the problem of vanishing gradients. This has underscored the need for more intricate network designs, which have subsequently been explored in later works on single image spectral super-resolution.

- **Residual Network**

Vanishing gradient is a problem in many deep learning models where the gradients of the loss function become very small. This occurs during the backpropagation process used to update the weights, especially in deep networks with many layers. When the gradients become too small, the weights in the early layers of the network are updated very slowly, or not at all, during training. This means that these nets do not effectively learn from the training data, and the model's performance can be poor as a result. To address the vanishing gradient problem, He *et al.* (2016) first proposed the residual network, which adds shortcut paths in parallel with convolution blocks, forcing the blocks to learn residuals. In a residual network (or block), during backpropagation the gradients can flow directly through these shortcut connections backwards from later layers to earlier filters, providing a

clear path for the gradient to propagate all the way through the network. This alleviates the vanishing gradient problem, allowing the network to be trained effectively even when it's very deep. The benefits of residual networks are not only their ability to mitigate the vanishing gradient problem but also their capacity to be easily trained with conventional methods, their ability to handle more complex functions and their improved performance with increasing depth.

Can and Timofte (2018) presented a residual network-based approach (HSRNET) that includes a main network and a residual convolutional layer. The main network progressively reconstructs the RGB inputs into spectra. Two residual blocks were introduced within the main network to capture more complex features. A skip connection between the input and output of the network work is added to recover the residual of the main network. To better extract information on different scales, different kernel sizes are used within the main network (5×5) and the skip connection (7×7) is different. The input for the main network was 17×17 image patches. Thanks to skip connections and multi-scale feature extraction, HSRNET achieves higher reconstruction accuracy compared to sequential networks.

Kaya *et al.* (2019) further improved HSRNET by considering colour consistency (HSRNET+) as an extra constraint. In the proposed method, a camera sensitivities estimation block was added, which minimizes the difference between the original RGB and reproduced RGB. The reproduced RGB was converted from the recovered spectra and the estimated camera function. However, limited by the accuracy of the estimated camera function, considering colour consistency doesn't further improve the performance on the CAVE dataset.

EDSR stands for Enhanced Deep Residual Networks for Single Image Super-Resolution. This model was proposed by Lim *et al*. (2017), for spatial super-resolution. The EDSR represents an improvement over the original residual network by removing unnecessary modules. Compared to ordinary residual blocks, the batch normalization is removed in EDSR, simplifying the architecture. In addition, the authors also introduced a multiscale residual block that is designed to extract information from different scales, further enhancing the network's capabilities. Cai *et al*. (2022) adopted EDSR for spectral super-resolution as a benchmark in their spectral reconstruction toolbox.

The MIRNET (Zamir *et al.*, 2020), Multi-Image Restoration Networks, is a model for high-quality image restoration tasks. MIRENT starts with a low-level feature extraction layer followed by a series of recursive residual groups. Each group includes several multiscale residual blocks which consist of a set of parallel branches of different dilation rates, which can capture features at various scales. The outputs of these branches are then concatenated and further processed to produce the final output. The parallel branches also allow information exchange across parallel streams in order to consolidate the large-scale features with the small-scale features. Finally, the processed features are used to reconstruct the output image by a convolutional layer. In addition, a skip connection is

added from the input layer to the output layer. This algorithm was adapted by Cai *et al.* (2022) for spectral super-resolution problems due to its capacity to extract contextualized representations.

Leveraging spatial super-resolution models, both EDSR and MIRNET have showcased their capability in extracting spatial features. We can reasonably assume that their skill in spatial feature extraction enhances their ability to resolve metamerism by using spatial information as context. Thanks to its more complex structure and multi-scale feature extraction, MIRNET outperforms EDSR.

- **U-Net**

The U-Net was first introduced by Ronneberger *et al.* (2015) for biomedical image segmentation. The U-Net accomplishes this through its unique architecture, which is symmetric and has a '**U**' shape as shown in Figure 9. It comprises two paths: the contracting path (encoder) and the expanding path (decoder). The contracting path uses convolutional and pooling layers to extract features from the input image, while the expanding path employs deconvolutional layers to up-sample the feature maps and restore the spatial resolution of the image. For spectral super-resolution tasks, the depth of the data increases as it moves through the expanding path. U-Net also incorporates several skip connections between the encoder and decoder paths that directly link corresponding layers.

Both U-Net and residual networks use skip connections to address the vanishing gradient problem. The skip connections in U-net are used to pass information directly across the network from each level in the encoding path to the corresponding level in the decoding path. These connections enable the U-Net to learn both low-level and high-level features in the input image, allowing it to utilize both local and global spatial information as context when performing spectral reconstruction.

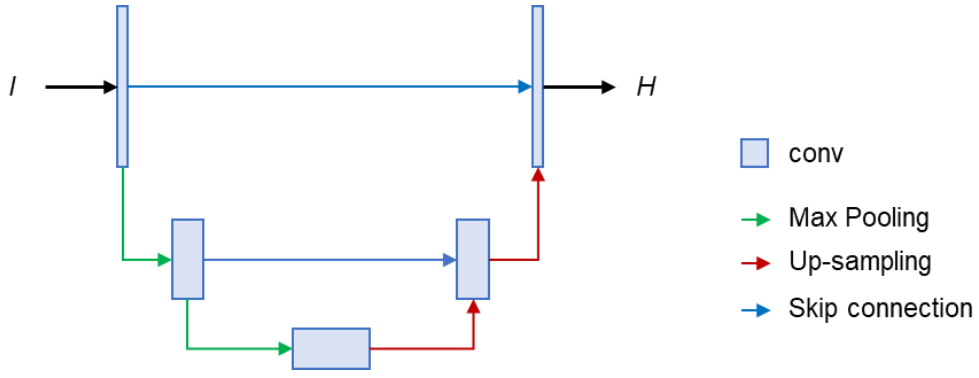


Figure 9. A glimpse of a U-network.

Stiebel *et al.* (2018) first used U-Net in spectral super-resolution (HSUNET), seeking the ability of U-Net to extract spatial information from an image. Compared with the original U-Net, the pooling layer in this work has been removed to avoid the loss of useful information. The approach employed in this study involves utilizing the nearby spatial information as context to reconstruct the spectral

reflectance. During the training phase, the network was fed with image patches that had a size of 32×32. The authors also pointed out that simply adding depth to the network won't increase performance; instead, 5 layers with 128 filters were reported to be ideal. Besides, this method was trained to minimize a combination loss including the colour difference ($\Delta E2000$) between the recovered spectrum and ground-truth after both were converted to RGB colour space. The $\Delta E2000$ was calculated based on an arbitrarily chosen white reflectance assumed to hold for the entire dataset.

Yan *et al.* (2018a) also proposed a multi-scale U-Net-based spectral reconstruction method (MSCNN). Different from Stiebel, the convolution block was followed by a max-pooling layer to avoid overfitting on the encoding path. Since these two networks were trained with the same database, it was interesting to have two opposite views on the pooling layer. A possible explanation could be that the patch size used in Yan's work is 64×64. A bigger patch size could involve local and non-local information during reconstruction, but it also increases the risk of overfitting. This also underscores the lack of interpretability in deep models, leaving researchers with limited insights into how the model functions. Compared to Stiebel's work, Yan's research yields higher error rates in both RMSE and MRAE. A potential explanation for this could be that Stiebel incorporated colour consistency as an additional constraint, which can help improve the accuracy of spectral reconstruction. On the other hand, Yan's approach predominantly leverages spatial information as context.

Banerjee and Palrecha (2020) put forth a real-time spectral super-resolution method built on the U-Net structure, referred to as MXRUNET. In their work, they experimented with several XResnet models (He *et al*., 2019) serving as auto-encoders. A blur layer and a self-attention layer were introduced on the decoder path, which considerably increased the complexity of the model compared to Stiebel's work. However, the reported reconstruction accuracy was lower than that of Stiebel's. Interestingly, the authors noted that during their model testing, the reconstruction accuracy did not correlate with the complexity of the model. This observation hinted at the potential overfitting of the network.

Chen *et al*. (2021) introduced HINET, a U-Net-based architecture that incorporates a Half Instance (HI) Normalization Block for feature extraction during down-sampling. In the HI block, the feature extracted with a convolutional layer is divided into two parts; the first part undergoes instance normalization, while the second part is retained as context. The reason for using instance normalization is to maintain consistent normalization procedures during both training and testing. Different from the original U-Net, residual blocks are added in the up-sampling path (decoder). Initially developed for spatial super-resolution and image enhancement, HINET was later modified for spectral super-resolution by Cai (2022) as a benchmark in their spectral reconstruction toolbox. Compared to other models in the toolbox, HINET achieves moderate reconstruction accuracy.

- **Dense Net**

A dense net that connects each layer with several other layers is designed to achieve richer feature extraction. In a dense block, each layer receives input from all previous layers and passes its features to all following layers. This design enables the network to extract richer and more diverse features from the input image. Additionally, the use of skip connections allows for better gradient flow during training, which can help to reduce the problem of vanishing gradients (Huang *et al.*, 2017). Several spectral super-resolution methods based on dense networks have been presented.

Galliani *et al.* (2017) used a dense network, in their spectral super-resolution algorithm (HSDNET). The network is structured in a way that it first contracts to encode the input data and then expands to decode and recover the spectrum. For each scale level, a dense block is used, which includes four densely connected convolutional layers. In this work, the network takes a 64×64 image patch as input with a multi-scale structure that ensures the network receives extra information from the whole input image patch.

Shi *et al.* (2018) improved HSCNN by replacing the linear up-sampling step with a 1×1 convolutional layer and then added several dense blocks. They named their work HSCNN-D. In HSCNN-D, the network takes 50×50 image patches as input and is trained to minimize the MRAE. Furthermore, the authors also proposed a residual net (HSCNN-R), which adds several residual blocks to the HSCNN. The reconstruction accuracy of the HSCNN-D outperformed the reviewed sequential and residual networks. One possible explanation for this could be that dense blocks in HSCNN-D reuse features directly from all preceding layers, making the model more parameter-efficient. However, compared to more recent models that employ attention blocks, HSCNN-D shows no advantage in terms of reconstruction accuracy. Since the work was open-sourced, HSCNN-D has been frequently used as a comparison group and has been cited by numerous subsequent researchers.

- **Attention Networks**

Attention blocks are modules of neural networks that enable the model to focus on specific parts of the input that are more relevant to a particular task. In an attention network/block, the input is first processed by an encoder, which generates a sequence of hidden states. These hidden states are then used by an attention mechanism to calculate a set of weights that indicate the importance of each input component. The weights are then multiplied by the corresponding components, producing a context that captures the most relevant information in the input. The context is then used as input for the decoder, which generates the output. For spectral super-resolution, attention blocks help the network to better extract spatial-spectral information from local and non-local areas.

Li *et al*. (2020) introduced an innovative spectral super-resolution method based on an Adaptive Weighted Attention Network (AWAN). This method utilizes a sequential network as its backbone, bolstered by several residual attention blocks for intensive deep feature extraction. To avoid vanishing gradients, a global residual connection is also adopted. Each residual attention block consists of a residual module and additional paired convolutional operations with kernels in different scales to extract features in different scales. In addition, a self-channel attention block is placed for more powerful feature correlation learning, the attention block works by adjusting the weights of each feature channel. A unique aspect of this work is the inclusion of colour consistency as an additional training constraint, which is achieved by incorporating an associated loss function. This comprehensive approach allows AWAN to outperform its predecessors in NTIRE 2020.

Zhao *et al.* (2020) proposed a 4-level hierarchical regression network (HRNET), in which, at each level, the spatial size was halved compared with the upper level. A down-sampling and up-sampling strategy (Shi *et al*., 2020), designed for spatial super-resolution, was adopted in HRNET. This hierarchical structure could reduce noise in the recovered spectral image and keep high-frequency information. Each level includes two convolutional layers and several residual dense blocks. Additionally, at the end of each level, a residual global attention block is used to extract information from a pixel that is distant from the target. Skip connections are also added between each level. This network was trained with image patches sized 256×256. Although bigger patch sizes (320×320 and 384×384) were also tested, the results showed that training the network with a larger patch size would not necessarily increase reconstruction accuracy. From the analysis in Chapter 5, we discovered that HRNET focuses on local spatial information even though it was trained on larger image patches. This explains why increasing the size of the training image patch does not further improve reconstruction accuracy. One possible assumption is that the global attention block causes HRNET to focus on local areas. However, due to the 'black box' nature of networks, we can only speculate about their inner workings.

Li *et al.* (2020) proposed an attentional U-shape network (RAAUNET). Different from the U-net introduced in the previous section, the proposed U-net used residual blocks to replace the plain convolutional units. The skip connection was also replaced by spatial attention blocks to focus on more relevant regions. What is interesting in the work is the author first presented a boundary-aware constraint to better recover spectral information around edges. The edge was detected by a Gaussian filter followed by a Prewitt operator. If the reconstruction error around the edge is high, the boundary-aware constraint will guide the network to focus on the edge. However, since the MRAE used in this work could only present an average measurement of the whole image, the benefit from edge awareness became insignificant. The network took 64×64 image patches as inputs when training.

Since existing spectral reconstruction methods treat pixel-wise features almost equally, relevant and irrelevant image parts would receive equivalent focus, which is not efficient. Peng *et al.* (2020) introduced a residual pixel attention network (RPANET) that could focus on different parts of an image patch. The network starts with a convolutional layer and is sequentially followed by several residual spatial attention blocks. The output of each attention block is then fused to another convolutional layer for final reconstruction. A global skip connection is also added between the first and last convolutional layers which skips all attention blocks. The network took 64×64 image patches as inputs when training.

Cai *et al.* (2022) proposed a mask-guided spectral-wise transformer (MST) to recover spectral information from RGB inputs. This work was adjusted to the NTIRE 2022 database as 'MST++' and won first place in that competition (Arad *et al*., 2022). In the proposed network, spectral self-attention blocks were used to extract spectral-wise similarity. The attention was guided by a mask-guided mechanism sensitive to spatial regions with high-fidelity spectral representation, which allowed the network to learn information from long-range pixels. Therefore, the network was trained with image patches sized 256×256. However, as the network begins to use increasingly larger spatial features as context, it becomes essential for the network to select the most relevant features from the image scene. Using the entire patch as a reference could lead to overfitting. A more effective approach would be to use a mask that guides the network's focus, which could explain the superior performance of MST++. When developing our model, it's also crucial to select the most relevant features from the image to serve as a context for resolving the metamerism issue.

The above attention blocks focus either on spatial or spectral information. Hu *et al.* (2022) proposed a dual-domain attention network (HDNET) that used two mechanisms to focus on both. The extracted spectral and spatial information were connected with a feature fusion block with several depth wise-separable convolution layers. Additionally, this network considered spatial frequency domain similarity to better recover spectral details, which transforms the image into the Fourier domain to reconstruct more high-frequency detail. HDNET was trained on image patches sized 256×256. However, HDNET doesn't outperform MST++ or MIRNET. The lack of interpretability makes it difficult to explain why HDNET doesn't have an advantage, even with its dual attention mechanism.

A common point that the above attention networks share is that they are all interested in non-local spatial information that is away from the target position. As a result, these networks require larger image patches as input, and some of the networks have benefited from this. However, currently, there is no evidence to suggest that the non-local spatial information is correlated with the spectral features of the target position. Therefore, there is a potential risk that these networks may be overfitting. Before trusting the network, it is necessary to analyse the role local and non-local spatial information plays in the spectral super-resolution process.

- **Multi-branch Networks**

Different from all the above networks, which only have a single stream, the multi-branch network has multiple independent branches. The aim of using multiple branches is to better extract different features from inputs. After that, the extracted features will be fused for the final spectral reconstruction.

Nathan *et al.* (2020) presented a dual-branch reconstruction network that harnesses the power of both dense blocks and multi-scale hierarchical feature extraction. The process begins with the extraction of spatial information through a convolutional block, as described by Liu *et al*. (2018). The extracted spatial information is then separated into two equally important feature extraction branches. One of these branches employs a dense block for the processing of spatial data, while the other leverages a multi-scale hierarchical feature extraction block to capture and learn features at multiple levels of detail. Despite their different approaches, both branches aim to generate image patches of the same size. Ultimately, the image patches reconstructed by both branches are merged, resulting in a recovered patch that encapsulates the strengths of each branch. The network is designed to work with image patches of size 64×64.

Lei *et al.* (2020) noticed that pixels located in different positions of an image require different-sized fields to do the reconstruction, and this is also true for pixels that correspond to different materials. Therefore, the author proposed a network with several parallel blocks (PFMNET). For each block, features with different scales are extracted. The reconstructed image patch from each block is fused to provide the recovered spectra. This network took image patches sized 64×64 as input.

To avoid the influence of overlapped camera spectral functions, He *et al.* (2021) took the camera spectral function as guidance to group the spectral bands (SRFNET). For example, the spectral bands for the CAVE dataset could be separated into three groups which correspond to blue, green-blue, and green-red, respectively. Then the extracted features would be forwarded to three independent mechanisms, each including a channel attention module, spatial-spectral module, and feature extraction module. After that, a channel attention module with two pooling layers was added before the outputs from each group were fused to generate the final reconstruction result. This network was trained on image patches sized 128×128. Unlike other deep methods, which are trained on minimizing the MRAE or RMSE, this work used SAM as the loss function.

Despite their complexity, multi-branch networks do not appear to offer any advantages in reconstruction results compared to previously reviewed models. Given their computational demands and design intricacies, these networks may not be well-suited for tasks related to spectral super-resolution. Additionally, multi-branch architectures risk extracting redundant features and further complicate the interpretability of the model.

- **Generative Adversarial Networks (GANs)**

GANs consist of two parts: a generator and a discriminator. The role of the generator is to produce data that closely resembles real data, with the objective of causing the discriminator to incorrectly classify its outputs as real. The discriminator's job is to differentiate between real and fake (generated) data. Given a set of data, the discriminator must identify which samples are real (from the dataset) and which are fake (produced by the generator). The generator starts by producing data that looks nothing like what it's supposed to. But over time, as it receives feedback from the discriminator, it gradually gets better at generating data that closely resembles the real dataset. For spectral reconstruction, the generator, which is a convolutional neural network, recovers the spectra from RGB input, while the discriminator judges whether the form of the generated spectra is distinguishable from the form of the original spectra. These two parts compete against each other until the recovered spectra are similar enough to the original spectra.

Alvarez-Gila *et al.* (2017) proposed a GAN-based spectral reconstruction method. In the proposed work, a U-Net that measures local spatial context was used as the generator. While the loss function was replaced by the PatchGan (Pathak *et al.*, 2016) based discriminator. This network took image patch sizes of 256×256 as input. Then the entire image would be reconstructed with non-overlapping patches.

Liu and Zhao (2020) further improved the GAN-based method, by adding residual and scale blocks (Li *et al*., 2018) for feature extraction. Liu *et al.* (2022) introduced a conditional GAN-based spectral information reconstruction network for textile RGB images. However, like all GAN-based methods, the listed network faces the same limitations, these include the fact that GANs are difficult to train. Additionally, the interpretability of these models poses a huge challenge. The complexity of GAN architectures and the intricate interplay between the generator and discriminator layers often obscure the underlying reasons for specific outputs. Given the impressive results GANs have demonstrated in spatial super-resolution, it's reasonable to expect that they might also excel in spectral super-resolution, even though only a few studies have been published on this topic. However, when it comes to interpretability, GANs present even greater challenges.

2.5.3. Summary

The previous section provided a detailed review of the state-of-the-art deep convolutional neural network-based methods for single-image spectral super-resolution. In this section, we aim to summarize the listed works based on four key aspects to provide a more holistic understanding of the current progress in this field.

- **Deep vs. Traditional Methods**

Compared with the traditional methods introduced in Section 2.2, deep convolutional neural network-based methods exhibit higher reconstruction accuracy. Apart from the general features of deep methods introduced earlier in this section, the superiority of deep methods can be explained by the following. Firstly, the number of parameters in deep methods is much higher than that in dictionary-based methods. Deep methods can easily have over 1 million parameters, while the over-complex dictionary in Arad's work only has 1,000 atoms. The complex network structure enables deep methods to represent complex, nonlinear patterns in spectral data. Secondly, deep methods are trained on larger datasets, whereas dictionary-based methods are trained on selected spectral data. Thirdly, all the listed deep methods use spatial information as additional information to reduce ambiguity when reconstructing spectral information, which is rare in dictionary-based methods.

- **Training Strategy**

All of the networks listed here are examples of supervised learning, which involves end-to-end learning of the mapping from RGB to the spectrum. Although there has been a recent increase in the availability of spectral reflectance data, it remains less accessible compared to RGB image data. Thus, adapting unsupervised learning techniques for single-image spectral super-resolution could allow trained models to handle more general tasks. Fubara *et al*. (2020) incorporated an unsupervised approach into their work, utilizing colour consistency as a loss function. However, the authors made an assumption that maintaining colour consistency implies that the reconstructed spectrum is nearly identical to the original spectrum. This assumption is not necessarily true due to metamerism, where different spectral power distributions can produce the same perceived colour. This suggests that discovering an effective strategy for applying unsupervised learning to spectral super-resolution could be a valuable research direction.

- **Network Architecture**

In chronological order, the network structure used in single image spectral super-resolution which started with a sequential structure becomes more and more complex. There are also mixtures in the listed structures. The U-net structure enables the network to effectively extract information from a larger area. The residual and dense block can alleviate the vanishing of gradients, as a result, the network could better learn the mapping from RGB to spectrum. The attention networks with relatively complex structures achieve the best accuracy at this stage, for example, the MST++ won the NTIRE 2022 competition. Attention mechanisms help the network to focus on important spatial and spectral features.

As the network structures for spectral super-resolution become more intricate, there is a noticeable trend towards involving a larger spatial area in the network. In contrast to earlier models that only captured local spatial information, recent models can extract spatial information from almost the

entire image. However, a challenge remains in focusing on the relevant information from such a large area, which has led to the incorporation of attention mechanisms. Nonetheless, there is currently no consensus on the optimal spatial area size required for this task.

- **Error Measurements and Loss Function**

The commonly used loss functions for existing deep methods include RMSE, MRAE & SAM. The authors chose loss functions with different purposes. However, none of the listed loss functions could describe spectral and spatial accuracy at the same time. In addition, the average measurement of the whole image does not provide a good indication of performance in evaluating the reconstruction accuracy on specific targets. Currently, no error measurement could directly measure the performance facing the metamerism problem. As a result, there is no report on how the existing deep methods specifically perform on metamer data. Therefore, in Chapter 5, we compare existing deep methods for reconstructing selected metameric samples.

2.5.4. Limitations of CNN-Based Methods

Although deep neural network-based methods can provide some encouraging results, they still face the following challenges (Arad *et al*., 2020).

- **Colour Consistency of Results**

The hyperspectral and RGB images are physically related. Therefore, for accurate reconstruction, the recovered spectra should be able to be converted into the input RGB under the same conditions (illumination and camera sensitivity). However, because the deep method commonly learns end-to-end mapping from RGB to spectra without considering the camera function, the recovered spectral reflectance cannot be converted to the input RGB even under the same conditions. Colour consistency can be used as a constraint to increase the reconstruction accuracy. Considering this, Lin and Finlayson (2020) improved HCSNN+ by adding a constraint that ensures the recovered spectral can be converted into the input RGB. Kaya *et al.* (2019) used an additional block to estimate the camera function within the network stream. Deep learning models, particularly those used for end-to-end mapping from RGB to spectrum, have often overlooked constraints such as the camera function. However, it's worth noting that knowing the camera function is not an unreasonable assumption, as this information can typically be obtained from the camera's datasheet or through measurements.

- **Dependence on Spatial Information**

All the above deep methods rely on spatial information, and when the structure of the network becomes complex, it requires a larger image patch as input. From one aspect, it could be understood

as the network having the ability to extract information from a very huge area. However, on the opposite aspect, the relation between a long-range pixel and the spectral of the target remains unknown. There is a risk that the network is overfitting when it requires information from the whole image. This dependence on spatial information also limits the performance of the networks on single-pixel reconstruction.

- **Lack of Discussion on Resolving Metamerism**

Metamerism is the fundamental challenge in computational spectral reconstruction. In the absence of metamerism, there would be a direct and unique mapping from RGB to spectrum. However, due to the limited spectral resolution of RGB sensors in the real world, the mapping from RGB to spectrum is ambiguous. The ability of deep learning methods to model complex non-linear datasets has made them effective in improving the accuracy of spectral reconstruction. However, most studies focus on learning end-to-end mapping from RGB to spectrum without addressing the issue of metamerism within the network. As a result, few studies have discussed the solution of metamerism. Metamerism is a big challenge that hinders spectral super-resolution in real-world applications, as it can significantly affect the accuracy of classification, which is one of the primary applications of hyperspectral imaging.

- **Black Box Problem**

Like all deep neural network-based methods, the deep spectral super-resolution methods also suffer from the black box problem. This refers to the challenge of understanding how these models arrive at their predictions. While the architecture and parameters of the network are known, it can be difficult to follow how the input features are being processed and combined to generate the output.

This means the proposed method lacks interpretability. The complex network does not bring an extra understanding of the problem; the improved accuracy benefits from increased network complexity making the understanding more complex. How to effectively use the spatial information from patches remains unknown. Even though deep neural network-based methods rely on patches, questions like what kind of features are actually used within a patch and how can those features be used more efficiently still wait to be answered.

Compared with deep neural network-based methods, prior-based or shallow network-based methods often exhibit lower accuracy. Influenced by the NTIRE competition, the research community has increasingly prioritized the performance of proposed models over their interpretability. However, while deep neural network-based methods receive more research attention, this does not mean that prior-based methods have no potential for improvement. Currently, the most common metric for evaluating algorithms is the MRAE, which is the basis for many deep methods. In contrast, regression-based methods are trained to optimize a generic RMSE and are then tested using MRAE.

Lin *et. al.* (2021) proposed a regression-based spectral super-resolution method that was trained on MRAE and achieved results comparable to those of DNN-based methods. Moreover, prior-based methods are more interpretable than deep methods, which allows for a greater understanding of the problem at hand. Some researchers have attempted to combine the ideas of these two methods. Fubara *et al.*, (2020) proposed a method that separates end-to-end mapping from RGB to spectral reflectance into two networks, one for learning basis functions and another for learning the weights.

2.6. Research Targets

In this chapter, we have highlighted the limitations of both existing traditional prior-based methods and deep neural network approaches to single image spectral super-resolution. Dictionary-based methods are constrained by their dictionary size, which, compared to deep models, gives them less capacity to learn the mapping from RGB to spectrum. Particularly in pixel-wise methods, there's an inability to resolve metamerism, given their lack of contextual information to clear ambiguities in metamerism samples. Consequently, dictionary-based methods reported thus far exhibit lower reconstruction accuracy than deep models. However, deep models are not readily explainable. Additionally, few existing methods have factored in colour consistency as a constraint. Therefore, to overcome these limitations in state-of-the-art approaches, we propose the following research objectives.

The primary **goal** of this research is to develop an interpretable spectral super-resolution method based on data prior. We aim to recover spectral information from RGB input by training models using spectral data prior. Unlike existing opaque deep methods, the proposed model will prioritize interpretability, with a focus on creating a method that is both accurate and comprehensible.

The proposed spectral super-resolution model should satisfy the following criteria:

- **Colour Consistency**

The proposed model should ensure the recovered spectrum produces the same RGB value under identical conditions (for the given camera model and exposure).

- **Ability to Resolve Metamerism**

This study will explore what kind of information can resolve the ambiguity associated with metamerism in spectral super-resolution. While other researchers have used deep neural networks to tackle the one-to-many mapping caused by metamerism in reconstructing the spectrum, no discussion has been provided on how existing methods perform with metamerism data. We also aim to compare existing deep methods in reconstructing metamerism samples.

- **Explainable**

Compared to existing deep methods, the proposed model should be more explainable. This entails a less complex and low-dimensional model. Furthermore, the steps within the proposed model should be clear, necessitating an explicit understanding of the features used. Therefore, this study will investigate how spatial information—referring to pixel position and arrangement in an image—can be integrated into the process of reconstructing spectral information.

Chapter 3. RGBPQR Colour Space

In the preceding section, we reviewed the state-of-the-art for single image spectral super-resolution. For traditional prior-based techniques, accurately modelling spectral reflectance is a critical step towards creating an explainable spectral super-resolution method. As previously noted, existing models often fail to maintain colour consistency—a discrepancy where the reconstructed spectrum and the original spectrum appear with different colours. Ensuring colour consistency could serve as an effective constraint when learning the mapping from RGB to spectrum. To address this issue, we have utilised a modification of the LabPQR colour space as our spectral reconstruction model. This is the first time such a model has been applied to single image spectral super-resolution. The advantages of employing LabPQR are elaborated upon in the sections that follow.

3.1. LabPQR Colour Space

As mentioned in Chapter 1, hyperspectral data could be accurately represented in a few dimensions. Colour science research has long used an interim connection space to represent the most important features of multi- or hyperspectral pixels in only a few dimensions. Derhak and Rosen (2006) introduced a six-dimensional interim connection space called LabPQR to represent high-dimensional spectral data. The first three dimensions are CIELAB colourimetry values under a specific viewing condition connected to linear XYZ tristimulus values, and the last three dimensions (PQR) are spectral reconstruction dimensions representing the difference between the original spectrum and the spectrum reconstructed from the colourimetric value (also referred to metameric black).

Spectral reflectance S can be represented by the LabPQR colour space as:

$$S = TN_c + VN_p \qquad 3\text{-}1$$

T is an n by 3 reconstruction matrix that up-samples the XYZ colourimetric tristimulus vector, N_c to the n wavelength samples of the spectrum. V is an n by 3 matrix of the PQR spectral bases.

Similarly, N_p are the PQR weights (coefficients). The transformation matrix T is determined from a data prior by (assuming sufficient samples are provided so that $N_c N_c^T$ is invertible) using least squares analysis.

$$T = SN_c^T(N_c N_c^T)^{-1} \quad 3\text{-}2$$

where S is the measured spectral reflectance from a prior spectral database. This determines T to minimise the square error of the reconstruction while maintaining colour consistency. The residual between the up-sampled spectrum and the original is then obtained by:

$$E = S - TN_c \quad 3\text{-}3$$

The PQR bases, V, are derived from E using principal component analysis (PCA). To keep the most significant spectral feature with a few dimensions, only the first three eigenvectors are preserved as the PQR bases:

$$V = (v_1, v_2, v_3) \quad 3\text{-}4$$

V is orthogonal to the CIE XYZ colour-matching function and therefore does not affect the represented spectrum's colour. T and V in the LabPQR space are both derived from prior data. As the LabPQR colour space relies on the CIE XYZ colour-matching functions, the resulting colour is independent of the device (camera) used. LabPQR colour space has been used in applications like spectral gamut mapping (Tsutsumi *et al.*, 2007), and spectral colour reproduction (Wu *et al.*, 2014).

In Lin's (2020) work, spectral data is represented in a similar way to LabPQR which also includes two parts. However, the key difference lies in the reconstruction function used to up-sample RGB values into a spectrum. Lin utilized the camera model to derive the reconstruction function, without optimizing it with any prior data. As a result, the reconstruction function can be unsmooth, leading to up-sampled spectra with high errors. To estimate the residual between the up-sampled spectra and the original spectra, Lin retrained the HSCNN+ using the residual as the response.

3.2. RGBPQR Colour Space

The colour image in single image spectral super-resolution is usually in device-dependent RGB. To suit the spectral super-resolution tasks, we have adapted the LabPQR colour space by replacing the CIEXYZ colour-matching function with the RGB camera model. We name this new spectral model RGBPQR colour space. The RGBPQR colour space follows the same concept as the LabPQR colour space with a detailed derivation presented in the rest of this section. Note that it is

assumed that the camera RGB spectral sensitivities are known. This will also be the case for converting RGB to device independent CIEXYZ.

From Chapter 1, the RGB value of a given camera can be modelled as:

$$i_k = \int c_k(\lambda) l(\lambda) s(\lambda) d\lambda \qquad 3\text{-}5$$

and can be represented in a matrix form as:

$$\mathrm{P} = CLS \qquad 3\text{-}6$$

where P are the sensed RGB values, C is the camera spectral response's function, L is a diagonal matrix that represents the illumination and S represents the spectral reflectances. The combined LS represents the spectral radiance (R), enabling 3-6 to be simplified to:

$$\mathrm{P} = CR \qquad 3\text{-}7$$

Similar to LabPQR, the RGBPQR colour space aims to have the reconstruction function T to minimise the error between the up-sampled spectral radiance $\hat{R}(= TP)$ and the original radiance R in the least squares manner:

$$T = argmin\|R - T\mathrm{P}\|^2 \qquad 3\text{-}8$$

Maintaining colour consistency requires the reconstructed spectra to produce the original RGB values (i.e., $C\hat{R} = CR$), therefore, the reconstruction function T needs to satisfy the following constrains:

$$CT = I \qquad 3\text{-}9$$

where I is the 3×3 identity matrix. From this, the transformation (up-sampling function) T can be derived from the pseudo inverse of P:

$$T = R\mathrm{P}^T(\mathrm{P}\mathrm{P}^T)^{-1} \qquad 3\text{-}10$$

The up-sampled spectra are denoted as $\hat{R} = T\mathrm{P}$, any residuals between the two spectra, can be expressed as a set of spectral differences:

$$E = R - T\mathrm{P} \qquad 3\text{-}11$$

For the residual to represent metameric black, this requires $CE = 0$:

$$C(R - T\mathrm{P}) = CR - CT\mathrm{P} = (I - CT)\mathrm{P} = 0 \quad \text{3-12}$$

The residual is represented by bases V derived from the principal component analysis (PCA) of E, where N_p is the weight for each basis and μ is the mean of the residual E. With all eigenvectors, the spectral reflectance can then be losslessly represented as:

$$R = T\mathrm{P} + VN_p + \mu \quad \text{3-13}$$

The RGBPQR colour space uses the first three eigenvectors of the residual as PQR bases. Employing this equation as the spectral data model transforms the spectral reconstruction task into a problem of estimating the PQR coefficients, N_p.

To determine how many eigenvectors are sufficient to accurately represent the spectral data, we will measure the representation error of spectral data with different numbers of eigenvectors. For i eigenvectors, the spectral data is approximated as:

$$\hat{R} = T\mathrm{P} + V^{(i)}N_p^{(i)} + \mu \quad \text{3-14}$$

we explore the model's capacity to represent hyperspectral data in limited dimensions.

3.2.1. Experiment Data

The largest hyperspectral database to date is the NTIRE 2022 (Arad *et al*., 2022), which forms the basis of this experiment. The published dataset has a spatial size of 482×512 pixels and includes corresponding RGB images. The RGB values were generated using known camera sensitivities (shown in Figure 10) but were scaled arbitrarily to simulate different exposure levels. Additionally, random noise was added to the RGB images to replicate real-world applications. However, for the specific focus of this experiment, we simplified the process of generating RGB values. We used the same camera functions but in contrast to the NTIRE RGB values, the RGB values in this experiment were normalized by the global maximum, and no extra noise was added. This corresponds to the assumption that all RGB values are taken with the same exposure. The converted RGB value was quantized into 8 bits (0 to 255). This normalization approach was adopted in NTIRE 2018 and 2020 for the clean training track. Using a global maximum to normalize RGB values removes the need to account for variance in exposure, allowing a concentrated study on metamerism without the added complexity of additional parameters.

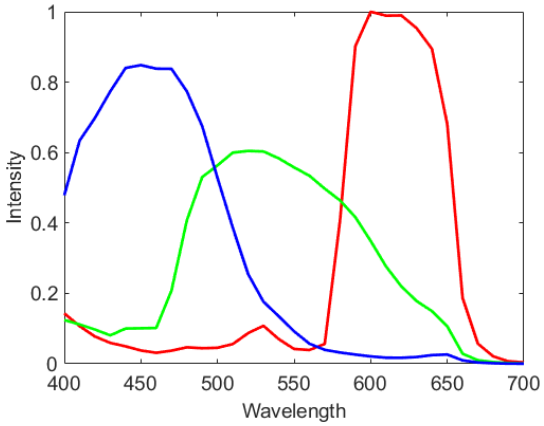


Figure 10. Camera response functions of the RGB camera sensor from NTIRE 2022.

Randomly collecting RGB samples from the database may introduce bias based on the proportion of RGB values in the database. Consequently, RGB values that are rare in the database may have a lower likelihood of being collected. To make the collected samples represent the whole dataset as much as possible, 1,000 spectral samples and their corresponding RGB samples were chosen from each image. Rather than select the sample randomly, samples were collected to represent the gamut of each image. To achieve this, the samples were first converted into the Lab colour space, which is known for its ability to detect subtle colour differences.

Figure 11 illustrates an example image in both RGB and a^*b^* space. Pixels from each image were then grouped into 20 clusters by K-means clustering as illustrated in Figure 12 utilizing the Euclidean distance metric in the a^*b^* space. 50 samples were randomly selected from each cluster to give 1,000 representative samples for each image. In total, 900 spectral images from the NTIRE 2022 database were used as prior data, yielding a total of 900,000 training samples in this experiment.

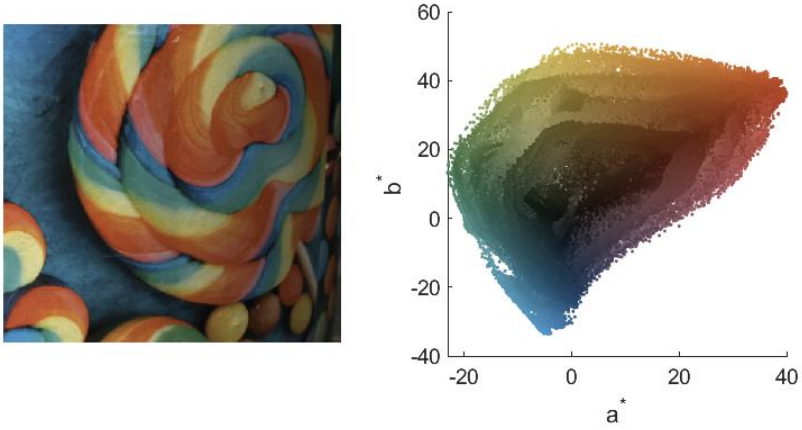


Figure 11. Image in RGB and the samples in a^*b^* space colour mapped by the corresponding RGB values.

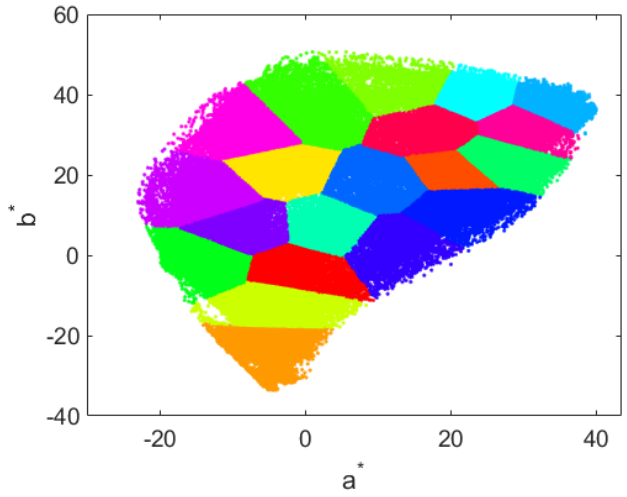


Figure 12. Clusters in a^*b^* space, 50 samples are randomly collected from each cluster. The clusters would guarantee the collected samples cover the gamut of the image.

Figure 13 compares the gamut of the collected dataset using clustering with the gamut of 1,000 randomly collected samples from each of the same 900 images. It is evident that using clustering for sample selection gives a broader gamut coverage and provides a more comprehensive representation of the entire database. It's important to note that our sample collection approach is designed to capture a wider representation of the dataset, with the primary aim of evaluating the RGBPQR model's capability in representing spectral data with metamerism. Consequently, when compared to a randomly collected dataset, the proposed data collection method might provide a less accurate representation of the proportion of the database, potentially resulting in a slightly higher average reconstruction error.

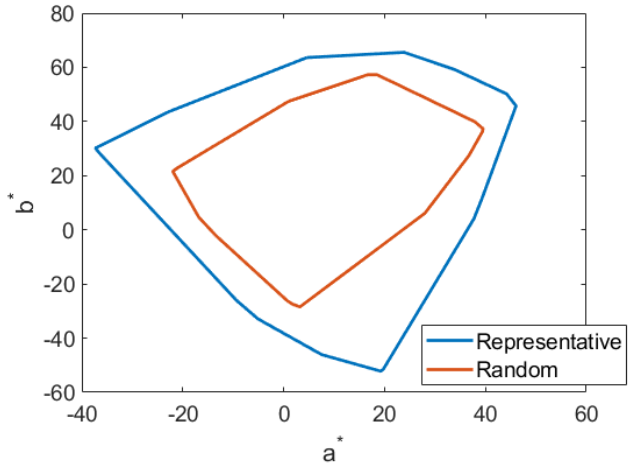


Figure 13. The gamut formed from the convex hull of collected datasets in Lab space. The representative dataset is sampled using clustering, whereas the random dataset is randomly collected.

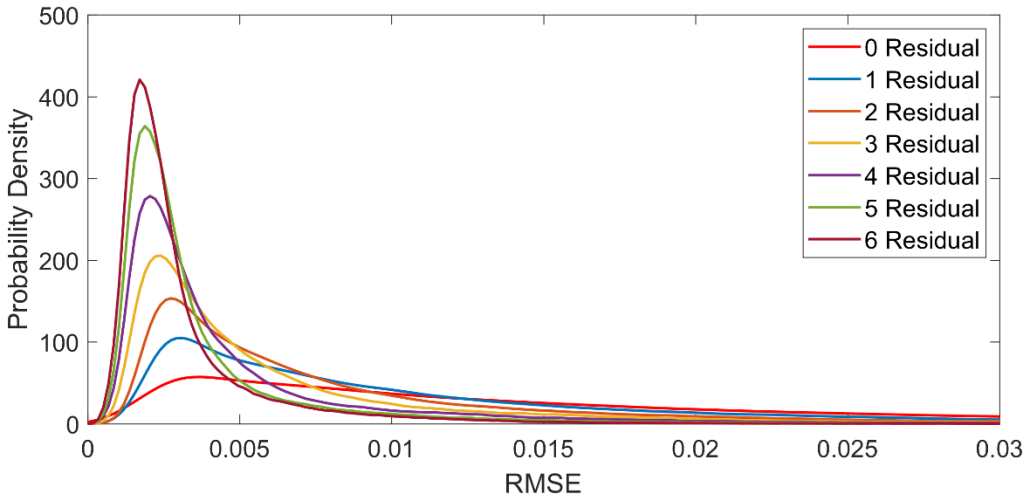
3.2.2. Representing Spectral Data

In this section, the effectiveness of utilizing the RGBPQR colour space for representing the spectral reflectance data is evaluated.

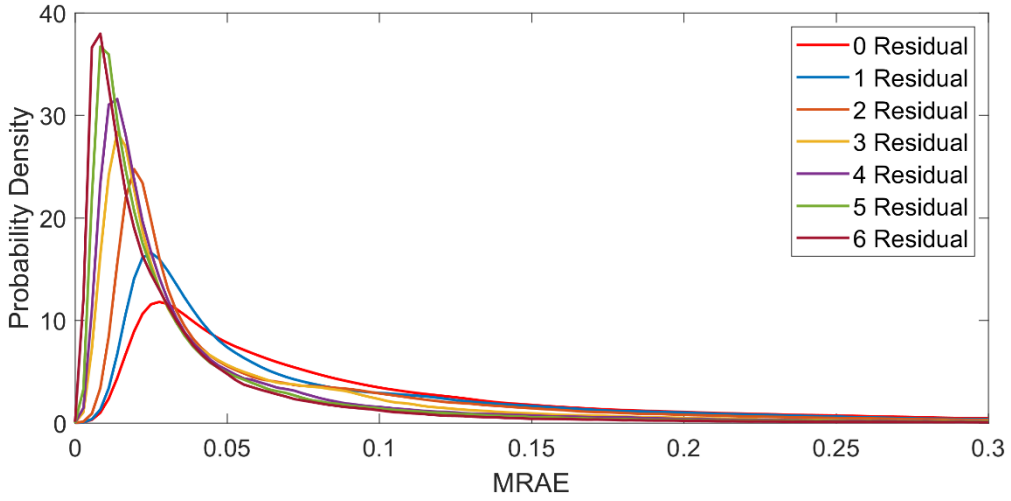
Table 4 compares the accuracy achieved when representing spectral samples using a specified number of components. Notably, the RGBPQR model demonstrates similar representation accuracy to PCA in terms of RMSE. In the RGBPQR model, the RGB values are used as the first three components. Fixing the first three components to the RGB terms does not appear to have a significant impact on representation accuracy. The P-value between the error distribution of adding the 2nd and 3rd is 0.08, indicating a significant difference, while the P-value between adding the 4th and 5th is less than 0.05. This suggests that adding high-order components has less influence on the representation accuracy. However, when using six dimensions to represent spectral data, RGBPQR exhibits slightly higher values for SAM and MRAE compared to PCA. This difference can be attributed to the fact that both RGBPQR and PCA optimize RMSE, leading to some variations in other accuracy metrics.

Table 4. Representation accuracy of spectral data with a different number of components. The highlighted column compares RGBPQR with 3 residual components (6 in total) with 6 PCA components.

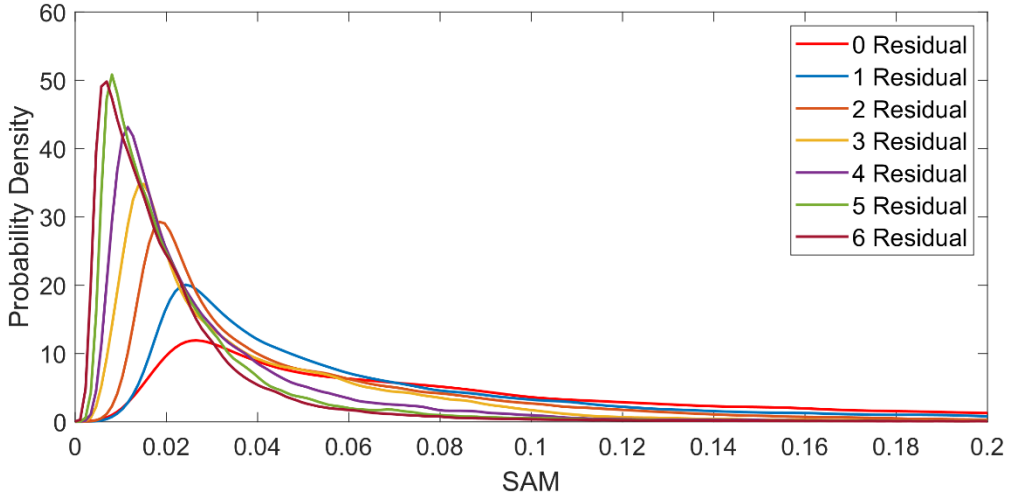
Total Components		1	2	3	4	5	6	7	8	9
RGBPQR Mean Error	RMSE			0.032	0.017	0.012	0.009	0.007	0.006	0.004
	MRAE			0.139	0.118	0.091	0.064	0.055	0.046	0.038
	SAM			0.109	0.069	0.053	0.045	0.031	0.025	0.026
PCA Mean Error	RMSE	0.068	0.038	0.022	0.017	0.012	0.009	0.007	0.005	0.004
	MRAE	0.546	0.276	0.135	0.122	0.085	0.057	0.047	0.034	0.025
	SAM	0.277	0.145	0.094	0.071	0.053	0.037	0.024	0.018	0.014
RGBPQR 95% Error	RMSE			0.068	0.035	0.026	0.020	0.016	0.011	0.009
	MRAE			0.049	0.399	0.303	0.213	0.192	0.157	0.127
	SAM			0.318	0.194	0.143	0.101	0.085	0.069	0.060
PCA 95% Error	RMSE	0.151	0.085	0.047	0.034	0.025	0.019	0.015	0.011	0.009
	MRAE	1.811	1.056	0.461	0.428	0.286	0.193	0.170	0.123	0.089
	SAM	0.647	0.375	0.257	0.197	0.146	0.095	0.071	0.050	0.040



a. Error distribution of RMSE



b. Error distribution of MRAE



c. Error distribution of SAM

Figure 14. Error distribution when reconstructing spectral data with a different number of residual components.

Figure 14 presents the error distributions from representing spectral data with various numbers of residual bases, utilizing three different error measurements. Considering both Table 4 and Figure 14, we can see that the accuracy of the represented spectra improves significantly with the addition of the first three residual components. Adding the first three residual components yields a significantly different error profile, with more samples demonstrating lower error rates. The inclusion of the 4th to 6th components further reduces the error, as evidenced in Table 4. However, relative to the addition of the first three components, the impact on error distribution is successively smaller.

Figure 15 displays the up-sampling function derived from the training data, and the first six eigenvectors obtained from the residual, along with the total variance explained by the corresponding component.

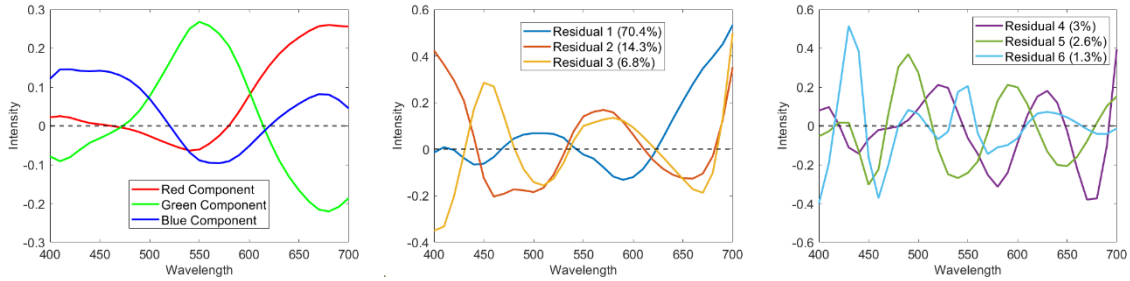


Figure 15. The up-sampling functions and the first six eigenvectors are derived from the residual with the total variance of the residuals explained by that eigenvector in parentheses.

The first three residual components appear to be relatively smooth compared with the 4th to 6th components. The first component displays a strong intensity in the red range, while the second and third components exhibit similarities in the green-to-red range but show opposing characteristics in the blue range. The high intensity of residual components at both ends is due to the low sensitivity of the camera function in the corresponding range, resulting in a relatively high residual. The 4th to 6th components illustrate detailed spectral features that appear unsmooth with successively higher frequency variation. The first three components can account for over 91.2% of the total variance of the residual. In terms of spectral representation, when using the first three eigenvectors to represent the residual, the accuracy of the RGBPQR colour space is equivalent to a 6-dimensional PCA basis model. Therefore, in this study, we assume that using three eigenvectors to represent the residual (metameric black) is sufficient to achieve high accuracy.

Figure 16 illustrates a randomly selected example of spectral reflectance depicted with various numbers of residual bases. When using more than three bases, the difference between the shape of represented spectra is relatively small, the first three residual components could represent most of the spectral detail of the example.

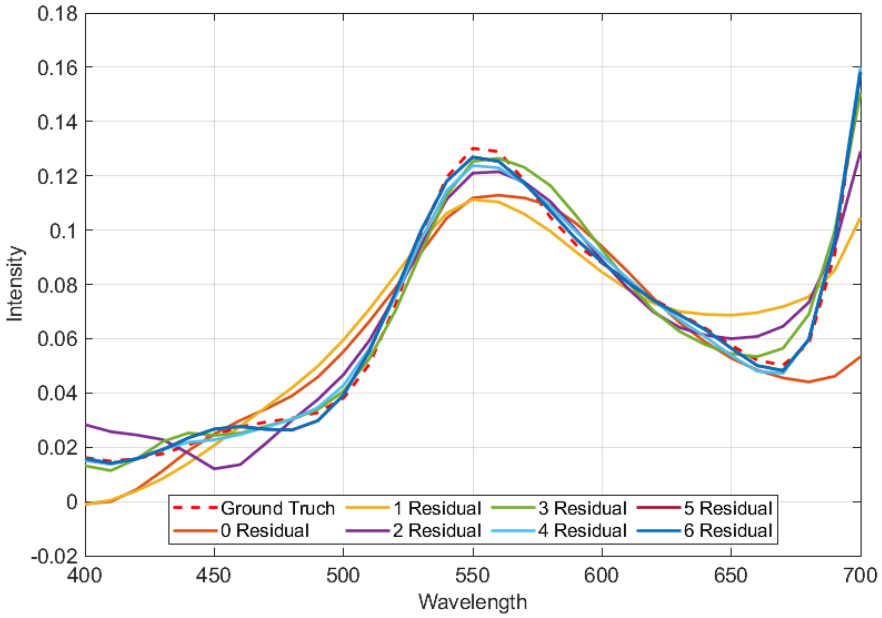


Figure 16. Example spectrum represented with different numbers of residual basis.

To summarize, the RGBPQR colour space is capable of representing spectral reflectance data in a low-dimensional manner. In the following section, we will explore if this model is better suited for spectral reconstruction tasks than other linear spectral reflectance models.

3.2.3. Advantages of the RGBPQR Colour Space

- **Fewer Components Require Estimation**

The first 3 components are given directly from the input RGB values. Only the residual components (PQR) need to be estimated to give the spectrum. For the spectral super-resolution problem, reconstructing spectral reflectance becomes a problem of estimating the coefficients of the residual (PQR) components. The rest of this thesis discusses how to effectively estimate the PQR coefficients.

- **Colour Consistency**

In comparison to other linear spectral models discussed in Chapter 1, RGBPQR inherently adheres to colour consistency as a constraint. Colour consistency or RGB consistency is defined as the ability of the reconstructed spectrum to give the original RGB values using the known camera response functions. Compared with other spectral reflectance models, the RGBPQR colour space is colour-consistent by design. The RGB components up-sample the base spectrum and the PQR

components are orthogonal to the camera response functions, so will not affect the reconstructed RGB values (metameric black). Therefore, when reconstructing with RGBPQR, there is no requirement to include an additional constraint for colour consistency.

- **Suitable to Resolve the Metamerism Problem**

The RGBPQR colour space is a representation of spectral data that has two parts. The first part, RGB, gives the solution from possible metamerism spectra that minimises the least squares error. The second part, PQR, represents a metameric black that recovers the residual between the least squares solution and the original spectrum. To reconstruct a spectrum, one must select the most probable spectrum from a metamer by deciding the weights for the residual basis (PQR). When three residual bases are used, the RGBPQR colour space simplifies the metamerism problem by estimating a 3-dimensional weight vector, which is much more straightforward than other spectral models.

3.4. Reconstruction with no Residual

Before introducing the PQR coefficients estimation, we first look at the reconstruction accuracy with the up-sampled spectrum from RGB value without the residual. In the experiment, the RGB image is generated in the same manner as the RGB value in prior data. In addition, we compare this approach with Arad's sparse dictionary method (Arad & Ben-Shahar, 2016). Arad's model was created using code published by the author and was learned from the same number of samples from the training images. Arad's dictionary comprised 1000 bases after training, and the reconstructed spectrum was represented by 28 atoms in the dictionary. The camera function that transformed spectral data into RGB was identical for both models, while Arad’s original model used CIE XYZ colour-matching functions as the camera model.

To assess the effectiveness of the trained model, a separate testing dataset with 50 images from the NTIRE 2022 database was employed. Using the RGBPQR model, with no residual PQR weights, the complete spectral image was reconstructed from the RGB inputs. Table 5 compares the accuracy of the RGBPQR model with Arad's method, revealing that the RGBPQR model achieved higher accuracy even without estimating the residual weights. Note that, the error reported in Table 5 is lower than that reported in the NTIRE 2022 competition since the RGB images used in the competition were affected by noise and were generated with different exposure settings. Therefore, it is not appropriate to compare the results from Table 5 with those of the NTIRE competition.

Table 5. The reconstruction accuracy of the up-sampled spectrum on the NTIRE 2022 database.

	Arad	Up-sampled
RMSE	0.064	0.027
MRAE	0.224	0.132
SAM	0.272	0.123

Figure 17 compares the mean relative absolute error (MRAE) of the reconstructed spectrum in the 450 nm, 550 nm, and 650 nm channels. The results demonstrate that, in most cases, the RGBPQR model outperformed Arad's method in terms of the recovery spectrum. However, a few pixel areas with relatively high errors were compensated by areas with low errors, resulting in an overall low average error in the table (More results can be found in Appendix 1).

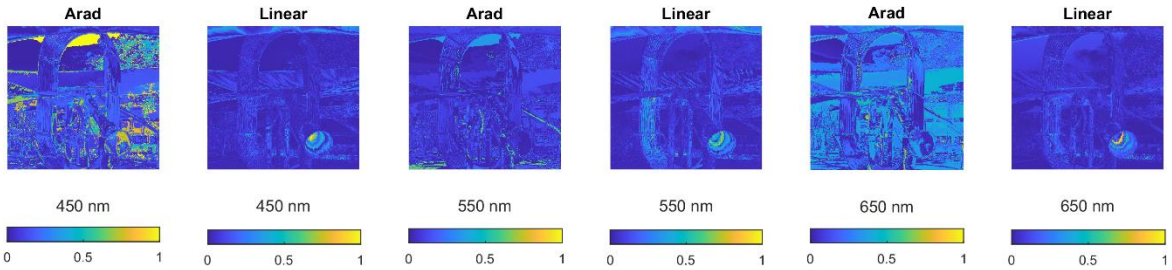


Figure 17. MRAE of the reconstructed spectrum in 450, 550 and 650 nm of the NTIRE 2022, in most cases, the linearly up-sample spectrum appears with less error.

The superior performance of the RGBPQR model over Arad's work can be attributed to the fact that the RGBPQR model assumes knowledge of the camera's spectral responses. However, this assumption is not a significant one as the camera spectral function can be obtained from the datasheet published by the manufacturer. Therefore, the knowledge of camera spectral function does not pose a hindrance to the practical application of the RGBPQR model in real-world problems. To further increase the reconstruction accuracy, it is necessary to recover the residual between the up-sampled spectrum and the original spectrum by estimating the PQR coefficients.

3.5. Shannon Entropy of the PQR Coefficients

Up-sampling from RGB minimises the global error using a linear mapping. However, in general, the mapping will not be linear, so part of the residual would reflect this non-linear relationship. i.e., the residual (PQR) may have a local correlation with measured RGB values. Therefore, we are interested in whether is it possible to estimate the PQR weights from the corresponding RGB value. To answer this question, we look at the Shannon entropy, which is a measure of the amount of uncertainty or randomness in a given set of data.

The Shannon entropy of a discrete random variable X with a probability mass function $p(x)$ is defined as:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \qquad 3\text{-}15$$

The unit of Shannon entropy is bits, and it represents the minimum number of bits needed to represent the information contained in the random variable X. Intuitively, entropy measures the amount of uncertainty or surprise associated with the outcome of a random variable. If the probability distribution is highly skewed and concentrated on a few outcomes, the entropy is low, indicating that there is little uncertainty. Conversely, if the probability distribution is more evenly spread out, the entropy is higher, indicating that there is more uncertainty.

In the RGBPQR model, the original spectral radiance R can be approximately represented $\hat{R}$ as the following equation:

$$\hat{R} = T\mathrm{P} + VN_p + \mu \qquad 3\text{-}16$$

In 3-16 T, V and μ are derived from the prior data. With a given RGB dataset P, recovering the corresponding spectra becomes estimating the corresponding N_p (PQR weights). Then the $H(N_p)$ can be used to define the ambiguity of this problem.

The PQR weights are converted into 8-bit values because the Shannon entropy is based on discrete outcomes. The entropy is estimated by measuring the density of samples located in each bin. For the collected prior dataset, when representing each of the PQR weights with 8 bits, the entropy of the PQR weights is 13.7 bits. While a uniformly distributed three-dimensional variable would have an entropy of 24 bits.

When given the RGB value, the remaining ambiguity of the PQR weights can be represented as the conditional entropy $H(N_p|\mathrm{P})$ which can be estimated by the chain rule for Shannon entropy as:

$$H(N_p|\mathrm{P}) = H(N_p, \mathrm{P}) - H(\mathrm{P}) \qquad 3\text{-}17$$

where $H(N_p, \mathrm{P})$ is the joint entropy of the RGB values and PQR weights. For the prior data, $H(N_p|\mathrm{P})$ equals 8.3 bits, which is less than the entropy of PQR weights, $H(N_p)$. Therefore, knowing the RGB values could reduce the ambiguity of the PQR weights. This reflects the additional information that can be obtained from the non-linear relationship. However, it is important to note that the PQR weights represent metameric black, so relying solely on RGB values cannot resolve metamerism. As a result, there is remaining ambiguity with the PQR weights. To further resolve the ambiguity

caused by the metamerism, additional information is required to achieve a more precise estimation of the PQR weights. A more detailed discussion of the entropy in the PQR coefficients will be presented in Chapter 4.

3.6. Estimating PQR Weights by Linear Regression

In the previous sections, we have demonstrated that the RGBPQR colour space can provide a good representation of spectral data in a low-dimension manner. In this section. We first test using linear regression to estimate the weights for PQR bases from the RGB input.

Assume there is a linear transform from the RGB values (P) to the weights of the residual components as 3-18:

$$N_p = M\mathrm{P} + b \qquad 3\text{-}18$$

where M is the transform matrix and b is a bias. When estimating 3 residual components, M is a 3×3 matrix. With the prior data M and b can be estimated, in this experiment the 'regress' function in MATLAB was used to learn this mapping. With the estimated M and b, the PQR weights can be estimated as:

$$\hat{N}_p \approx M\mathrm{P} + b \qquad 3\text{-}19$$

Then the recovered spectra from RGB data can be represented as:

$$\hat{S} \approx R\mathrm{P} + V\hat{N}_p + \mu \qquad 3\text{-}20$$

Table 6 compares the reconstruction accuracy achieved using different numbers of residual bases. Estimating the weights of residual bases through linear regression does not result in any significant enhancement in reconstruction accuracy. This result is explained since the up-sampling function already optimises the linear solution based on the RGB values. Therefore, the residual between the up-sampled and original spectrum cannot be estimated using a linear function of RGB.

Table 6. Reconstruction accuracy of spectral data with a different number of residual bases.

	0	1	2	3
RMSE	0.027	0.027	0.027	0.026
MRAE	0.132	0.132	0.132	0.131
SAM	0.123	0.123	0.122	0.122

3.7. Estimating PQR with Non-linear Regression

To achieve higher reconstruction accuracy, it is necessary to use a non-linear model to estimate the PQR weights from RGB values.

3.7.1. Dictionary Learning

Dictionary-based methods have traditionally been employed to address nonlinear mapping problems. In this section, we introduce the estimation of PQR weights based on a dictionary that has been learned from prior data. The training dataset used in this section is the same as that in the previous sections, consisting of 900,000 spectral samples. The learned dictionary functions as a lookup table, with mappings from RGB values to PQR weights. To learn this dictionary, the 8-bit RGB values are further quantised to give 8 to 64 bins per channel to make the data less sparse in the high-dimensional space. Specifically, we divided the 8-bit training RGB data into sub-cubes, and the corresponding PQR weights for each sub-cube were obtained as the mean of the weights of the training residual samples within that sub-cube. When reconstructing, the input RGB value is used as the index to provide the corresponding PQR values. When building the dictionary, the PQR weights of bins without training samples were set to zero. However, it is rare for an RGB value from the testing data set to be out of the training dataset especially when the RGB value has been compressed.

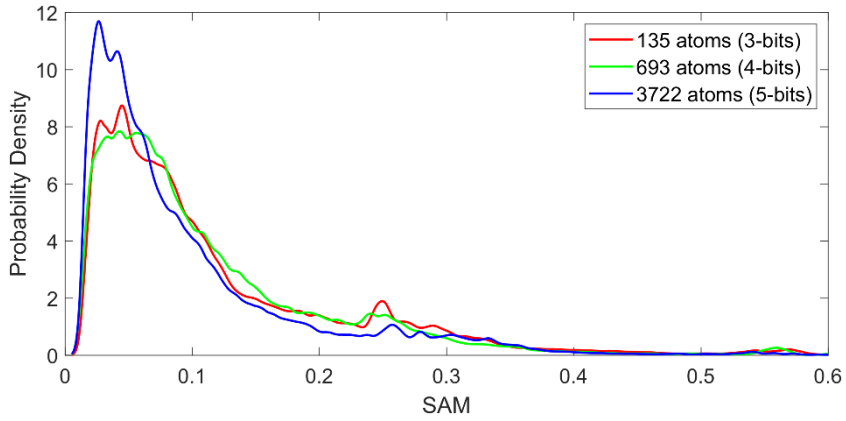
Table 7 compares the reconstruction accuracy when using the learned dictionary to estimate the PQR weights and reconstruct the spectra of NTIRE 2022. 5 levels of compression have been tested which include 3-8 bits of RGB space (shown as number of bins). The testing dataset was the same as in the previous section.

Using a 3-bit RGB dictionary results in a total of 512 sub-cubes (atoms). However, due to the gamut limitations of the training dataset, only 135 of these atoms have training samples. Despite this, utilizing the 3-bit dictionary to estimate PQR weights does not largely improve accuracy. This could be attributed to the fact that 135 parameters are insufficient to represent the RGB to PQR weight mapping accurately. To achieve better accuracy, a more complex dictionary is necessary.

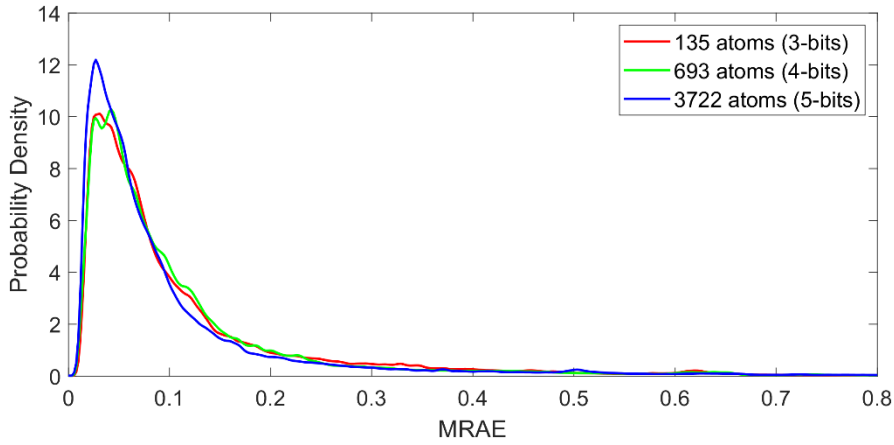
Table 7. The construction accuracy when using the learned dictionaries to estimate the PQR weights and reconstruct the spectra.

	Linear	8 bins	16 bins	32 bins	64 bins	256 bins
RMSE	0.027	0.026	0.024	0.024	0.024	0.025
MRAE	0.132	0.134	0.127	0.122	0.119	0.118

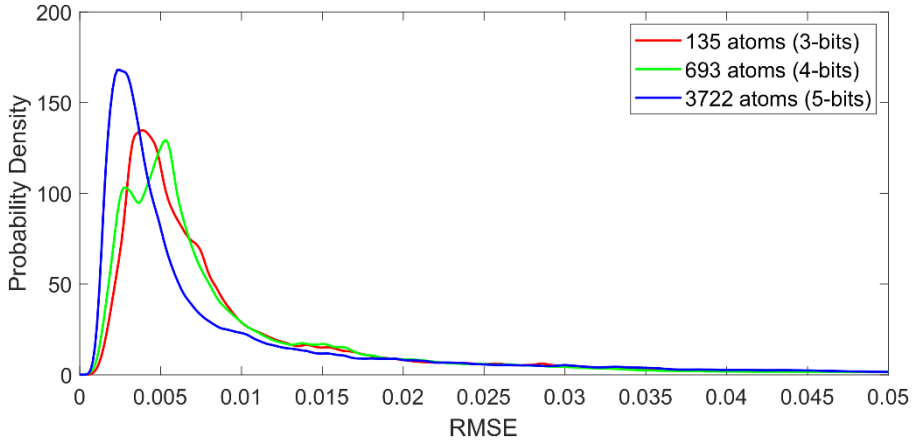
SAM	0.123	0.119	0.112	0.108	0.103	0.098



a. Error distribution of SAM



b. Error distribution of MRAE



c. Error distribution of RMSE

Figure 18. Error distribution when reconstructing spectral images with dictionaries with different complexity.

Alternatively, when compressing the RGB values into 4 bits, the reconstructed spectra exhibit a small but noticeable improvement. This is due to the availability of 693 valid atoms in the 4-bit RGB

dictionary, which better represents the mapping from RGB to PQR. However, when further increasing the complexity of the dictionary the RMSE reconstruction error didn't show any increase while the other two error measurements still decreased. This is because the up-sampling function is designed to minimize the least squares error, which results in the residual represented as RMSE becoming relatively small. The decrease in MRAE and SAM suggests that a more complex model can better map RGB to PQR in finer detail, but the improvement obtained by using a complex dictionary is insignificant. Increasing the number of bins for each channel to 256 doesn't lead to improved reconstruction accuracy. Despite the enhanced granularity, the persistent error primarily comes from metamerism. To truly enhance the reconstruction precision, there's a need to learn a specific mapping from RGB to individual metamerism samples, rather than the average of all such samples.

Figure 18 shows the distribution of SAM, MRAE and RMSE when the number of bins from each colour channel of the dictionary is 8, 16, and 32, respectively. In Figure 18, the distribution plots share a similar appearance, with some minor variations. When using a dictionary, the probability of obtaining a low error score increases as the complexity of the dictionary decreases. However, increasing the number of bins in the plot does not provide a significant improvement, indicating that the primary cause of the errors in these cases is due to metamerism. This type of problem cannot be solved by using a more complex dictionary.

3.7.2. Shallow Neural Network

Neural networks are able to model non-linear relationships, so can be used for non-linear regression. Shallow neural networks with varying widths were used to estimate PQR weights from the RGB value inputs. The experiments utilized the MATLAB regression learner toolbox. All networks in the experiment had a single fully connected hidden layer, followed by a sigmoid activation layer. The experiment tested different sizes of the full-connection layer. The trained network used RGB values as input and estimated the corresponding P, Q, and R weights. The same training dataset was used as in the previous section, and a 5-fold cross-validation was employed during training. Table 8 compares the reconstruction accuracy for the 50 test images.

Table 8. The construction accuracy when using the trained shallow network to estimate the PQR weights and reconstruct the spectra.

	Linear	Dictionary	Nodes in the hidden layer		
			10	25	100
RMSE	0.027	0.024	0.024	0.024	0.024
MRAE	0.132	0.127	0.125	0.120	0.122

SAM	0.123	0.112	0.107	0.110	0.106

Table 8 indicates that a narrow network, when employed to estimate PQR weights, can achieve comparable reconstruction accuracy to the 4-bit dictionary learning-based approach. However, increasing the network's width does not lead to large improvement in reconstruction. This implies that we are reaching the limit of estimating PQR weights from RGB values in the presence of metamerism.

Without additional information to resolve the metamerism, it is impossible to further enhance the accuracy of estimating the PQR weights solely from RGB values.

3.8. Summary

In this chapter, we have adapted the LabPQR colour space into the RGBPQR colour space to better accommodate the task of single image spectral super-resolution. We've demonstrated that the RGBPQR model can effectively represent spectral data in a low-dimensional context. Based on this model, the spectral reconstruction task from an RGB image becomes an exercise in estimating the coefficients of PQR (residual) components. However, due to metamerism, relying solely on RGB values does not eliminate all ambiguities related to the PQR coefficients. Therefore, to accurately estimate the PQR coefficients, it is crucial to incorporate additional information. The remainder of this thesis will discuss how to effectively estimate the PQR coefficients in a way that is easily interpretable.

Chapter 4. Quantifying Contextual Information on Resolving Metamerism

In the previous chapter, the RGBPQR colour space was introduced as a model for spectral super-resolution. With the model, reconstructing spectral reflectance from RGB value became a problem in estimating the weights of the residual PQR components. However, using RGB values alone to estimate these weights led to suboptimal reconstruction accuracy due to the issue of metamerism. To resolve metamerism, additional information is required.

Addressing this challenge, convolutional neural network-based techniques incorporate information from neighbouring pixels as contextual data to mitigate metamerism. According to the reported reconstruction accuracy, using spatial information surrounding the target pixel shows potential as a source of context to resolve metamerism. Yet, how this spatial information is utilized by networks remains unclear, given their "black box" nature. To design our explainable spectral super-resolution method, it is crucial to examine which types of spatial information from the neighbouring area are most suitable for this specific task. Currently, however, no method exists that can compare the metamerism resolution ability of different potential additional information sources.

However, the RGBPQR colour space also offers an opportunity to quantitatively measure the ability to resolve metamerism. This is achieved by comparing the reduction of ambiguity from the PQR coefficients, which represent metamerism. In this chapter, we aim to investigate how different types of local spatial information can disambiguate the PQR weights and improve their accuracy, ultimately leading to more precise spectral reconstruction.

4.1. Using Shannon Entropy to Estimate the Additional Information

4.1.1. Introduction

As mentioned in section 3.15, the ambiguity of reconstructing the spectra can be represented by the entropy of the PQR weights. With the conditional entropy, we have also demonstrated that with the given RGB value the ambiguity of the corresponding PQR weights can be reduced. However, because of metamerism, it is impossible to remove all the ambiguity in the PQR weights without additional information. In this section, we will explore how conditional Shannon entropy can be used to compare the reduction in ambiguity provided by different types of spatial information. Our objective is to identify an appropriate type of spatial information that can be effectively incorporated into our explainable spectral super-resolution methodology.

- **Shannon Entropy**

Shannon entropy, also known as information entropy, is a measure of the uncertainty or randomness associated with a random variable. It was introduced by Claude Shannon in 1948 as a way to quantify the amount of information contained in a message.

The entropy of a discrete random variable X with possible values $\{x_1, x_2, \dots, x_n\}$ and corresponding probabilities $\{p_1, p_2, \dots, p_n\}$ is given by the formula:

$$H(X) = -\sum p_i \log_2 p_i \qquad 4\text{-}1$$

where the logarithm is usually taken to base 2, and the entropy $H(X)$ is measured in bits. The formula can be interpreted as the expected value of the information content of X, where the information content of an event with probability p_i is given by $-\log_2 p_i$.

The entropy of a random variable is a measure of the degree of uncertainty associated with it. A random variable that is highly unpredictable or random will have a high entropy, while a variable that is highly predictable will have a low entropy. The maximum possible entropy of a random variable is $\log_2 n$ bits, where n is the number of possible outcomes, and this is achieved when all outcomes are equally likely.

In the RGBPQR colour space, assume the original spectra S can be approximately represented $\hat{S}$ as:

$$\hat{S} = T\mathrm{P} + VN_p + \mu \qquad 4\text{-}2$$

where T, V and μ are derived from the prior data. For a given RGB, P, recovering the corresponding spectra becomes estimating the corresponding N_p (PQR weights). Therefore $H(N_p)$ can be used to define the whole ambiguity of this problem. $H(N_p)$ can be estimated as:

$$H(N_p) = -\sum_{i=1}^{n} p\left(N_{p_i}\right) \log_2 p\left(N_{p_i}\right) \quad \text{4-3}$$

To calculate the Shannon entropy, it is necessary to convert the PQR weights (N_p) into a discrete variable, where the number of bins (n) represents the number of partitions of the dispersed N_p. However, using 8-bit data makes the collected sample too sparse compared to the number of bins, which cannot provide an accurate representation of the sample's distribution in the high dimensional space. Therefore, in this study, we have chosen to set the number of bins for each dimension to 64 (6 bits), as this allows for a manageable number of bins while still providing sufficient resolution for our analysis. With three dimensions for PQR, the total number of bins (n) equals $64^3 = 262{,}144$. While increasing the number of bins gives better resolution, introducing additional information increases the number of dimensions, with the available samples becoming sparser, making it difficult to estimate the distribution of samples accurately. Therefore, by using 6 bits, we strike a balance between the accuracy and manageability of the dataset, making it easier to estimate the distribution of samples in the high-dimensional space accurately.

In 4-3, $p\left(N_{p_i}\right)$ represents the probability that a PQR weight falls in the ith bin, which can be calculated from:

$$p\left(N_{p_i}\right) = \frac{N_{N_{p_i}}}{N} \quad \text{4-4}$$

where $N_{N_{p_i}}$ is the number of samples within the ith bin, and N is the total number of samples in the dataset.

- **Conditional Entropy**

Conditional entropy is a measure of the uncertainty or entropy of a random variable given the information provided by another random variable, i.e., that other random variable is known. The reduced uncertainty can be understood as the information provided by that other random variable.

In our problem, when given the RGB value, the remaining ambiguity of the PQR weights can be represented as the conditional entropy $H(N_p|\mathrm{P})$, which can be estimated by the chain rule of Shannon entropy as:

$$H(N_p|\mathrm{P}) = H(N_p, \mathrm{P}) - H(\mathrm{P}) \quad 4\text{-}5$$

where $H(N_p, \mathrm{P})$ is the joint entropy of the RGB values and PQR weights which can be estimated from:

$$H(N_p, \mathrm{P}) = -\sum_{i=1}^{n} p\left(N_{p_i}, \mathrm{P}\right) \log_2 p\left(N_{p_i}, \mathrm{P}\right) \quad 4\text{-}6$$

When estimating the entropy of RGB values ($H(\mathrm{P})$), the RGB value has also been converted into 6 bits to reduce the influence of the increasing dimensionality.

The information provided by the RGB value on the PQR weights can be represented as:

$$I_{rgb} = H(N_p) - H(N_p|\mathrm{P}) \quad 4\text{-}7$$

As discussed in section 3.5, this represents non-linearities in the relationship between the spectrum and the RGB values. If a one-to-one correspondence between RGB values and PQR weights existed, knowledge of the RGB value would fully determine the PQR weight, resulting in zero conditional entropy. In other words, the RGB value would eliminate any uncertainty about the PQR weight. However, due to metamerism, there is no such one-to-one mapping. This study assumes that additional information reduces the uncertainty in the PQR weights and therefore represents the ability to resolve metamerism.

When additional spatial information is added, the conditional entropy of the PQR weights with a given RGB value and the additional information can be represented as:

$$H(N_p|\mathrm{P}, A) = H(N_p, \mathrm{P}, A) - H(\mathrm{P}, A) \quad 4\text{-}8$$

where A represents the additional variables, it could be neighbour RGB values or extracted feature vectors from the local image patch. The extra information provided by the additional spatial variable can be represented as:

$$I_{add} = H(N_p, \mathrm{P}) - H(N_p|\mathrm{P}, A) \quad 4\text{-}9$$

I_{add} could be used to measure the amount of ambiguity removed from adding additional spatial information while the RGB value of the target is known.

4.1.2. Methodology

- **Dataset in This Study**

In order to address the issue of sparsity resulting from adding dimensions, a larger spectral dataset was collected in this study. This was accomplished through the collection of three datasets, two of which were randomly collected, while the third was an edge dataset. Having two randomly collected datasets serves the purpose of ensuring that the collected data adequately represents the entire database. On the other hand, having the edge dataset is motivated by the observation that edge pixels are more likely to differ from their immediate neighbours, potentially making them more informative. All spectral data was obtained from the NTIRE 2022 spectral database, and the RGB values were generated using the same methodology as described in Chapter 3. A separate RGBPQR model was trained for each dataset.

The two random datasets in this study were collected using the same approach as in the previous chapter, but with a collection of 5,000 samples per image taken from all 900 training images, rather than just 1,000 samples per image. The gamut for these new samples is similar to that of the dataset in the previous chapter. The increased redundancy in the new random samples was collected to overcome sparsity.

The edge dataset was created by first applying a Canny edge detection filter to the grayscale image and then collecting samples from the edges. Similar to the random datasets, 5,000 samples were collected from each image.

Additionally, a neighbour patch size of 21×21, centred on the target pixel, was collected for all samples as additional spatial information, in addition to the target's RGB values.

- **Different Types of Spatial Information Investigated**

To investigate the potential for reducing ambiguity in PQR weights by involving local spatial information, this study explores the use of additional spatial information sourced from local image patches centred on the target sample. This additional information is categorized into two groups:

1. Neighbour RGB pixels of the target pixel including the pixels immediately left, right, above and below.
2. Simple statistical measures of the image patch, including the minimum and maximum values of the RGB patch, as well as the mean, standard deviation, skewness, and kurtosis of RGB values within the image patch. The statistical measures were applied to each of the RGB components within the patch.

In order to mitigate the impact of higher dimensionality on the accuracy of entropy estimation, we have confined all additional information to a 3-dimensional vector.

4.1.3. Results

Table 9 displays the calculated Shannon entropy of the PQR coefficients $H(N_p)$, along with the conditional entropy when using information from the local image patch. When compared with data collected randomly, the entropy of the PQR coefficients is higher when the target sample is situated on an edge. This could be attributed to the fact that samples collected from edges are more likely to originate from a variety of materials. Since the entropy from the two randomly collected datasets are similar, it suggests that our collected dataset is a good representation of the original NTIRE 2022 database. However, knowing the RGB value on the edge removes less ambiguity in the PQR coefficient, because edge spectral samples are more likely to be a mixture of multiple endmembers, reducing the correlation between the RGB value and the spectrum.

The results from the two randomly collected datasets were similar. For both, introducing the neighbouring pixel as context offered a similar degree of information for resolving metamerism. These findings suggest that incorporating neighbouring pixels can slightly reduce the ambiguity of the PQR weights. However, when compared with the randomly collected samples, adding neighbouring pixels from an edge sample yielded more supplementary information. This is attributed to the fact that the additional information in this study is able to provide a more unique context for the PQR weights. As neighbouring pixels from an edge sample are more likely to differ from the target pixels, the supplemental information obtained when adding such neighbouring pixels appears to be greater.

Upon comparing neighbouring pixels from various positions, we observed that the pixel above and below imparts slightly more additional information than the left and right pixels across all three datasets. This trend suggests that vertical pixels within a given area are more likely to exhibit different values compared to their horizontal counterparts. This phenomenon could be attributed to the NTIRE dataset encompassing scenes featuring buildings and other manmade structures, which often present horizontal edges.

To ensure the estimation is not affected by overfitting, we conducted a noise test. We introduced a 1-bit noise, a one-dimensional variable comprising random zeros and ones, and used this as our "additional information" during the conditional entropy estimation of the PQR weights. For a 6-bit sampling of the PQR weights, the $H(N_p, noise)$ yielded a value of 8.18 for our initial random dataset which is 1 bit larger than $H(N_p)$. This indicates that the estimation did not consider the noise as a form of genuine information. However, with 8-bit sampling, the scenario shifted. While $H(N_p)$ was 12.45, $H(N_p, noise)$ stood at 13.33, suggesting that the estimation had indeed factored in the noise

as information. This anomaly can be attributed to the sparsity when representing the PQR using 8 bits enabling the noise bit to provide disambiguation.

Table 9. Additional information when adding neighbour RGB pixel. The entropy of the PQR weights is highlighted in red. The conditional entropy values, denoted in blue, correspond to when various types of contextual information are added, while the associated additional information is highlighted in green.

	Variables	Random data 1	Random data 2	Edge data
Entropy of PQR weights	$H(N_p)$	7.17	7.19	7.26
RGB	$H(N_p, \mathrm{P})$	13.19	13.19	13.51
	$H(\mathrm{P})$	9.14	9.14	9.18
	$H(N_p \mid \mathrm{P})$	4.05	4.05	4.33
	I_{rgb}	3.12	3.14	2.93
Neighbour RGB pixels				
RGB with left neighbour RGB	$H(N_p, \mathrm{P}, A_{left})$	11.82	11.82	12.88
	$H(\mathrm{P}, A_{left})$	14.80	14.79	15.72
	$H(N_p \mid \mathrm{P}, A_{left})$	2.98	2.96	2.84
	I_{left}	1.08	1.08	1.49
RGB with right neighbour RGB	$H(N_p, \mathrm{P}, A_{right})$	11.82	11.82	12.87
	$H(\mathrm{P}, A_{right})$	14.79	14.78	15.71
	$H(N_p \mid \mathrm{P}, A_{right})$	2.97	2.96	2.84
	I_{right}	1.08	1.09	1.49
RGB with above neighbour RGB	$H(N_p, \mathrm{P}, A_{above})$	11.97	11.97	13.04
	$H(\mathrm{P}, A_{above})$	14.89	14.88	15.84
	$H(N_p \mid \mathrm{P}, A_{above})$	2.92	2.91	2.79
	I_{above}	1.13	1.14	1.54
RGB with below neighbour RGB	$H(N_p, \mathrm{P}, A_{below})$	11.98	11.98	13.03
	$H(\mathrm{P}, A_{below})$	14.89	14.88	15.83
	$H(N_p \mid \mathrm{P}, A_{below})$	2.91	2.91	2.79
	I_{below}	1.13	1.14	1.54

Table 10 shows the conditional entropy and the additional information when involving simple statistical measurements of the image patch as additional context. The maximum and minimum in the table is measured as the maximum and minimum R, G and B value separately within the 21×21 patch, the value is not necessarily from one particular pixel. The mean, standard deviation, skewness and kurtosis measure the distribution of the R, G and B values separately within the image patch. The measured statistical results have also been converted to 6 bits.

Table 10. Additional information when adding statistical information from a 21×21 neighbourhood.

	Variables	Random data 1	Random data 2	Edge data
Entropy of PQR weights	$H(N_p)$	7.17	7.19	7.26
RGB	$H(N_p, \mathrm{P})$	13.19	13.19	13.51
	$H(\mathrm{P})$	9.14	9.14	9.18
	$H(N_p \mid \mathrm{P})$	4.05	4.05	4.33
	I_{rgb}	3.12	3.14	2.93
Simple statistical measures				
Maximum RGB from the patch	$H(N_p, \mathrm{P}, A_{max})$	15.66	15.66	15.86
	$H(\mathrm{P}, A_{max})$	17.32	17.31	17.52
	$H(N_p \mid \mathrm{P}, A_{max})$	1.66	1.64	1.66
	I_{mean}	2.39	2.40	2.67
Minimum RGB from the patch	$H(N_p \mid \mathrm{P}, A_{min})$	13.67	13.67	13.57
	$H(\mathrm{P}, A_{min})$	16.03	16.01	16.13
	$H(N_p \mid \mathrm{P}, A_{min})$	2.37	2.35	2.56
	I_{max}	1.69	1.70	1.77
Mean RGB from the patch	$H(N_p, \mathrm{P}, A_{mean})$	13.04	13.04	13.34
	$H(\mathrm{P}, A_{mean})$	15.55	15.54	16.96
	$H(N_p \mid \mathrm{P}, A_{mean})$	2.51	2.49	2.63
	I_{min}	1.54	1.55	1.70
Standard deviation of the patch	$H(N_p, \mathrm{P}, A_{std})$	14.73	14.74	14.86
	$H(\mathrm{P}, A_{std})$	16.71	16.70	16.91
	$H(N_p \mid \mathrm{P}, A_{std})$	1.98	1.96	2.05
	I_{std}	2.07	2.08	2.28
Skewness of the patch	$H(N_p, \mathrm{P}, A_{skew})$	14.52	14.46	14.78
	$H(\mathrm{P}, A_{skew})$	17.05	17.00	17.36
	$H(N_p \mid \mathrm{P}, A_{skew})$	2.53	2.54	2.58
	I_{ske}	1.52	1.51	1.75
Kurtosis deviation of the patch	$H(N_p, \mathrm{P}, A_{kur})$	11.45	11.42	15.43
	$H(\mathrm{P}, A_{kur})$	14.98	14.95	13.51
	$H(N_p \mid \mathrm{P}, A_{kur})$	3.53	3.53	3.67
	I_{tur}	0.52	0.52	0.66

Compared with adding a neighbour RGB value as additional information, adding the minimum or maximum RGB value within a patch could remove more ambiguity from the PQR weights, especially for the maximum value. One possible explanation for this is that the target sample is not likely to have the maximum or the minimum RGB value from within the image patch, so adding the maximum or the minimum RGB value could make the combined 6-dimensional vector more unique. The maximum or the minimum process selects more unique information from the image patch than using any pixel as additional information.

The additional information gained by incorporating the mean RGB of the image patch is larger than when adding a neighbouring pixel, but smaller than when adding the maximum. Given the size of the image patches (21×21), the samples within the patch are unlikely to belong to a single material. Consequently, the mean RGB value is likely to differ from the central pixel. Similarly, adding the standard deviation of the 21×21 image patch as additional context provides more additional information than adding neighbouring pixel values.

We further examined the spatial information derived from a highly localized area. In particular, we incorporated basic statistical measurements from RGB values within a 3×3 image patch (details included in Appendix 2). We discovered that the mean value of these highly local pixels provides less additional information compared to adding neighbouring pixel values, let alone statistical context from larger patches. This can be attributed to the likelihood of the neighbouring pixels being similar to the central pixel within such a small patch. The averaging process tends to dilute unique values within the patch, consequently reducing the combined vector's potential distinctiveness when the averaged RGB value is used as additional context. In contrast, the skewness and kurtosis of the 3×3 image patch could significantly decrease the conditional entropy of the PQR coefficient. However, determining high-order distribution descriptors with a small sample size is not reliable, and the reduction in entropy may be due to noise.

Using the skewness and kurtosis from the 21×21 image patch as additional context does not remove markedly more ambiguity in the PQR weights compared with other types of contexts. This implies that the distribution of local pixel values may differ in their average values and spread, but share similar characteristics in their shape and distribution symmetry.

4.1.4. Summary

From the result, since the RGB values of neighbouring pixels are often similar to the target pixel, simply adding neighbour pixel values is not likely to be particularly effective to distinguish metamer RGB values. Compared with a very local area (3×3), extracting information from a larger area increases the ability to resolve metamerism. However, the very local information is still important in

the spectral reconstruction process, since they are highly correlated to the target sample which could reduce the influence of noise on the pixel value of the target.

Reconstructing the spectra of edge samples is more challenging since they have higher ambiguity than randomly collected samples, even when additional information is involved, the conditional entropy (remaining ambiguity) is still higher. The high ambiguity may result from the sample being on the boundary between two materials, resulting in a mixture of multiple spectra that increases the challenge during reconstruction. Therefore, when designing any reconstruction method, the edge sample should be given higher attention.

Compared with adding a particular RGB value as additional context, adding simple statistical measurements provides more additional information on the PQR coefficients. Therefore, the distribution descriptor and even a texture descriptor of an image patch should be considered as context information when resolving metamerism.

This experiment has several limitations, the addition of spatial information in the form of a three-dimensional vector resulted in an increase to 9 dimensions. Despite down sampling both RGB values and PQR weights into 6 bits, the resulting samples remain sparse in the 9-dimensional space. This sparsity means that many bins in the high-dimensional space receive only zero or one sample, which makes it difficult to obtain an accurate estimation of the distribution. The curse of dimensionality presents a challenge, as it limits our ability to test more complex (higher dimensional) additional information, for example, texture descriptors such as Gabor filters. However, understanding the potential of high-dimensional data to provide context information is essential to this study. Another limitation of estimating the Shannon entropy with bins is the result could be influenced by noise. With the increasing dimensionality when adding extra context, the samples became sparse in the high dimensional space. As a result, the measurement could take noise as additional disambiguating information, and as a result, the sample is overfitted to noise which gives a misleading low conditional entropy.

4.2. Differential Entropy by Modifying the Bin Size

4.2.1. Estimating Joint Entropy by Rescaling the Bin Size

The PQR coefficients are not uniformly distributed and are constrained by the number of samples that can be collected. Consequently, certain bins may contain only 0 to 1 samples, making the entropy estimation unreliable. To mitigate this issue, we propose using a scaled bin size within the PQR space to estimate the distribution of the PQR coefficients and, subsequently, the differential entropy of these coefficients.

Differential entropy is a concept stemming from information theory that broadens the principle of entropy—traditionally a measure of uncertainty or randomness—to apply to continuous probability distributions. In the context of continuous random variables, differential entropy is defined as:

$$h(X) = -\int f(x)\log f(x)dx \qquad 4\text{-}10$$

where $f(x)$ is the probability density function of the random variable X and the integral is taken over the whole range of X. One important thing to note about differential entropy is that, unlike the entropy for discrete variables, it can be negative. This is because, for continuous variables, the probability density function can be greater than 1 at some points, making the corresponding logarithmic term positive. When the differential entropy exists and is finite, it follows the chain rule which can be used to calculate the conditional entropy. Additionally, the mutual information between continuous variables can be estimated as $I(X,Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$ which is the same as for discrete entropy.

In this experiment, the differential entropy is estimated by adjusting the width of bins to estimate the probability density function of the PQR coefficients as a piecewise function. Taking the P coefficient as an example, the differential entropy of the P coefficient is estimated by:

$$h(P) = -\int \frac{p_i}{w_i}\log\left(\frac{p_i}{w_i}\right) \approx -\sum_{i=1}^{n} p_i \log\left(\frac{p_i}{w_i}\right) \qquad 4\text{-}11$$

Where n is the number of bins, and w_i is the width of the ith bin, Figure 19 shows the estimated probability density function of the P coefficient with uniform bin width. Although there are 64 bins, the probability is very low outside the range of -0.5 to 0.5. When estimating the entropy of the PQR coefficients, samples would be located in a 3-dimensional space defined by the bins from each variable.

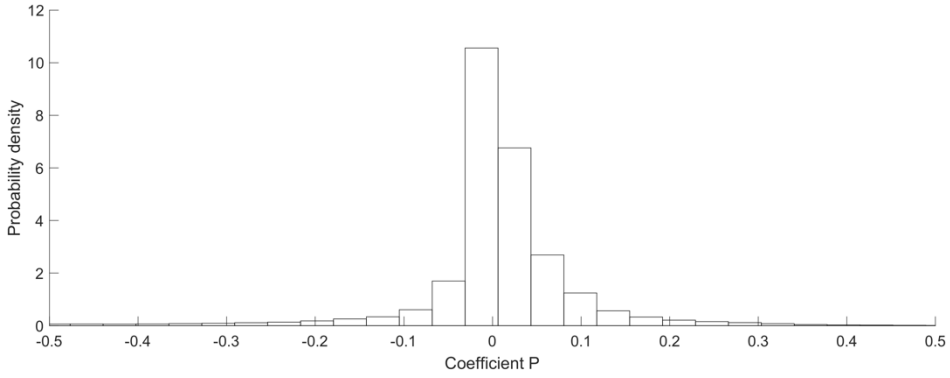


Figure 19. This represents the distribution of P coefficients with a uniform bin width, where the range of P has been constrained between -0.5 and 0.5.

However, limited by the number of bins, Figure 19 can only approximate the probability density function, which limits the accuracy of the estimated entropy. In this section, we propose a histogram distribution estimation method with scaled bin width. The basic idea behind the method is to adjust the width of bins, so that bins with high probability density are narrower, while bins with low probability density are made wider. The bin width is adjusted by making the number of samples allocated to each bin approximately similar (effectively using histogram equalisation). Figure 20 shows the estimated probability density function of each PQR coefficient separately using 64 bins. Compared with using uniform bin size, the estimated histogram better represents the distribution with the same number of bins. Building on the scaled bin width, we can more accurately estimate the joint entropy of the PQR coefficients, as well as the conditional entropy when additional information is incorporated.

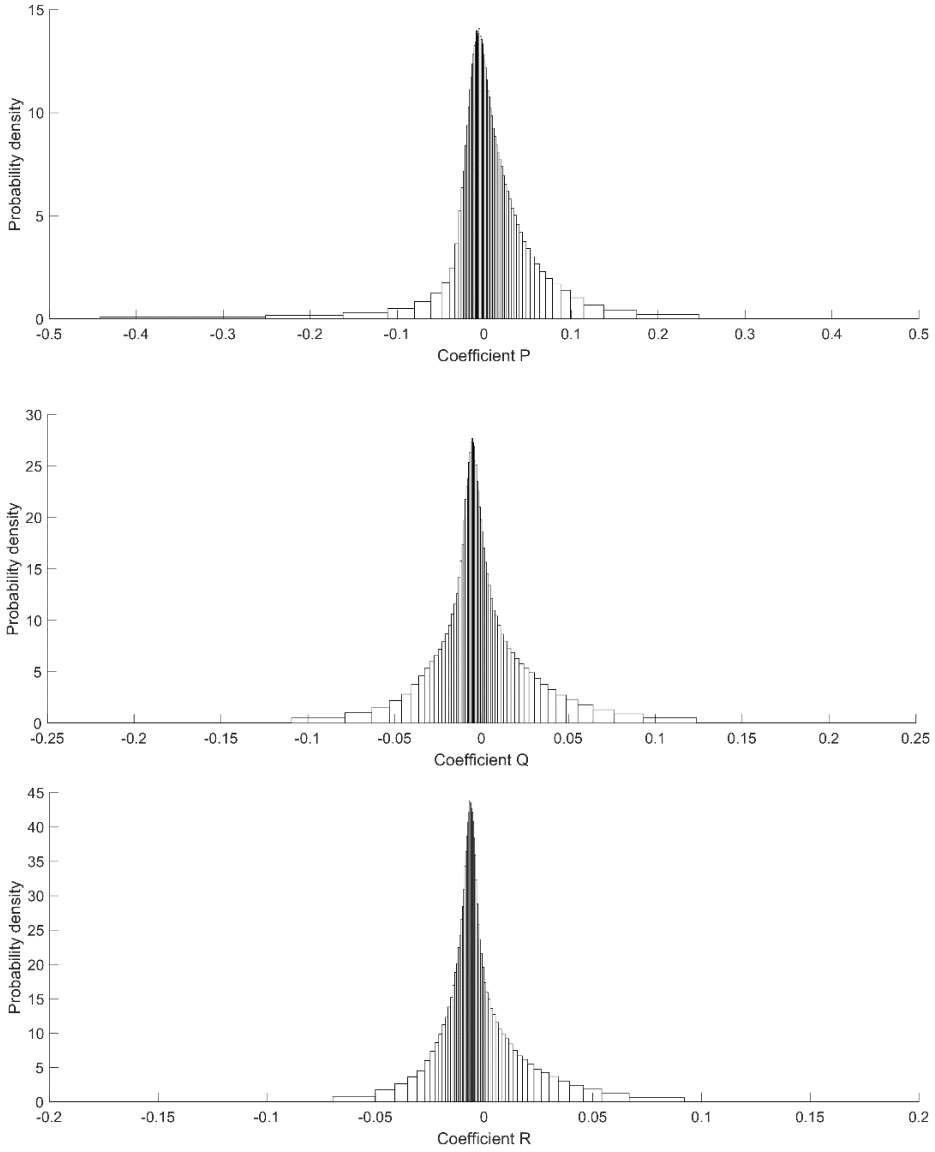


Figure 20. The distribution of PQR coefficients with adjusted bin width can more accurately estimate the distribution than a histogram with uniform bin width.

Table 11. The estimated joint entropy from different methods. The proposed scaled bins method estimates the joint entropy of the PQR coefficient with a similar result to the k-NN-based methods.

	P	Q	R	PQR
Uniform bins entropy	-1.30	-1.91	-2.22	-6.33
Scaled bins entropy	-1.35	-1.99	-2.25	-6.89
KL entropy	-1.35	-2.01	-2.26	-7.06
Copula entropy	-1.35	-2.01	-2.26	-6.98

Table 11 compares the estimated differential entropy based on our scaled bin method with the estimated entropy derived from uniform bins. Furthermore, we have also presented results from two entropy estimators for comparison: the Kozachenko–Leonenko (1987) differential entropy estimator and the copula estimator (Ariel & Louzoun, 2020). These two methods estimate the joint entropy by assessing the k-nearest neighbours (k-NN) and have been proven to deliver reasonably accurate results. According to the data, our proposed scaled bins method offers an estimation that is more closely aligned with these trustworthy estimators than the uniform bin-based method. We opted not to base our experiment on the neighbour-based entropy estimators because they consider repeated information as additional information. For instance, when an RGB value identical to the target RGB is added, the joint entropy $H(N_p, \mathrm{P}, \mathrm{P})$ estimated by our method would be equivalent to $H(N_p, \mathrm{P})$. In these cases, the context provides no extra information to uniquely identify the metamerism samples. In contrast, in k-NN based methods, the estimator would perceive the repeated context as information since it impacts the distance between samples. Therefore, our proposed method, which concentrates on identifying unique contextual information to distinguish metamerism samples, is more suited to this particular problem.

4.2.2. Results from the Differential Entropy

Utilizing the scaled bin entropy estimation, we repeated the experiment outlined in section 1, with the results presented in Table 12.

Note that since the entropy measurement method is different, the numerical values from Table 12 cannot be directly compared with previous tables. However, the results demonstrate a similar trend. Simple statistical measures of the local image patch provide more information for resolving metamerism than just using neighbouring RGB values as context. The maximum RGB value from a neighbouring region showed a greater advantage in resolving metamerism than other statistical methods. However, the kurtosis of the neighbouring pixels didn't contribute much additional

information, which could be explained by the fact that kurtosis doesn't vary markedly across different small image patches. Compared to randomly collected samples, neighbouring pixels of the target sample along edges can provide more information for uniquely distinguishing metamerism samples. Therefore, in designing our method, local features such as edges should be considered as potentially useful features upon which to focus.

Table 12. Additional information resulting from local context when using differential entropy estimated from the scaled bin method.

	Variables	Random data 1	Random data 2	Edge data
Entropy of PQR weights	$h(N_p)$	-6.84	-6.89	-6.77
RGB	$h(N_p, \mathrm{P})$	-16.78	-16.93	-16.69
	$h(\mathrm{P})$	-6.31	-6.29	-6.28
	$h(N_p \mid \mathrm{P})$	-10.48	-10.64	-10.42
	I_{rgb}	3.68	3.75	3.65
Immediate neighbour RGB pixels				
Left	I_{left}	1.51	1.50	2.07
Right	I_{right}	1.51	1.50	2.07
Above	I_{above}	1.60	1.59	2.17
Below	I_{below}	1.60	1.59	2.17
Simple statistical measures within a 21×21 window				
Max	I_{max}	3.67	3.63	3.81
Min	I_{min}	2.16	2.15	2.34
Mean	I_{mean}	2.49	2.48	2.47
STD	I_{std}	3.12	3.09	3.22
Skewness	I_{skew}	2.80	2.72	3.01
Kurtosis	I_{kur}	1.09	1.05	1.26
Local Binary Pattern histograms within a 21×21 patch				
LBP	I_{lbp}	5.24	5.16	5.35

Several times we have made the observation that texture measures are likely to be of benefit. To test this, we explore a widely used abstract texture descriptor, the Local Binary Pattern (LBP), as contextual information. The LBP is extracted from the converted grayscale image patches with a radius of 1, starting with the top-left pixel. The histogram of LBPs within a 21×21 window provides a texture descriptor. The first three coefficients learned by PCA from the extracted LBP histograms are used as the contextual vector to estimate the joint entropy. The results in the last row of Table 12 show that using the LBP-based feature descriptor as additional information can significantly

reduce the ambiguity in metamerism samples. Thus, local texture can serve as a valuable source of additional information to address metamerism, which is sensible given that texture features are often associated with specific materials. Identifying the texture of a surface can aid in material identification, and by extension, the determination of the material's reflectance.

4.2.3. Verifying the Results

In this section, we compare the reconstruction accuracy achieved when representing the spectral samples through dictionary learning, using the additional information tested in Table 12 as the contextual index. In contrast to the dictionary discussed in Chapter 3.7.1, which solely used RGB values as indices, the present test dictionary incorporates both RGB and the tested context for indexing. To build this dictionary, we compress each RGB value and context component into 6 bits to mitigate overfitting. Specifically, we divide the samples into sub-cubes, and the corresponding PQR weights for each sub-cube are determined as the mean of the PQR weights of the residual within that sub-cube. During the reconstruction process, the input RGB value and context serve as the index to retrieve the corresponding PQR values.

The dictionary learning experiment was conducted using Random Dataset 1. Table 13 compares the spectral reconstruction errors when utilizing the various contexts to construct the dictionaries. The results presented represent the accuracy achievable with a dictionary-based method, reflecting the context's ability to resolve metamerism. This accuracy aligns with the trend observed from the previous entropy analysis. Therefore, our proposed quantification technique effectively measures the contextual information's capacity to address metamerism. The findings suggest that leveraging local textural data yields superior reconstruction accuracy compared to other forms of additional information. This leads us to hypothesise that local textural features, as exemplified by the LBP histogram hold promise in resolving metamerism.

Table 13. Representation error via dictionary learning. The initial row designates the kind of additional information used, while the first column itemizes the type of error observed.

	Solely RGB	Left	Right	Above	Below	Max	Min	Mean	STD	Skew	Kur	LBP
RMSE	0.019	0.015	0.015	0.015	0.015	0.101	0.012	0.012	0.011	0.014	0.017	0.009
SAM	0.113	0.105	0.105	0.105	0.105	0.089	0.101	0.099	0.095	0.102	0.110	0.083
95% RMSE	0.039	0.031	0.031	0.030	0.030	0.022	0.025	0.026	0.023	0.030	0.037	0.019
95% SAM	0.314	0.303	0.303	0.302	0.302	0.272	0.296	0.291	0.284	0.296	0.309	0.260

4.3. Summary

In this chapter, we examined potential additional information sources capable of reducing the ambiguity in PQR coefficients by measuring their conditional entropy. Both Shannon and differential entropy utilizing a scaled bin size were tested. The results show that the RGB value of the target sample can reduce the ambiguity of the PQR weights. However, it cannot resolve metamerism, leaving some residual ambiguity. To further diminish this ambiguity and discover a unique context that distinguishes metamerism samples, additional information is needed. Our results suggest that local textural features could serve as a promising choice for contextual information to effectively address the issue of metamerism.

Nonetheless, the entropy estimator utilized in this study has certain limitations. It is affected by noise and the sampling becomes less accurate with increasing dimensionality, which restricts the type of additional information that can be assessed in this study. While these limitations could potentially be overcome by advanced high-dimensional multivariable entropy estimation methods, that is not the main focus of this study. However, the findings presented here are sufficient to act as a reference in designing feature extraction methods aimed at resolving the metamerism issue.

Chapter 5. Analysing How Deep Models Resolve Metamerism

The previous chapter highlighted the potential of very local spatial information, specifically neighbouring pixels, to provide contextual information that reduces ambiguity in PQR weights. However, given the constraints of entropy estimation techniques, assessing spatial information over a broader range or in a more complex manner remains challenging. Nonetheless, for the creation of our explainable method, identifying the right contextual information to tackle metamerism remains essential.

Meanwhile, deep convolutional neural network-based single-image spectral super-resolution methods have demonstrated impressive accuracy in reconstructing spectral data, as discussed in Chapter 2. The increased accuracy is associated with the deep models' ability to resolve metamerism by utilizing spatial information during the reconstruction process. However, like all deep neural network-based methods, spectral super-resolution models suffer from the 'black box' issue. It remains unclear how spatial information contributes to the spectral recovery process, and the required range of spatial information to resolve metamerism remains undefined.

To bridge these gaps, this chapter will analyse existing deep CNN-based spectral super-resolution methods for their solution to resolving metamerism. Our objective is to examine current deep neural network-based methodologies and understand how they utilize spatial information in two aspects. **First**, we aim to determine the necessary spatial range that provides sufficient information for the neural networks to carry out their reconstruction. **Second**, we aim to identify valuable contextual information for resolving metamerism by analysing the potential spatial information utilized by deep models. These results can then guide the development of our explainable method.

We begin this chapter with a case study that analyses the types of information that a deep model uses when determining the shape of its output during the reconstruction of metamer samples. Through this analysis, we aim to demonstrate the importance of local texture in the reconstruction process.

5.1. Case Study of Local Textures on Resolving Metamerism

5.1.1. Introduction

Since different materials often exhibit unique textures, texture can serve as valuable contextual information for characterizing corresponding materials. As different materials are associated with specific spectral properties, determining the material can aid in identifying the corresponding spectrum. Therefore, we hypothesize that deep models partially utilize local textures as contextual information to resolve metamerism. To test this hypothesis, we investigate how local textures influence the decision-making process of the HSCNN+, which has demonstrated the ability to distinguish a group of metamer samples. Our approach involved generating various textures and merging them with a pure colour background. Subsequently, we used the HSCNN+ model to reconstruct these generated image patches. Our primary focus is on identifying scenarios in which the network reconstructs the same RGB value into different spectral shapes. Essentially, this study provides crucial insights into how the network utilizes spatial information in its decision-making process.

5.1.2. Experiment Data

This case study is based on a group of metamerism samples from the NTIRE 2018 spectral dataset which the original HSCNN+ was trained on. In Figure 21, two images containing metamerism samples are presented. The image on the right displays green samples from a billboard that appear to have a similar RGB value to the grass on the left image. However, the painting on the right image has a different spectral radiance (S_p) to the grass (S_g) as shown in Figure 22.

Figure 21. Images that contain metamerism samples. The metamerism samples are the grass from the left image and the painting on the right image.

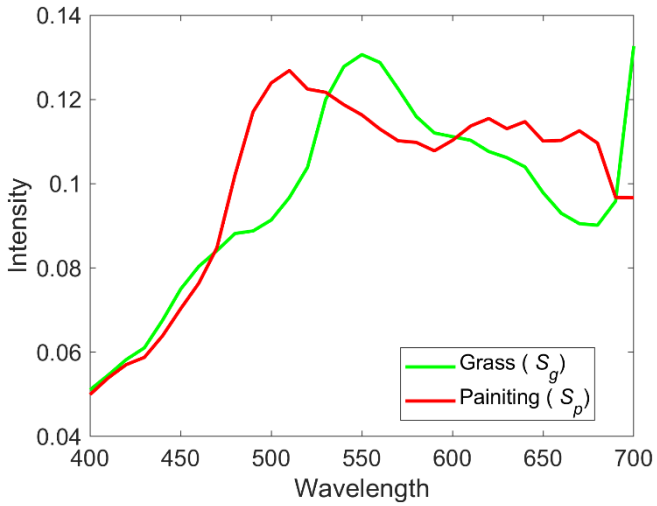


Figure 22. The spectral radiance of the two materials in the metamerism set differs. These two different-shaped spectra share the same RGB values.

The HSCNN+ could successfully reconstruct the metamerism sample from each scene into the correct shape. Therefore, we are wondering what contextual information the network used to determine the shape of its reconstruction result.

- **Visualisation of Spectral Similarity**

For visualisation, it is useful to distinguish those reconstructed spectra that are more similar to S_p or S_g. So given 2 spectra S_1 and S_2, the spectral angle similarity can be defined as:

$$SAS(S_1, S_2) = \frac{2}{\pi}\sin^{-1}\left(\frac{\sum_\lambda S_1(\lambda)S_2(\lambda)}{\sqrt{\sum_\lambda S_1^2(\lambda)}\sqrt{\sum_\lambda S_2^2(\lambda)}}\right) \quad 5\text{-}1$$

Equation 5-1 is similar to SAM but uses $\sin^{-1}$ instead of $\cos^{-1}$, and gives a value of 1 when $S_1 = S_2$. Use of the $\sin^{-1}$ will amplify small differences between similar spectra. Let the spectra for the patch reconstructed by HSCNN+ for each pixel be $S(x, y)$. Let 5-2

$$SAS_{min} = \min\left(SAS\big(S(x,y), S_p\big), SAS(S(x,y), S_g)\right) \quad 5\text{-}2$$

be the minimum similarity among all samples from the reconstructed patch when compared to either the grass or the painting reference spectrum. A visualisation can then be formed by using the red channel to show the similarity to the painting spectrum, S_p; and the green channel to show similarity to the grass spectrum, S_g:

$$R(x,y) = \frac{SAS\big(S(x,y), S_p\big) - SAS_{min}}{1 - SAS_{min}}$$

$$G(x,y) = \frac{SAS\big(S(x,y), S_g\big) - SAS_{min}}{1 - SAS_{min}} \qquad 5\text{-}3$$

$$B = 0$$

5.1.3. Reconstructing Result Based Solely on RGB Values

We begin this case study with all pixels of the input image patch having the same RGB value, devoid of textural details. Through this test, we can observe how the network reconstructs the spectra solely based on RGB values.

- **RGB Value from the Plants**

First, we tested the mean RGB value of an image patch cropped from the grass image, as shown in Figure 23.

Figure 23. The image patch from the grass.

The generated image patch is displayed in Figure 24 (a). All pixels in the input image patch have the same RGB value [27 29 19]. The input image patch is sized 140×140. Note that the billboard in the right image of Figure 21 also contains this RGB value, even though their corresponding spectral radiance is different.

In this scenario, the colour mapped visualisation clearly shows that the reconstructed data all exhibit a grass shape, since the training data consists of a significantly higher proportion of grass compared to painting samples, with a proportion of over 90%, consequently, in the absence of additional spatial information, the neural network will rely on the training data's proportion and output spectra with a grass shape.

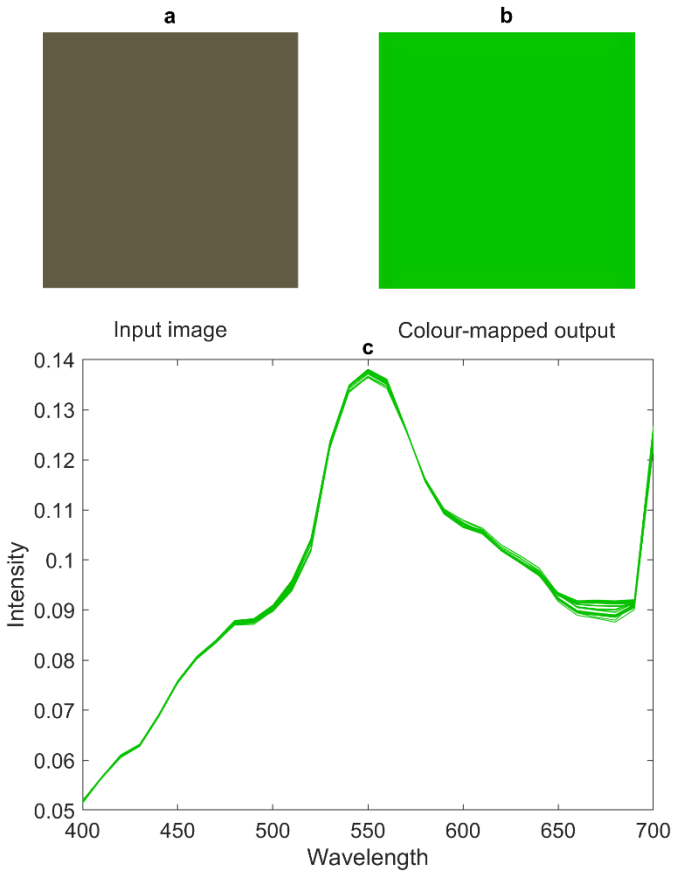


Figure 24. HSCNN+ reconstruction results when the input image patch lacks additional information. When the input RGB value corresponds to both grass and painting, all spectra have been recovered to the grass shape. (a) presents the generated input image patch. (b) displays the colour-mapped visualisation of the patch. (c) illustrates the colour-mapped reconstructed spectra from the image patch, where colour is calculated according to Equation 5-2.

- **RGB Value Purely from the Painting**

Since the shape of the reconstructed spectrum is related to the proportion of samples in the training dataset, in this test, we generate an image patch with an RGB value that can only be found in the painting scene.

Figure 25 displays the outcome of reconstructing an input image consisting solely of pixels with values [22 25 16]. The resulting visualisation depicts all pixels in a red colour, resembling the shape of S_p. This outcome was expected, as there is a one-to-one mapping between this RGB value and the spectrum in the training dataset, indicating that the HSCNN+ is capable of accurately reconstructing the spectrum to the S_p shape.

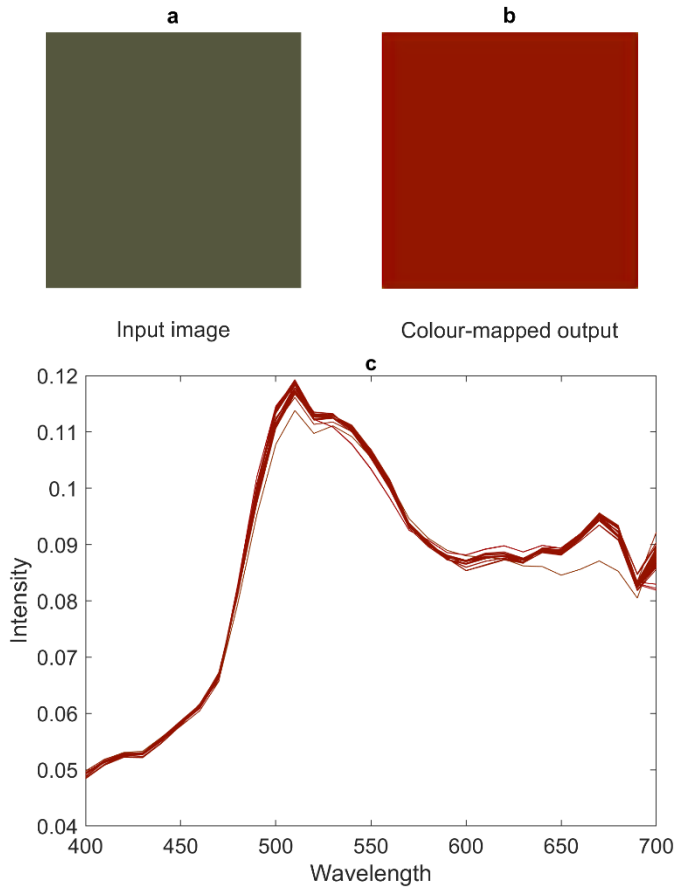


Figure 25. HSCNN+ reconstruction results when the input image patch lacks additional information. When the input RGB value purely corresponds to the painting, all spectra have been recovered to the painting shape.

- **RGB Value from the Painting**

Figure 26 shows one cropped patch from the man-made painting scene, while Figure 27 displays the reconstruction results when the input RGB values are all set to [21 23 16], which corresponds to the mean RGB value of an image patch. This specific RGB value can also be found in the grass with a higher representation in the training dataset. Consequently, the reconstructed image exhibits a grass shape.

Figure 26.Cropped patch from the painting.

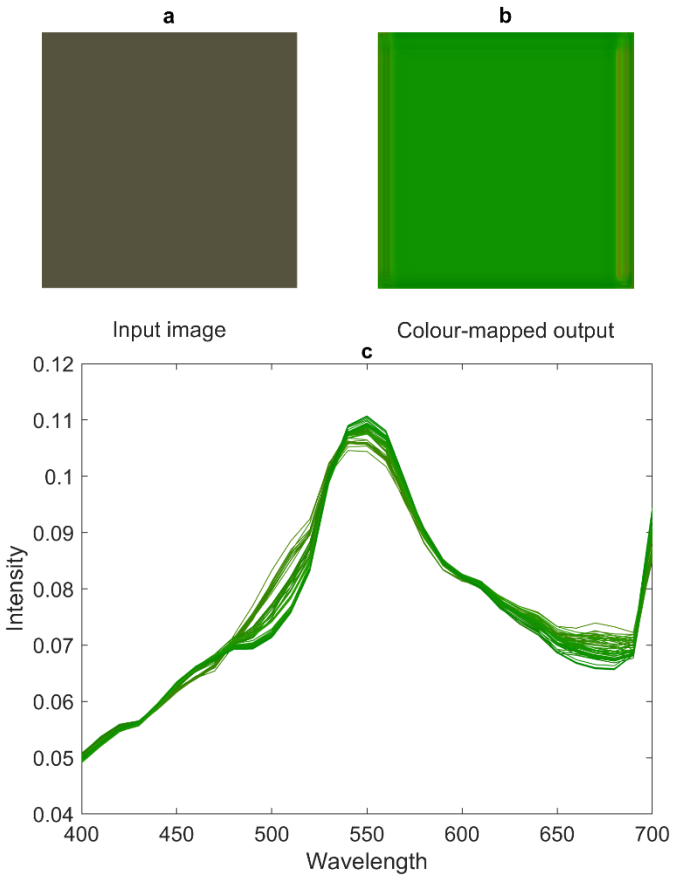


Figure 27. HSCNN+ reconstruction results when the input image patch lacks additional information. When the input RGB value corresponds to both grass and painting, all spectra have been recovered to the grass shape.

5.1.4. Reconstruction with Textural Information

When reconstructing the image patch shown in Figure 26 using HSCNN+, the resulting spectra output would exhibit a painting shape. This indicates that HSCNN+ relies on additional spatial information from the image patch to determine the shape of its output. In order to determine what specific spatial information the network is using, this test generates image patches containing two metamer RGB values. The average RGB value from the grass patch [27 29 19] is used as the background and the average RGB value from the painting [21 23 16] is used to provide additional textural features introduced into the background. We are interested in what type of spatial information would cause HSCNN+ to reconstruct the metamerism samples into the painting shape.

- **Random Samples**

The first test randomly introduces 500 samples to the background within the central 100×100 region to avoid the influence from the image edge. The resulting spatial information lacks any distinct texture.

From the results in Figure 28, none of the spectra were restored to the painting shape. The reconstructed spectra can be divided into two groups with similar shapes but differing scales, corresponding to the intensity of the different RGB values. This test indicates that randomly adding metamerism samples with no distinct texture does not lead the network to reconstruct any spectra into the painting shape.

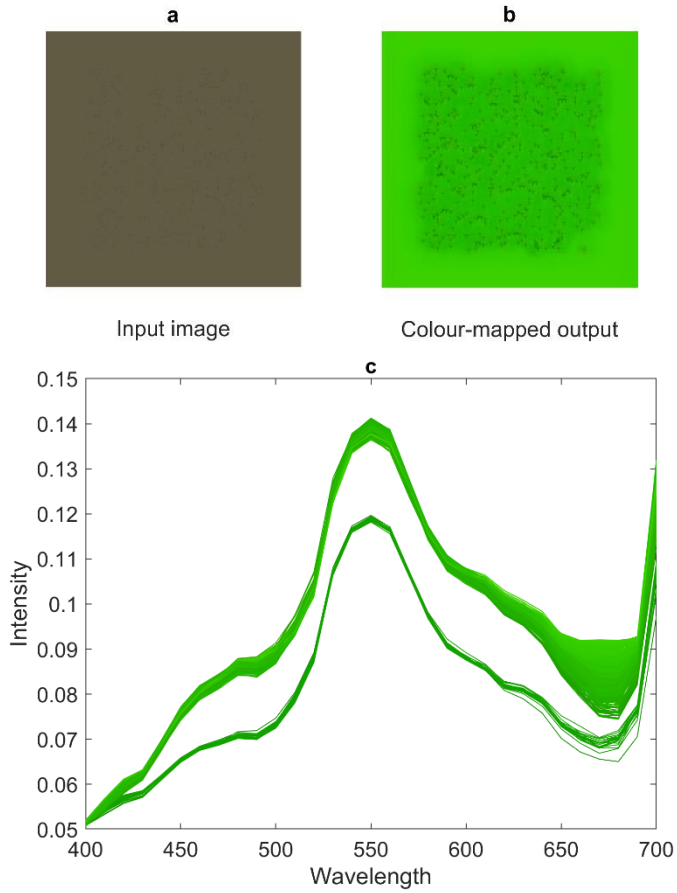


Figure 28. Reconstruction results upon adding additional information in the form of random samples with metamerism RGB values. All reconstructed spectra display a grass shape.

- **Lines**

In this test, additional information is introduced in the form of a grid. The grid lines have a width of 1 pixel and a length of 100 pixels. The RGB values in the generated image patch remain the same as in the previous test.

The reconstructed result in Figure 29 shows that **no** spectra have been recovered into the painting shape. Therefore, the straight lines as additional information is insufficient to reconstruct the spectrum into the painting shape.

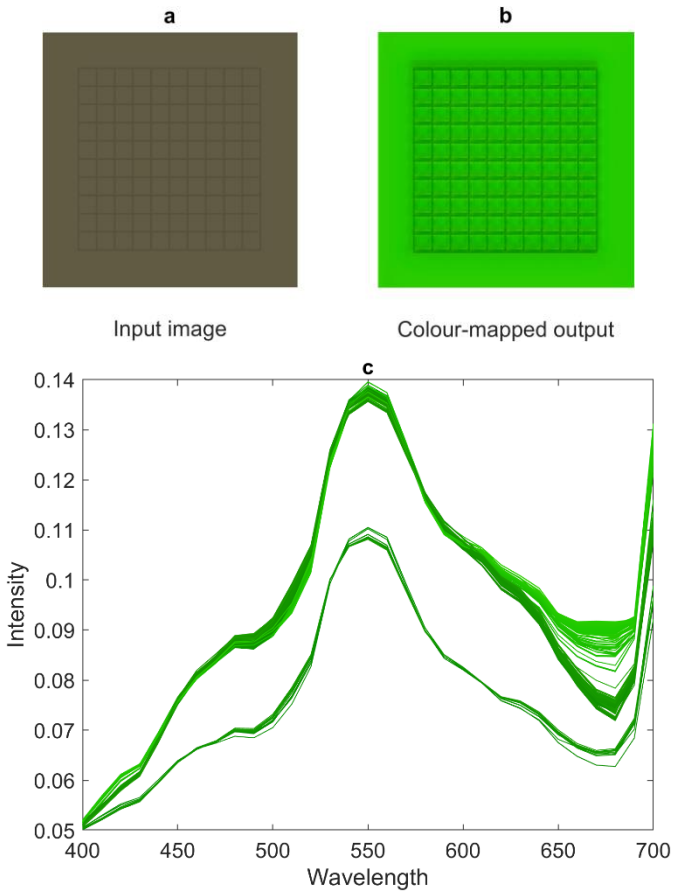


Figure 29. Reconstruction results upon adding additional information in the form of grids with metamerism RGB values. All reconstructed spectra display a grass shape.

- **Circle**

In this experiment, we introduced additional information in the form of a circular pattern with a width of 1 pixel and a radius of 30 pixels. The RGB values in the generated image patch remained consistent with the previous test.

The reconstruction results, as depicted in Figure 30, reveal that samples located along the diagonal of the circle were reconstructed to resemble the shape of the painting. This test indicates that HSCNN+ relies on textures with specific orientations as contextual cues to guide its output when tackling the metamerism problem.

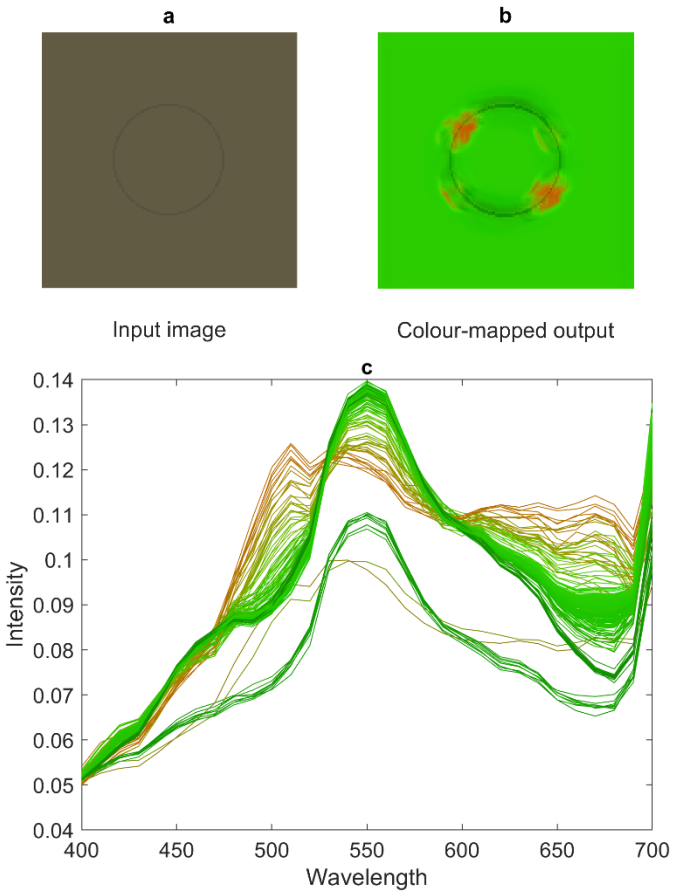


Figure 30. Reconstruction results when additional information is in the form of a circle.

- **Edges**

The edges in the painting patch are detected by a Canny filter with the default threshold in MATLAB, as shown in Figure 31. We crop the edge image to 100×100 and use the edge pixels to set the introduced pixel value as texture, as depicted in Figure 32 (a).

Figure 31. Detected edge from the painting image patch.

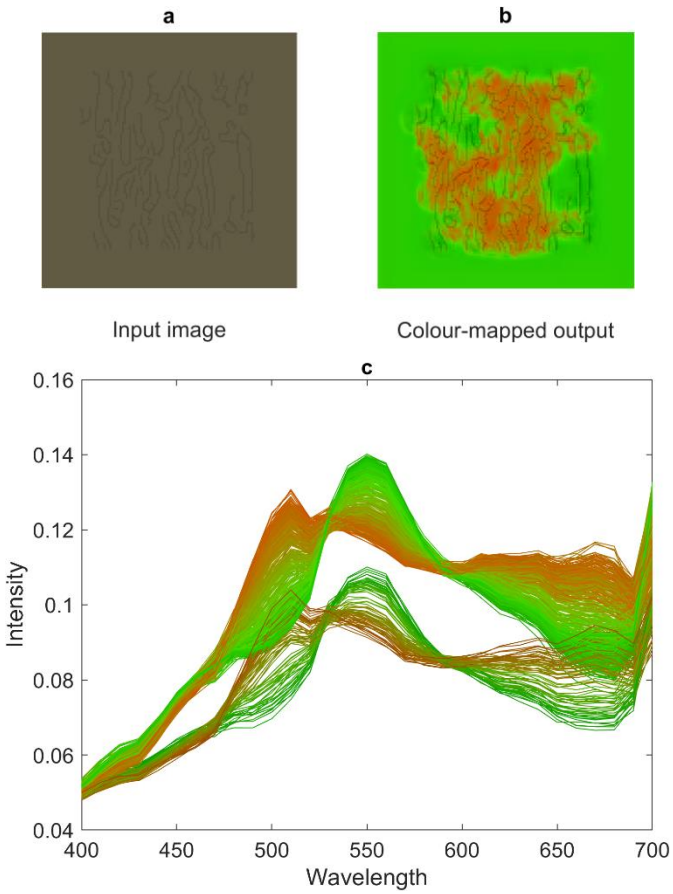


Figure 32. Reconstruction results upon adding additional information in the form of edges with metamerism RGB values. Sample with particular edge features have been recovered into the painting shape.

The reconstructed output can be seen in Figure 32. The red-coloured recovered spectra in Figure 32 (c) correspond to those which resembled the shape of the painting. Figure 32 (b) displays the location of pixels that have a painting shape. The top-left and bottom-right corners of Figure 32 (b) have been reconstructed into a grass shape, where the textural information is primarily in the form of vertical lines. Considering the previous test, this particular texture led HSCNN+ to reconstruct the metamerism sample into the grass shape. Where there is more variation of orientation the samples are reconstructed into the painting shape. This experiment suggests that HSCNN+ uses textures as context to determine the shape of its output. In this example, the network utilized texture as additional spatial information to resolve metamerism.

- **Edge with RGB Values Purely from the Painting**

In this test, we used [22, 25, 16] which is exclusively present in painting image as the RGB value of the introduced edges. The reconstructed result is displayed in Figure 33, and it reveals that almost all background pixels have been reconstructed into the shape of the painting. Since the introduced RGB value only corresponds to the painting in the training dataset, HSCNN+ uses local spatial information as context to determine the shape of the reconstructed spectrum for background pixels.

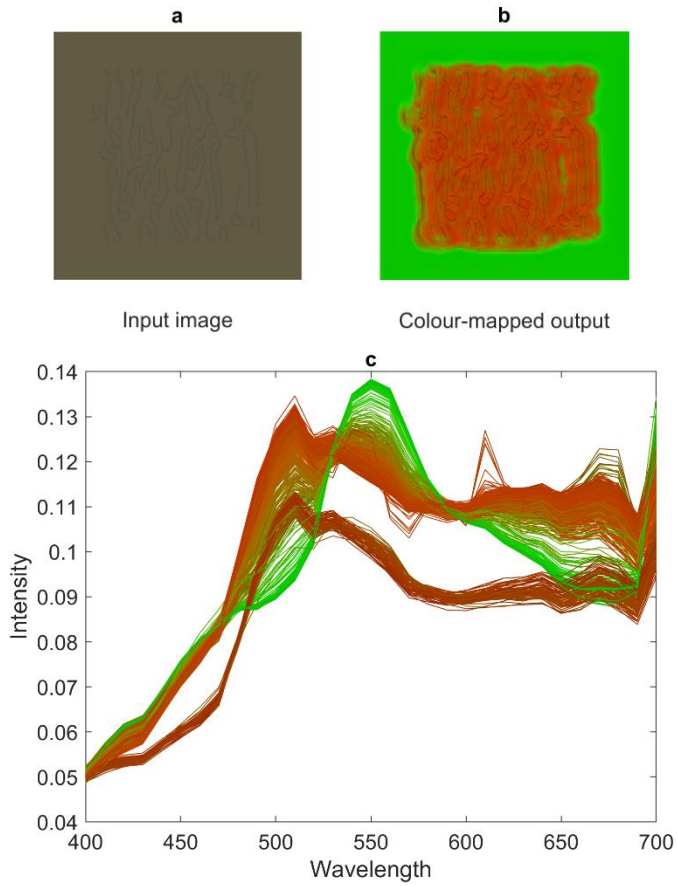


Figure 33. Reconstruction results upon adding additional information in the form of edges with RGB value purely from the painting. Most samples near edge features have been recovered into the painting shape.

5.1.5. Case Study Conclusion

In addition to the tests described above, we have also evaluated other textures as additional information for HSCNN+. Detailed results of these additional tests can be found in Appendix 3. The results from these tests further support our finding that HSCNN+ relies on specific textures to determine the shape of the reconstructed spectrum.

This experiment investigated some of the factors influencing a network's output selection when dealing with metamerism. Specifically, we examined the conditions under which the network would generate an output spectrum resembling the 'Painting' spectral shape, which was underrepresented in the training dataset. Our findings suggest that the network relies on both RGB values and spatial information (texture) to resolve metamerism. RGB values corresponding to specific materials played a crucial role in the process, while textures provided contextual cues for determining the output shape.

However, it's worth noting that this study was limited to a single metamerism sample. To gain a more comprehensive understanding of the role of spatial information, particularly texture, in resolving metamerism, we present a broader study involving multiple metamerism sets and multiple neural networks in the following sections.

5.2. Sensitivity Analysis of Spatial Information

5.2.1. Introduction

In the previous section, we examined how different textural features of the input image patch influence the spectrum recovered by HSCNN+. Although this case study yielded insightful findings, it does not provide a comprehensive overview of how networks use spatial information. To explore this more fully, we analyse the sensitivity of spatial information in single-image super-resolution neural networks.

Sensitivity analysis, in the context of machine learning and computational modelling, helps discern how the output variability of a model can be assigned to different input sources (Hooker, 2004). We investigate the impact of removing or disturbing spatial information from the input image on the reconstruction result. These modifications are called ‘attacks’ in this study. If a network is sensitive to a particular attack, it suggests the network used the feature associated with that attack in its reconstruction process.

One of the focal points of our investigation is the size of spatial information exploited by these neural networks. Knowing the necessary size of the spatial area is critical in designing our spectral-super-resolution method for selecting the neighbourhood size for examining local texture.

In addition, we want to know what features these networks predominantly rely on. Networks typically extract a wide variety of features from the input data, ranging from simple colour and intensity details to more complex textures and shapes. Analysing the potential features used by existing deep models could help us to choose appropriate features when designing our method.

Therefore, this sensitivity analysis has two parts. The first part varies the untouched neighbour sizes and aims to analyse the size of the neighbourhood used by neural networks during reconstruction. The second part with a fixed-size neighbourhood analyses the influence of different features on the networks’ performance.

5.2.2. Dataset and Networks

- **Dataset**

This experiment uses images from the NTIRE 2022 dataset, currently the largest published hyperspectral image dataset available. The RGB images used in this experiment are generated differently from the RGB values in the previous chapter. Since the neural networks tested have been trained on this dataset, we use the original RGB value provided with the dataset (Arad *et al*., 2022). The RGB values in the NTIRE 2022 dataset aim to simulate a scenario where the camera function is known but not fully controlled. When generating the RGB image, the image is scaled to a typical exposure by setting the mean RGB value to 0.18. Additionally, Poisson noise has also been added during the generation of the RGB data to simulate noise within the image capture process.

To give a diversity of colours, the metamerism datasets were gathered by utilizing the average RGB value from each coloured patch of a colour checker (with the exception of the grey and white blocks). As shown in Figure 34, one image from the NTIRE 2022 dataset displays the colour checker. To account for noise and quantization, the metamerism sets were collected selecting samples with similar colours based on the angle between RGB values. Samples from the entire NTIRE dataset within a colour angle below the threshold (0.02) with the reference RGB were included in the metamerism set. Ultimately, 18 metamerism sets, each with 20 spectral images, were collected. Since samples from the same image were likely to have a similar spectrum, each image contributed only a limited number of samples to the dataset.

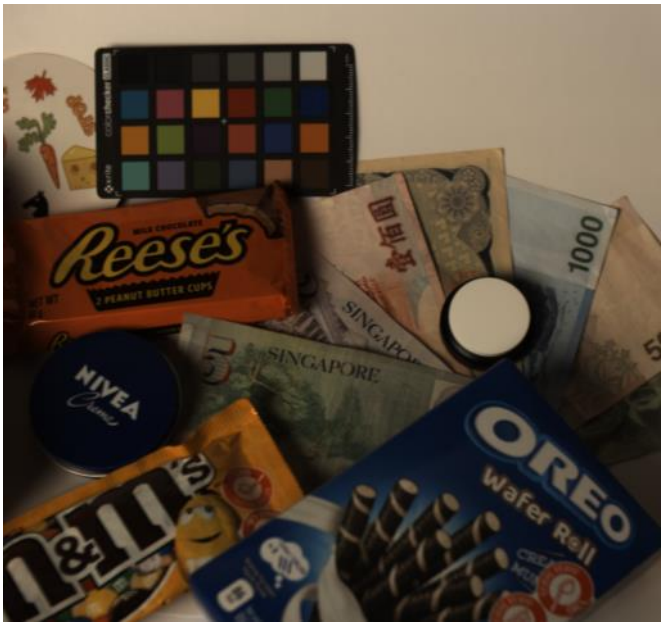


Figure 34. An RGB image from the NTIRE 2022 competition, in which the colour checker served as a reference for the collection of metamerism sets.

It is important to note that we used the RGB image from the NTIRE 2022 database in these experiments. We chose not to generate a 'clean' RGB image and retrain all the networks for a few reasons. First and foremost, our interest lies in understanding the types of information that these neural networks might be utilizing. Given that these networks were trained with noisy data, they possess the capability to extract useful information even from a disturbed image. Since our interest lies in understanding what networks have learned to resolve metamerism, applying sensitivity analysis based on the training data makes sense in this context. Secondly, retraining all networks would require a huge investment of time and resources. It also carries a degree of risk, as we do not have precise knowledge of the hyperparameters used during the original training of these networks. We are conducting this test using the training dataset because our aim is to understand what these models have learned to resolve metamerism. Therefore, testing their performance on the training data is not a concern in this context.

- **Pretrained Neural Networks**

The networks selected for the sensitivity tests are: HSCNN+ (Xiong *et al*., 2017), EDSR (Lim *et al*., 2017), HRNET (Zhao *et al*., 2020), MIRNET (Zamir *et al*., 2020), HINET (Chen *et al*., 2021), HDNET (Hu *et al*., 2022), and MST++ (Cai *et al*., 2022). All the listed networks are from the model zoo published by the author of the MST++, and all the algorithms have been pre-trained by the author and their team with the NTIRE 2022 dataset. A detailed review of the listed networks can be found in Chapter 2.

5.3. Varying Neighbour Size

In this experiment, various attacks will be introduced to the input image. These attacks are specifically designed to manipulate different types of image information. Since our focus is mainly on how spatial information can be used to resolve metamerism, hence, the pixel representing a metamerism sample (the target pixel), will be left untouched.

The image is first attacked by keeping a varying-sized circular area centred on the target pixel untouched while executing attacks on the remaining surrounding areas. The objective here is to observe how the size of the untouched area affects the networks' ability to accurately reconstruct the target sample. We are interested in the smallest size of the untouched neighbouring area that can yield a negligible difference in reconstruction accuracy. Additionally, we are interested in understanding how global information impacts the reconstruction results, especially given that newly developed methods are increasingly utilizing larger areas for analysis.

In this study, we explore untouched areas ranging from a radius of 0 pixels to a radius of 100 pixels. The maximum radius of 100 pixels was chosen considering the original image size of 512×482 and

the largest training image patch size of 256×256 used by the networks. A radius of 100 pixels covers the majority of the area a network might use for image reconstruction and reduces the chance of extending beyond the image boundaries. For some networks which could extract information from a large area, we have extended the maximum distance to 300 pixels. To give more focus to local features and because local features have a higher correlation to the target pixel, we employed a step of 1 pixel from 0 to 15 pixels. For radii from 20 to 100 pixels, we increased the step to 10 pixels. This was a strategic decision to strike a balance between experiment detail and the time required to conduct the study.

The remainder of this section introduces the attacks that have been used in this study, including removing all neighbouring spatial information, blurring the image, adding noise, and attacks in the colour channels.

In our analysis, we operate under the assumption that a significant increase in the average reconstruction error, resulting from attacks, is defined as an error greater than the sum of the average error observed with the untouched image and its associated standard deviation. Furthermore, we will be closely studying the distribution of these reconstruction errors when the input image has been subjected to attacks. This detailed examination will help elucidate the effects of our various attacks on the performance of the image reconstruction process.

5.3.1. Remove All Information Except the Neighbour Pixels.

In this test, all pixel values outside the untouched area centred on the target pixel are set to zero, effectively eliminating all additional information. Figure 35 provides an example of an attacked image where the untouched neighbouring region has radii of 20 and 100 pixels. In this scenario, all information except the pixels adjacent to the target sample has been removed.

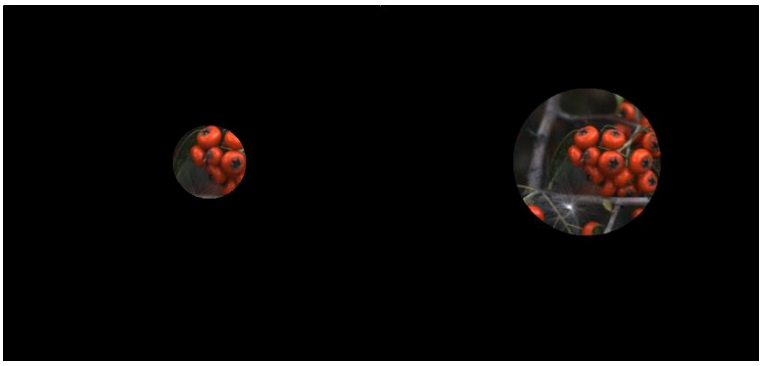


Figure 35. Images after the attack, where all information except for the areas surrounding the target with radii of 20 and 100 is represented in black.

We aim to investigate the minimum necessary spatial size that these networks utilize during their reconstruction process. We give two hypothesises, **first** as we increase the size of the untouched area, the reconstruction accuracy of the networks will improve. **Second**, we assume there is a minimum untouched spatial range the network requires to achieve a reconstruction accuracy similar to that achieved when the whole image is untouched (best performance). The results from our sensitivity analysis will help validate or refute these assumptions.

We use the results from HRNET as an example to analyse how the size of the untouched neighbouring region impacts the reconstruction accuracy. Given that HRNET was trained on 256×256 image patches, it potentially can extract information from a relatively large area within the image. This study specifically uses the SAM to measure reconstruction error since it is sensitive to the shape of the reconstructed spectrum. Results based on other error measurements are presented in Appendix 4.

Figure 36 displays the mean SAM derived from all samples from collected metamerism sets as a function of the radius of the untouched image area. The black line represents the average SAM when the input image is untouched (the best performance HRNET achieved). The dotted line indicates the average SAM plus one standard deviation when the image is untouched.

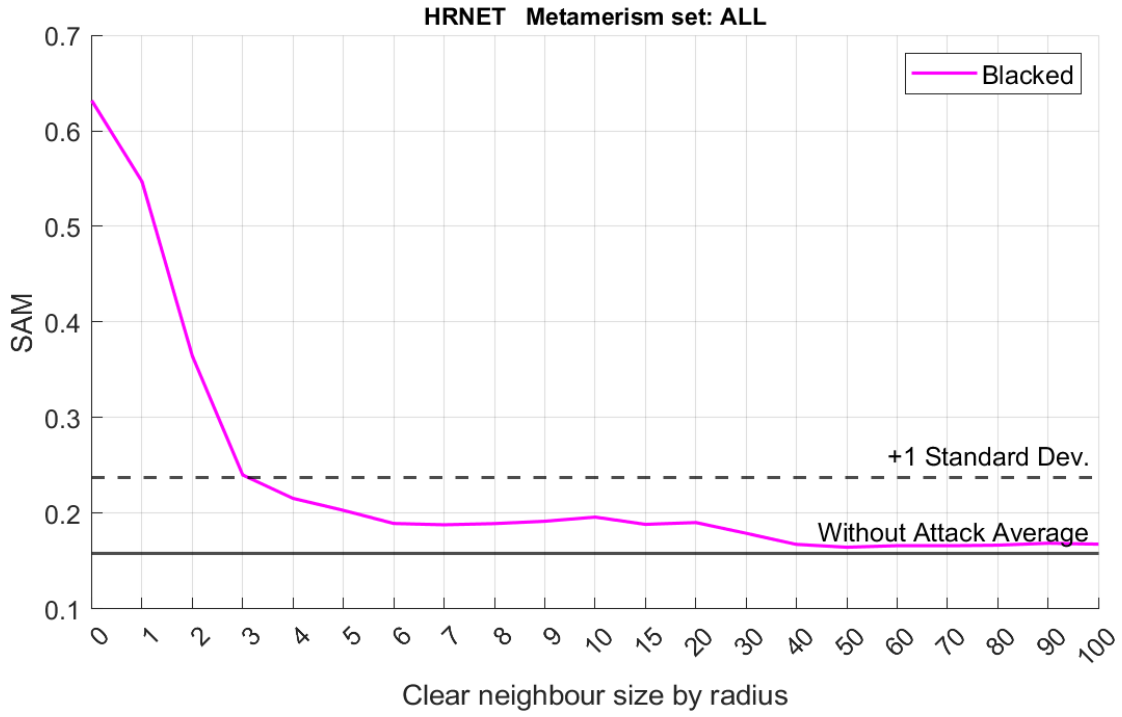


Figure 36. The average SAM from HRNET as a function of the untouched neighbour size when all information has been removed except the neighbour area. HRNET is sensitive to close neighbour pixels.

As the size of the untouched neighbouring region increases, the reconstruction error decreases. The plot slope shows a sharp decline when the untouched neighbouring region is less than 3 pixels in radius. However, when the untouched neighbouring region exceeds 3 pixels, the plot begins to flatten until the radius surpasses 20 pixels. Once the untouched neighbouring region expands beyond 40 pixels, the mean SAM of the altered image virtually overlaps with the accuracy when the input image is untouched.

Given that all information outside the untouched neighbouring region was set to black in this study, it suggests that for HRNET, an untouched image patch with a radius of 40 pixels provides the necessary information to accurately reconstruct the centre pixel. Additionally, HRNET exhibits high sensitivity to nearby pixels. Since closely located pixels usually share similar attributes, they are highly correlated, especially in images with smooth gradations of colour or intensity. Therefore, it makes sense for a network to have higher sensitivity of close neighbour pixels of the target.

Figure 37 illustrates the mean SAM as a function of the radius for HSCNN+. It is evident that HSCNN+ achieves its best performance with a smaller radius than HRNET. This observation suggests that HSCNN+ also relies on local spatial information to resolve metamerism.

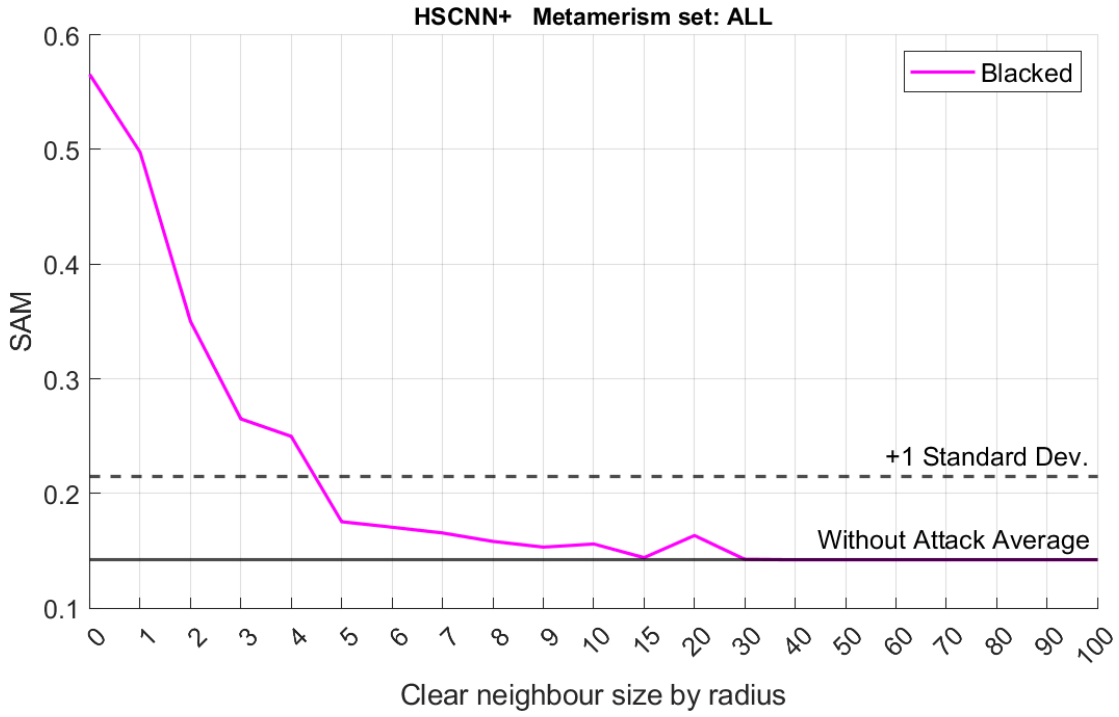


Figure 37. The average SAM from HSCNN+ as a function of the untouched neighbour size when all information has been removed except the neighbour area. HSCNN+ is sensitive to close neighbour pixels.

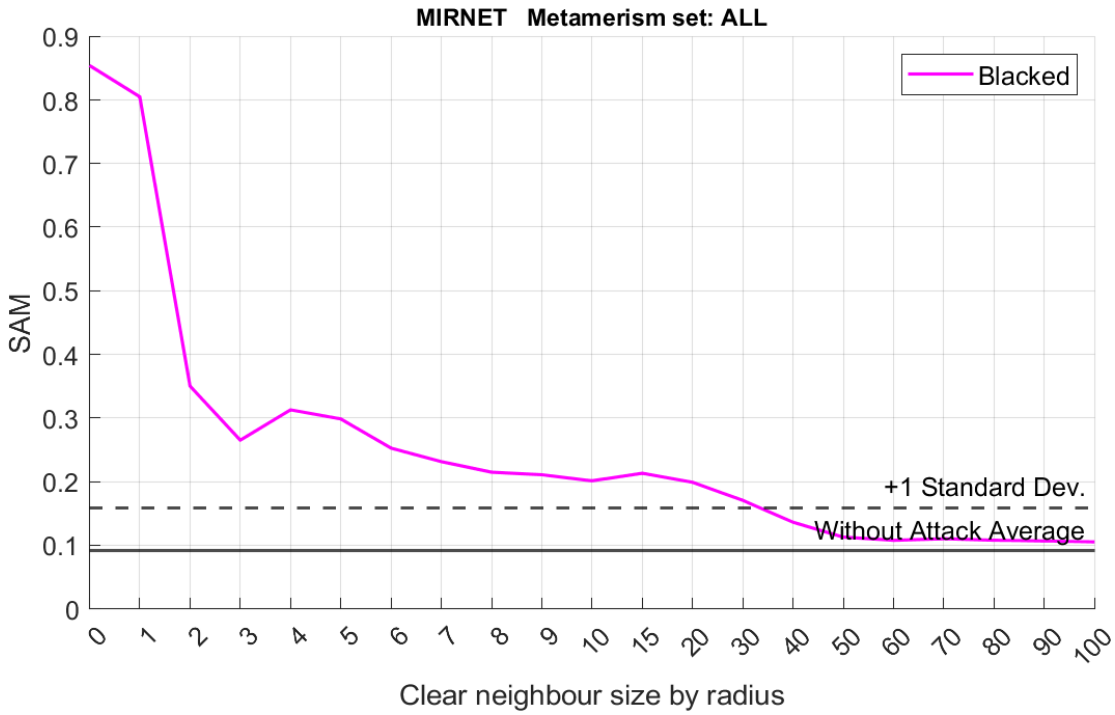


Figure 38. The average SAM from MIRNET as a function of the untouched neighbour size when all information has been removed except the neighbour area. MIRNET is sensitive to a larger range.

Figure 38 illustrates the mean SAM as a function of the radius for MIRNET, a network capable of extracting global spatial information during the reconstruction process. In comparison to HRNET and HSCNN+, MIRNET exhibits relatively higher sensitivity to the attack, resulting in a higher error compared to its best performance. However, it's important to note that the error plot follows a similar trend to that of the local networks. This suggests that even though MIRNET can extract global information, local spatial information remains important for its reconstruction accuracy. Results from other listed networks are presented in Appendix 4, they all have a similar trend with higher sensitivity to attacks in the local range.

5.3.2. Blurring the Image Outside the Neighbour Area

High-frequency spatial information encompasses edges, corners, and fine details. Notably, these edges and corners in the spatial domain often correspond to boundaries between materials. Consequently, these edges and corners can correspond to abrupt changes in the spectral domain, a factor that introduces complexity to the spectral reconstruction process. This relationship between changes in the spatial and spectral domains highlights the importance of these features during spectral reconstruction. Moreover, textures, which represent pixel intensity variations, could correspond uniquely to specific materials. Such a correlation suggests they might provide context

that aids networks in resolving metamerism during reconstruction. The texture could be associated with different frequencies.

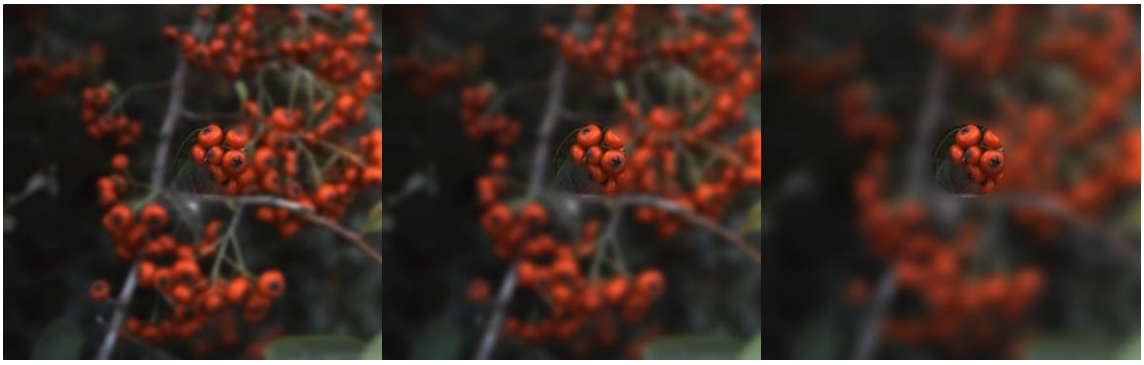


Figure 39. Image after the attack, where all information except for the areas surrounding the target with radii of 50 pixels is blurred with a gaussian filter with σ equal to 3, 5 and 10 pixels respectively.

Given this, we applied several low-pass filters to the original images, aiming to evaluate the performance of the networks when high-frequency information is eliminated. The attacked images underwent three levels of blurring, using Gaussian filters with standard deviations of 3, 5, and 10 pixels. Figure 39 illustrates an attacked image with a clear neighbour size of 50 pixels, from left to right lower frequency spatial information is removed from the rest of the image in the attack. By applying low-pass filters to the images, we also remove high spatial frequency noise. This step allows us to analyse how such noise affects reconstruction accuracy. If an algorithm is seriously overfitted, it may even treat noise as meaningful information. Consequently, if a network does not respond to attacks that remove very high frequencies, this could suggest that the network is robust to noise in a certain respect.

Figure 40 depicts the reconstruction accuracy as a function of the untouched neighbouring area size when the rest of the image is blurred. The plots are similar to the shape observed in Figure 36. When the untouched neighbouring area is larger than 30 pixels, HRNET can achieve a reconstruction accuracy like its best performance.

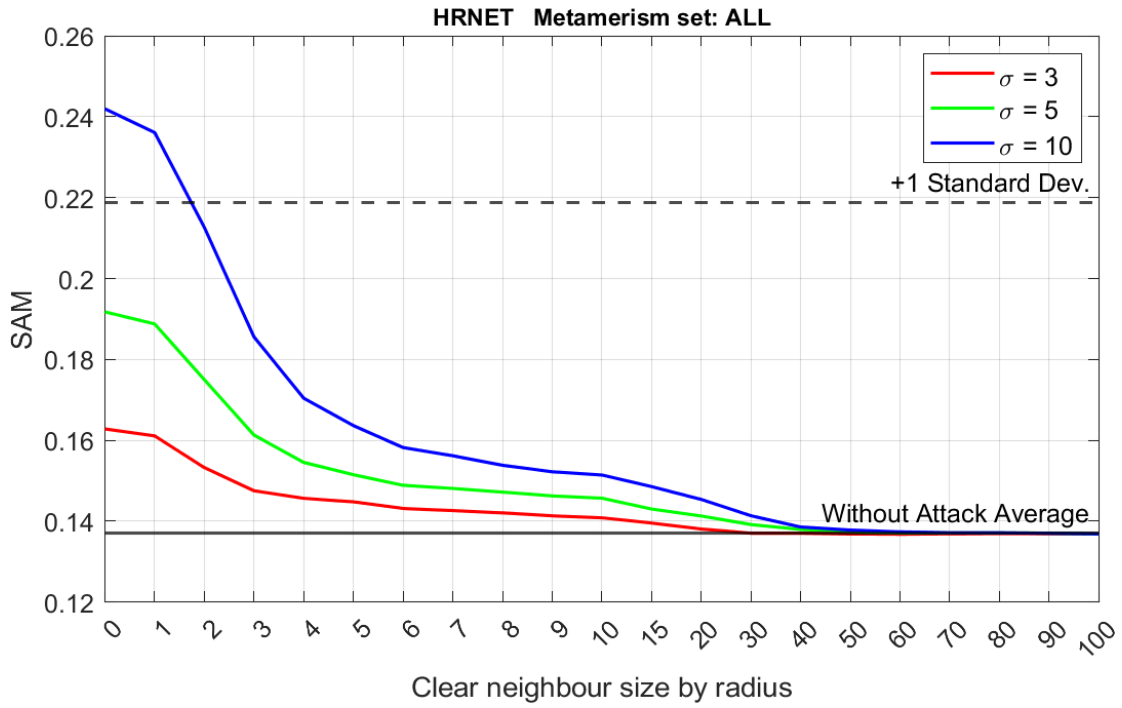
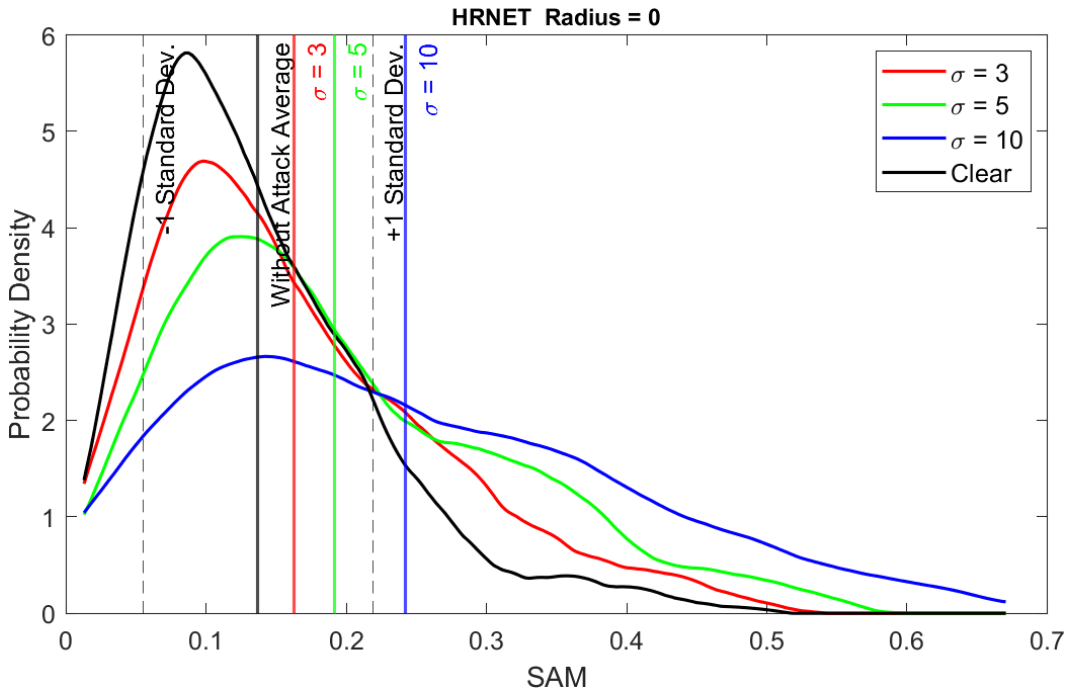


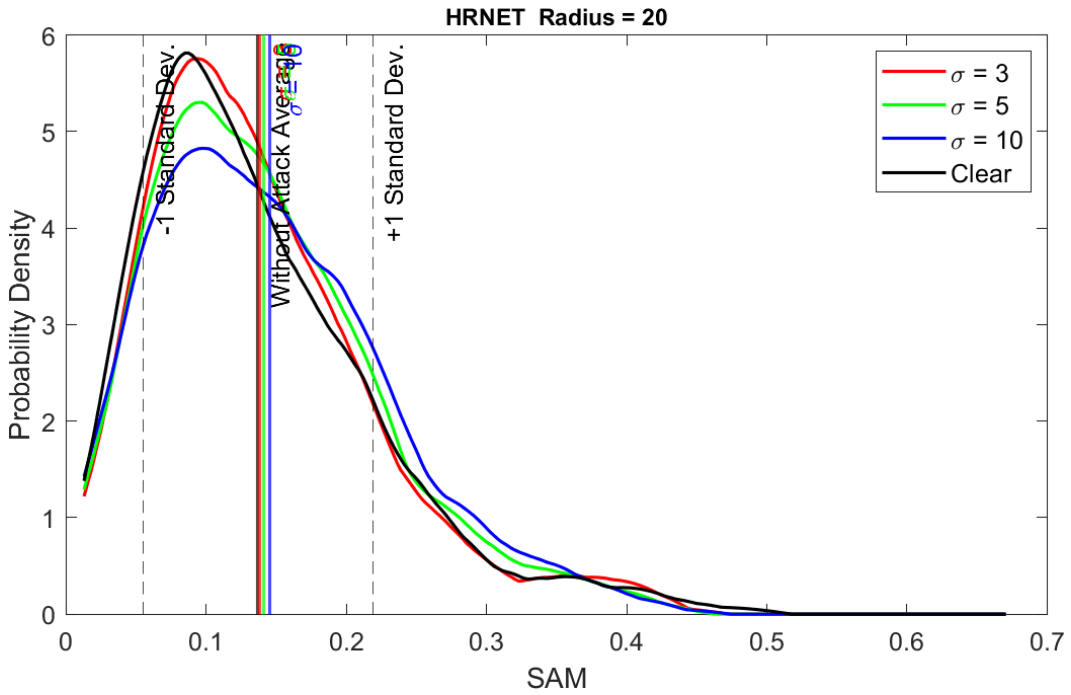
Figure 40. The average SAM from HRNET as a function of the untouched neighbour size when the rest of the image is blurred. When the size of the clear neighbour is larger than 30 pixels, HRNET has sufficient information for reconstruction.

When the neighbouring area is significantly blurred, the reconstruction has a higher error than when the image is slightly blurred. A larger error implies that the information the network relies on has been suppressed. The higher error resulting from the use of a Gaussian filter with a larger standard deviation (σ) can be explained by the fact that lower spatial frequency generally contains higher information energy. Therefore, using a filter with a larger σ increases the likelihood of removing useful information within the low to intermediate frequency band. The influence from attacking the image with a Gaussian filter with $\sigma = 3$ causes a relatively slight influence on the reconstruction accuracy, which indicates that HRNET relies less on high-frequency spatial features, and more on low to intermediate frequency.

Figure 41 (a) and (b) provide a clearer illustration of how the reconstruction accuracy shifts with changes in the size of the untouched neighbouring area.



(a)



(b)

Figure 41. Distribution of SAM of all collected metamerism samples when the untouched neighbour size is 0 and 20 respectively. When the untouched neighbour size is larger than 20, the error distribution is close to the best performance of HRNET.

The black curve in both figures represents the distribution of the SAM of HRNET's original performance on the selected samples, with the associated mean and standard deviation of SAM

displayed as vertical black lines. The colour curves and lines depict the distribution and mean of the SAM when the remainder of the image is blurred. When the untouched neighbouring area size is 0, the SAM distribution diverges significantly from HRNET's best performance. However, when this area extends to 20 pixels, the SAM distribution for the blurred remainder of the image largely overlaps with the original performance. This suggests that, without detailed textural information from the neighbouring area, it is challenging for HRNET to accurately recover spectra. Conversely, when the untouched neighbouring area spans 20 pixels, the local information within that region appears sufficient for HRNET to reach its best performance.

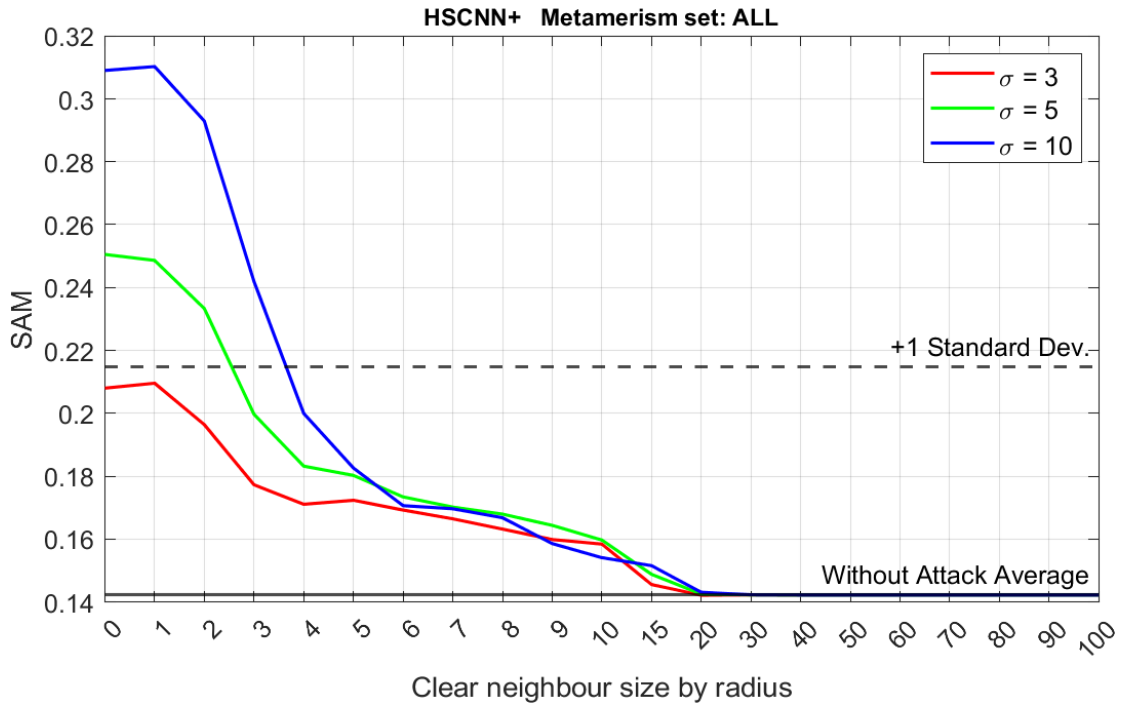


Figure 42. Responses of HSCNN+ (which extracts information in a relatively small area) when the image is blurred. When the attack happens out of the feature extraction area, the network has not been influenced by the attack (HSCNN+ extracts information from a 50×50 area).

In addition to the HRNET, we also assessed other networks to analyse the requisite size of the untouched neighbouring region to maintain comparable reconstruction accuracy when the whole image is not altered. Here, we present two examples. The first example involves HSCNN+ shown in Figure 43, which extracts information from a relatively small 50×50 area. If the untouched neighbouring area is larger than this feature extraction zone, any alterations to the remainder of the image do not affect the reconstruction accuracy of HSCNN+. In alignment with HRNET, HSCNN+ is particularly sensitive to alterations in the close neighbouring pixels (radius less than 3). When the distance from the target pixel exceeds 5, gradually expanding the size of the untouched neighbouring area incrementally enhances reconstruction accuracy, until approximating the accuracy attained when the image is untouched. If the radius of the untouched neighbouring region

is greater than 15 pixels, the distribution of the reconstruction error is close to that of the untouched result.

Unlike HRNET and HSCNN+, some networks rely on global information. Figure 43 illustrates how MIRNET responds to changes in the size of the untouched neighbouring area. Even though MIRNET was trained on image patches of size 128×128, it needs almost the entire image (untouched size larger than 300 pixels) to be unaffected to achieve its best performance. MST++ and HINET also require a substantial untouched area to perform optimally (presented in Appendix 4).

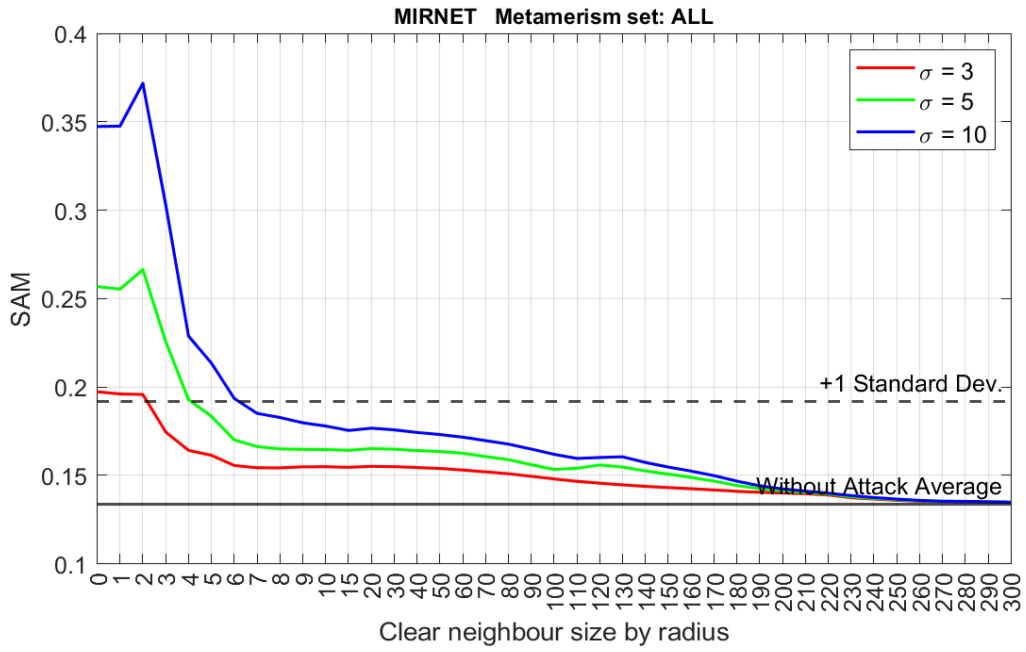


Figure 43. Responses of MIRNET (which extracts global information) when the image is blurred. The reconstruction gets close to the best performance when almost all image is untouched. (MIRNET was trained on image patches sized 128×128).

MIRNET also demonstrates high sensitivity to the close surrounding pixels, more than the pixels farther from the centre. This is indicated by a sharp drop in error when the untouched neighbouring area is larger than 5 pixels. Despite needing almost the entire image to reach its peak performance, MIRNET's reconstructed spectral shape changes slightly as the size of the untouched area increases. Figure 44 presents the reconstructed spectrum of a sample point as the untouched neighbouring area varies. Upon analysis of other samples, it is observed that increasing the untouched neighbouring area, when it's already larger than 30 pixels, has a minimal effect on the shape of the spectrum, but it does influence the scale. The right-hand plot in Figure 44 right displays the rescaled recovered spectrum when the untouched neighbouring area is 20 pixels in size. It can be observed that the shape of this rescaled spectrum nearly overlaps with the spectrum derived when the untouched size is as large as 300 pixels. It seems that MIRNET requires global information

to estimate a suitable scaling factor for the reconstructed spectrum. However, the shape of the reconstructed spectrum can be determined within a relatively local area. This suggests that while global information is important for scaling, the local textural information is critical in shaping the output spectrum in MIRNET's processing.

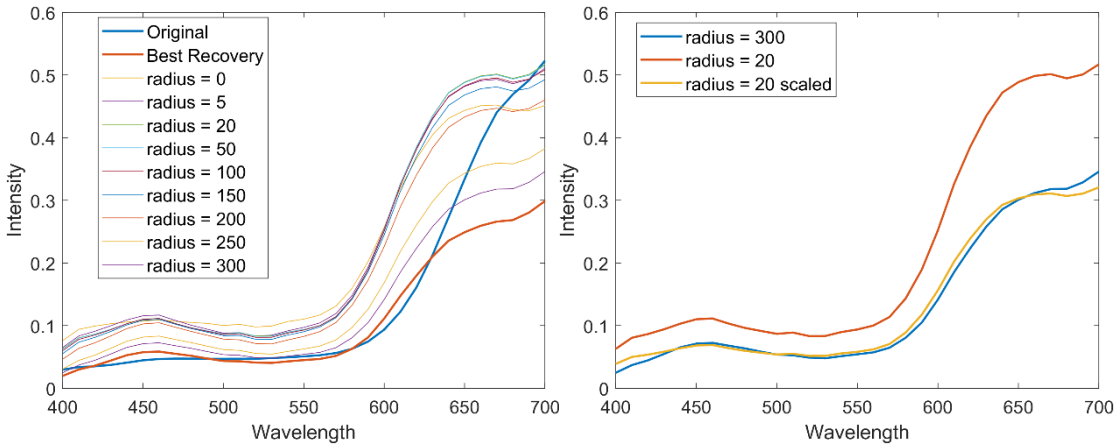


Figure 44. Left: This shows an example of a recovered spectrum that changes based on varying sizes of untouched neighbouring areas. Right: Here, we see a rescaled spectrum when the untouched neighbouring area is 20 pixels. As the untouched neighbouring size increases, it impacts the scale of the recovered spectrum but influences the shape to a lesser extent.

5.3.3. Adding Noise to the Image Outside the Neighbour Area

The influence of noise on images, leading to colour distortion, texture alterations, and loss of detail, can severely impact the accuracy of spectral image reconstruction. To understand the resilience of spectral super-resolution methods against such disturbances, this study introduces noise of various frequency characteristics into the images and investigates their impact on the reconstruction process.

Figure 45. The images after the attack when noise is added except for the areas surrounding the target with radii of 50 pixels; from left to right, each contains white, 'pink,' and 'blue' noise.

In our experiment, we added white Gaussian noise with a mean of 0 and a standard deviation of 10 (/255) to the image. An example of such a noised image, except for a circular area centred on the

target pixel, is shown on the left side of Figure 45. Beyond white noise, we also considered noise at different spatial frequencies. To simulate this, we transformed the white noise to create signals that imitate pink and blue noise. More specifically, we applied a low-pass filter to the white noise, creating a signal with more power at lower frequencies, similar to pink noise. Conversely, we used a high-pass filter to generate a signal that carries more power at higher frequencies, akin to blue noise. Figure 45 illustrates examples of 'pink' and 'blue' noised images in the middle and right panels, respectively. The inclusion of noise at different spatial frequencies allows us to evaluate the networks' sensitivity to disruptions in spatial information and to identify if there's a specific type of spatial information associated with a frequency that these networks prefer during the reconstruction process.

Figure 46 presents the SAM as a function of the size of the untouched image area when noise is added to the neighbour area. Compared to other forms of attacks, adding noise to the image seems to have less impact on HRNET's performance. This could be due to the fact that HRNET was trained on a noisy dataset, which has likely enhanced its robustness against noise. However, the impact becomes more significant as the noise level increases. Considering that lower spatial frequencies carry high energy, introducing decomposed low-frequency noise (pink noise in this study) only slightly impacts the signal-to-noise ratio in the low-frequency domain, resulting in an insignificant impact on HRNET's performance.

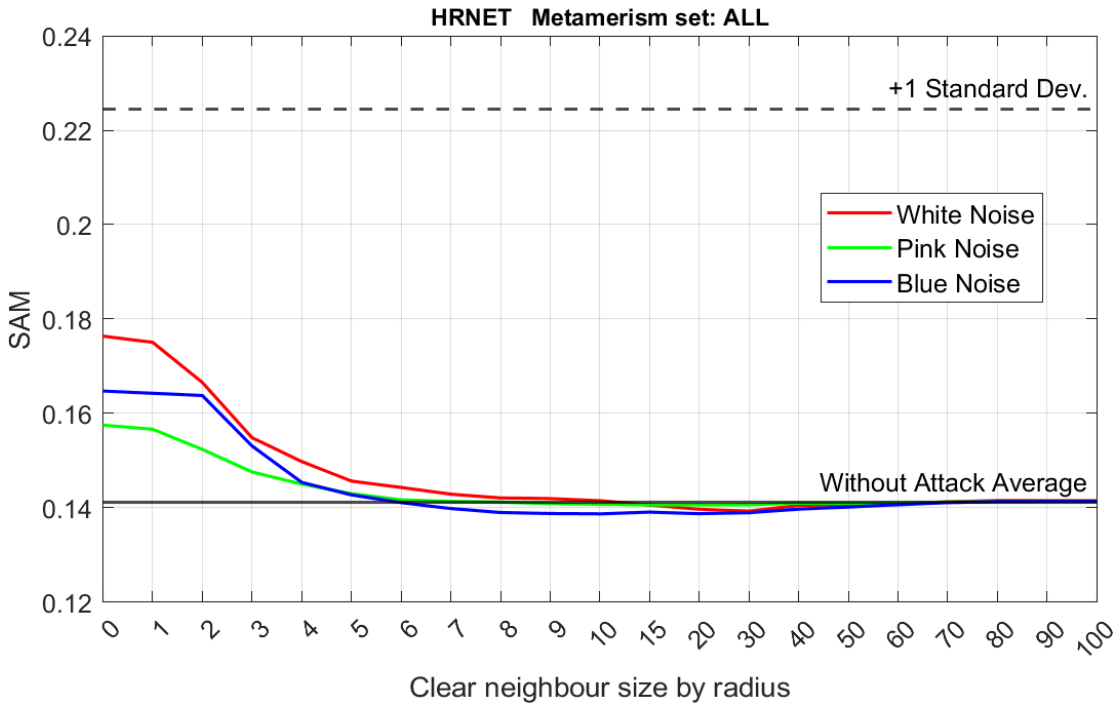


Figure 46. The SAM for HRNET as a function of the untouched neighbour size when the rest of the image is noised. When the untouched neighbour size is larger than 10 pixels, HRNET gets sufficient information for reconstruction.

The plot in Figure 46 shares the same overall shape as previous plots. This suggests that HRNET is most sensitive to its close neighbours, and when the area of clear neighbours is larger than 10 pixels, the network has sufficient information to perform optimally. This further underscores the importance of the immediate surrounding area in influencing the performance of HRNET. Results from other listed networks are presented in Appendix 4. All the networks in the list are more sensitive to noise in the local area, even those considered global networks that require a larger untouched area to achieve their best performance.

5.3.4. Attacks in the Colour of the Input Image

In this experiment, the attacks are designed to specifically target the colour information, which corresponds to the spectral data of the image. Rather than directly modifying the RGB values, we convert the RGB image into the YCbCr colour space and launch the attacks on the converted image separately in the Cb and Cr channels. The purpose of conducting the attack in the YCbCr space instead of the RGB space is to preserve the luma information intact. The colour of neighbour pixels can carry important spectral information, so we hope studying the impact of colour attacks can provide insights into how colour information contributes to reconstruction accuracy.

Figure 47. The images after the attack, with an untouched neighbour sized at 50 pixels, are displayed in pairs. The top two images show the attacked image when the Cb and Cr channels are blurred, while the bottom two images show the image when noise is added to the Cb and Cr channels.

Four different attacks were considered (see Figure 47): applying a Gaussian filter with a standard deviation of 10 pixels separately to the Cb and Cr channels; and adding white Gaussian noise ($\sigma = 10$) separately to the Cb and Cr channels. Then, the image in the YCbCr space is converted back into the RGB space.

In Figure 48, when an attack alters the spectral information in the neighbouring area within the YCbCr colour space, it significantly affects HRNET's reconstruction results. This suggests that HRNET not only extracts structure information from the neighbouring area but is also sensitive to the colours of neighbouring pixels.

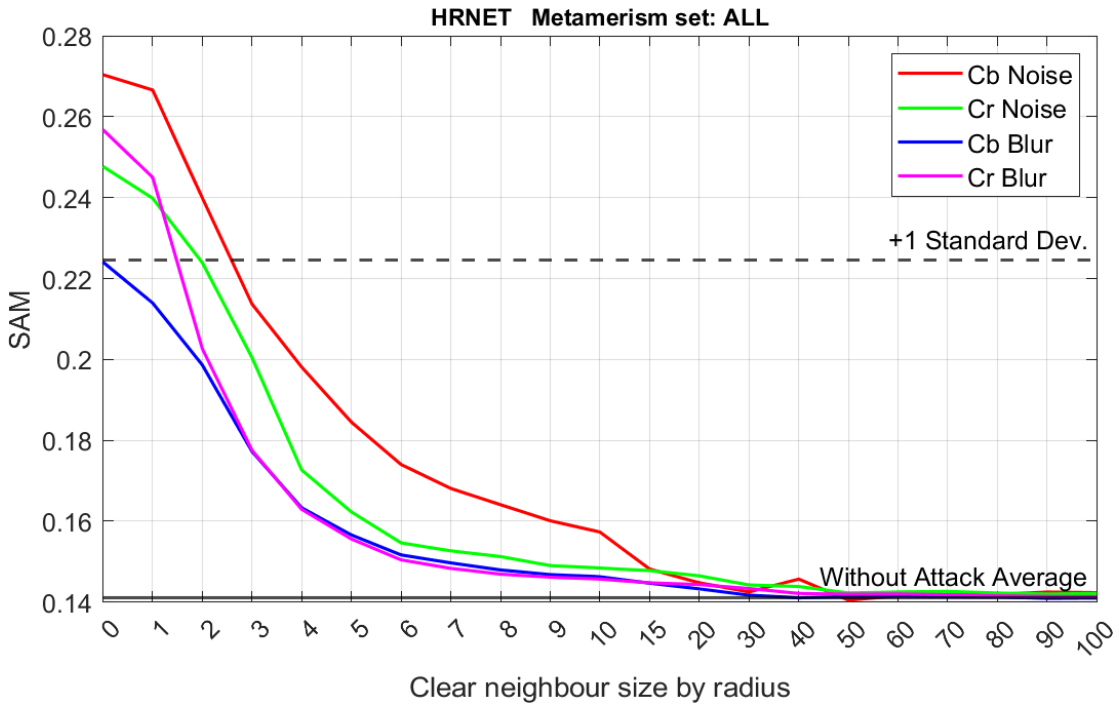


Figure 48. The SAM from HRNET as a function of the untouched neighbour size when the rest of the image is attacked in YCbCr colour space.

Similar to previous results, the immediate neighbouring pixels have the most significant impact on the reconstruction outcome. Additionally, once the size of the untouched neighbouring area exceeds 30 pixels, increasing this area further does not influence the reconstruction result, indicating that HRNET can extract adequate information from a local vicinity for effective spectral reconstruction. Results from other listed networks are presented in Appendix 4, all the networks in the list are more sensitive to attacks in the local area.

5.3.5. Discussion

Based on the previous analysis, it's evident that the EDSR, HRNET, and HSCNN+ primarily focus on local information, while HINET, HDNET, MIRNET, and MST++ depend on both local and global information to resolve metamerism. Results not shown here can be found in Appendix 4. However, all networks demonstrated high sensitivity to very local information. The impact on the shape of the recovered spectrum from neighbouring pixels decreases with increasing distance from the target pixel.

For local networks, an untouched area with a radius of 20-30 pixels provides the networks with sufficient information to recover the spectrum. For networks relying on global information, local information dictates the shape of the recovered spectrum, while global information primarily influences the spectrum's scale. Consequently, when designing our explainable feature extraction method, we could concentrate on the local area—for example, a 31×31 image patch centred on the target pixel, or even smaller for efficiency considerations.

However, the information in the local area is still complex. To efficiently extract useful information to resolve metamerism, choosing the right texture features is crucial. To address this, we have further analysed the sensitivity with a fixed neighbour size.

5.4. Different Textures with Fixed Neighbour Size

In this experiment, we standardized the attack area to be a circle with a radius of 20 pixels, centred around each target pixel and kept the pixels outside the attacked area untouched. As local spatial information has been demonstrated to be essential during the reconstruction process, our focus in this section shifts towards understanding the type of local spatial information that neural networks are likely to utilize to resolve metamerism. By selectively affecting certain types of spatial information and observing the networks' ability to resolve metamerism under these conditions, we can gain deeper insights into the kind of information that is important for spectral reconstruction. As demonstrated in 5.1, local textural features hold promises as contextual information when resolving metamerism. Therefore, we design the attack by Fourier Domain decomposition of the local patch and modify spatial information (especially local textures) corresponding to a certain frequency and orientation. The goal is to identify some of the features that are important to networks reconstructing the spectrum, thereby guiding us in designing our explainable method.

5.4.1. Attack the Image with Band Stop Filters.

We start this analysis by filtering the image with a group of band-stop filters, which can help us determine whether there is a preference for particular spatial frequencies in neural networks. Figure 49 depicts the masks used in this experiment to partition the Fourier domain based on spatial frequencies. These masks are applied in the Fourier domain to remove information corresponding to the highlighted regions. The left mask acts as a low-pass filter, while the three masks on the right function as band-stop filters.

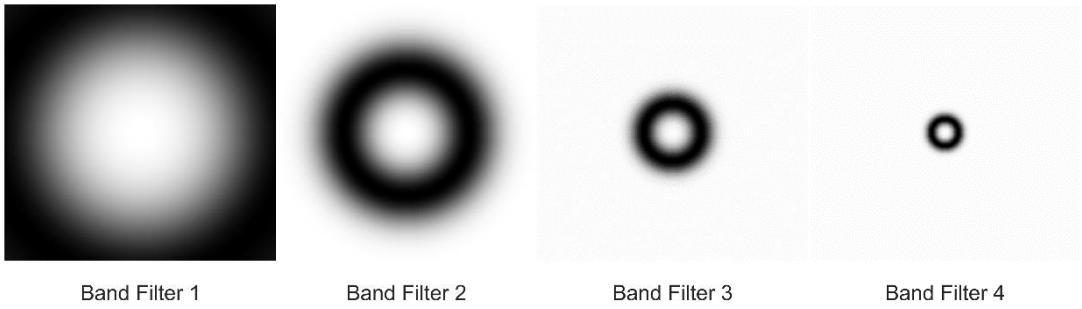


Figure 49. Band stop filter masks in the Fourier domain. The Information corresponding to the highlighted part will be removed from the attacked image.

The parameters of filters (masks) were designed to decompose the Fourier domain of an image sized 512×512. The magnitude of the highlighted part of the mask follows a Gaussian distribution in the radial direction. Figure 50 provides a visual representation of the filter magnitudes as a function of their distance from the centre of the image (frequency) within the Fourier domain. The parameters of each band stop filter are shown in Table 14.

Table 14. The parameters of the band stop filters are shown in Figure 49.

	Centre	50% width
Filter 1	182.15	-
Filter 2	82.66	31.63
Filter 3	36.90	14.12
Filter 4	16.47	6.30

To ensure smooth transitions between neighbouring masks, the half-peak magnitude (0.5) of the filter responses is designed to intersect. This careful design is implemented to ensure a comprehensive and accurate representation of the image's frequency information.

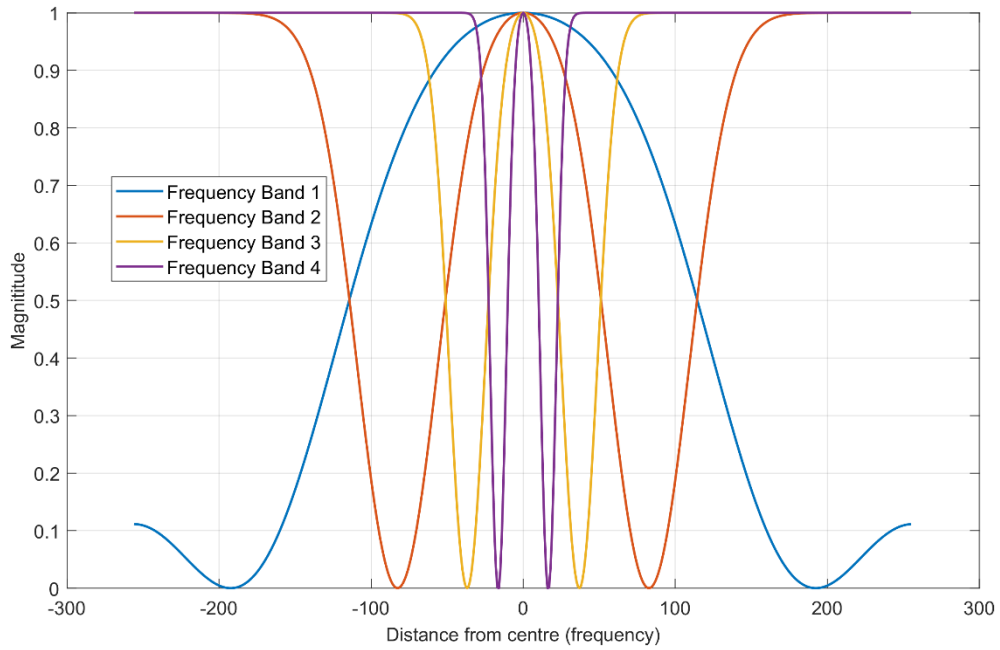


Figure 50. The magnitude of band-stop filters as a function of the distance from the image centre.

When applying the filter, it is necessary to pad the RGB image to 512×512 pixels. The padded image is transformed into the Fourier domain attacked by multiplying by the corresponding filter function, then transformed back to the image domain.

The "filtered" image is finally combined (fused) with an untouched image to generate the final attacked image, with only the area centred around the target sample being filtered. To avoid introducing additional edges, the boundary between the "untouched" and "filtered" areas has been smoothed by applying a Gaussian weight matrix, minimizing the introduction of extra edges in the final image. This process is illustrated in Figure 51.

The rationale behind separating the information in the image by their spatial frequencies is that each frequency band is typically associated with different scales of textures. A texture with a fine scale, such as sand or fine-grained wood, has high-frequency components because there are many small details that change rapidly in a small area. Conversely, a texture with a coarse scale, such as large bricks or wide-grained wood, has low-frequency components, as the details change less rapidly over the same area. These textures will result in energy that is concentrated closer to the centre of the Fourier image. By systematically eliminating different frequency bands from the image and observing how the networks respond, we can gain a better understanding of which frequencies (and therefore, which textures) are important for the networks' ability to accurately reconstruct the image and resolve metamerism.

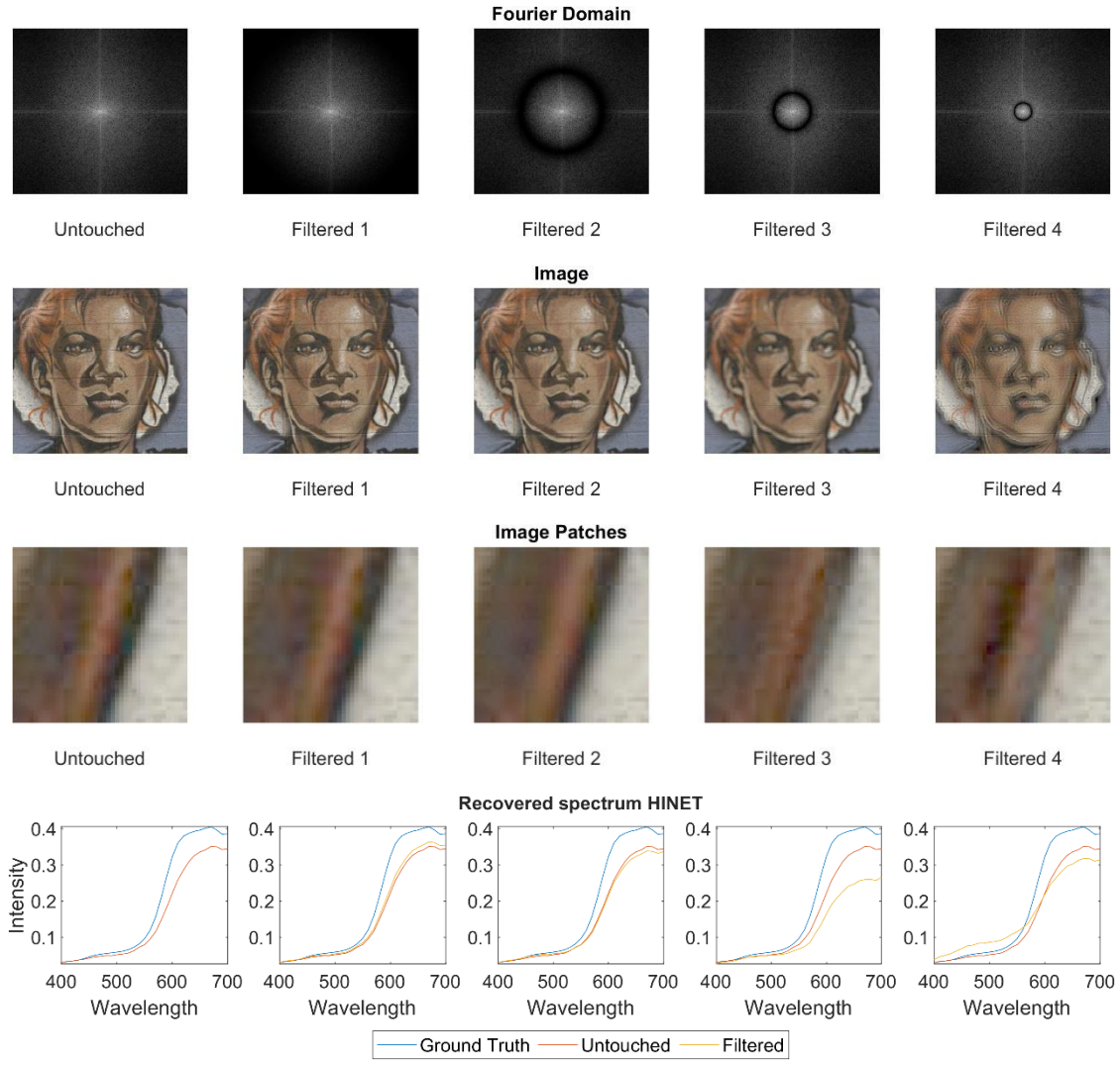
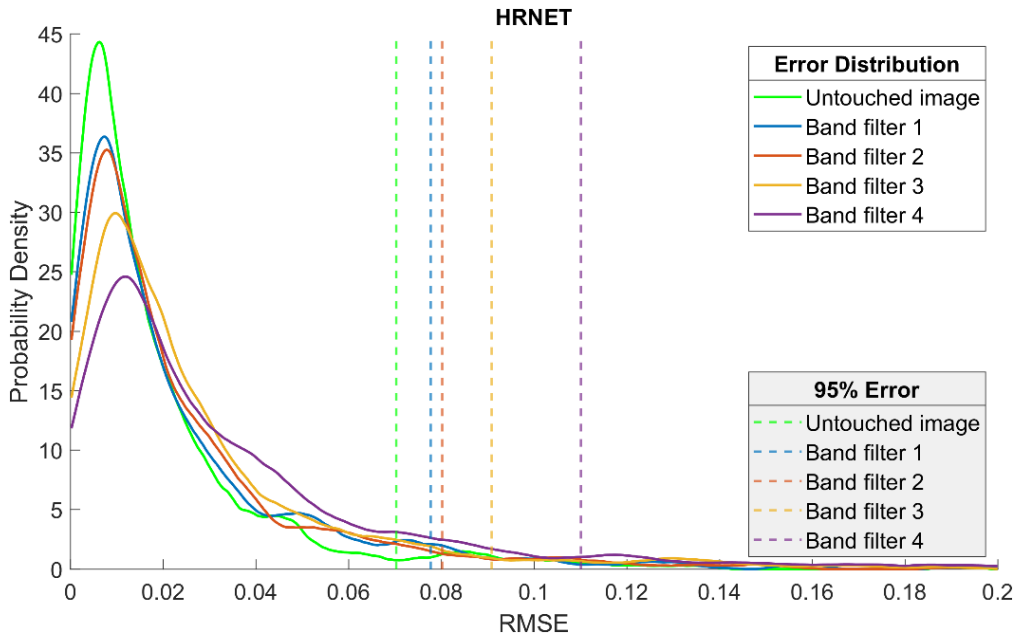


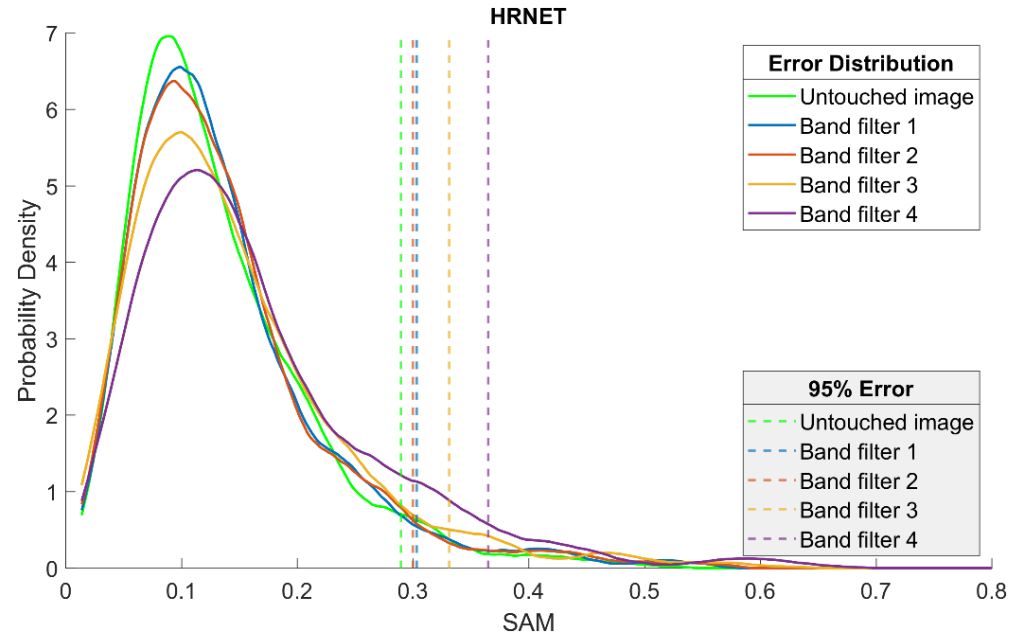
Figure 51. A comparison of the untouched and manipulated images. The top row displays filtered images in the Fourier domain; magnitudes in the Fourier domain are depicted using a logarithmic scale. The middle two rows present the filtered image and patch after applying band filters. The last row provides a direct comparison between the untouched image patch and the attacked patch reconstructed by HINET.

- **Results**

Figure 52 and Figure 53 demonstrate the reconstruction accuracy for HRNET and HINET after the neighbouring area is attacked by band-stop filters.



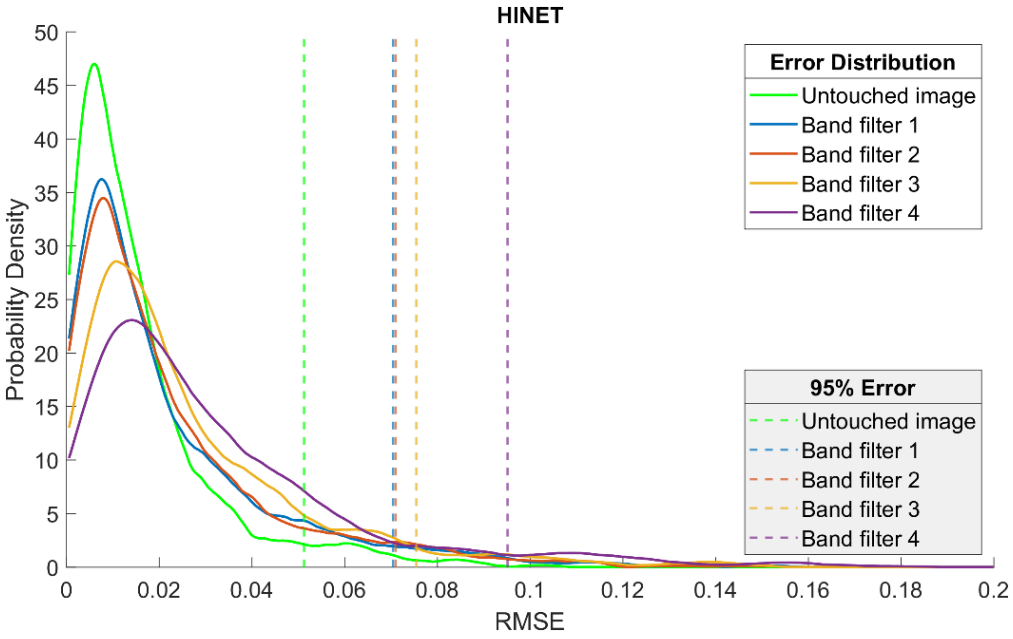
(a)



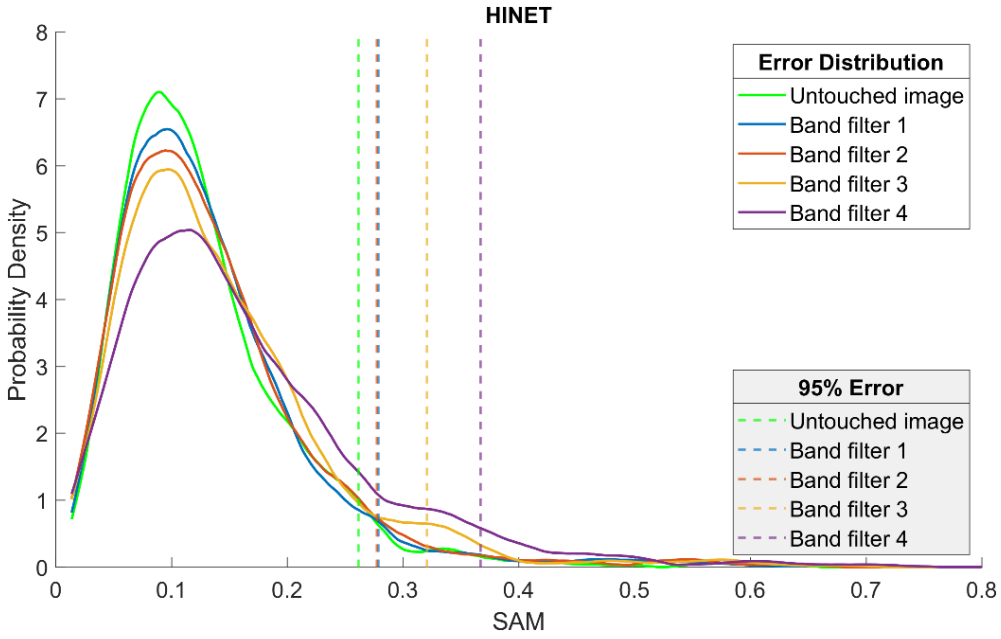
(b)

Figure 52. The error distribution and 95% error from HRNET when the image is attacked by band-stop filters.

The plots display the RMSE and SAM error distributions between the recovered and actual spectra and the 95% error (dotted line). HRNET utilizes local information, while HINET incorporates both global and local information. Additional results from other networks are provided in Appendix 5.



(a)



(b)

Figure 53. The error distribution and 95% from HIENT error when the image is attacked by band-stop filters.

In these plots, it is evident that all the frequency bands have an effect on the reconstruction. As the frequency of the removed band decreases, the reconstruction error increases. This is reflected in the changing error distribution and the rightward shift of the 95% error. It is clear that the tested neural networks rely on local spatial information to resolve metamerism and reconstruct spectra into the right shape. It is worth noting that there are individual instances (although rare) where the reconstruction accuracy improves compared to the untouched image.

- **Discussion**

1. *Consistency between networks.*

It is observed that all 6 networks compared exhibit similar responses to attacks, despite minor differences. These networks demonstrate robustness in the face of information loss in the high spatial frequency domain. This resilience can be attributed to the characteristics of the NTIRE 2022 dataset, which comprises a multitude of natural and manmade scenes with large patterns or objects. Consequently, instances of rapid change (fine details) are relatively rare. Considering that random noise has been introduced to the training RGB data, it appears that the networks might possess an inherent ability to mitigate the impact of this noise. Hence, the further application of a low-pass filter, referred to as 'Band 1', exerts only a minor influence on the reconstruction results.

2. *Responses are related to the energy of information in the Fourier domain.*

The accuracy of image reconstruction decreases when band-stop filters are applied to regions of low spatial frequency. This decline can be attributed to the higher energy levels typically associated with larger-scale features. When low spatial frequency information is removed, significant alterations in the image background, primary objects, and overall textures occur. These modifications can considerably impact the surrounding region centred on the target sample. The networks do not display any particular bias or preference towards any specific band. Instead, they operate impartially, guided by the information energy present within those respective bands.

3. *Global networks are also sensitive to changes in the local information.*

Global networks (HINET, MIRNET, MST++) display similar responses to local area attacks as the local networks. Consequently, despite their capacity to extract information from a substantial image area, local spatial information surrounding the target pixel remains crucial. Findings from the study on varied neighbour sizes suggest that local texture and patterns play an important role in the reconstruction processes of these networks. In the absence of local features as context, networks have a considerably higher tendency to inaccurately recover the spectral shape in metamerism instances.

4. *Networks may be sensitive to texture with a particular scale.*

By dividing the Fourier domain into multiple bands, we can analyse the sensitivity of networks to local textures at different scales. Generally, the influence from different scales is related to the associated energy. However, there are cases where networks are sensitive to texture features at a particular scale. Figure 54 shows an example where networks are sensitive to the attack associated with filter 3. This finding suggests that networks can extract textures at different scales and can then use a combination of information from these different scales as context to counter metamerism. The

network's ability to extract multi-scale textures is enhanced by architectures such as ResNet and U-Net, which include skip connections that assist in capturing multi-scale information.

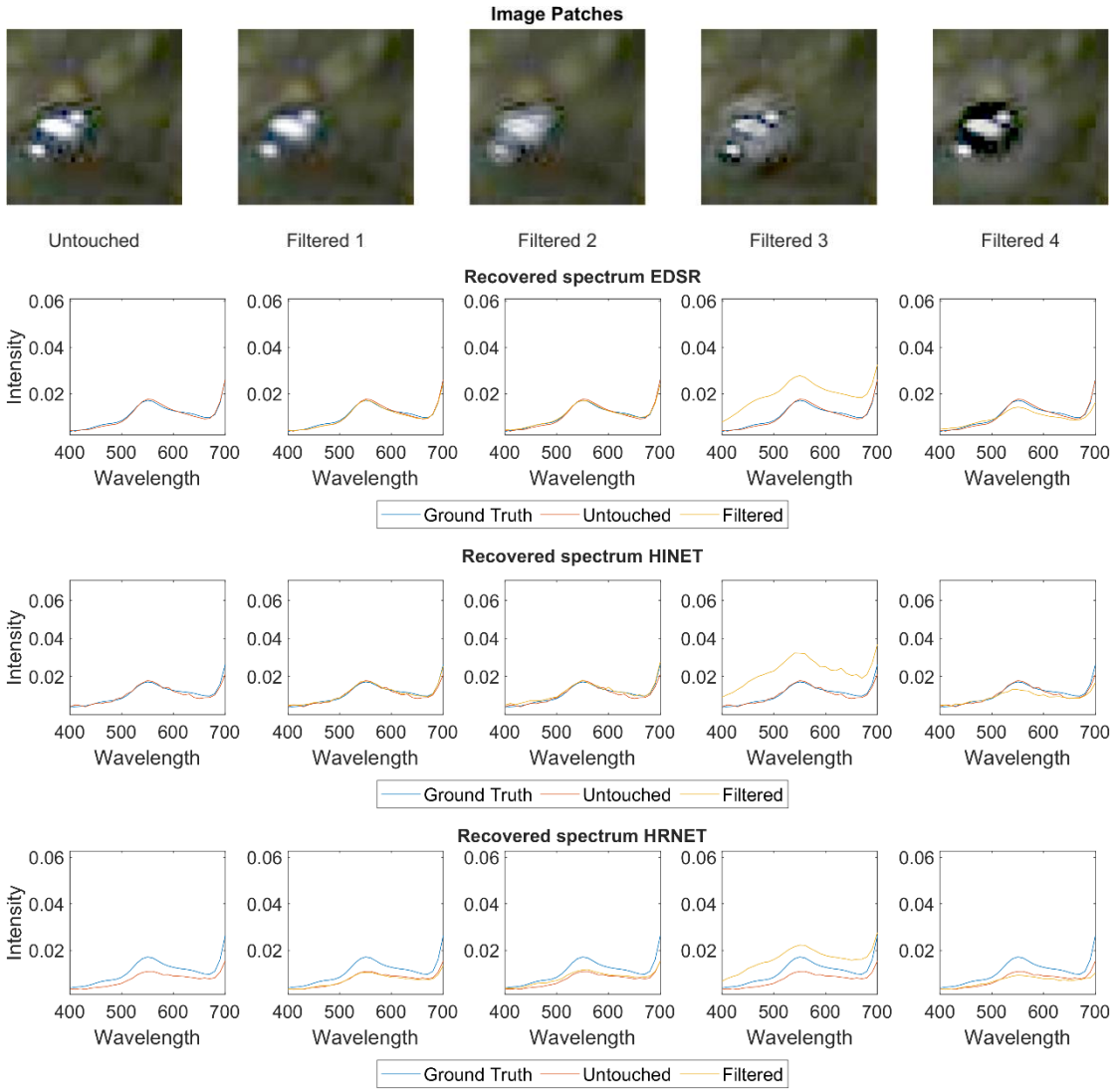


Figure 54. Untouched image patch and image patches when different scales of textures have been removed, reconstruction results from networks. Networks are only sensitive to the attack associated with filter 3.

5. *There are cases when smoothing the image increases the reconstruction accuracy.*

There are instances where the removal of high-frequency information from the neighbourhood by 'Band Filter 1' leads to an increase in reconstruction accuracy. Considering that the publisher added random noise to the training dataset, this increase in accuracy could be interpreted as the result of applying a low-pass filter to the image, which reduces the high-frequency noise. Figure 55 provides an example where smoothing an image patch can increase the reconstruction accuracy while removing textures on a larger scale can reduce it.

In Figure 55, there is consistency in how different neural networks react to similar attacks, suggesting that these networks rely on local spatial information as context to determine their output shape. When this context varies, networks display similar responses to the changed context. However, given the complexity of these networks, we can only hypothesize that they might share a similar strategy in utilizing local spatial information.

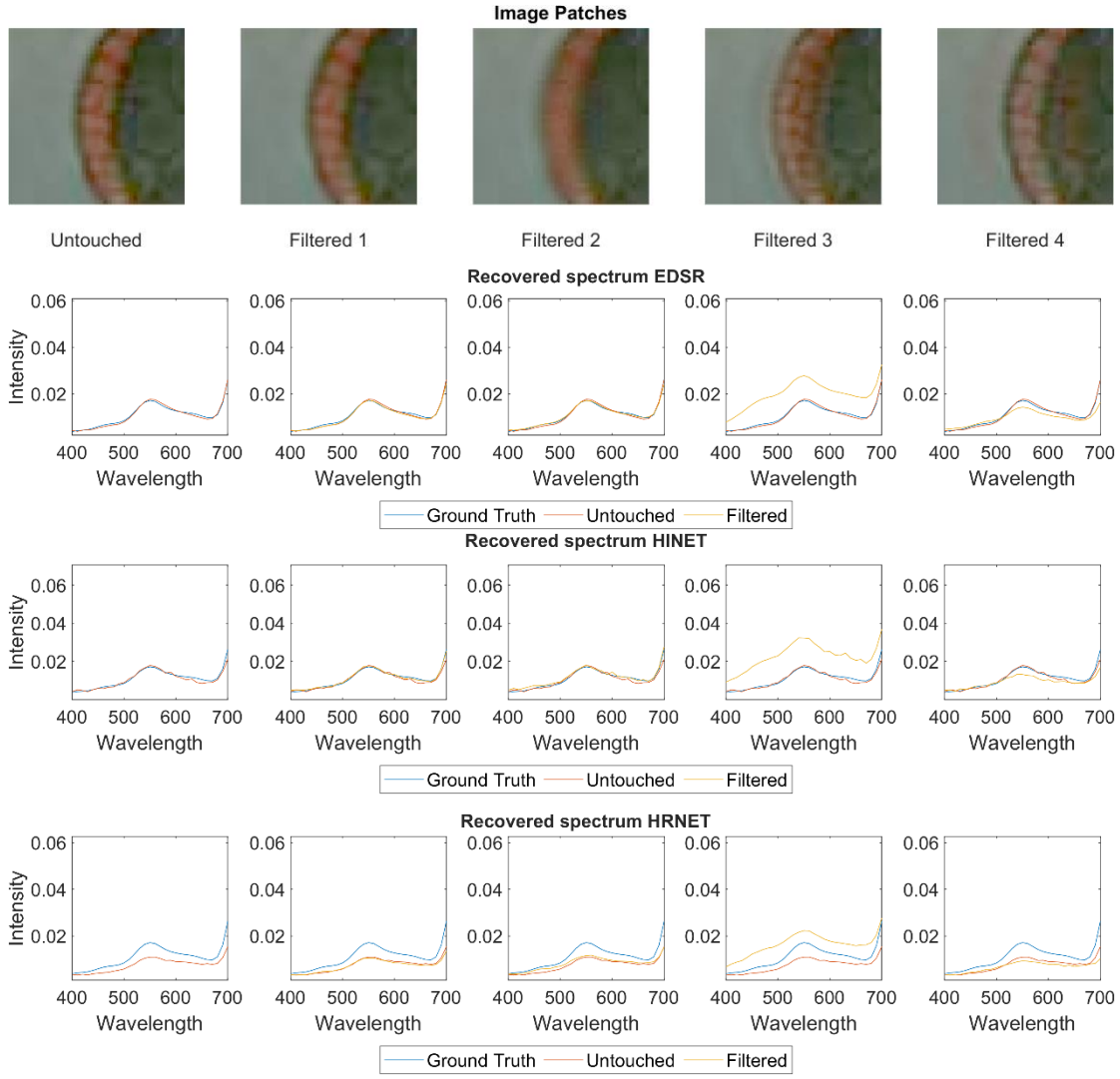


Figure 55. Untouched image patch and image patches when different scales of textures have been removed, reconstruction results from networks. In this case, smoothing the image increases the reconstruction accuracy.

- **Summary**

From the previous analyses, it is evident that different networks tend to exhibit similar responses to the same attack. This suggests that these networks all use local spatial information as context to provide their output. For instance, the band-stop filter affects texture from the attacked local area, and the networks are sensitive to changes in local textures. Therefore, these networks are potentially

utilizing local texture as context to resolve metamerism. Consequently, it may also be feasible for us to use local textural features as context when designing our explainable spectral super-resolution method. Regarding local textures at different scales, it's apparent that networks are sensitive to variations in scale. Despite the energy associated with each band, it's important to note that networks are more sensitive to attacks in low spatial frequency. However, fine spatial details cannot be neglected during the reconstruction process.

5.4.2. Attack the Image with Gabor Filters

- **Introduction**

In the previous subsection, we examined the effects of removing features of a particular scale. It was also observed in Figure 30 that some reconstructions are sensitive to orientation. As an extension of that, in this experiment, we will explore the networks' response to both scale and orientation. This is achieved by removing features extracted by Gabor filters which are specifically designed to decompose the Fourier domain into distinct regions considering both scale and orientation.

The Gabor filter is a linear filter commonly used for texture analysis, feature extraction, and edge detection. The Gabor filter is essentially a Gaussian kernel modulated by a sinusoidal plane wave. The filter can be adjusted for different orientations and scales (frequencies) by varying the parameters of the Gaussian and sinusoidal functions. This flexibility allows the Gabor filter to effectively detect and analyse features in images with varying orientations and scales.

Since the image size in the NTIRE 2022 is 482×512, we used 512×512 as the Fourier domain size to design the Gabor wavelet. Figure 56 shows the Gabor decomposition with a phase step of $\pi/4$. The circle in Figure 56 illustrates the half-peak magnitude (0.5) of the filter responses in the Gabor filter dictionary. To ensure smooth transitions between neighbouring masks, the half-peak magnitude of the filter responses is designed to intersect. The filter magnitudes as a function of their distance from the image centre are the same as what is shown in Figure 50. Please note that the colour scheme used in both Figure 56 and Figure 50 corresponds to the frequency of the filters, thereby offering a clearer understanding of the filter's characteristics and their relative positioning in the frequency domain.

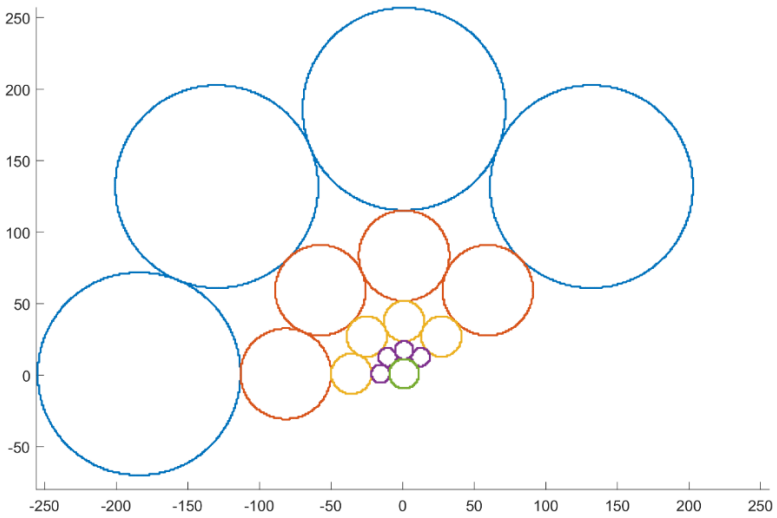


Figure 56. The contours indicate the half-peak magnitude of the filter responses in the Gabor filter dictionary. To ensure smooth transitions between neighbouring masks, the half-peak magnitude (0.5) of the filter responses is designed to intersect.

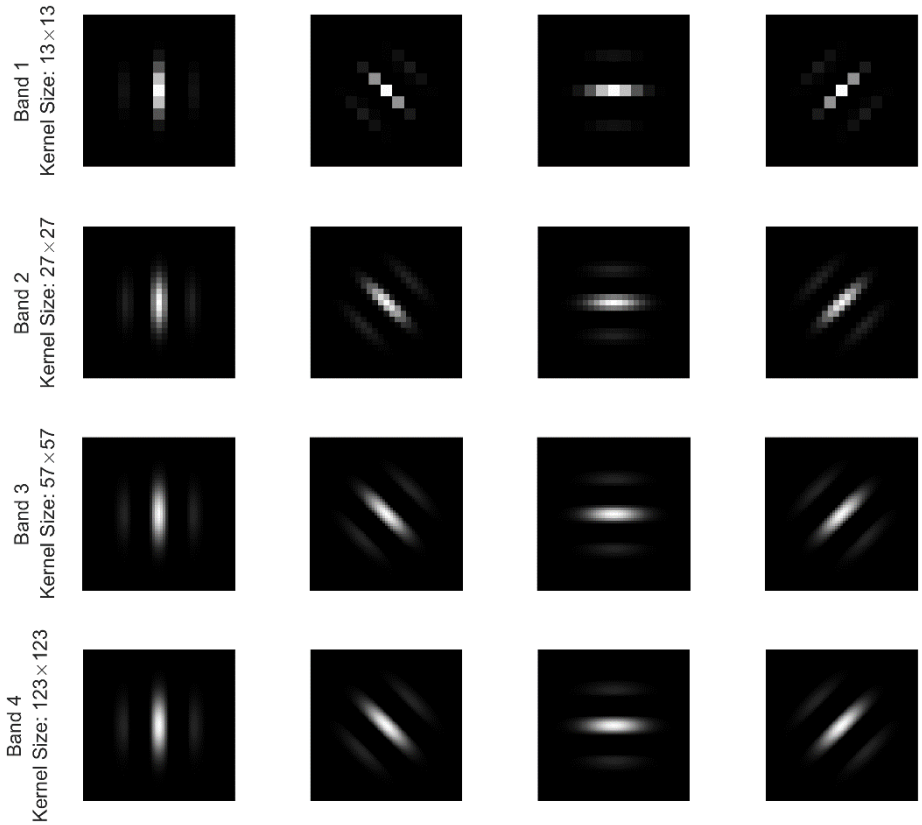


Figure 57. Gabor filter dictionary used in this analysis. The kernel size of each filter is given for each row.

Figure 57 shows the Gabor filter dictionary. In practice, Gabor filters are often applied as a bank of filters, each with different orientations and scales, to analyse an image comprehensively. The

responses from these filters can then be combined to extract features, identify patterns, or analyse the texture in the image. Then, the impact of removing these textures on the reconstruction process will be assessed by comparing the performance of networks.

- **Result**

Table 15 displays the reconstruction error, measured in SAM, of HINET when subjected to attacks by Gabor filters. On average, the removal of textures at the same scale but with different orientations does not yield noticeable differences. Given that textural features may appear in any orientation within a natural scene, networks do not seem to exhibit a preference for textural features from a specific orientation. However, this could primarily be because averaging diminishes the significance of individual responses to texture orientation.

Table 15. Reconstruction error from HINET in SAM when the input image is attacked by Gabor filters.

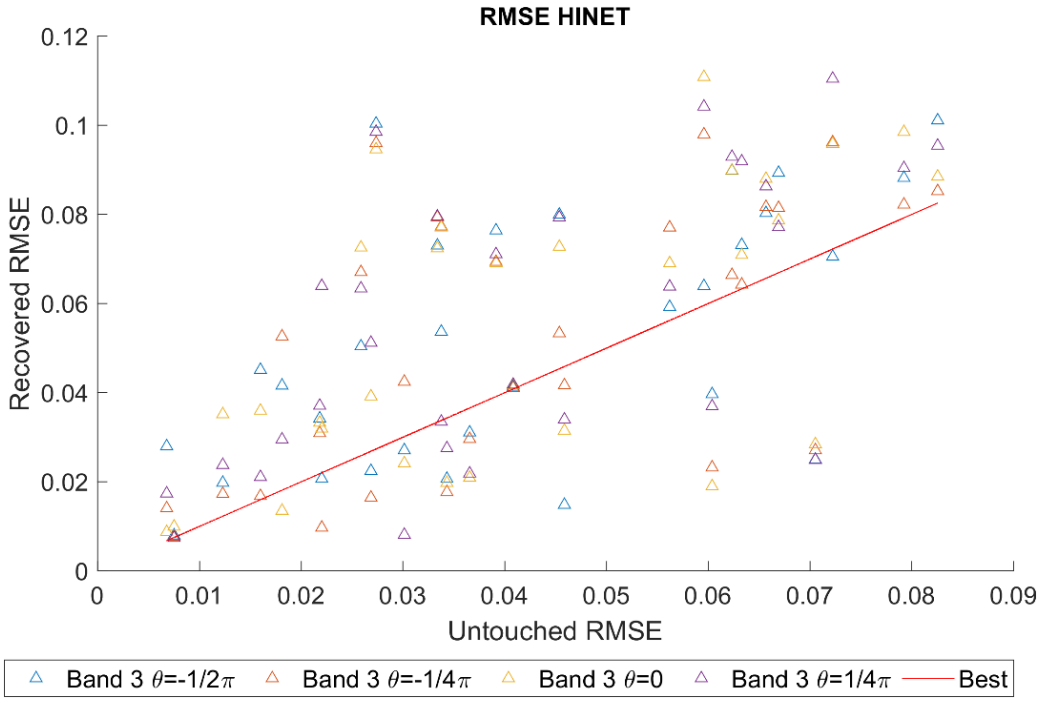
(a) SAM

	-1/2π	-1/4 π	0 π	1/4 π
Band 1	0.141	0.141	0.141	0.141
Band 2	0.142	0.142	0.141	0.142
Band 3	0.147	0.146	0.148	0.147
Band 4	0.159	0.158	0.157	0.157

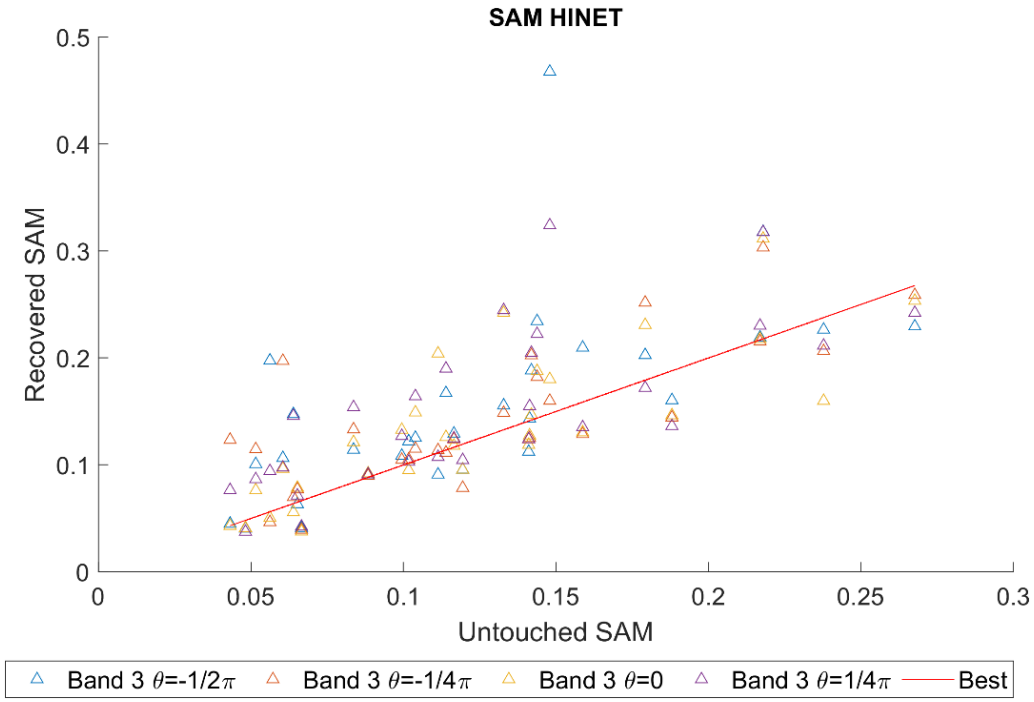
(b) 95% Error measured in SAM

	-1/2π	-1/4 π	0 π	1/4 π
Band 1	0.302	0.303	0.301	0.304
Band 2	0.308	0.309	0.308	0.309
Band 3	0.323	0.323	0.331	0.319
Band 4	0.352	0.348	0.354	0.347

However, for individual samples neural networks appear to be sensitive to attack on local texture in different orientations. Figure 58 (a) and (b) illustrates the RMSE and SAM of the recovered samples from a metamerism set when local textural features are removed across different orientations (distinguished by colour) at the same scale (Band 3). HINET responds differently to the attacks on different orientations removed from the input image. Consequently, HINET, along with other networks shown in Appendix 5, appears to be sensitive to the orientation of local spatial textures at the same scale.



(a)



(b)

Figure 58. Reconstruction accuracy is affected by removing local texture features in varying orientations but same scale. The red line represents untouched samples. HINET is sensitive to local textural features in different orientations, appearing with the change in the reconstruction accuracy of each spectral sample.

Figure 59 presents an example of a local area under attack by Gabor filters with different orientations but the same scale.

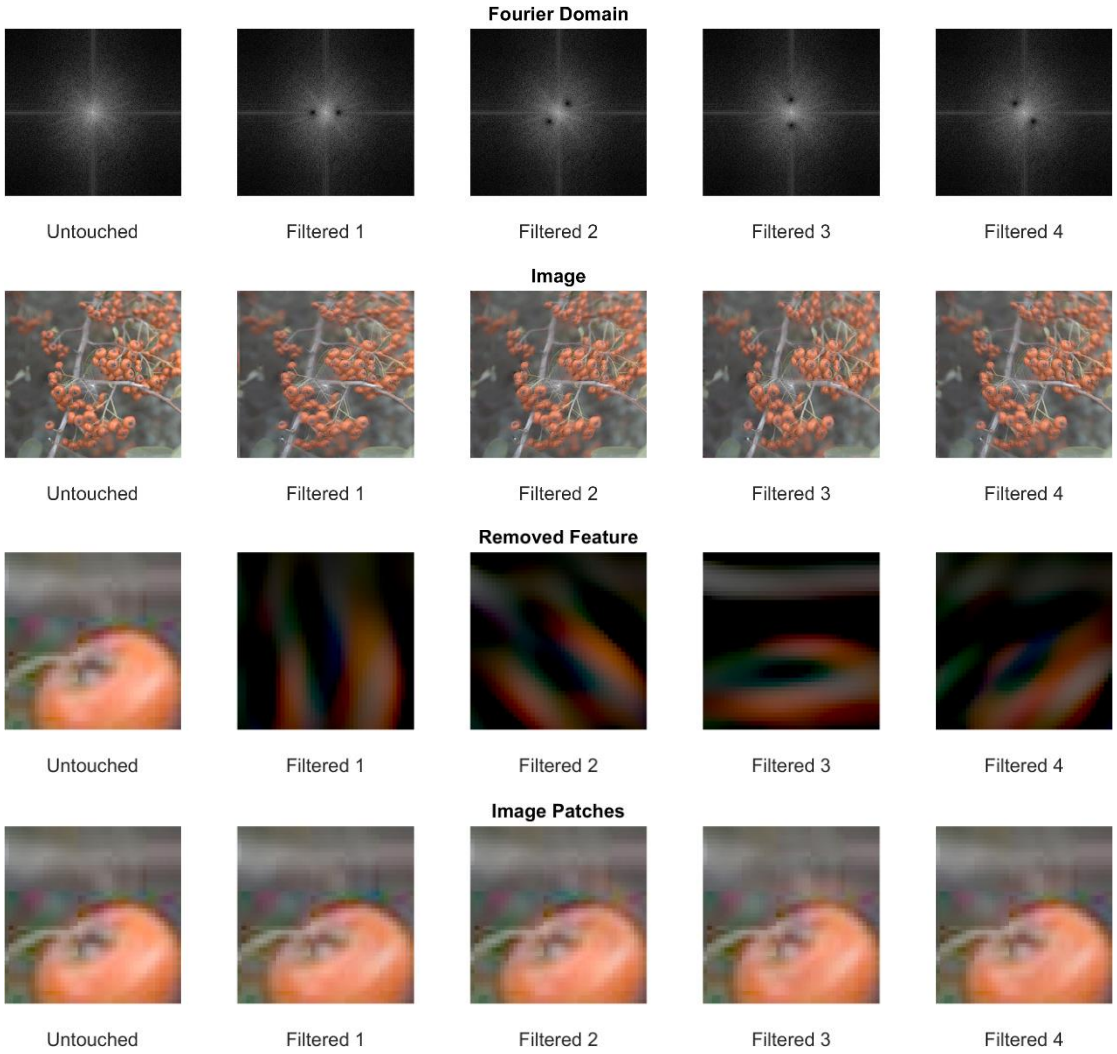


Figure 59. The top row shows the Fourier domains after the attack, the magnitudes in the Fourier domain are scaled logarithmically; The second row shows the image after applying band filters; the third row shows the removed feature by corresponding attacks; The last row shows the attacked patch.

Figure 60 displays the reconstructed spectrum from six networks for both the attacked images in Figure 59. In the example here, the listed neural networks are sensitive to the orientation of the removed textural features giving a change in the reconstruction spectrum. Since the different networks show consistency in their response to the attacks, the underlying textural feature removed by the filter is important to determine the shape of the reconstructed spectrum.

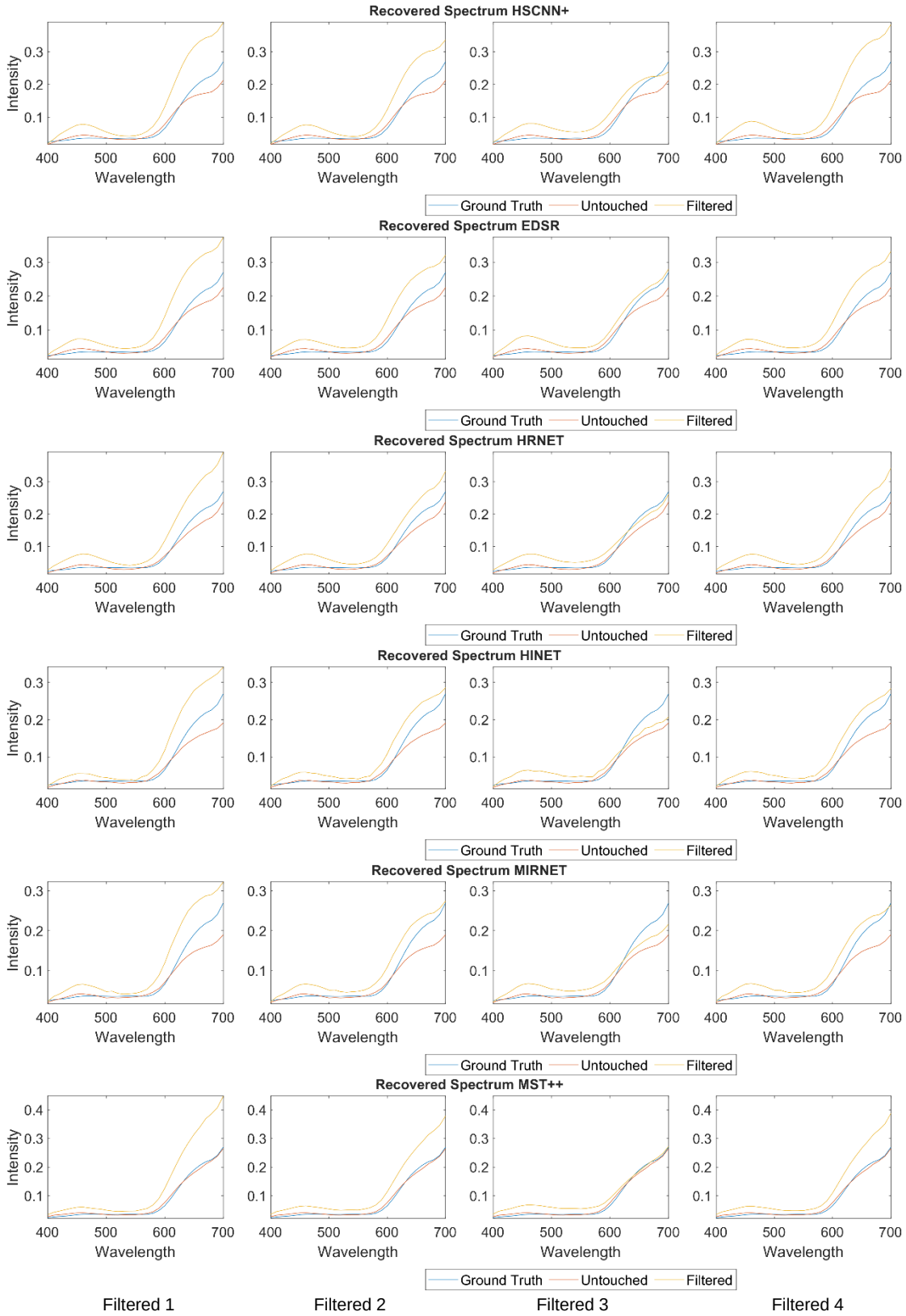


Figure 60. The reconstructed spectrum from six networks of the attacked images.

- **Discussion**

Given the fact that the RGB value of the target pixel remains untouched, a plausible explanation for the similar network responses to the attack could be that those networks utilize the same local textural features as contextual information to determine the structure of the recovered spectrum. As such, for our explainable method, we could employ a local texture descriptor to generate a coded texture vector. This vector could then provide supplementary context to more precisely estimate the PQR weights.

It is evident that networks display sensitivity not only to the scale of textural features but also to their orientation. In combination with the results from the previous subsection, there is no specific scale or orientation that networks particularly favour. Therefore, in designing our feature extraction method, a general texture descriptor may be more suitable than a specialized feature detector, for example, a series of Gabor filters or a multi-scale local binary pattern.

5.4.3. Summary

In this chapter, we sought to address two research questions: how large the spatial context needs to be, and what sort of texture features are important. Each neural network exhibits the highest sensitivity to pixels closest to the target, those within a radius of fewer than 5 pixels. Features that are 30 pixels away from the target sample have a slight influence on the shape of reconstructed spectra. Hence, it can be inferred that local spatial information is adequate to serve as a context for resolving metamerism in most cases.

Furthermore, the sensitivity of these neural networks to ‘attacks’ in local textures demonstrates their capacity to leverage texture information across different scales and orientations. Different samples are sensitive to different textural features, implying that it is important to be able to capture and represent a range of features. Consequently, when we create our own feature extraction method, we should opt for a general texture descriptor, given its suitability to this application.

However, due to the limitations in the way we 'decompose' local spatial information, the sensitivity analysis could only provide general information on the type of features useful for resolving metamerism. In the next section, we intend to allow the network to ‘display’ its sensitivity to spatial information through a gradient analysis.

5.5. Gradient Analysis of Spectral Super-Resolution Networks

5.5.1. Introduction

In the previous section, we have demonstrated that deep models use spatial context to resolve metamerism. When various texture features are removed from the input image, the accuracy of the reconstructed spectrum deteriorates. Additionally, we have also demonstrated that the local spatial information is essential for the listed networks. The influence diminished with increasing distance away from the target sample. In this section, we will investigate the same question in a different way by analysing the gradient of the reconstruction accuracy of target samples to the input RGB pixels. Automatic differentiation in PyTorch will be employed to measure the gradient. The gradient analysis enables us to examine how the network processes local spatial information during the reconstruction process.

To emphasize this experiment is going to address the following research questions:

1. What is the necessary spatial information size for resolving metamerism during spectral reconstruction using deep convolutional neural networks?
2. What type of spatial information is utilized by the network during the reconstruction process?

5.5.2. Methodology

- **Automatic Differentiation**

Automatic Differentiation is a computational technique used to efficiently and accurately compute the derivatives of functions, especially in the context of complex, high-dimensional functions (Baydin *et al.*, 2018). It is a valuable tool in many applications, including optimization, machine learning, and numerical simulations. Unlike numerical differentiation, which approximates derivatives using finite differences, and symbolic differentiation, which computes derivatives by symbolically manipulating mathematical expressions, automatic differentiation calculates derivatives by decomposing functions into a sequence of elementary operations and applying the chain rule of calculus to these operations. Automatic differentiation has gained popularity in machine learning due to its ability to compute gradients efficiently and accurately, which is crucial for the optimization of complex models like deep neural networks.

In this experiment, the automatic differentiation is facilitated with the auto-gradient function in PyTorch. The backward mode is used to measure the influence of the input RGB image on the accuracy of the reconstructed target spectrum. The backward mode, also known as reverse accumulation, starts the computation from the output variables and propagates the derivatives backwards, towards the input variables. The backward mode is particularly useful when the number

of output variables is small compared to the number of input variables, as it is often the case in many machine learning applications, such as training neural networks using gradient-based optimization algorithms.

- **Networks and Dataset**

In this experiment, we utilized five pre-trained networks: EDSR, HRNET, MIRNET HINET and HDNET. These networks have been introduced in the previous sections. We discovered that extracting the gradient of the input RGB image demands substantial computer memory. Due to the limitations of our device (64GB RAM), we restricted the number of networks involved in the experiment to five. The dataset employed in this study is the same as that used in the previous experiments in this Chapter, comprising 18 metamerism sets, with each set containing 20 images.

- **Measuring the gradient of the metamerism samples**

We computed the gradient of the reconstruction accuracy as measured by SAM as it allows us to specifically evaluate the impact on the shape of the reconstructed spectrum. The measured gradient reflects how much influence each RGB pixel in the image has on the accuracy of the reconstructed spectrum. To assess the networks' consistency in handling similar metamerism samples, we also measured the gradient of immediately neighbouring pixels and compared the results with the target sample.

5.5.3. Average Gradient Analysis of All Samples

Our analysis of the measured gradient results started with a broad examination of all the metamerism samples. The goal was to explore the connection between the gradient values of neighbouring pixels with respect to the target sample and their distance from the target pixel. This approach provides insights into the size of the spatial area that networks are sensitive to, and can potentially extract information from.

Figure 61 presents the mean absolute gradient for the five networks under investigation. For each sample, we took the absolute RGB gradient and then averaged the three channels. The final average patch is computed using all samples from the 18 collected metamerism sets. The displayed 2D and 3D patches are 40×40 pixels centred on the target pixel, even though the maximum measured distance was 80 pixels. The reason for limiting the displayed patch size is that the gradient becomes relatively flat and small beyond this size.

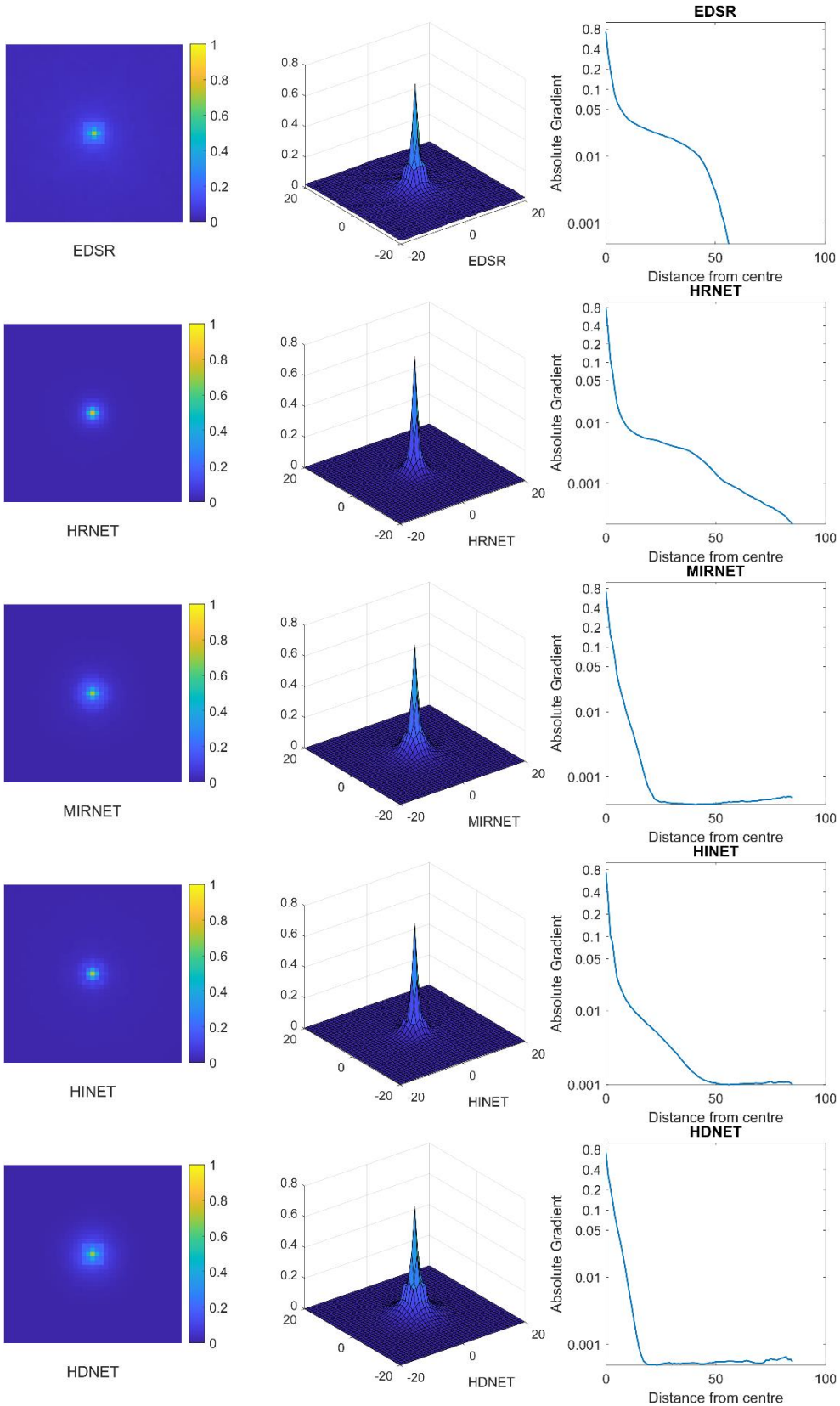


Figure 61. Left: the average absolute gradient of the listed five networks in a colour-mapped patch; middle: the average absolute gradient of the listed five networks in a 3D colour-mapped patch; right: the average absolute gradient as a function of the distance from the target pixel.

The average gradient values of the five listed networks exhibit a degree of consistency with minor variations. The target pixel, which is the pixel of interest, demonstrates the highest gradient values. This suggests that the RGB value of the target pixel has the most substantial impact on shaping the reconstructed spectrum. As the distance from a neighbour pixel to the target pixel increases, the influence of the neighbour pixel on the reconstructed spectrum diminishes. In the 2D plots, the pixels immediately above, below, to the left, and to the right of the target pixel, which form a cross pattern, display the highest gradient values among all neighbour pixels. However, when the neighbour pixel is located 10 or more pixels away from the target, the gradient plot becomes relatively flat, and the gradient values approach zero. This indicates that these distant pixels individually have a minimal impact on the shape of the reconstructed spectrum. Therefore, for the networks considered in this analysis, local spatial information from a 20×20 patch appears to exert the most significant influence on the shape of the reconstructed spectrum. This suggests that these networks primarily rely on spatial information from a 20×20 patch, even though some of them may have the ability to extract information from the entire image.

There are subtle differences in the sensitivity of neural networks to local pixels. EDSR stands out with a higher gradient for intermediate distance pixels, around 0.05, when the neighbour is 10 pixels away from the target. In contrast, other networks maintain a gradient of less than 0.01 at the same distance. HRNET and HINET exhibit a sharp decrease in gradient with increasing distance from the neighbour pixel to the centre, indicating that these two networks prioritize very local information to reconstruct most spectra. EDSR and HDNET manifest a small square with a high gradient around the target pixel, while MIRNET and HINET display a small circle around the target pixel in their gradient profiles. These variations suggest that different networks have distinct sensitivities to local spatial information.

When examining the gradient as a function of distance to the target pixel, a common pattern emerges: the gradient initially decreases significantly when the distance is less than 5 pixels, and then levels off. However, there are distinctions among the networks. EDSR exhibits a higher gradient for intermediate distance pixels, and the gradient drops to zero when the distance exceeds 60 pixels. This indicates that EDSR is particularly sensitive to local information within a fixed range. MIRNET and HDNET show similar patterns, with the gradient flattening out when the distance surpasses 20 pixels. Despite all these networks being sensitive to very local pixels, they utilize potential spatial information differently, with varying degrees of detail.

In summary, the investigation has highlighted the importance of local spatial information in the spectral reconstruction of metamerism data. Additionally, we observed that the five networks we examined share a common sensitivity to very local pixels but diverge in their utilization of spatial information further from the target pixel.

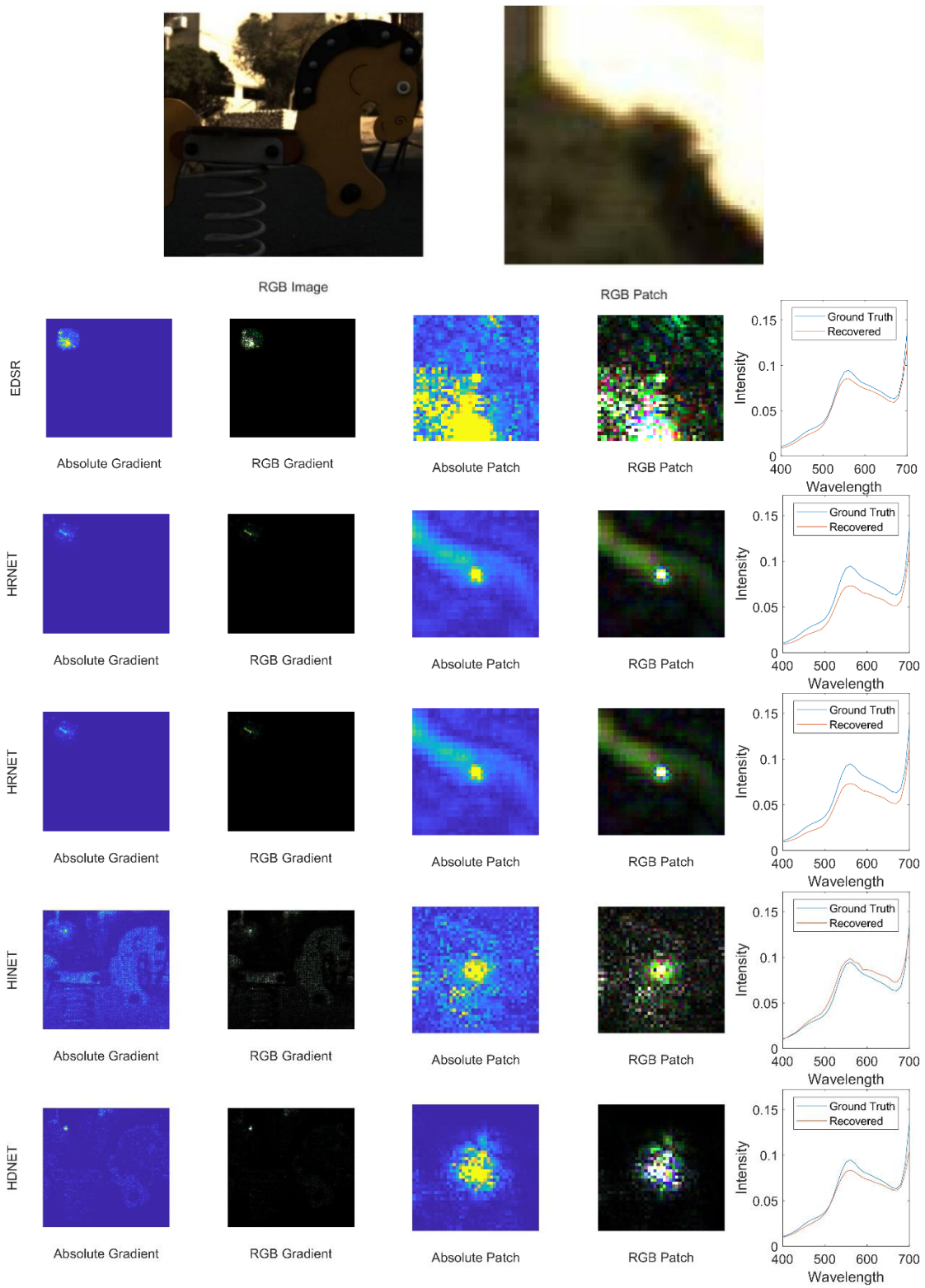


Figure 62. Example gradient image and recovered spectrum from the listed five networks. At the top we see the original RGB image and the 41×41 image patch centred around the target pixel. The left two columns display the colour-mapped absolute gradient image and the 'RGB' gradient image. The third and fourth columns display the absolute gradient patch and the RGB gradient patch centred around the target pixel. The right column compares the ground truth spectrum and the recovered spectrum from each network. Please note the absolute gradient has been scaled 5 times to show a clear pattern.

Although we identified some shared characteristics among the networks in their use of local spatial information, it's important to note that these networks exhibit distinct preferences for either local or global information in specific scenarios. To illustrate this point, Figure 62 shows an example where the networks' treatment of spatial information varies significantly. The purpose of showing the RGB gradient image is to investigate potential imbalances in the weighting of colour channels during reconstruction.

The EDSR and HRNET have extracted spatial information from a local area, while EDSR focuses on a square area, HRNET does not use any particular shape of area. The shape of the reconstructed spectrum from MIRNET, HINET and HDNET is affected by the spatial information from the whole image. All these three networks show attention to the horse from the image. However, the target spectrum is from plant leaves behind which have no relation to the horse, there is a risk these three networks are overfitted to the most particular object of the image. Since the target spectrum is stronger in green, the gradient result of all five networks appears to be more sensitive to the green colour channel of the input image.

This analysis provides insight into the distinct approaches networks take when extracting spatial information during the reconstruction of a single pixel. By examining individual cases, we can better understand the specific strategies they employ when dealing with different types of spatial information. Interestingly, networks that prioritize local information seem to focus on the texture within the image patch, whereas those that extract information globally show less interest in local textures. For instance, in the provided example, EDSR and HRNET appear to be sensitive to local textures. EDSR distinguishes between the plant leaf and background, while HRNET extracts information from the edge between the plant and background. In contrast, the other three networks show no distinct patterns related to local textures. It's important to note that the example presented is specific, and more results need to be analysed to draw general conclusions. Therefore, in the following sections, we will delve into how these networks use global and local information when reconstructing metamerism samples in individual cases.

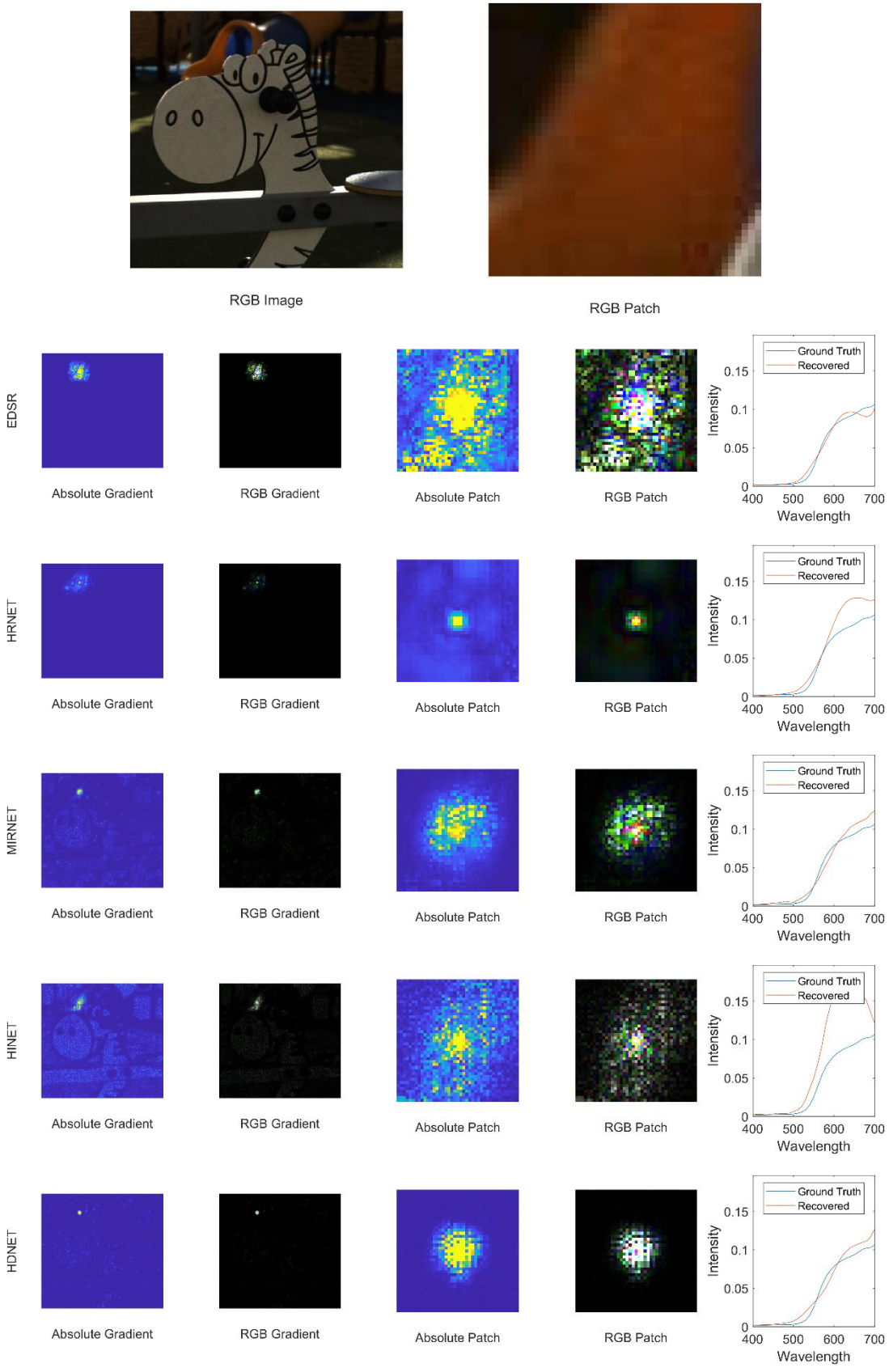
5.5.4. Gradient Analysis of Networks that Extract Global Information

Global spatial information encompasses the broader structure and larger context of an image, capturing relationships and patterns that extend across the entire image. In contrast, local spatial information focuses on smaller regions or neighbourhoods within the image. Networks can extract global spatial information in various ways, such as through pooling, attention mechanisms, and multi-scale fusion for example as used by auto-encoder and U-net type networks.

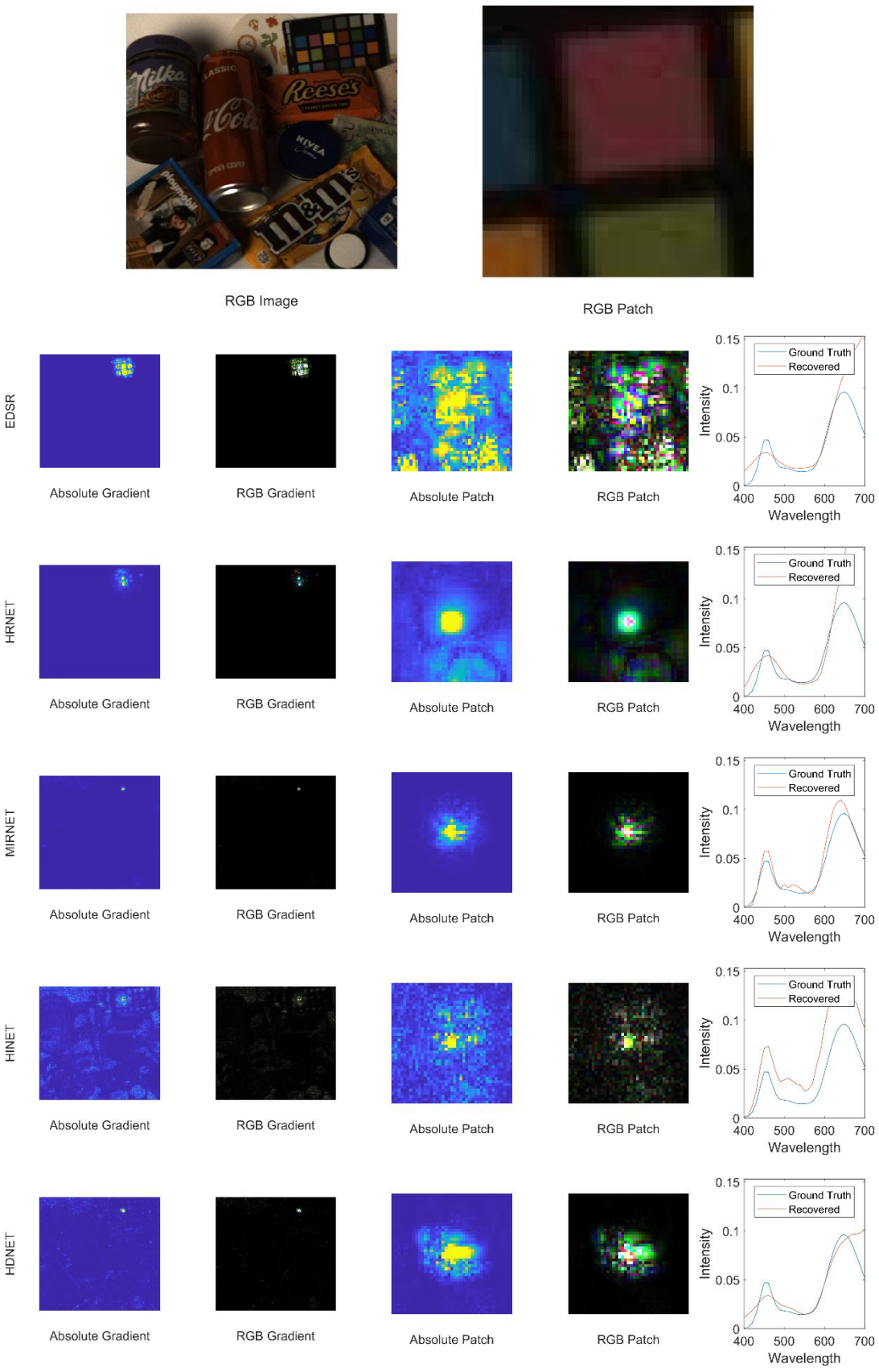
As discussed in previous sections, a trend in newer spectral super-resolution networks is their ability to extract spatial information from larger areas. However, it's important to note that only a small number of samples from the collected metamerism dataset exhibit noticeable gradients associated with global information. In most cases, samples display noticeable gradients primarily in the vicinity of the target pixel. This observation underscores the significance of understanding how networks preferentially utilize global spatial information when reconstructing individual pixels.

To determine the percentage of samples that rely on global information within our entire collected sample set, we've designed a simple test. In this test, we seek samples that contain pixels far from the target pixel which exhibit a large gradient. Specifically, we're looking for instances of pixels that display a gradient value exceeding 1% of the gradient on the target pixel, and these pixels are located outside the 64×64 square area centred around the target pixel. The rationale behind selecting a 64×64 size is that it represents the smallest image patch size on which many of the networks were trained. Moreover, all the networks involved in this test were trained on image patches that are larger than this size. The results show that 15% of MIRNET samples, 43% of HINET samples, and 16% of HDNET samples contain pixels that, while being distant from the target pixel, exhibit significant gradients. These findings provide an approximation of the proportion of instances where global information is utilized during the reconstruction process. Similar to what we learned from this study, HINET makes more of use global information, while the other two networks only require global information in a low proportion of less than 20%. Therefore, in most cases, those networks primarily rely on local information to do the reconstruction.

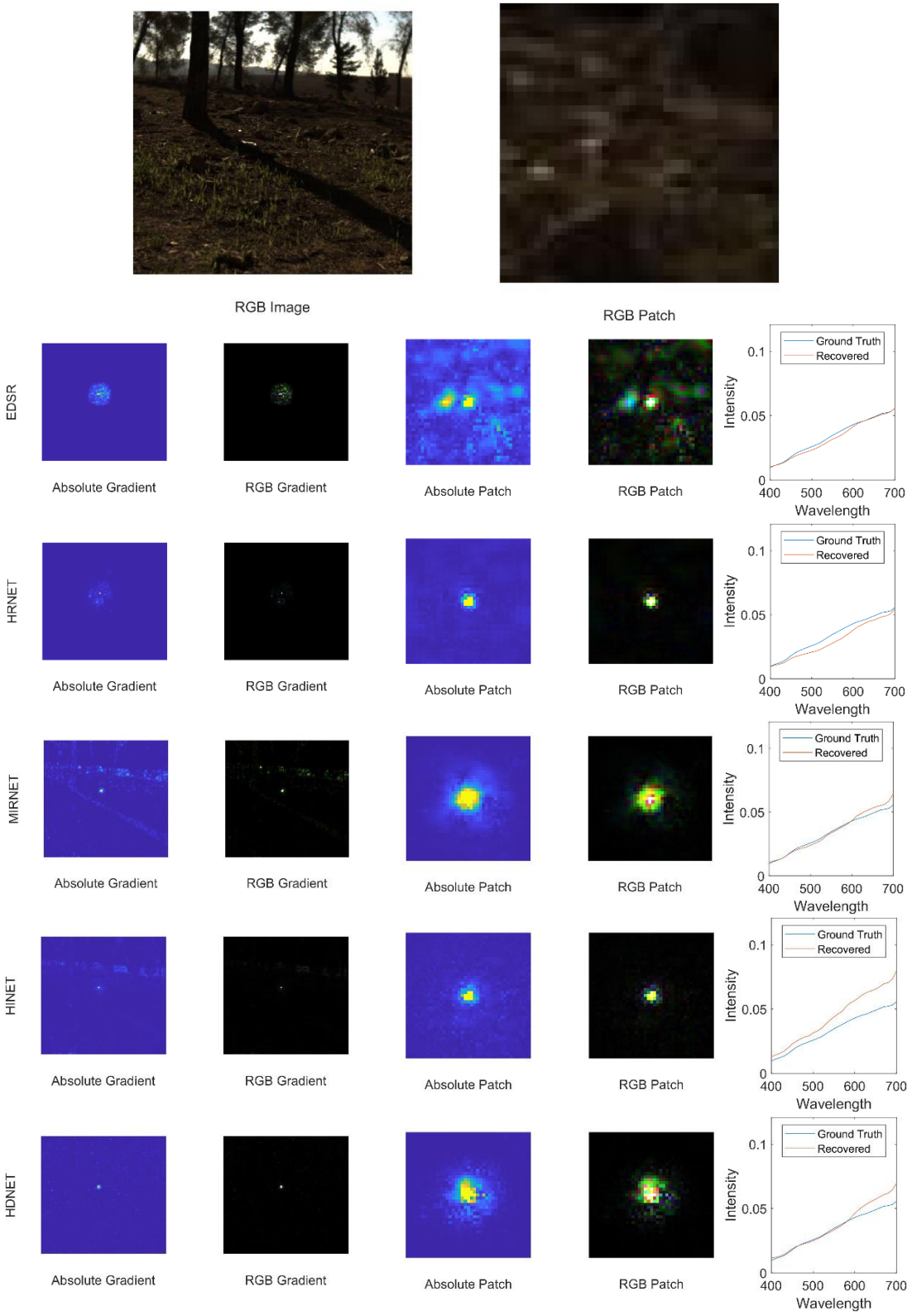
Figure 63 illustrates how certain networks appear to be sensitive to the global information within the input RGB image. Similar to the earlier example, EDSR and HRNET mainly concentrate on local spatial information, demonstrating an interest in image textures within the local area. A more detailed discussion of local textural information will follow in the next section.


RGB Image
RGB Patch
EDSR
HRNET
MIRNET
HINET
HDNET
Absolute Gradient
RGB Gradient
Absolute Patch
RGB Patch
Ground Truth
Recovered
Intensity
Wavelength
0
0.05
0.1
0.15
400
500
600
700

(a)


RGB Image
RGB Patch
EDSR
HRNET
MIRNET
HINET
HDNET
Absolute Gradient
RGB Gradient
Absolute Patch
RGB Patch
Ground Truth
Recovered
Intensity
Wavelength
0.15
0.1
0.05
0
400
500
600
700

(b)



(c)

Figure 63. The gradient of the input image that the listed networks are sensitive to global information. (a) shows an example where networks are sensitive to objects; (b) shows an example where networks are sensitive to image scenes; (c) shows where networks are sensitive to tree shadow.

Notably, the global information does impact the reconstruction results of networks like MIRNET, HINET, and HDNET. These networks seem to focus on similar global information, which includes the most representative objects in each image from a human perception standpoint. For instance, this can be observed with the horse, graffiti, and the wooden toy in Appendix 6. The objects potentially used as spatial information by these networks exhibit some common characteristics:

1. They form one complete object, rather than being fragmented or divided into smaller parts.
2. These objects occupy a significant portion of the image scene, making them more noticeable and influential in the overall context.
3. The objects are unique to the specific image scene and cannot be found in other images from the training dataset, making them distinctive and potentially useful to separate metamerism samples from each other.

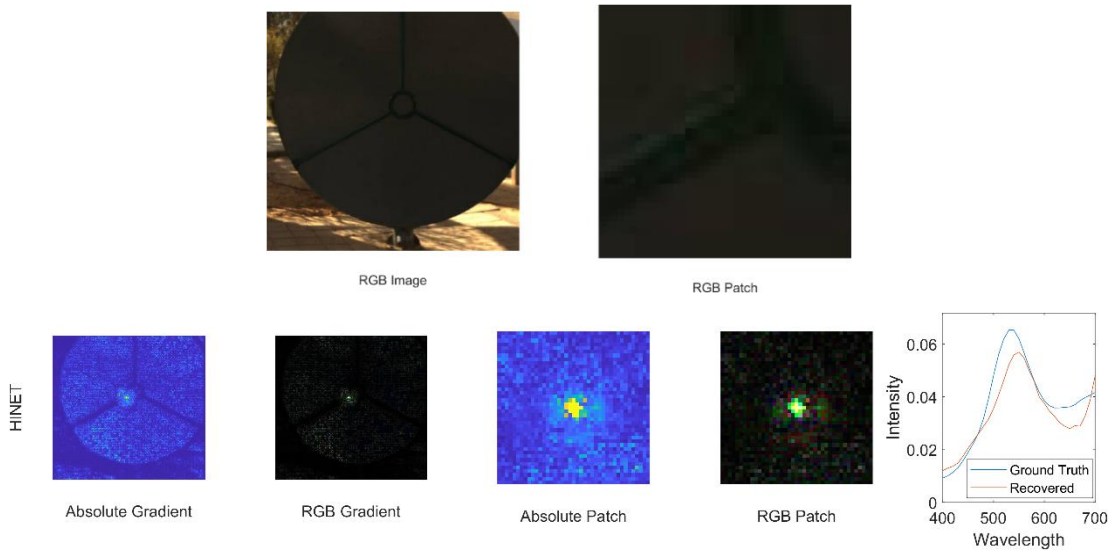
Assuming we employ a dictionary-based method to address metamerism, such unique objects can be utilized as context information to identify the training image, and subsequently, determine the spectra associated with the image. However, as these objects may only be present in one image from the training data and are not necessarily related to the material of the target, there is a risk of overfitting the network. On the other hand, when local texture alone cannot definitively identify a material, global information such as the object itself could serve as a useful contextual reference to resolve metamerism. However, the internal mechanism of networks does not operate in a manner similar to a dictionary-based method. How the network transforms global object information into context remains unclear.

In addition to objects, networks can also extract scene information from an image, as shown in Figure 63 (b) MIRNET is sensitive to the tree and the shadow. There are other examples where networks show interest in other image scenes such as the blue sky as shown in Appendix 6. Besides, there are also results showing the networks are sensitive to the shape of an object. This demonstrates that networks can identify and respond to various types of visual information in an image.

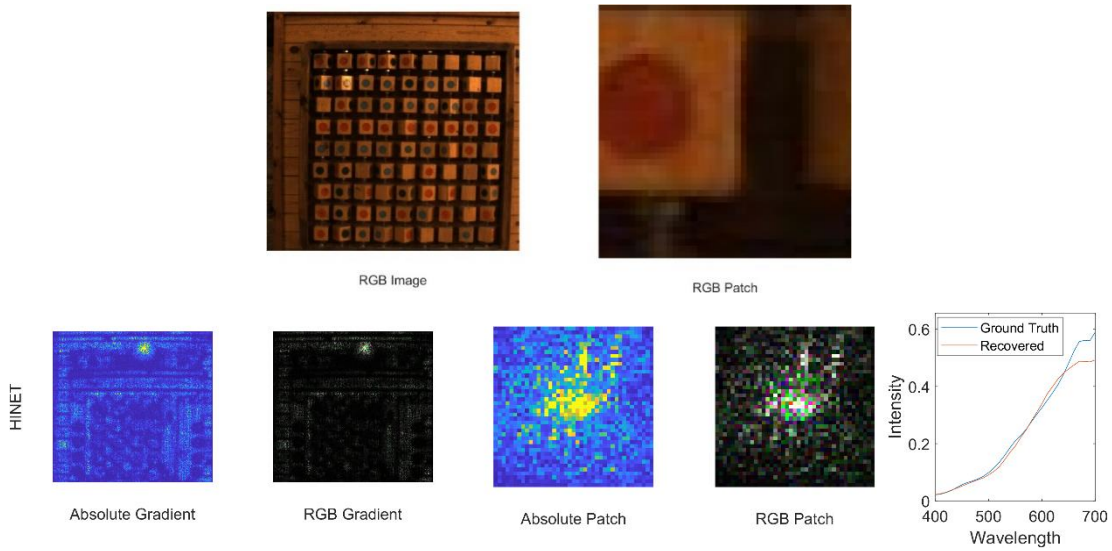
In terms of local spatial information, the gradient results of MIRNET, HINET, and HDNET show a weaker correlation with the image texture in the local area. Fewer samples from the global neural networks exhibit a clear structure in the local area compared to the local networks. These three networks appear to be sensitive to a circular local area centred around the target sample, and this finding holds true for other samples beyond the examples shown, as can be found in Appendix 5.

HINET, which is based on a U-net structure, has skip connections that enable it to capture global spatial information efficiently. As a result, HINET shows the highest sensitivity to global spatial information compared to other networks. Figure 64 displays some gradient results where only HINET

is sensitive to global information. However, HINET does not outperform the other networks, and the reconstruction accuracy of HINET is even lower than that of EDSR, which has a relatively simple structure and is based on local information. One possible explanation is that HINET is overfitted to high-level global features such as objects.



(a)



(b)

Figure 64. The gradient of input image where HINET is sensitive to global information.

In summary, MIRNET, HINET and HDNET could extract high-level features like objects, scenes, and attributes. These features capture the content and context of the image rather than the low-level, pixel-based information. The extracted high-level features could be used as context information to identify a particular image and further uniquely distinguish a spectral sample from a metamerism set. Therefore, using the high-level global spatial information in spectral super-resolution tasks may

increase the ability of a network to resolve metamerism. However, there are also risks and challenges associated with it:

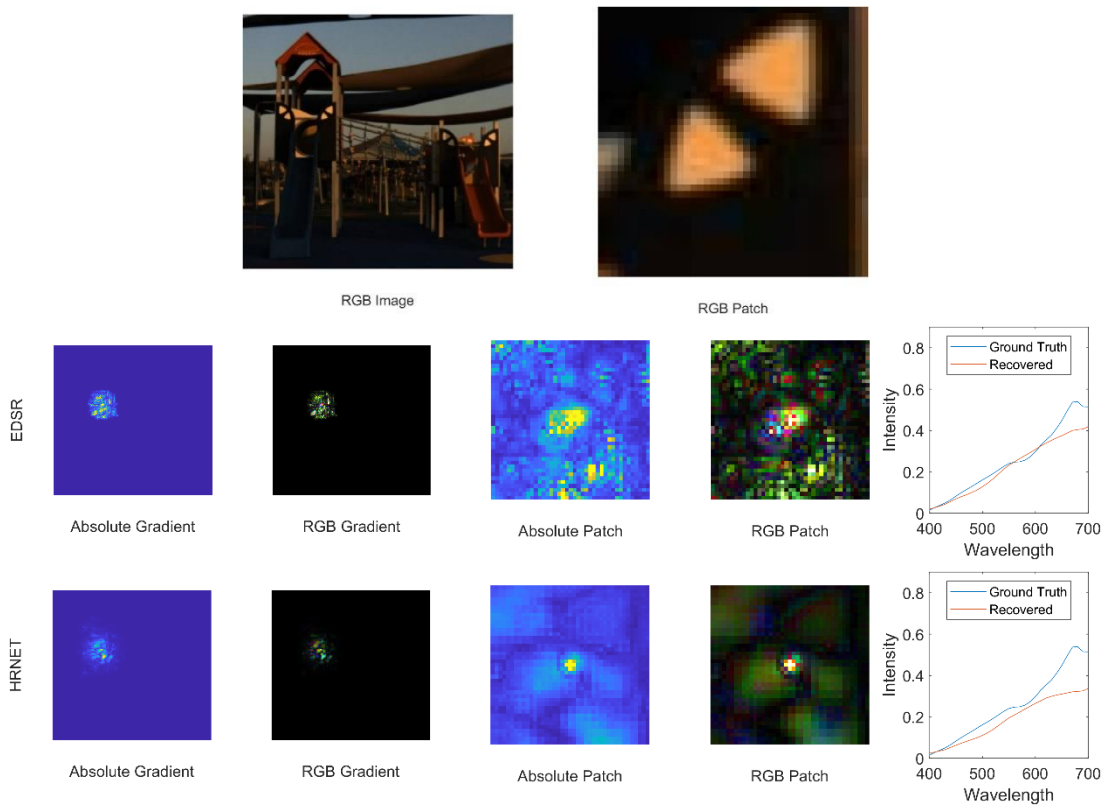
- Overfitting: When a network heavily relies on global spatial information, as is the case with the HINET, it increases the risk that the model may be overfitted to the training data.
- Complexity: Incorporating global spatial information often increases the complexity of the network, resulting in more layers, parameters, and computation. This can lead to longer training times, increased memory requirements, and difficulties in optimizing the network.
- Loss of local details: When a network focuses too much on global spatial information, it may miss or blur important local details, leading to artefacts or loss of fine-grained information in the reconstructed image.
- Interpretability: High-level features, especially those learned by deep learning models, can be difficult to interpret and understand, as they are often represented as abstract, high-dimensional vectors. This lack of interpretability can be a challenge when trying to explain the model's behaviour or debug potential issues.
- Difficulty in disentangling global and local information: In some cases, it may be challenging to separate the effects of global spatial information from local spatial information, particularly when the global context is strongly correlated with local details.

To mitigate these risks, it is essential to carefully design the network architecture, regularization techniques, and training strategies to balance the use of global and local spatial information effectively.

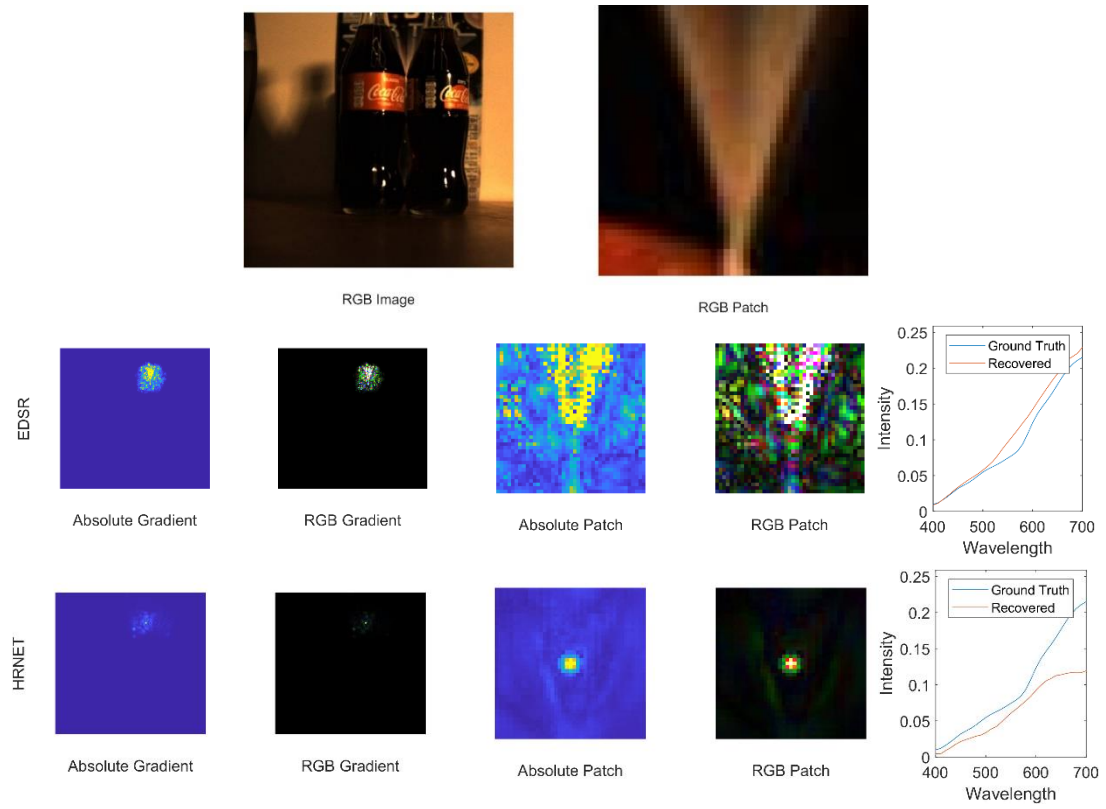
Utilizing global information in image processing can be beneficial, as it allows the network to focus on pertinent parts of an image rather than the entire image. In future research, exploring methods to guide the network's attention towards relevant regions may prove to be a valuable approach. From the literature, MST++ which uses a mask mechanism to guide the focus of networks, has been shown to have higher reconstruction accuracy.

5.5.5. Gradient Analysis of Networks that Extract Local Information

In the context of spectral super-resolution, local information refers to the spatial details and structure within a small neighbourhood of pixels around a target pixel. It includes edges, corners, and textures that are important for preserving the fine details. Since the MIRNET, HINET and HDNET mostly extract local spatial information as a circular area centred by the target pixel and show less relation to features like texture, in this discussion we are going to focus on the EDSR and HRNET only. Figure 65 presents several examples of gradient results for EDSR and HRNET.



(a)



(b)

Figure 65. The gradient of input image where networks are sensitive to local information.

As discussed earlier, EDSR is sensitive to a square area centred around the target pixel, whereas HRNET is sensitive to an arbitrary-shaped area within a specific range. When tackling the same reconstruction task, these networks appear to be sensitive to distinct features within the image patch. In the case illustrated by Figure 65 (a), HRNET is successful in effectively extracting the pattern of the target, while EDSR fails to delineate a clear shape of the patterns. However, neither network is able to learn the shape of the target spectrum, this result could be attributed to the rarity of the material in the training dataset. This assumption is supported by the fact that deep models also fail to learn the spectral shape, as detailed in Appendix 6. In Figure 65 (b), EDSR demonstrates a notable sensitivity to the edge feature, which results in assigning different weights to each side of the boundary. This differentiation becomes evident in the variations observed in EDSR's gradient across the boundary. Conversely, the gradient results produced by HRNET reveal a weak correlation to the edge information within the image patch. More gradient results of all five networks using the local spatial information can be found in Appendix 6.

In this study, it has been observed that EDSR and HRNET adopt different strategies for utilizing local spatial information. The HRNET tends to be highly sensitive to the pixels surrounding the target pixel and demonstrates an affinity for extracting local patterns within the region of interest. Conversely, EDSR exhibits a heightened sensitivity to low-level features, particularly edges, within local spatial information.

This ability of EDSR to discern detailed local structures and patterns could potentially account for its superior performance over HRNET. Through this investigation, we have underscored the importance of local spatial information in spectral reconstruction. It can serve as critical context information, aiding in the precise determination of particular spectra. Extracting spatial information from a fixed local area presents a promising approach for harnessing local spatial information effectively. By refining these strategies, we can further develop our method that uses local spatial information as context to estimate PQR weights.

5.5.6. Gradient Analysis of the Neighbour Pixels of the Target

In this section, we analyse the gradients of the immediate four neighbouring pixels around the target sample (above, below, left, and right). The rationale for studying neighbouring pixels is to assess the consistency of network-based spectral reconstruction. Analysing the gradients of neighbouring pixels, which are typically similar or closely related to the target sample, can provide valuable insights into the strategies employed by these networks. It is important to note that the gradients in this analysis were extracted based on the accuracy of reconstructing neighbouring pixels instead of the target pixel. Figure 66 presents an instance where the gradient outcomes for the immediate neighbouring pixels as similar to that of the target pixel.

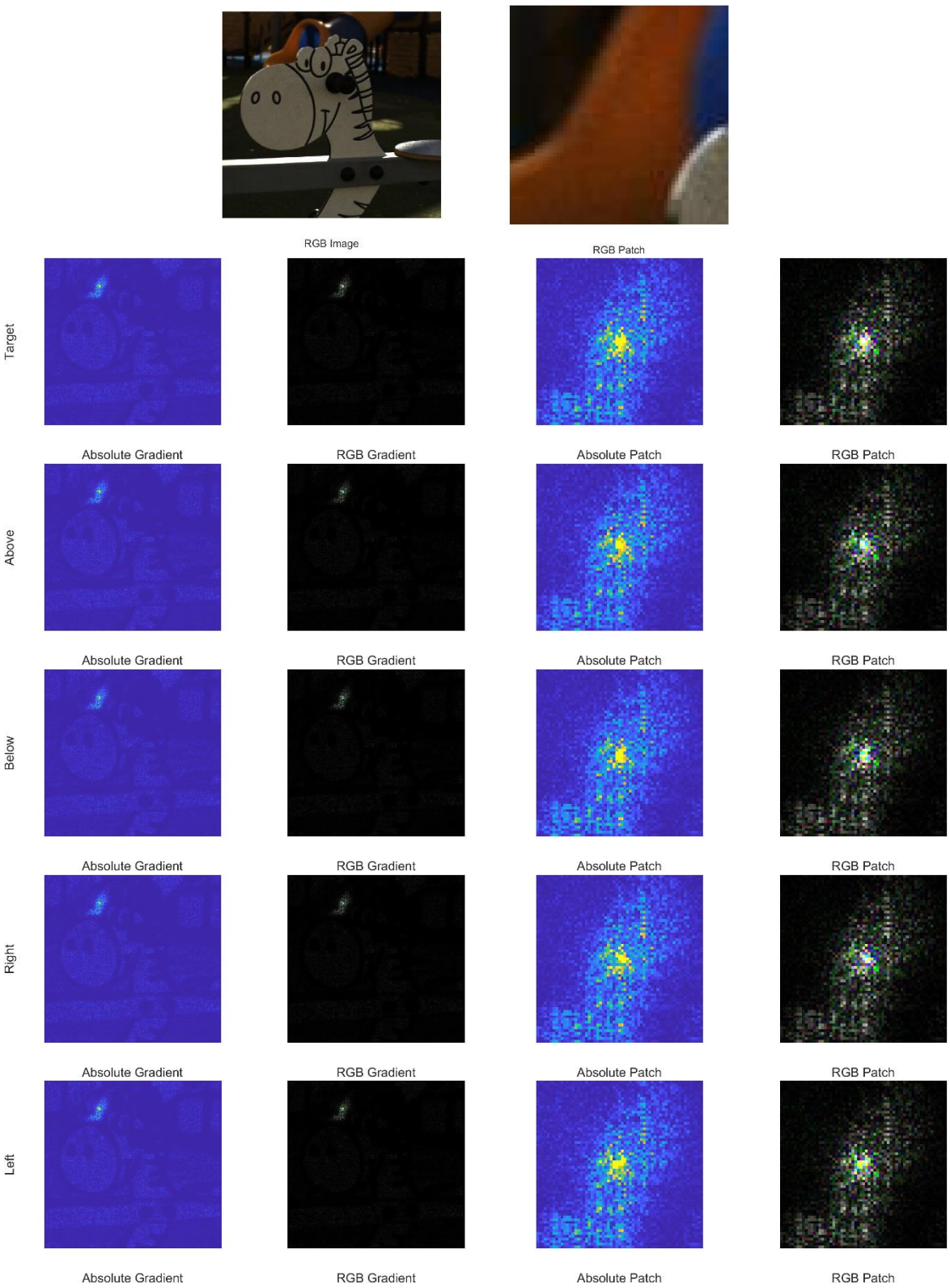


Figure 66. Examples of gradient of neighbour pixels which are similar to the target pixel.

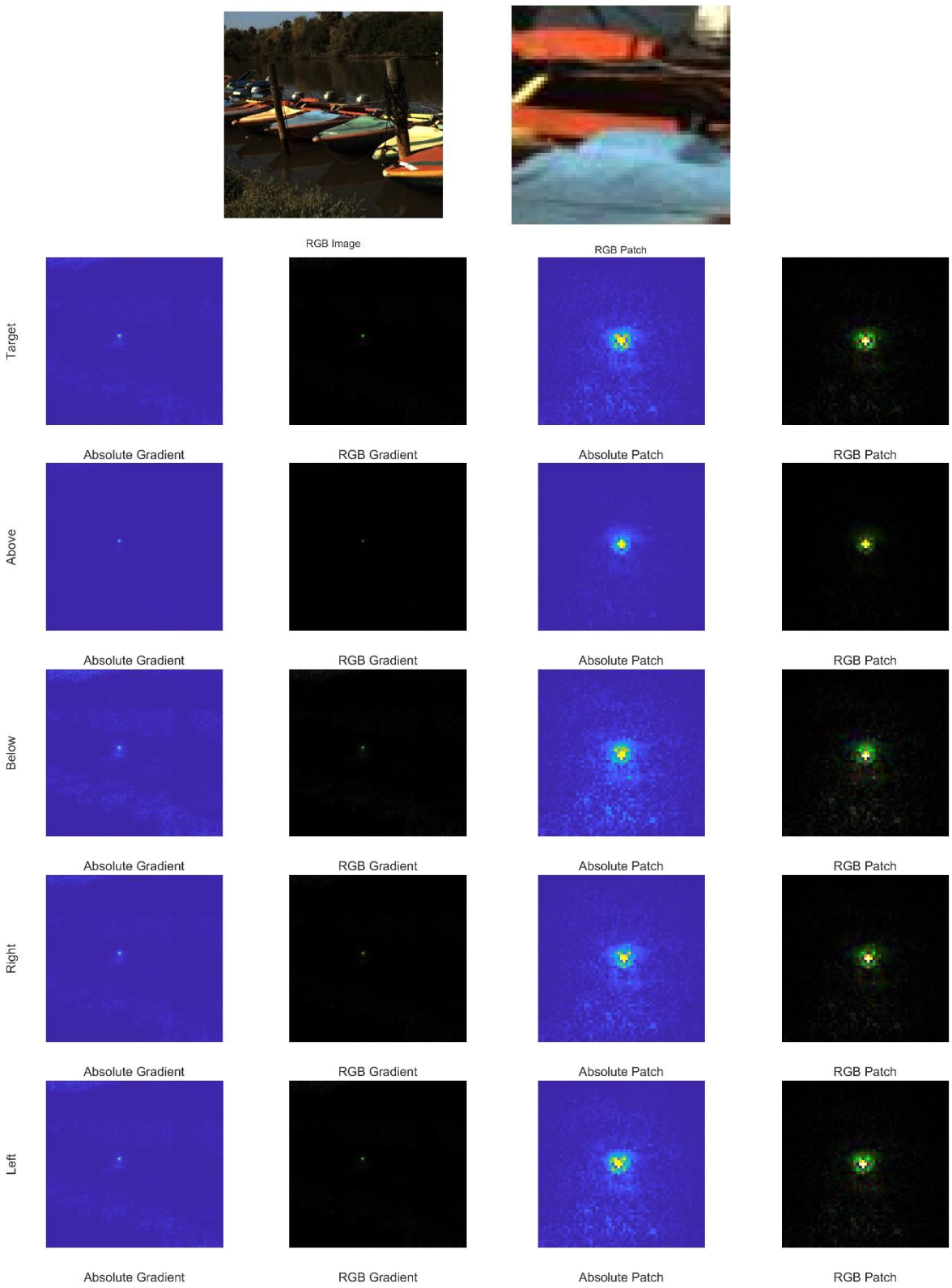


Figure 67. Example of gradient of neighbour pixels which are different from the target pixel.

Upon examining all gathered gradient results from neighbouring pixels, it becomes apparent that when these neighbouring pixels are similar to the target (as shown in Figure 66), the information extracted by a network also exhibits global and local congruency. This insight suggests that networks are adaptive in capturing consistent patterns and structures when neighbouring pixels have comparable characteristics to the target pixel.

Figure 67 showcases instances when neighbouring pixels greatly differ from the target pixel. Such situations arise when the target pixel is near an edge or when the image contains complex patterns. Consequently, neighbouring pixels and the target pixel may originate from distinct materials, or the spectral of the edge pixel may be a mixture of different end-members. In these scenarios, even when the local scene remains consistent, networks extract different local features.

From these observations, one insight can be derived: networks can extract different features from similar local scenes. Moreover, the spatial information extracted bears a relation to the RGB value of the pixel under reconstruction. This suggests that networks dynamically adapt their feature extraction strategies based on the specific characteristics of the pixel being reconstructed.

5.5.7. Conclusion

Based on an evaluation of the gradient results from the five networks, their sensitivity to spatial information reduces significantly when the distance from the target pixel exceeds 5-10 pixels. As such, the area from which these networks predominantly extract spatial information can be approximated to a 20×20 pixels neighbour or even smaller, even though more recent networks possess the capacity to encompass larger regions. This conclusion is further supported by a close observation of the gradient results. Instances, wherein networks exhibit sensitivity to global information, are relatively infrequent among the evaluated samples. Thus, despite potential capabilities for global feature extraction, networks appear to prioritize local spatial information within a limited range around the target pixel.

In the analysis of the local gradient results, it has become evident that networks exhibit sensitivity to the local structural elements such as lines, corners, and edges among others. Additionally, networks can extract textural elements from within local areas. Networks that exclusively lean on local spatial information, particularly EDSR, demonstrate a heightened capability in deriving local spatial information. Considering the relatively superior spectral reconstruction accuracy exhibited by EDSR, it could be inferred that local spatial information can be employed as supplemental data to alleviate metamerism.

When it comes to global spatial information, there are limited instances indicating that networks exhibit sensitivity to high-level global information, such as objects. Although such high-level global

information can serve to uniquely identify a training image and its associated spectral reflectance, it also carries a potential risk. HINET which has the highest sensitivity to global information, didn't provide a higher reconstruction than other networks. However, from the literature, MST++, which uses a masking mechanism to guide the attention of the network on extracting global information, appears with relatively higher reconstruction accuracy. In future designs of deep learning models for spectral super-resolution methods, it is advisable to steer the network's focus towards more relevant areas within the image scene.

5.6. Summary and Conclusion

At the beginning of this chapter, our experiments revealed that local texture information can impact HSCNN+'s ability to reconstruct the shape of the spectrum. However, it's important to note that this case study was based on a single specific metamerism sample and hence, might not be broadly representative.

To draw inspiration from existing neural networks for our explainable method, we conducted two sensitivity analyses and leveraged automatic differentiation to visualize areas of the image that impacted reconstruction accuracy. The experiments were tailored to answer two research questions: the required spatial size to resolve metamerism, and what type of spatial information can be utilized. Both sensitivity analysis and gradient analysis suggest that networks are most sensitive to pixels close surrounding the target pixel, with sensitivity decreasing as the distance from the target pixel increases. Sensitivity analysis further indicates that networks primarily rely on local information to determine the shape of the reconstructed spectrum. Within the local area, according to the gradient analysis, networks are sensitive to structure and texture; removing textural information impacts the shape of the reconstructed spectrum.

In conclusion, our findings suggest that networks predominantly rely on local spatial information to resolve metamerism. A circular region with a radius of 20 pixels, or even less, could provide adequate information for networks to ascertain the shape of the recovered spectrum in most cases. Tested neural networks are sensitive to local texture features, and removing these features would eliminate the contextual information necessary for neural networks to resolve metamerism. However, neural networks do not show a preference for any local texture in a particular scale or orientation. Different samples were sensitive to different scales and different orientations. Therefore, when designing our feature extraction method, a general texture descriptor would be beneficial.

Based on these findings, in the next chapter, we will present an explainable spectrum super-resolution method that uses features extracted by multiple scale Local Binary Pattern (LBP) as context to estimate the PQR coefficients.

Chapter 6. Single Image Spectral Super-Resolution Using RGBPQR

In the previous sections, we introduced the RGBPQR colour space and illustrated its advantages in spectral reconstruction tasks. The RGB components of a standard RGB colour image directly make up the first three components of the colour space. These directly weight corresponding spectral bases that ensure colour consistency. The residuals are then represented using PCA components as PQR. Estimating the PQR residual weights directly from the RGB values without supplementary information leads to limited accuracy due to metamerism. From analysing deep convolutional neural network-based methods, it has been shown that spatial information, particularly local texture, holds promise as contextual information for addressing the metamerism issue. Consequently, in this section, we will explore the use of the RGBPQR colour space for single image super-resolution, while employing local texture as context to tackle metamerism in an interpretable manner.

6.1. Using Local Binary Patterns to Resolve Metamerism

6.1.1. Introduction

In the last chapter, by analysing the existing deep convolutional neural network-based methods, we demonstrated that local textural features from a limited area are essential in resolving metamerism. Surfaces of particular materials often exhibit distinctive textural characteristics. Therefore, by characterizing the spatial features of these surfaces, including their texture, it is possible to distinguish the materials and further resolve the metamerism problem. In the last section, we demonstrated that networks are sensitive to textures detected by Gabor filters. However, Gabor filters face a challenge in converting results from a filter bank into a context vector. Unlike Gabor filters, Local Binary Patterns (LBPs) do not have this issue, so we chose LBPs as our texture descriptor in this chapter.

LBPs are a simple and efficient texture descriptor first introduced by Ojala *et al*. (1994). The LBP operator works by comparing the intensity of a central pixel with its surrounding neighbours in a local neighbourhood. It then generates a binary code based on whether the intensity of the neighbouring pixels is greater than or equal to the intensity of the central pixel. A histogram of LBP values can be calculated for an entire image or specific regions within the image. The LBP histograms can then be used as feature vectors for various tasks, such as classification, recognition, or comparison. The key benefits of LBP include its computational efficiency, robustness to monotonic changes in illumination, and its ability to capture local texture information effectively (L. Liu *et al.*, 2017). Given that LBPs have shown remarkable performance in texture classification tasks, there is a strong basis for our hypothesis that the feature vector extracted by LBPs could be used to distinguish metamerism samples. Before testing the LBPs to resolve metamerism in spectral reconstruction, in this section, we will test whether it is possible to use LBP histograms to distinguish the metamerism samples.

6.1.2. Dataset and LBP Histogram Extraction

- **Dataset**

In this experiment, we collected a new dataset based on the NTIRE 2022 database. We produced the RGB values in the same manner as described in Chapter 3, which differs from the original approach in NTIRE 2022. To investigate the metamerism issue in depth, we compiled this dataset with a focus on metamer RGB values. A total of 5,000 unique RGB values were utilized as benchmarks to identify associated spectral data from the NTIRE 2022 database. These RGB references were randomly selected, fitting within the gamut span of the NTIRE 2022 dataset. Depending on the occurrence frequency of each reference RGB value within the database, we collected anywhere from hundreds to thousands of samples per value. Notably, a single RGB reference value could be associated with multiple materials appearing with varying spectral shapes. In total, our comprehensive collection process resulted in over 1.5 million spectral samples.

While collecting the training spectral data, the RGB values from a 31×31 image patch centred on each target sample have also been collected. In Chapter 5, we have demonstrated that a clear neighbour with a radius of 20 pixels shows an insignificant difference compared with the original image; while the gradient analysis also shows that pixels that are more than 10 pixels away from the target sample show insignificant influence on the reconstruction result. Therefore, in this experiment, we limited the size of local spatial information to a 31×31 patch for each sample.

- **Local Binary Patterns**

In this experiment, we implemented two distinct Local Binary Pattern (LBP) methods, both of which are computed from the grayscale image. The first method examines the eight neighbouring pixels surrounding each central pixel in a clockwise sequence, beginning with the top-left pixel. The centre pixel is used as the threshold. The resulting 8-bit local binary values range in value from 0 to 255. From here, we create an LBP feature vector, generated by the histogram of local binary pattern values derived from the 31×31 image patch centred by the metamerism sample. This results in a 256-dimensional feature vector.

As inferred from Chapter 5, networks make use of local textures from multiple scales to resolve metamerism. Therefore, we introduce the Multi-Scale LBP (MLBP) as the second method. Unlike the single-scale LBP, which only assesses neighbouring pixels, MLBP measures pixels at three distinct scales: 1, 5, and 9, enabling it to capture both fine and large-scale textures. Figure 68 illustrates the pixels evaluated at different scales. The left side of Figure 68 is the same as the single-scale LBP, which measures the immediate neighbouring pixels. In the middle of Figure 68, the pixels assessed are approximately 5 pixels distant from the target pixel. On the right side of Figure 68, the evaluated pixels are approximately 9 pixels away from the target. Compared to LBP, MLBP captures texture features over a range of scales. When constructing the feature vector, each scale contributes its own histogram. These three histograms are then concatenated to produce the final feature vector, which has 768 dimensions.

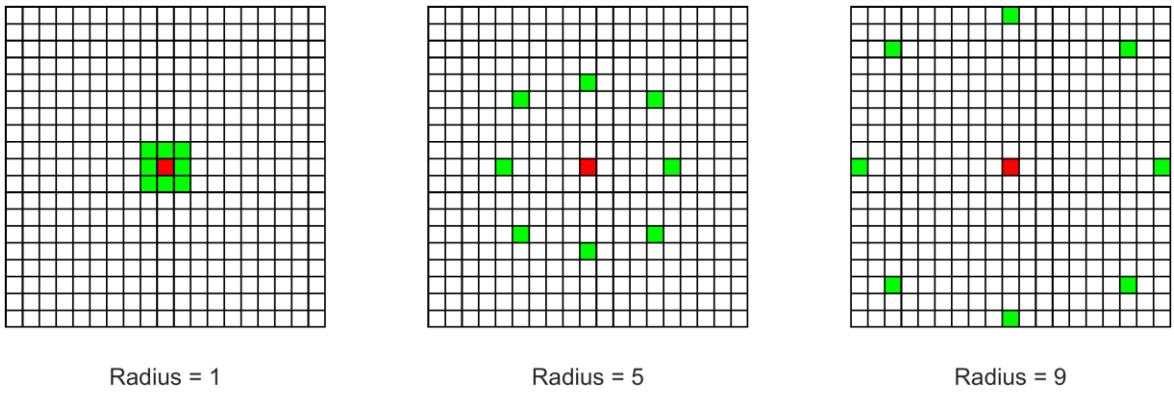


Figure 68. MLBP with different scales, the red pixel is the target pixel, while the green pixels show the pixel that has been compared to generate the LBP code of the target pixel.

Figure 69 displays an example image patch along with the extracted MLBPs at three different scales. From the results, we observe that the LBPs with a radius of 1 can extract fine details, while the LBPs with radii of 5 and 9 capture larger-scale features.

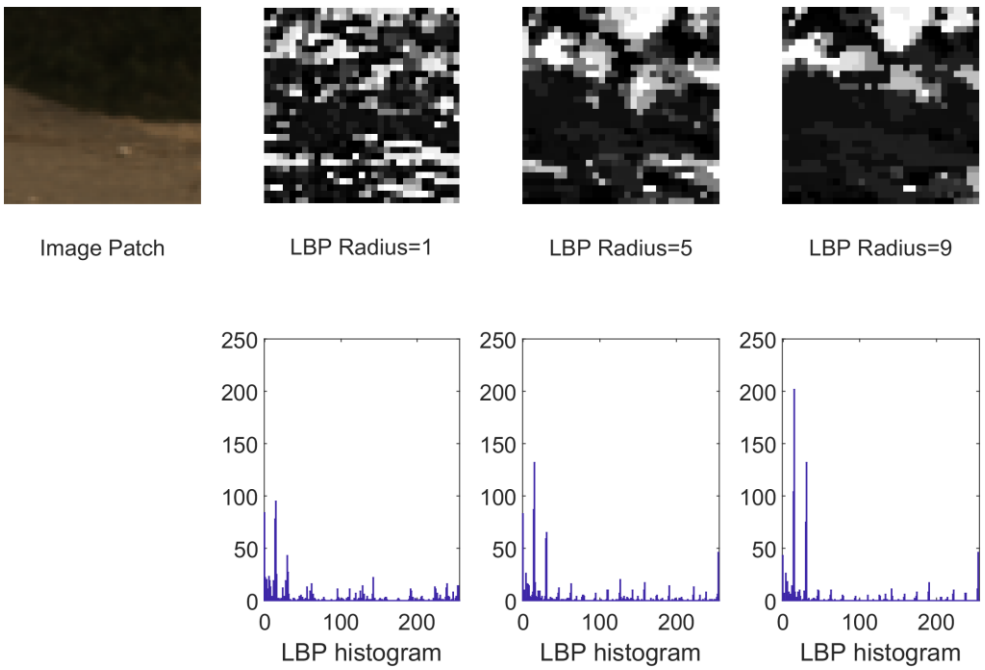


Figure 69. Examples of extracted features on different scales, and their corresponding histogram.

6.1.3. Using LBP Context to Distinguish Metamerism Samples.

- **Single-Scale Local Binary Pattern**

Figure 70 depicts the spectra of an example metamerism set, constituting two distinctive spectral groups (Group 1 and Group 2).

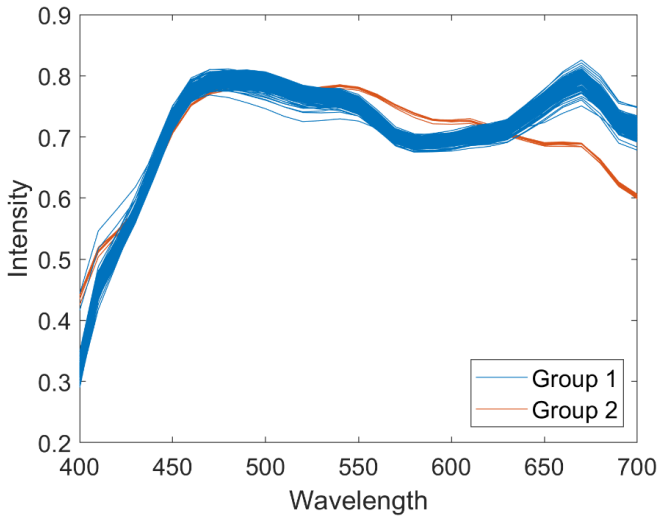


Figure 70. Spectral reflectance from example metamerism set.

Figure 71 presents the image patches from both spectral groups, along with their corresponding spectra and LBP histograms. The left three columns represent image patches related to Group 1, while the right three columns display patches from Group 2.

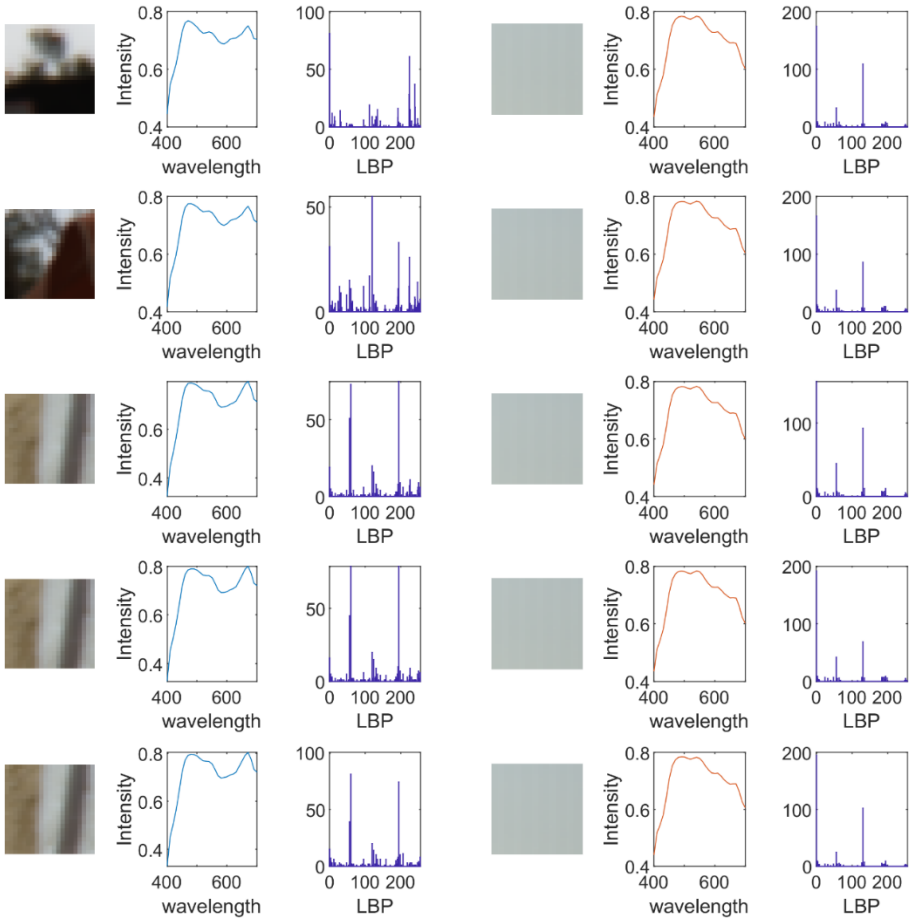


Figure 71. Example image patches from both spectral groups together with their corresponding spectra and LBP histograms.

In this example, image patches from Group 1 exhibit varied patterns, while the image patches from Group 2 are all flat which corresponds to the sky. Remarkably, LBP histograms between the two groups, given their distinct patterns, are noticeably different. In this context, it is evident that the LBP vector has the potential to differentiate metamerism samples. However, an important observation to note is the variability in the appearance of image patches within Group 1. This suggests that a single material could have multiple local structures. This variability increases the complexity of capturing the relationship between a material's local texture and its associated spectrum. It also highlights the difficulty of using the LBP descriptor in encoding such variations.

To assess the viability of using the LBP histogram for distinguishing metamerism samples, we employed t-SNE to visualize the clusters or groupings (data structure) within the data of these high-dimensional vectors in a simplified 2-dimensional format. The left of Figure 72 presents the t-SNE plot of the 256-dimensional LBP histograms from the example metamer samples. Spectral samples from Group 2 form a distinct cluster, clearly separate from Group 1. This result suggests the potential of leveraging LBP histograms, representing local spatial information, to differentiate metamerism samples. However, the samples in Group 1 also appear with multiple clusters. This variation can be attributed to the unique local textures of the corresponding samples. A crucial aspect to acknowledge is that a single material could correspond to multiple types of local textures, thereby leading to a complex relationship between the local texture and the spectra.

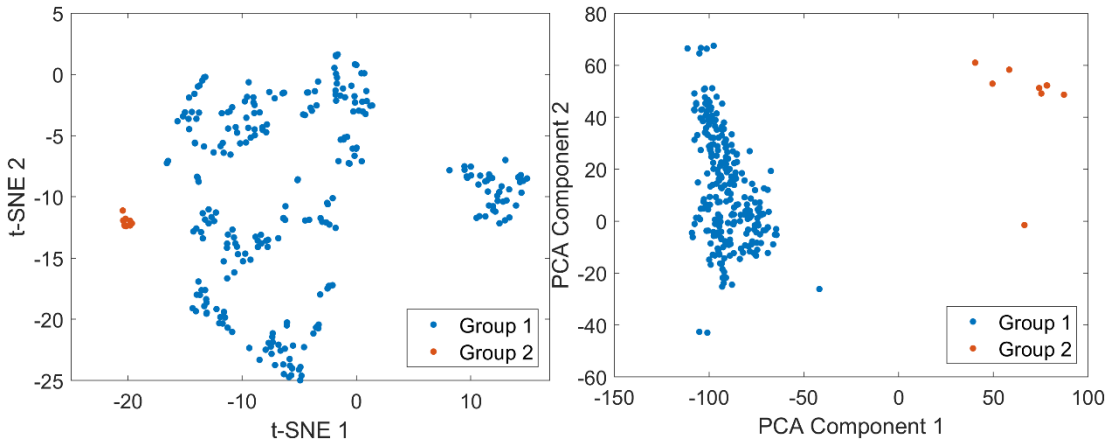


Figure 72. Left: the t-SNE result of the 256-dimensional LBP vector; Right: Sample distribution in PCA space defined by the first two components. Sample dots are colour-mapped based on their corresponding spectral group.

However, using all 256 dimensions of the LBP vector as context can significantly increase the model's complexity and raise the risk of overfitting. To address this, we applied PCA to the LBP histogram vectors, revealing that the first three components accounted for 87.7% of the total variance. When considering the first six components, this explanation of variance rises to 91.9%. It's important to note that these LBP histograms discussed herein originate from all 5,000 metamerism sets. The right part of Figure 72 uses the weights of the first two PCA components to cluster the example metamer samples. This clearly shows that the first two components could distinguish the Groups for this metamerism set.

Figure 73 depicts another example of a metamerism set. This set, like the previous one, consists of two types of spectra which exhibit differences primarily in the red wavelengths. However, compared to Example 1, the local structures within this example appear to be considerably more complex for both spectral groups. Correspondingly, there is also notable variation within the LBP histograms of each group.

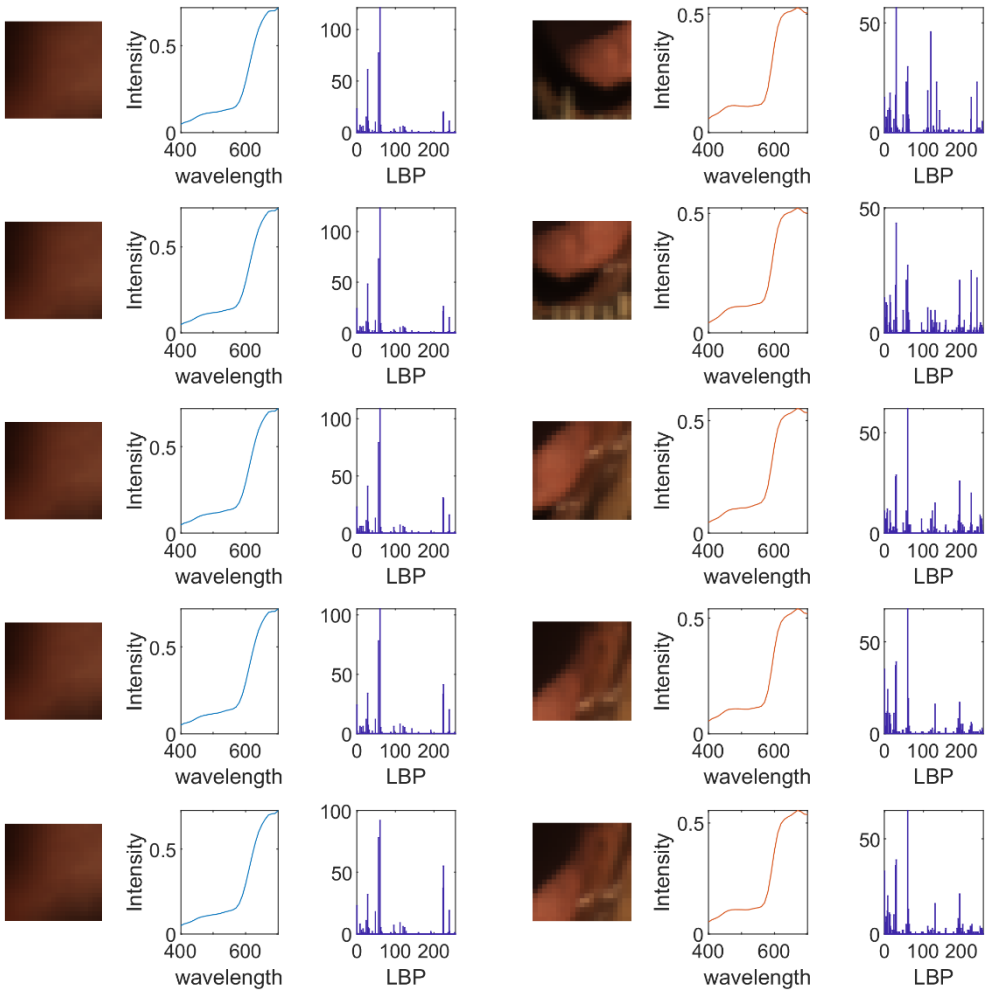


Figure 73. Image patches from both spectral groups together with their corresponding spectra and the LBP histograms.

The left panel of Figure 74 shows the t-SNE plot when using the 256-dimensional LBP histogram vector as a reference, it is clear the two spectral groups have been separated. The right panel of Figure 74 illustrates that in cases such as this, where the local structure is relatively complex, the first two components are insufficient to fully differentiate between the metamerism samples. However, when using the first six components as a reference the metamerism samples could be separated. Figure 75 shows the t-SNE plot of the weights from the first 6 PCA components.

From the previous illustrations and the result from analysing the other metamerism samples, the local spatial information represented by LBP histograms can be used to distinguish metamerism samples. However, since one material could be associated with multiple types of local structure, the relation between the local structure and the corresponding spectral reflectance is relatively complex. In some cases, detailed spatial features are required to separate metamerism samples.

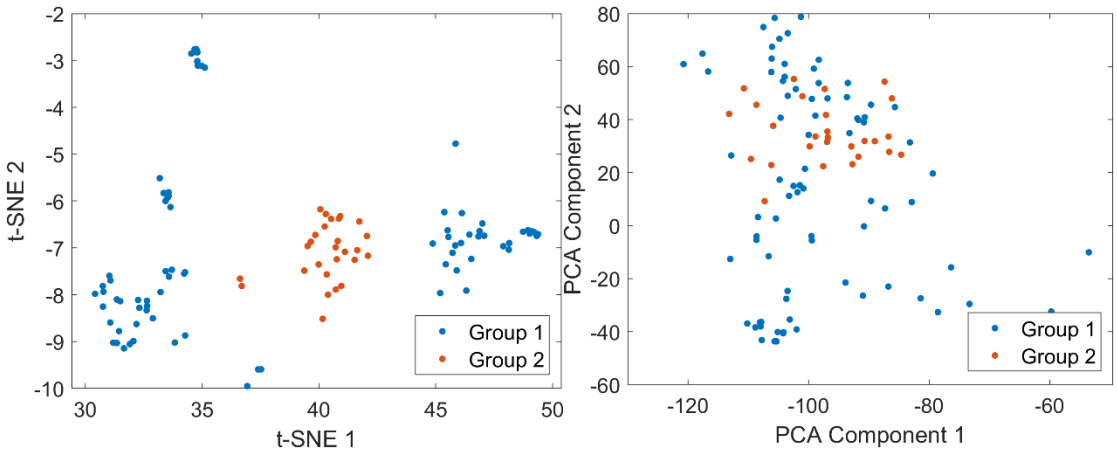


Figure 74. Left: the t-SNE result of the 256-dimensional LBP vector; Right: Sample distribution in PCA space defined by the first two components. Sample dots are colour-mapped based on their corresponding spectral group.

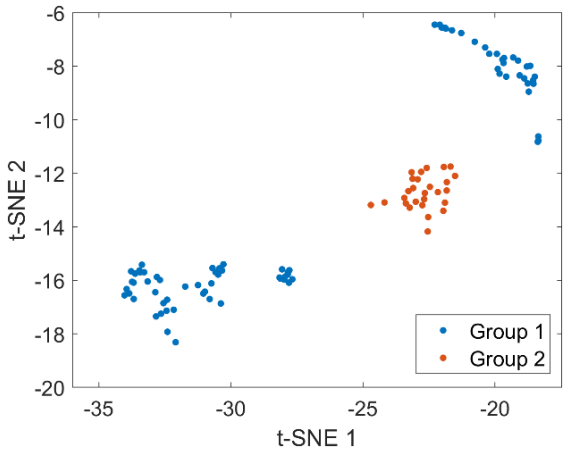


Figure 75. t-SNE result of the first 6 PCA components, sample dots are colour-mapped based on their corresponding spectral group.

- **Multi-scale Local Binary Patterns**

In this experiment, we will also evaluate the effectiveness of MLBPs in distinguishing metamerism samples. Figure 76 displays the image patch, target spectra, and the MLBP histogram for the same metamerism set as depicted in Figure 73. The MLBP histogram consists of 768 bins, which combine the LBP histograms from three scales: 1, 5, and 9. The displayed samples have been chosen at random.

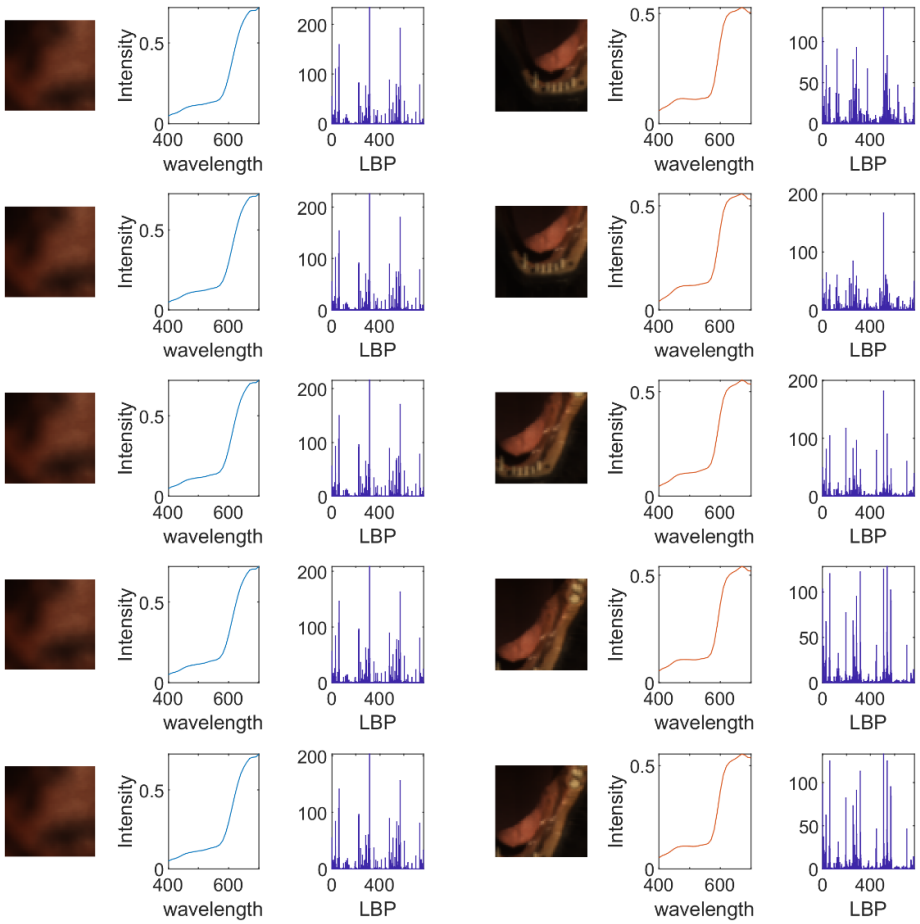


Figure 76. Image patches from both spectral groups together with their corresponding spectra and the MLBP histogram.

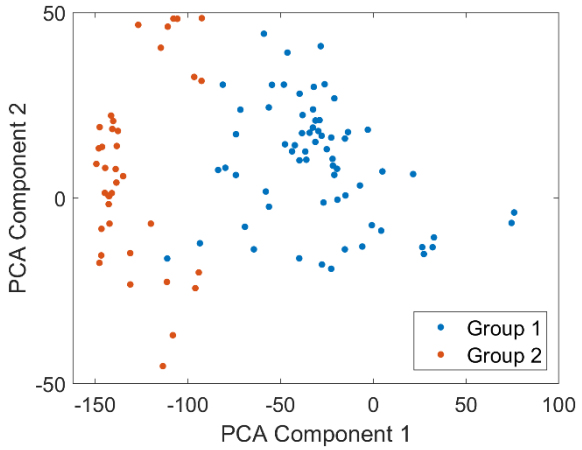


Figure 77. Sample distribution in PCA space is defined by the first two components of MLBPs. Sample dots are colour-mapped based on their corresponding spectral group.

Figure 77 shows that the weight of the first PCA components of the MLBP histograms can effectively distinguish the metamerism samples. Thus, incorporating multiscale textural features can aid in distinguishing metamerism samples with a low dimensionality in this case. However, given the complexity of the 768-dimensional histogram, the first three PCA components account for only 74.2% of the total variance, and even the first six PCA components explain only 83.9% of the total variance. Therefore, relying solely on a limited number of components might result in a loss of valuable detailed textural features derived from the MLBPs.

- **Single and Multi-scale LBPs on Distinguishing Metamer Samples**

The previous sections demonstrated the potential effectiveness of LBPs as contextual information for distinguishing metamer samples in various case studies. In this section, our goal is to assess how well LBPs can differentiate metamer samples when used as an index in a dictionary, considering all the collected samples. Table 16 provides a comparison of the accuracy achieved when representing spectral samples using a dictionary. This dictionary uses either RGB values with LBPs or MLBPs as indices for estimating 3 PQR weights. The context vector is composed of the weights of the first six PCA components of the LBP histogram vectors. The construction of the dictionary follows a similar methodology as described in Chapters 3.7.1 and 4.2.3. The present test dictionary incorporates both RGB and the tested context for indexing. To build this dictionary, we compress each RGB value and context component into 6 bits to mitigate overfitting. Specifically, we divide the samples into sub-cubes, and the corresponding PQR weights for each sub-cube are determined as the mean of the PQR weights of the residual within that sub-cube. During the reconstruction process, the input RGB value and context serve as the index to retrieve the corresponding PQR values.

Table 16. Comparing single and multiple scale LBP on resolving metamerism

	RMSE	SAM	95% RMSE	95% SAM
Up sampled	0.024	0.071	0.055	0.216
Best	0.007	0.019	0.014	0.054
LBPs	0.007	0.020	0.014	0.056
MLBPs	0.007	0.020	0.014	0.057

During the reconstruction process, the input RGB value and context serve as the index to retrieve the corresponding PQR values. The 'Best' row in the table reflects the representation error of the RGBPQR model using only three residual components and assuming the PQR weights are estimated accurately. Remarkably, when employing LBPs as context to build a dictionary, the representation accuracy closely approaches the best achievable accuracy. These outcomes indicate

that utilizing textural features from both LBP and MLBP can markedly distinguish metamer samples. It is essential to emphasize that the results presented in Table 16, which showcase the representation error, represent the best achievable performance of a dictionary-based method. However, given the memory consumption associated with using a huge dictionary, the remainder of this chapter focuses on developing a smaller yet effective model to learn the mapping.

6.1.4. Conclusion

In this section, we have showcased the effectiveness of both single and multi-scale LBP histogram vectors in distinguishing metamerism samples. The ability of LBPs to capture local texture variations could play a crucial role in addressing challenges related to metamerism in spectral super-resolution. The next challenge is to accurately estimate the PQR weights using RGB values, and the contextual information provided by LBPs.

6.2. Proposed Single Image Spectral Super-Resolution Method

In this section, we aim to leverage the RGBPQR colour space as a spectral model for recovering spectral information from individual RGB images. Concurrently, we utilize spatial information derived from LBPs to address the metamerism problem. The RGB value of the target pixel, combined with the spatial context represented by the LBP code, will be used to estimate the residual component weights (PQR) as a regression problem. Given the non-linear relationship between the PQR coefficients and the context vector, the regression method must be able to model such complex mapping. Therefore, we have chosen to employ neural networks to address this regression problem. The next section describes our proposed method.

6.2.1. Methodology

- **Dataset**

Similar to Chapter 3, the spectral data for this experiment was sourced from the NTIRE 2022 dataset (Arad *et al*., 2022). A total of 1,980,000 samples were collected from 900 images. These were divided into 1,800,000 training samples and 180,000 validation samples. The RGB values were generated in the same manner as in the previous chapters, which differs from the original NTIRE 2022 dataset. While collecting the training spectral data, the RGB values from a 31×31 patch centred around each target sample were also collected.

- **Training Neural Network to Estimate the PQR Weights from RGB Value and Feature Vectors.**

In this section, our objective is to develop a model capable of estimating PQR coefficients using RGB values, with the local spatial feature vector serving as context. Neural networks, known for their ability to learn complex nonlinear mappings, are our model of choice for this experiment. (Potentially any other non-linear regression mapping could be used.) To enhance the model's ability to represent the mapping while maintaining interpretability, our approach utilizes shallow networks consisting of three fully connected layers to deduce PQR coefficients from RGB values alongside feature vectors. Each layer contains 256 nodes and an associated ReLU activation layer. Various network structures were tested, and in this section, we report only the model that demonstrated outstanding accuracy. The MATLAB regression learner was used for network training, employing a 5-fold cross-validation strategy. The network's primary goal was to minimize the disparity between the predicted PQR coefficients and the benchmarked ones.

In the regression, the input is a concatenation of the RGB value and the spatial context vector. This spatial context originates from the LBP histograms extracted from a 31x31 image patch centred on the target pixel. Both single and multiple-scale LBP introduced in the previous section have been tested. The scores (weights) from the first six principal components serve as the context vector. Our objective in testing these varied LBP vectors is to identify the most effective spatial context.

This combined 9-dimensional vector is then used as input for the trained model to estimate the corresponding PQR weight. The network is trained to minimize the RMSE between the estimated PQR coefficients and the reference PQR coefficients, a higher estimation accuracy in the estimated PQR coefficient could lead to a better spectral reconstruction accuracy. When the residual components (PQR) weights have been predicted, the spectral reflectance can be recovered as described in Chapter 3.

6.2.2. Results and Discussion

- **Validation**

This section examines the training error based on the validation data collected from the training images. Conducting in-sample tests allows us to assess the model's ability to estimate the residual and mitigate the effects of metamerism. Moreover, comparing these results with the testing data will help ascertain if the model is prone to overfitting. Table 17 presents a comparison of the reconstruction errors for the training spectral samples. This includes the up-sampled spectrum discussed in Chapter 3, spectra where the residual is estimated solely from RGB values, and spectra

where the residual is predicated using the regression model. Mean error and 95% error associated with both single-scale LBP and multi-scale LBP are included in the comparison.

Table 17 indicates that including the spatial context from LBP enhances reconstruction accuracy. This demonstrates that the LBP-based spatial context can mitigate the PQR ambiguities related to metamerism. This marked increase in reconstruction accuracy demonstrates that the spatial insights from LBP effectively tackle the metamerism challenge within the training dataset.

Table 17. Comparison of reconstructing validation spectra by different ways of estimating the PQR weights.

	Up-sampled	PQR estimated from RGB	PQR estimated from RGB+LBP	PQR estimated from RGB+MLBP
RMSE	0.025	0.023	0.016	0.016
Mean SAM	0.136	0.126	0.105	0.106
95% RMSE	0.049	0.044	0.034	0.035
95% SAM	0.371	0.363	0.312	0.307

Figure 78 compares the error distributions measured by SAM when reconstructing the validation dataset using the trained models. It's evident that using LBP or MLBP as context could significantly reduce errors in the reconstructed spectra. The proportion of larger SAM (larger than 0.2) has been reduced from 24% from the up-sampled to 12% for both LBP and MLBP solutions. This improvement is particularly noticeable for samples that are affected by metamerism.

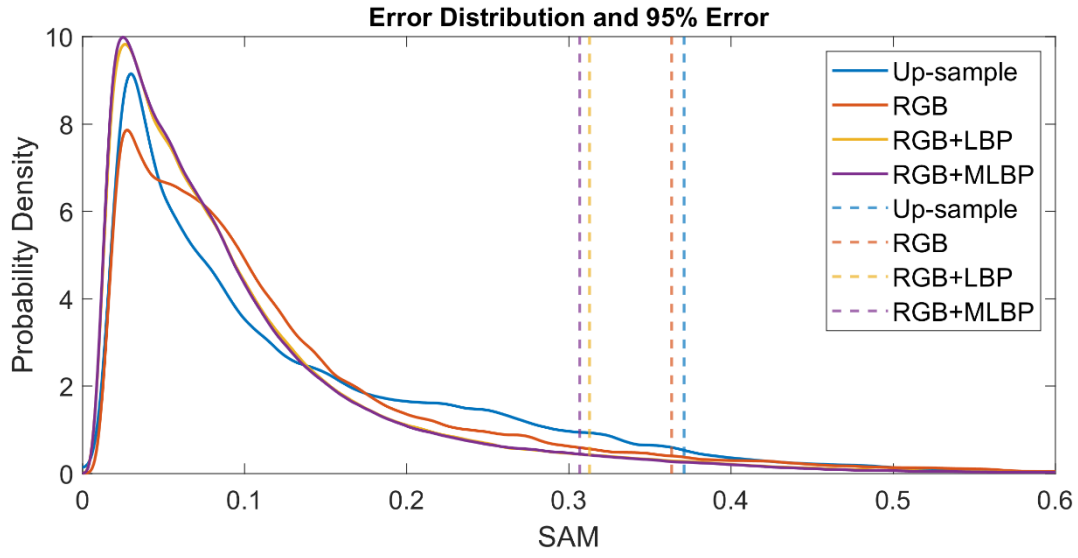


Figure 78. SAM distribution and the 95% error (dotted line), the legend shows how the residual spectra are estimated. From the error distribution adding spatial context extracted from LBP could reduce the influence from the metamerism.

- **Collected Testing Samples.**

To evaluate the trained model's efficacy, we sourced a test spectral dataset from an additional 50 spectral images supplied in the NTIRE 2022 dataset. From each image, 2,000 spectral samples were randomly selected, giving a total of 100,000 samples. The RGB values of the test samples were produced using the identical procedure employed for the training dataset.

The testing results shown in Table 18 indicate that the strategy utilizing RGB values, concatenating with local spatial details as context, consistently outperforms the others, as evidenced by both the mean and the 95% error metrics. This underscores the pivotal role of spatial information, represented here by the local binary pattern, in improving reconstruction precision by reducing the influence of metamerism.

Table 18. Comparison of reconstructing testing spectrum with different ways to estimate PQR weights in terms of two performance metrics.

	Up-sampled	PQR estimated from RGB	PQR estimated from RGB+LBP	PQR estimated from RGB+MLBP
RMSE	0.026	0.024	0.021	0.020
Mean SAM	0.129	0.118	0.100	0.098
95% RMSE	0.055	0.053	0.050	0.046
95% SAM	0.368	0.304	0.285	0.282

Figure 79 illustrates the error distribution and the 95% error of the testing dataset, represented as SAM. Our focus on SAM stems from its connection to the reconstructed spectrum's shape. It is observed that the error distribution appears with three variations, which correspond to the up-sampled spectra (without residual), estimating the residual solely using RGB, and using RGB plus spatial context to recover the residual. While the peaks of these error distributions are closely aligned, it's noteworthy that when the residual is derived from the spatial context, the proportion of samples with a pronounced SAM (exceeding 0.2) diminishes. Similarly, as seen in the validation results, the proportion of large SAM values has been reduced from 23% in the up-sampled method to less than 12% in the LBP-based methods. Additionally, the proportion of samples with a SAM less than 0.1 has increased by 10% with the introduction of textural context represented by LBPs. In comparison with the validation results, there is no evidence that the model has been overfitted. The MLBP-based method shows a slight improvement in the reconstruction accuracy compared to the single-scale LBP.

In comparison with the reconstruction accuracy when solely using RGB values to estimate the PQR weights, the increase in reconstruction accuracy is attributed to using local texture to resolve metamerism.

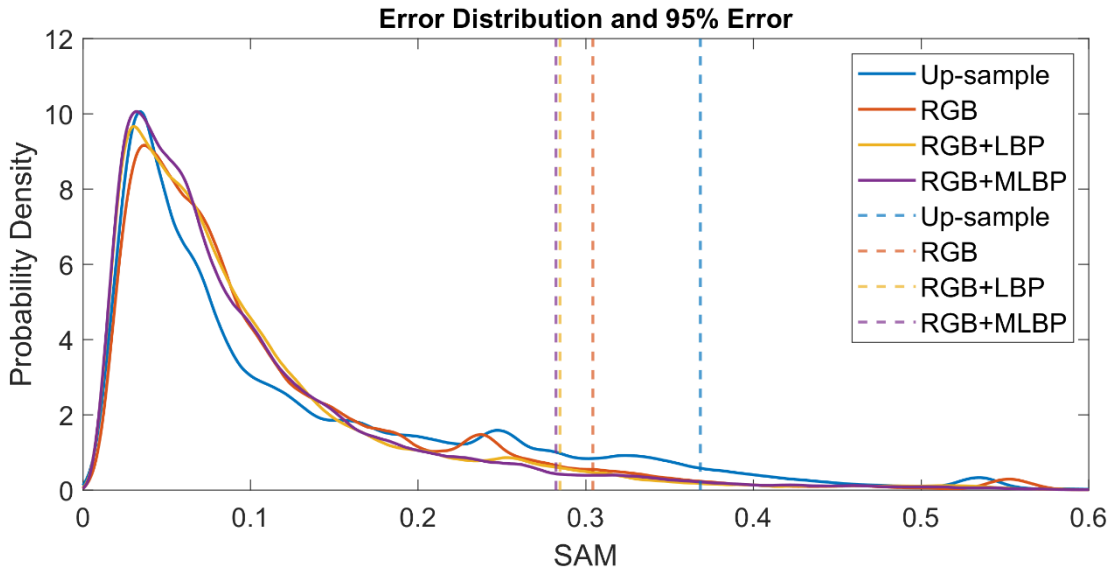


Figure 79. SAM distribution and the 95% error (dotted line). From the error distribution adding spatial context extracted from LBP could reduce the influence from the metamerism.

- **Quantization Error**

In this experiment, the RGB values were quantized to 8 bits in the same manner as NTIRE 2018, 2020. However, the quantised RGB values would introduce quantization errors, particularly for darker scenes. Normalizing the RGB values with a global maximum introduced higher relative quantization errors in darker scenes compared to brighter scenes. To assess the extent to which the reported reconstruction error has been influenced by quantization errors, we tested using the real RGB values to learn the RGBPQR model and train the regression neural networks. The real RGB values refer to linear RGB values calculated from spectral radiance and camera functions without quantization. Table 19 presents the mean and 95% reconstruction error for the testing data. It's important to note that these results are derived from the same testing spectral dataset presented in the previous sections, and MLBP was used as context.

It is evident that eliminating the quantization error in RGB values can lead to a slight improvement in reconstruction accuracy. This improvement can be explained by the fact that quantization introduces additional metamerism samples where, without quantization, the real RGB values are distinguishable. However, it's important to note that quantization errors do not alter the overall trend of reconstruction accuracy when using MLBP to estimate the PQR weights. To maintain alignment with real-world scenarios where RGB values are typically quantized, and to ensure consistency in our discussion, the remainder of this section will continue to be based on quantized RGB values.

Nevertheless, it's worth noting that it is possible to further enhance reconstruction accuracy by using real RGB values or higher precision, such as 12-bit RGB values.

Table 19. Comparing real RGB values with quantised RGB values for reconstructing the test sample.

	Real values		Quantised to 8 bits	
	Up-sampled	PQR estimated from RGB+MLBP	Up-sampled	PQR estimated from RGB+MLBP
RMSE	0.025	0.021	0.026	0.021
SAM	0.116	0.094	0.129	0.100
95% RMSE	0.052	0.048	0.055	0.050
95% SAM	0.340	0.265	0.368	0.285

- **Reconstruct Complete Image**

In this section, we evaluate the performance of the trained model on complete images, rather than collected spectral samples. We assessed the model using an additional 50 hyperspectral images from the NTIRE dataset. These images are distinct from those used for training. Due to the model's requirement for a 31×31 image patch to reconstruct the centre pixel, we only reconstruct pixels that are at least 15 pixels away from the image edge during our image reconstruction process. Nonetheless, this narrow edge constraint will not adversely impact our analysis of the reconstructed image results. The reason we didn't use a padding technique for full image reconstruction is that such padding would adversely distort the LBP used to represent patch textual information.

Table 20. Reconstructing complete images with different methods to estimate the PQR weights.

	Up-sampled	PQR estimated from RGB	PQR estimated from RGB+MLBP
RMSE	0.026	0.024	0.022
SAM	0.127	0.110	0.103
95% RMSE	0.049	0.045	0.041
95% SAM	0.260	0.232	0.213

The data indicates that incorporating spatial context, derived from multi-scale LBP, reduces the error in reconstructing real images, both in terms of mean error and the 95% error. Thus, for actual images, the proposed method effectively enhances reconstruction accuracy by partially resolving metamerism.

Figure 80 compares the reconstruction accuracy of an example image, as measured by the absolute error, at three selected wavelengths corresponding to blue, green, and red wavelengths. Results of the reconstruction of other hyperspectral images can be found in Appendix 7.

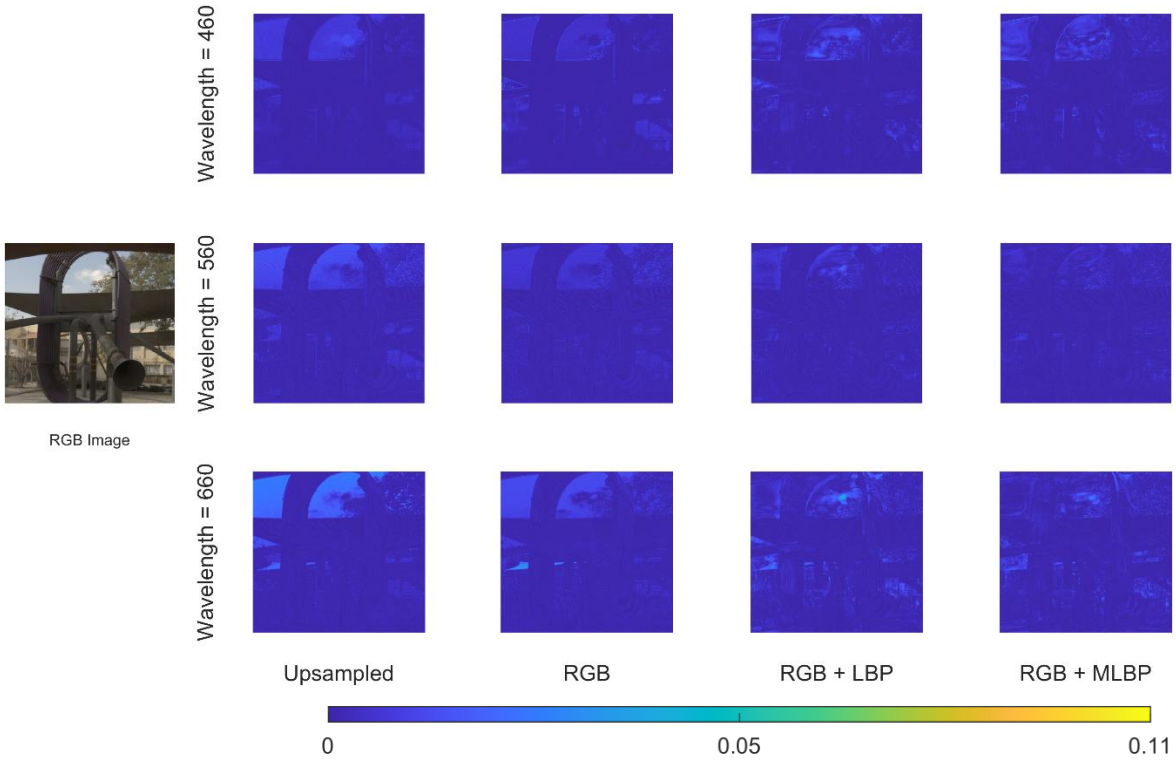


Figure 80. Error in the reconstructed image, from left to right: without estimating the residual; estimating residual from RGB value; estimating residual from RGB value and single scale LBP context and estimating residual from RGB value and multi-scale LBP context.

The findings suggest that estimating the residual from spatial context helps the model to recover spectral scenes more accurately. Nevertheless, in some instances, the trained model doesn't succeed in accurately predicting the PQR coefficients, leading to elevated errors. Several factors could contribute to these errors: the material may not have been represented in the training samples, or the model might lack the complexity needed to accurately predict PQR. Lastly, the local texture feature encapsulated by the LBP histogram may be inadequate in distinguishing between the spectra. Our emphasis on solely local textural characteristics might have overlooked other potentially valuable features within the image.

- **Comparison to Other Works**

In this subsection, we compare the proposed method against Arad's sparse coding method and several deep model-based methods which are the U-net, MIRNET and HINET. Arad's work is trained on the code published by the author with the same RGB image in our experiment. The deep models are also retrained on this dataset. Note that since the RGB image in this study is different from that

in NTIRE 2022, we can't directly compare the results presented in the NTIRE competitions. Table 21 compares the mean and 95% reconstruction errors among the discussed methods.

Table 21. Comparison of the proposed method and the existing methods.

	MLBP	Arad	U-net	MIRNET	HINET
RMSE	0.022	0.064	0.037	0.015	0.020
SAM	0.103	0.272	0.096	0.063	0.084
95% RMSE	0.041	0.129	0.077	0.029	0.038
95% SAM	0.213	0.485	0.218	0.135	0.176

In contrast to the sparse coding approach, our method provides superior reconstruction accuracy. Nevertheless, when compared with deep learning models, particularly those with complex architectures, our method doesn't fare as well. Four main reasons for this discrepancy are:

- The proposed model was trained solely on specific sample collections, while the deep learning models had the advantage of being trained on the entire database. This could mean our model had fewer opportunities to learn the intricate mappings from RGB to spectrum. There may also be additional instances of metamerism not seen during training. The reason for using selected data rather than the entire dataset is to account for the possibility that the number of parameters within our model may not be sufficient to learn all the mappings from the entire database. The selected samples were designed to maximize the ability to represent the entirety of the database.
- The inherent simplicity of our model, combined with its significantly fewer parameters, might limit its ability to model the mapping as comprehensively as the more complex deep models.
- In particular, PQR are only the principal components of the residual and do not account for all of the errors. Also, only a few principal components of the texture histograms could not account for all the detailed textures.
- Our model only leverages local texture as its contextual basis. In contrast, deeper models might potentially harness any available information from the image patch. Our method might be missing out on pertinent information that may be represented by other local features.

That said, our objective was to develop a model that is more interpretable, rather than a "black box." This focus on interpretability comes at the cost of some accuracy. While the results are aligned with our expectations, it underscores the need for further refinement to enhance the method's reconstruction accuracy in future works.

- **Explainability of the Proposed Method**

Compared to end-to-end deep models, our proposed method offers enhanced explainability. Rooted in the RGBPQR colour space, our approach transforms the spectral reconstruction task into a non-linear regression challenge, focusing on estimating the weight of a three-dimensional residual component. Through an analysis of prevalent deep models, we have demonstrated that local textural information can aid in resolving metamerism. However, due to the black-box nature of deep models, understanding the exact spatial features they utilize and their application methods remains challenging.

Our design leverages local textural information extracted from LBP, providing a clear outline of the type of contextual information used. Even though we do employ a shallow neural network for the regression, it plays a minimal role in our overarching method. The neural network utilized in the proposed method aims to address a non-linear regression problem, serving as a small component within the overall model. Its impact on the interpretability of the entire method is minimal. Several alternative solutions to NN exist, as the primary requirement is to address non-linear regression. The choice of NN is based on its effectiveness in resolving non-linear regression problems. Other viable options include regression tree, support vector machines, Gaussian process regression, and kernel approximation regression. The clarity of its purpose and function, in contrast to the black-box operations of deep models, underscores the superior explainability of our approach. Furthermore, our proposed model is inherently simpler than many deep models. With fewer parameters and dimensions, it not only streamlines operations but also enhances the model's transparency and interpretability. This simplicity is a key factor in enhancing its interpretability. The following section demonstrates the interpretability of the proposed model through a case study that analyses the reconstruction error in a single image.

Figure 81 shows the reconstruction error of a selected test image, with absolute errors presented in three channels. Using LBP or MLBP the reconstruction error in the white painting areas of the image is reduced. This indicates that the method successfully used the local context to resolve metamerism associated with the input RGB values.

Figure 82 shows an example of reconstructed spectra and the original spectrum from the 'letter E' in the image. The reconstruction with MLBP as the context has a similar shape to the original. Due to the limited number of components used in the RGBPQR model, very detailed spectral features cannot be recovered in this case (the basis functions in Figure 15 are smooth). Figure 83 displays an example image patch from the 'letter E' in the image, with LBP features at different scales. We searched for the most similar training sample in the training set based on the 9-dimensional code of the shown sample. The closest training sample is shown in Figure 84. The training spectrum is displayed on the left, which is the same as the target sample. Since the local textures are also similar, the model is based on the selected training sample to reconstruct the target spectrum.

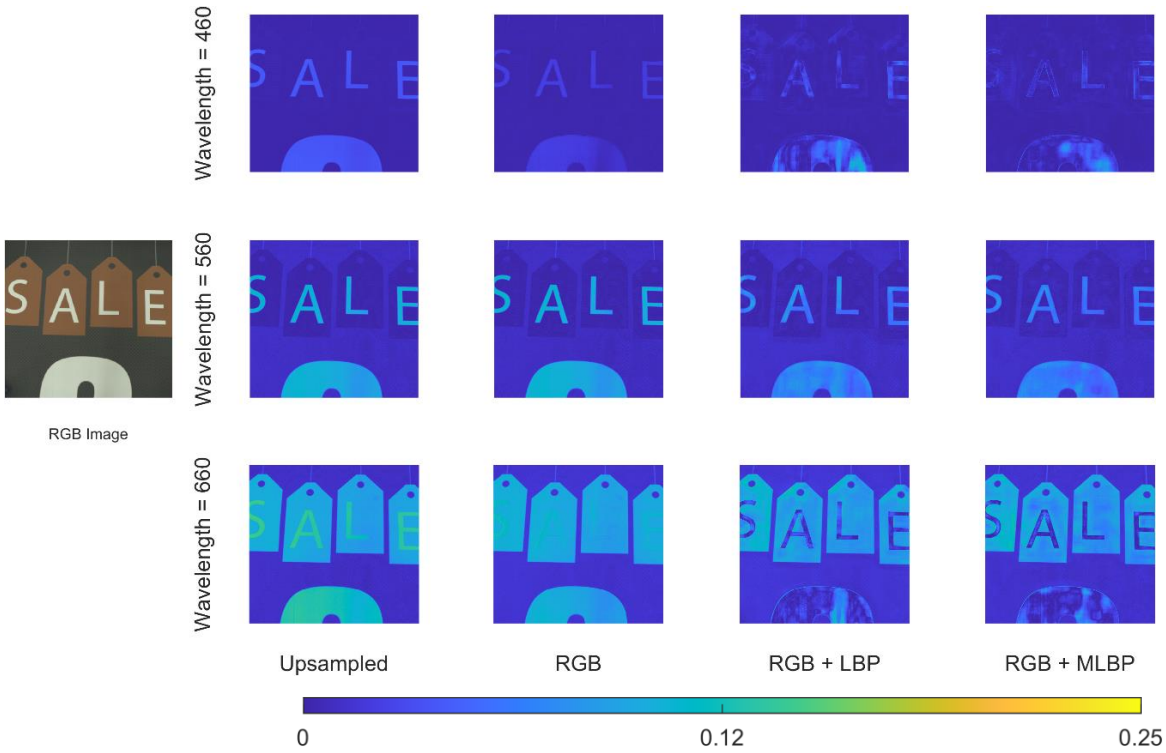


Figure 81. Error in the reconstructed image, from left to right: without estimating the residual; estimating residual from RGB value; estimating residual from RGB value and single scale LBP context and estimating residual from RGB value and multi-scale LBP context.

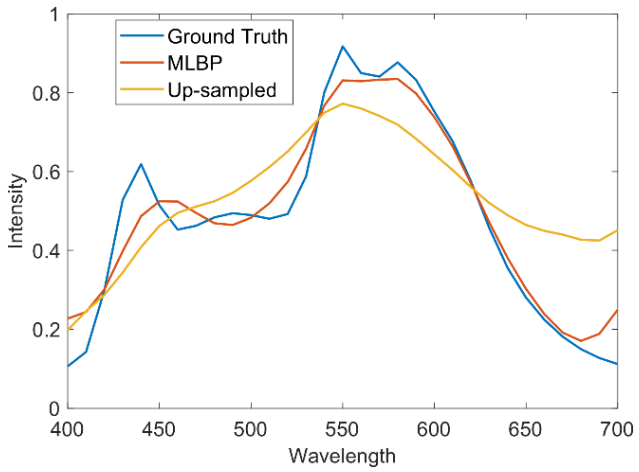


Figure 82. Reconstructed spectra and the original spectrum from the ‘letter E’ in the image.

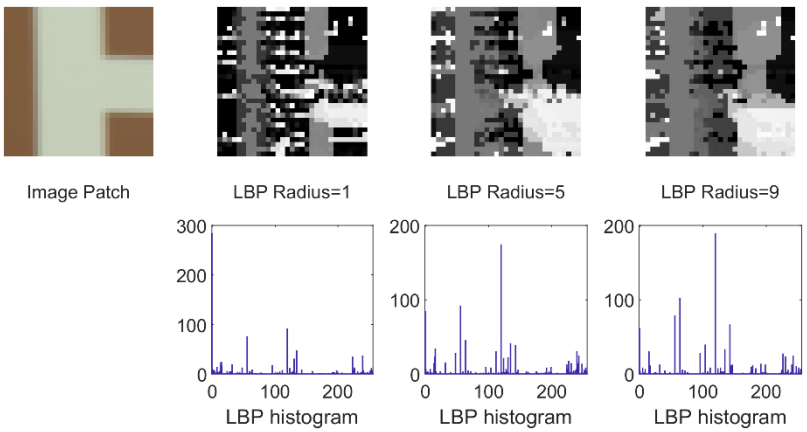


Figure 83. Examples of the target image patch and the extracted features on different scales.

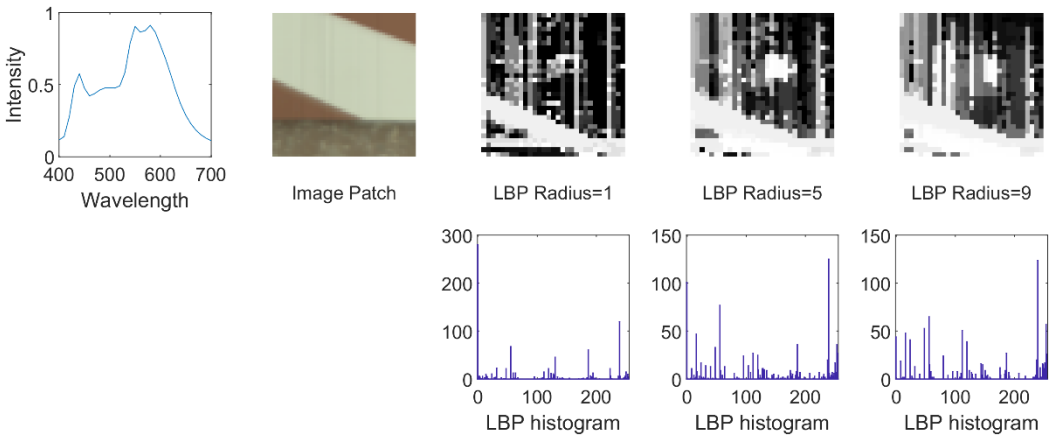


Figure 84. The closest sample to the ‘Letter E’ from the training dataset and the corresponding LBP features.

Figure 85 displays an example from the 'red pattern' in the image, where the proposed model failed to reconstruct the spectrum accurately. We selected a target pixel and displayed it in the image patch, along with the corresponding LBP features in Figure 86. Similarly, we also identified the closest training sample to the 'red pattern' from the training dataset, as shown in Figure 87. However, the closest training sample has a different shape spectrum than the target, which explains the incorrect shape of the reconstructed spectrum. We also searched for metamer samples of the 'red pattern' in the training set, but there were no samples with the same shape as the target sample. Therefore, we cannot expect the model to accurately recover the 'red pattern' since it was not learned during training.

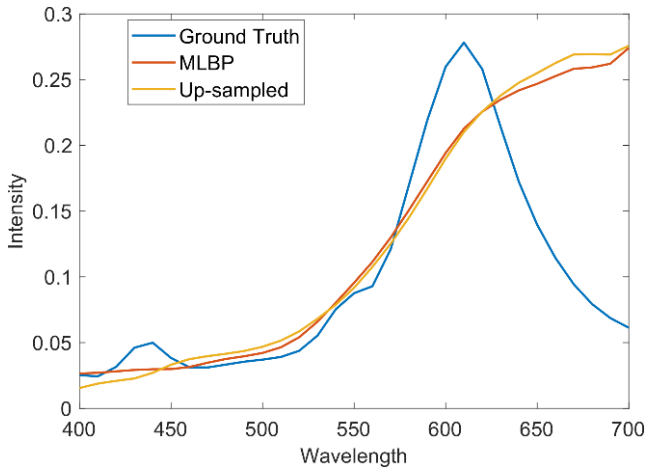


Figure 85. Reconstructed spectra and the original spectrum from the ‘red pattern’ in the image.

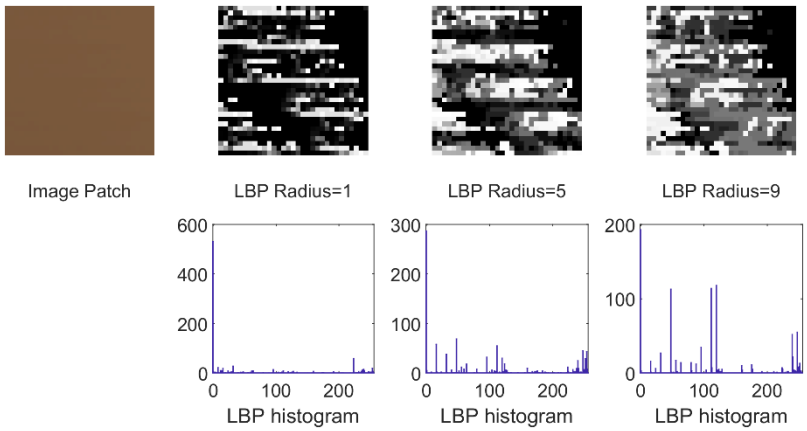


Figure 86. Examples of the target image patch and the extracted features on different scales.

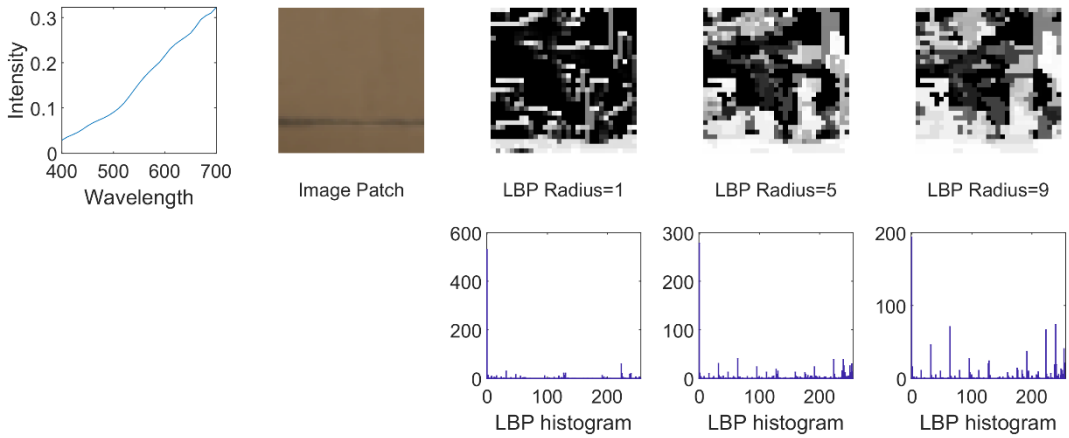


Figure 87. The closest sample to the ‘red pattern’ from the training dataset and the corresponding LBP features.

6.3. Conclusion and Subsequent Studies

In this chapter, we explored the feasibility of utilizing the RGBPQR colour space as a spectral model to achieve single spectral super-resolution. The final reconstructed spectrum results from the addition of an up-sampled spectrum derived from RGB values and an estimated residual (often termed as 'metamer black'). The linear up-sampling function is informed by prior data, while the residual is estimated by determining the weights of the learned residual components. For this process, we leveraged a shallow neural network to predict the weights of the residual component. This network uses RGB combined with a spatial context vector that is derived from PCA scores of the LBP histogram of a 31x31 image patch centred on the target pixel as its input. The network output is the learned residual (PQR) component.

In conclusion, our proposed method significantly enhances reconstruction accuracy compared to a strategy that exclusively uses RGB values to approximate the residual. The examination of the SAM distribution of samples demonstrates that our method reduces the proportion of samples with a high SAM (exceeding 0.2) caused by metamerism. As a result, our approach effectively mitigates the adverse effects of metamerism on reconstruction precision.

When compared to sparse code methods that cannot resolve metamerism, our proposed method offers higher reconstruction accuracy. The RGBPQR model has demonstrated its advantage as a spectral model in spectral super-resolution tasks. Additionally, local texture has been shown to be effective in resolving metamerism as contextual information.

However, it's important to note that our proposed model does not perform as well as existing deep convolutional neural network-based methods. Nevertheless, its significant advantage lies in its simplicity, rooted in a low-dimensional model and a shallow regression network. This simplicity enhances interpretability, making it a valuable approach in scenarios where explainability is crucial.

While the RGBPQR model isn't lossless, potential solutions for enhancing reconstruction accuracy include employing more components to represent the residual, beyond just the first three (PQR). Another possible way to further increase the accuracy is deploying a convolutional neural network to estimate the weights of the residual components. Such a network might marginally compromise on interpretability but could increase accuracy in estimating residual component weights, thus augmenting spectral reconstruction precision. When contrasted with an end-to-end single image spectral super-resolution network, a model predicated on the RGBPQR colour space retains superior interpretability.

Chapter 7. Conclusion and Future Works

7.1. Summary and Conclusion

In this study, we introduced a novel, interpretable single-image spectral super-resolution method based on a low-dimensional spectral model (RGBPQR).

Chapter 1 introduces the motivation behind low-cost hyperspectral imaging. It acknowledges the benefits and challenges of spectral super-resolution in the context of low-cost hyperspectral imaging. The research aims to conduct a thorough analysis of existing techniques, identifying areas for enhancement, and leveraging these insights to create an innovative spectral super-resolution method.

In Chapter 2, we provide an overview of low-cost hyperspectral imaging, with a specific focus on single-image spectral super-resolution. We identify two main categories of approaches: deep convolutional neural network-based methods and traditional machine learning approaches. Traditional machine learning methods struggle with metamerism, leading to lower reconstruction accuracy. In contrast, deep learning methods, while more accurate, lack interpretability. This has driven our objective to develop an interpretable solution that can effectively tackle metamerism. The proposed method should meet the following three objectives:

1. Colour Consistency

   The proposed model should ensure the recovered spectrum produces the same RGB value under identical conditions (for the given camera model and exposure).

2. Ability to Resolve Metamerism

   This study explores what kind of information can resolve the ambiguity associated with metamerism in spectral super-resolution. While other researchers have used deep neural networks to tackle the one-to-many mapping caused by metamerism in reconstructing the

spectrum, no discussion has been provided on how existing methods perform with metamerism data. We also aim to compare existing deep methods in reconstructing metamerism samples.

3. Explainable

Compared to existing deep methods, the proposed model should be more explainable. This entails a less complex and low-dimensional model. Furthermore, the steps within the proposed model should be clear, necessitating an explicit understanding of the features used. Therefore, this study will investigate how spatial information—referring to pixel position and arrangement in an image—can be integrated into the process of reconstructing spectral information.

Additionally, in Chapter 2, we delved into the error measurements and loss functions utilized in this research domain, and we introduced the 95^{th} percentile error that provides a more holistic representation of overall reconstruction performance. We also advocate for the analysis of the error distribution to gain a clearer understanding of reconstruction accuracy.

Driven by the objectives, in Chapter 3, we derived the RGBPQR colour space from the LabPQR colour space. The RGBPQR model provides a compact and accurate representation of spectral data. It consists of two main components: an up-sampled spectrum based directly on RGB values, generated using an adaptive up-sampling function, and residual components referred to as "metamerism black," obtained through PCA of the residuals. The RGBPQR represents spectral data in a concise, low-dimensional model. The proposed model maintains colour consistency by design by directly using the input RGB components to weight corresponding basis spectra and having the additional reconstruction spectra (the PQR components) orthogonal to the camera functions. These ensure that the weights for these vectors will not affect the RGB colour produced from the reconstructed spectra. The model therefore transfers the problem of resolving metamerism to a regression problem through estimating the orthogonal PQR components. This study underscores the benefits of the RGBPQR model in single-image spectral super-resolution tasks, including its simplicity, low dimensionality, and ease of interpretation. This model can serve as a fundamental tool for spectral super-resolution tasks.

We also made efforts to quantitatively analyse the capacity to resolve metamerism. Chapter 4 explores the need for additional information beyond RGB values to effectively address metamerism and enhance spectral reconstruction accuracy. Many existing deep convolutional neural networks leverage spatial context to overcome metamerism challenges. Building upon the RGBPQR model, this chapter introduces a bin-based conditional entropy estimation method, which allows for a quantitative assessment of the additional information provided by various contextual features. Results indicate that local spatial information can significantly diminish ambiguities in PQR weight estimation for spectral samples, which demonstrates the effectiveness of utilizing local texture as contextual information to address the issue of metamerism.

To resolve the metamerism in a more explainable manner, chapter 5 analysed existing deep models with two aspects: What spatial regions are crucial for metamerism resolution, and which features within these regions are important? To address these, we did two experiments, focusing on sensitivity and gradient analyses. Our findings suggest that networks predominantly rely on local spatial information to resolve metamerism. A circular region with a radius of 20 pixels, or even less, could provide adequate contextual information for networks to ascertain the shape of the recovered spectrum in most cases. Tested neural networks are sensitive to local texture features, and removing these features impaired the ability of the networks to accurately reconstruct the spectrum. This implies that such features are a source of the context necessary for neural networks to resolve metamerism. However, neural networks do not show a preference for any particular local texture in a particular scale or orientation. Different samples were sensitive to different scales and different orientations. Therefore, a general texture descriptor would be beneficial when designing our feature extraction method.

Building on the insights from Chapters 4 and 5, Chapter 6 introduces an innovative, explainable single image spectral super-resolution methodology. This approach combines RGB with local spatial features extracted from local binary patterns to resolve metamerism. Consequently, our proposed strategy reduces reconstruction errors stemming from metamerism. When compared with the sparse coding method our technique boasts superior accuracy. Moreover, in comparison with deep networks, our method has a lower dimensionality and a more streamlined structure. Given the transparency of each step in our process, our method is considerably more interpretable than deep network counterparts. This was demonstrated using the data to explain the spectra produced in cases where the spectrum was accurately recovered, and more importantly, in cases where the recovered spectrum was incorrect.

The proposed method has successfully achieved the research objectives. RGBPQR ensures that the reconstructed spectrum maintains colour consistency, metamerism is effectively resolved by utilizing local texture information extracted from LBP as context, and the method retains its interpretability through its simple structure and low dimensionality. Through our research, we acquired an understanding of the issues within spectral super-resolution from RGB images and provided a straightforward, low-dimensional, and interpretable solution.

7.2. Future Works

Current spectral super-resolution research is primarily driven by the NTIRE spectral reconstruction competition, which is held every two years. Researchers are provided with spectral and RGB data by the organizers to train models that recover spectral details from RGB images. However, the data published by NTIRE consists of general outdoor and indoor scenes, while the competition is judged

by MRAE, which focuses on general reconstruction accuracy. Consequently, researchers in this field tend to concentrate solely on recovering spectral images without utilizing the recovered spectral data in any practical applications. As a result, there is a lack of in-depth discussion regarding the usefulness of spectral super-resolution. It remains unclear whether the recovered spectral data could improve classification or regression results compared to using RGB imaging alone. We suggest that future researchers explore the potential of using the recovered spectral data in real-world applications. They should identify applications based on hyperspectral imaging and assess the capability of spectral super-resolution models to handle such tasks effectively. This approach would provide valuable insights into the practical utility of spectral super-resolution techniques.

The current focus of RGB image-based spectral super-resolution is on reconstructing spectral details within the visible range. This limitation arises because RGB camera sensors can only provide spectral measurements within this range. However, hyperspectral images encompass wavelengths beyond the visible range, including the near infrared. Despite the RGB sensor's inability to directly measure near infrared wavelengths, the texture features captured by RGB images may still exhibit correlations with near infrared wavelengths. Consequently, it may be feasible to utilize RGB images to estimate spectral intensities in the near infrared for one or a few wavelengths. This hypothesis also worth testing in future research. Also worth exploring is the incorporation of an infrared band, resulting in RGBI images, which could facilitate the reconstruction of a broader spectrum of wavelengths.

In our proposed method, the residual difference between the up-sampled spectrum derived from RGB values and the original spectrum is captured using three principal components from PCA. While these first three components account for more than 90% of the total variance in the residual, it is not a perfect representation. To enhance the accuracy of reconstruction, it would be worthwhile to consider estimating more than just three residual components.

The model was trained by using only a selection of samples from the available dataset. Consequently, some spectra were not represented in the model. Rather than simply selecting the training samples randomly an additional step could be added to ensure that all of the spectra within the dataset are represented in the training samples.

To explore the full potential of the RGBPQR colour space in terms of reconstruction accuracy rather than interpretability, a convolutional neural network can be employed to estimate the PQR coefficients. In this approach, image patches would serve as inputs, with the network outputting the PQR weights for each pixel within the input patch. This would enable the local texture features used to be optimised by training, at the expense of interpretability. Such a hybrid approach should require significantly fewer parameters than conventional deep learning approaches.

Furthermore, this research exclusively focuses on utilizing local textures extracted by LBP. However, there are other potential sources of local spatial information worth exploring. One particularly intriguing option is leveraging learned filters from trained neural networks. These networks are often trained for classification tasks, suggesting that their filters could serve as valuable additional information for resolving metamerism. Moreover, employing these filters could provide insights into the deep model by visualizing the filtered features.

References

Adão, T., Hruška, J., Pádua, L., Bessa, J., Peres, E., & Morais, R. (2017). Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry. *Remote Sensing, 9*(11), 1110. doi:10.3390/rs9111110

Aeschbacher, J., Wu, J., & Timofte, R. (2017). In defense of shallow learned spectral reconstruction from RGB images. *IEEE International Conference on Computer Vision*, 471-479. doi:10.1109/ICCVW.2017.63

Agahian, F., Amirshahi, S. A., & Amirshahi, S. H. (2008). Reconstruction of reflectance spectra using weighted principal component analysis. *Color Research & Application, 33*(5), 360-371. doi:10.1002/col.20431

Akhtar, N., & Mian, A. S. (2018). Hyperspectral recovery from RGB images using Gaussian Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(1), 100-113. doi:10.1109/TPAMI.2018.2873729

Akhtar, N., Shafait, F., & Mian, A. (2014). Sparse spatio-spectral representation for hyperspectral image super-resolution. *European Conference on Computer Vision*, 8695, 63-78. doi:10.1007/978-3-319-10584-0_5

Alvarez-Gila, A., Van De Weijer, J., & Garrote, E. (2017). Adversarial networks for spatial context-aware spectral image reconstruction from RGB. *IEEE International Conference on Computer Vision Workshops*, 480-490. doi:10.1109/ICCVW.2017.64

Arad, B., & Ben-Shahar, O. (2016). Sparse recovery of hyperspectral signal from natural RGB images. *European Conference on Computer Vision*, LNCS(9911), 19-34. doi:10.1007/978-3-319-46478-7_2

Arad, B., Ben-Shahar, O., Timofte, R., Lin, Y.-T., & Finlayson, G. (2018). NTIRE 2018 challenge on spectral reconstruction from RGB images. *Conference on Computer Vision and Pattern Recognition Workshops*. doi:10.1109/CVPRW.2018.00138

Arad, B., Timofte, R., Ben-Shahar, O., Lin, Y.-T., & Finlayson, G. D. (2020). NTIRE 2020 challenge on spectral reconstruction from an RGB image. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 1806-1822. doi:10.1109/CVPRW50498.2020.00231

Arad, B., Timofte, R., Yahel, R., Morag, N., Bernat, A., & Cai, Y. (2022). NTIRE 2022 spectral recovery challenge and data set. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 863-881. doi:10.1109/CVPRW56347.2022.00102

Ariel, G., & Louzoun, Y. (2020). Estimating differential entropy using recursive copula splitting. *Entropy, 22*(2), 236. doi:10.3390/e22020236

Ayala, F., Echávarri, J. F., Renet, P., & Negueruela, A. I. (2006). Use of three tristimulus values from surface reflectance spectra to calculate the principal components for reconstructing these spectra by using only three eigenvectors. *Journal of the Optical Society of America A, 23*(8), 2020-2026. doi:10.1364/JOSAA.23.002020

Ayhan, B., & Kwan, C. (2017). Application of deep belief network to land cover classification using hyperspectral images. *International Symposium on Neural Networks*, 269-276. doi:10.1007/978-3-319-59072-1_32

Banerjee, A., & Palrecha, A. (2020). MXR-U-Nets for real time hyperspectral reconstruction. *arXiv preprint*. doi:10.48550/arXiv.2004.07003

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., & Siskind, J. M. (2018). Automatic differentiation in machine learning: a survey. *Journal of Marchine Learning Research, 18*, 1-43. doi:10.48550/arXiv.1502.05767

Burns, S. A. (2020). Numerical methods for smoothest reflectance reconstruction. *Color Research & Application, 45*(1), 8-21. doi:10.1002/col.22437

Cai, Y., Lin, J., Hu, X., Wang, H., Yuan, X., & Zhang, Y. (2022). Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17481-17490. doi:10.1109/CVPR52688.2022.01698

Can, Y. B., & Timofte, R. (2018). An efficient CNN for spectral reconstruction from RGB images. *arXiv preprint*. doi:10.48550/arXiv.1804.04647

Cao, X., Du, H., Tong, X., Dai, Q., & Lin, S. (2011). A prism-mask system for multispectral video acquisition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(12), 2423-2435. doi:10.1109/TPAMI.2011.80

CBRNE TECH INDEX. Hyperspectral Imaging (HSI). Retrieved from https://www.cbrnetechindex.com/Chemical-Detection/Technology-CD/Molecular-Spectroscopy-CD-T/Hyperspectral-Imaging-CD-MS

Chakrabarti, A., & Zickler, T. (2011). Statistics of real-world hyperspectral images. *Computer Vision and Pattern Recognition (CVPR)*, 193-200. doi:10.1109/CVPR.2011.5995660

Chang, Y., Bailey, D., & Le Moan, S. (2021). A new coefficient estimation method when using PCA for spectral super-resolution. *36th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 1-6. doi:10.1109/IVCNZ54163.2021.9653296

Chen, L., Lu, X., Zhang, J., Chu, X., & Chen, C. (2021). HINET: Half instance normalization network for image restoration. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 182-192. doi:10.1109/CVPRW53098.2021.00027

Cheriyadat, A., & Bruce, L. M. (2003). Why principal component analysis is not an appropriate feature extraction method for hperspectral data. *IEEE International Geoscience and Remote Sensing Symposium*, 6, 3420-3422. doi:10.1109/igarss.2003.1294808

Chi, C., Yoo, H., & Ben-Ezra, M. (2010). Multi-spectral imaging by optimized wide band illumination. *International Journal of Computer Vision, 86*(2-3), 140. doi:10.1007/s11263-008-0176-y

Cohen, J. (1964). Dependency of the spectral reflectance curves of the Munsell color chips. *Psychonomic Science, 1*(1-12), 369-370. doi:10.3758/BF03342963

Derhak, M., & Rosen, M. (2006). Spectral colorimetry using LabPQR: an interim connection space. *Journal of Imaging Science and Technology, 50*(1), 53-63. doi:10.2352/J.ImagingSci.Technol.(2006)50:1(53)

Devassy, B. M., & George, S. (2020). Dimensionality reduction and visualisation of hyperspectral ink data using t-SNE. *Forensic Science International, 311*, 110194. doi:10.1016/j.forsciint.2020.110194

Fairman, H. S., & Brill, M. H. (2004). The principal components of reflectances. *Color Research & Application, 29*(2), 104-110. doi:10.1002/col.10230

Farrell, M. D., & Mersereau, R. M. (2005). On the Impact of PCA Dimension Reduction for Hyperspectral Detection of Difficult Targets. *Geoscience and Remote Sensing Letters, 2*(2), 192 - 195. doi:10.1109/LGRS.2005.846011

Foster, D. H., & Amano, K. (2006). Frequency of metamerism in natural scenes. *Journal of the Optical Society of America, 23*(10), 2359-2372. doi:10.1364/JOSAA.23.002359

Fu, Y., Zheng, Y., Zhang, L., & Huang, H. (2018). Spectral reflectance recovery from a single RGB image. *IEEE Transactions on Computational Imaging, 4*(3), 382-394. doi:10.1109/TCI.2018.2855445

Fubara, B. J., Sedky, M., & Dyke, D. (2020). RGB to spectral reconstruction via learned basis functions and weights. *Conference on Computer Vision and Pattern Recognition Workshops*, 480-481. doi:10.1109/CVPRW50498.2020.00248

Galliani, S., Lanaras, C., Marmanis, D., Baltsavias, E., & Schindler, K. (2017). Learned spectral super-resolution. *arXiv preprint*. doi:10.48550/arXiv.1703.09470

Gao, L., Kester, R. T., Hagen, N., & Tkaczyk, T. S. (2010). Snapshot image mapping spectrometer (IMS) with high sampling density for hyperspectral microscopy. *Optics Express, 18*(14), 14330-14344. doi:10.1364/OE.18.014330

Gardner, A. S., & Sharp, M. J. (2010). A review of snow and ice albedo and the development of a new physically based broadband albedo parameterization. *Journal of Geophysical Research: Earth Surface, 115*(1). doi:10.1029/2009JF001444

Geng, Y., Mei, S., Tian, J., Zhang, Y., & Du, Q. (2019). Spatial constrained hyperspectral reconstruction from RGB inputs using dictionary representation. *IEEE International Geoscience and Remote Sensing Symposium*, 3169-3172. doi:10.1109/IGARSS.2019.8898871

Glassner, A. S. (1989). How to derive a spectrum from an RGB triplet. *IEEE Computer Graphics and Applications, 9*(4), 95-99. doi:10.1109/38.31468

Hagen, N. A., & Kudenov, M. W. (2013). Review of snapshot spectral imaging technologies. *Optical Engineering, 52*(9), 090901. doi:10.1117/1.OE.52.9.090901

Hajipour, A., & Shams‐Nateri, A. (2017). Effect of classification by competitive neural network on reconstruction of reflectance spectra using principal component analysis. *Color Research & Application, 42*(2), 182-188. doi:10.1002/col.22050

Halicek, M., Fabelo, H., Ortega, S., Callico, G. M., & Fei, B. (2019). In-vivo and ex-vivo tissue analysis through hyperspectral imaging techniques: revealing the invisible features of cancer. *Cancers, 11*(6), 756. doi:10.3390/cancers11060756

Han, X.-H., Shi, B., & Zheng, Y. (2018). Residual HSRCNN: Residual hyper-spectral reconstruction CNN from an RGB image. *International Conference on Pattern Recognition (ICPR)*, 2664-2669. doi:10.1109/ICPR.2018.8545634

He, J., Li, J., Yuan, Q., Shen, H., & Zhang, L. (2021). Spectral response function-guided deep optimization-driven network for spectral super-resolution. *IEEE Transactions on Neural Networks and Learning Systems, 33*(9), 4213-4227. doi:10.1109/TNNLS.2021.3056181

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer vision and Pattern Recognition*, 770-778. doi:10.1109/CVPR.2016.90

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., & Li, M. (2019). Bag of tricks for image classification with convolutional neural networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 558-567. doi:10.1109/CVPR.2019.00065

Hooker, G. (2004). Discovering additive structure in black box functions. *Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 575-580. doi:10.1145/1014052.1014122

Hu, X., Cai, Y., Lin, J., Wang, H., Yuan, X., & Zhang, Y. (2022). HDNET: High-resolution dual-domain learning for spectral compressive imaging. *IEEE Conference on Computer Vision and Pattern Recognition*, 17542-17551. doi:10.1109/CVPR52688.2022.01702

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 4700-4708. doi:10.1109/CVPR.2017.243

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory, 14*(1), 55-63. doi:10.1109/TIT.1968.1054102

Imai, F., Taplin, L., & Day, E. (2002). Comparison of the accuracy of various transformations from multi-band images to reflectance spectra. *Technical Report, Rochester Institute of Technology*. Retrieved from https://scholarworks.rit.edu/article/923/

Imai, F. H., & Berns, R. S. (1999). Spectral estimation using trichromatic digital cameras. *International Symposium on Multispectral Imaging and Color Reproduction for Digital Archives*, 42, 1-8. Retrieved from https://api.semanticscholar.org/CorpusID:59782553}

Jia, Y., Zheng, Y., Gu, L., Subpa-Asa, A., Lam, A., & Sato, Y. (2017). From RGB to spectrum for natural scenes via manifold-based mapping. *IEEE International Conference on Computer Vision*, 4705-4713. doi:10.1109/ICCV.2017.504

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. *European Conference on Computer Vision*, 694-711. doi:10.1007/978-3-319-46475-6_43

Kamruzzaman, M., & Sun, D.-W. (2016). Introduction to hyperspectral imaging technology. In *Computer Vision Technology for Food Quality Evaluation* (pp. 111-139): Elsevier. doi:10.1016/B978-0-12-802232-0.00005-0

Kaya, B., Can, Y. B., & Timofte, R. (2019). Towards spectral estimation from a single RGB image in the wild. *International Conference on Computer Vision Workshop* 3546-3555. doi:10.1109/ICCVW.2019.00439

Khodr, J., & Younes, R. (2011). Dimensionality reduction on hyperspectral images: A comparative review based on artificial datas. *2011 4th International Congress on Image and Signal Processing*, 4, 1875-1883. doi:10.1109/CISP.2011.6100531

Kim, D. G., Burks, T. F., Qin, J., & Bulanon, D. M. (2009). Classification of grapefruit peel diseases using color texture feature analysis. *International Journal of Agricultural and Biological Engineering, 2*(3), 41-50. doi:10.3965/j.issn.1934-6344.2009.03.041-050

Kim, J., Lee, J. K., & Lee, K. M. (2016). Accurate image super-resolution using very deep convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1646-1654. doi:10.1109/CVPR.2016.182

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting, 32*(3), 669-679. doi:10.1016/j.ijforecast.2015.12.003

Kolassa, S., & Martin, R. (2011). Percentage Errors Can Ruin Your Day (and Rolling the Dice Shows How). *Foresight: The International Journal of Applied Forecasting*(23). Retrieved from https://ideas.repec.org/a/for/ijafaa/y2011i23p21-27

Koundinya, S., Sharma, H., Sharma, M., Upadhyay, A., Manekar, R., & Mukhopadhyay, R. (2018). 2D-3D cnn based architectures for spectral reconstruction from RGB images. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 844-851. doi:10.1109/CVPRW.2018.00129

Kozachenko, L. F., & Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii, 23*(2), 9-16.

Kulappurath, S. K., & Shamey, R. (2021). The effect of luminance on the perception of small color differences. *Color Research & Application, 46*(5), 929-942. doi:10.1002/col.22637

Lafon, S., & Lee, A. B. (2006). Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(9), 1393-1403. doi:10.1109/TPAMI.2006.184

Lee, Y.-K., & Powers, J. M. (2005). Comparison of CIE lab, CIEDE 2000, and DIN 99 color differences between various shades of resin composites. *International Journal of Prosthodontics, 18*(2).

Lei, Z., Zhiqiang, L., Peng, W., Wei, W., Shengcai, L., & Ling, S. (2020). Pixel-aware deep function-mixture network for spectral super-resolution. *The Thirty-Fourth AAAI Conference on Artificial Intelligence*. doi:10.1609/aaai.v34i07.6978

Li, J., Wu, C., Song, R., Li, Y., & Liu, F. (2020). Adaptive weighted attention network with camera spectral sensitivity prior for spectral reconstruction from RGB images. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 462-463. doi:10.1109/CVPRW50498.2020.00239

Li, J., Wu, C., Song, R., Li, Y., & Xie, W. (2020). Residual augmented attentional U-shaped network for spectral reconstruction from RGB images. *Remote Sensing, 13*(1), 115. doi:10.3390/rs13010115

Li, Y., Wang, C., & Zhao, J. (2017). Locally linear embedded sparse coding for spectral reconstruction from RGB images. *IEEE Signal Processing Letters, 25*(3), 363-367. doi:10.1109/LSP.2017.2776167

Lim, B., Son, S., Kim, H., Nah, S., & Mu Lee, K. (2017). Enhanced deep residual networks for single image super-resolution. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 136-144. doi:10.1109/CVPRW.2017.151

Lin, Y.-T., & Finlayson, G. D. (2020). Physically plausible spectral reconstruction. *Sensors, 20*(21), 6399. doi:10.3390/s20216399

Lin, Y.-T., & Finlayson, G. D. (2021). On the Optimization of Regression-Based Spectral Reconstruction. *Sensors, 21*(16), 5586. doi:10.3390/s21165586

Liu, L., Fieguth, P., Guo, Y., Wang, X., & Pietikäinen, M. (2017). Local binary features for texture classification: Taxonomy and experimental study. *Pattern Recognition, 62*, 135-160. doi:10.1016/j.patcog.2016.08.032

Liu, P., & Zhao, H. (2020). Adversarial networks for scale feature-attention spectral image reconstruction from a single RGB. *Sensors, 20*(8), 2426. doi:10.3390/s20082426

Liu, Y., Zhang, J., & Zhang, Y. (2022). Hyperspectral reconstruction from a single textile RGB image based on the generative adversarial network. *Textile Research Journal, 93*(1-2). doi:10.1177/00405175221118105

Lodhi, V., Chakravarty, D., & Mitra, P. (2019). Hyperspectral imaging system: Development aspects and recent trends. *Sensing and Imaging, 20*(1), 1-24. doi:10.1007/s11220-019-0257-8

Lu, G., & Fei, B. (2014). Medical hyperspectral imaging: a review. *Journal of Biomedical Optics, 19*(1), 10901. doi:10.1117/1.JBO.19.1.010901

Maloney, L. T. (1986). Evaluation of linear models of surface spectral reflectance with small numbers of parameters. *Journal of the Optical Society of America A, 3*(10), 1673-1683. doi:10.1364/JOSAA.3.001673

Manjunath, B. S., & Ma, W.-Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(8), 837-842. doi:10.1109/34.531803

Marimont, D. H., & Wandell, B. A. (1992). Linear models of surface and illuminant spectra. *Journal of the Optical Society of America A, 9*(11), 1905-1913. doi:10.1364/JOSAA.9.001905

Meng, J., Simon, F., Hanika, J., & Dachsbacher, C. (2015). Physically meaningful rendering using tristimulus colours. *Computer Graphics Forum*, 34(4), 31-40. doi:10.1111/cgf.12676

Muñoz-Huerta, R., Guevara-Gonzalez, R., Contreras-Medina, L., Torres-Pacheco, I., Prado-Olivarez, J., & Ocampo-Velazquez, R. (2013). A review of methods for sensing the nitrogen status in plants: advantages, disadvantages and recent advances. *Sensors, 13*(8), 10823-10843. doi:10.3390/s130810823

Murphy, R. J., Monteiro, S. T., & Schneider, S. (2012). Evaluating classification techniques for mapping vertical geology using field-based hyperspectral sensors. *IEEE Transactions on Geoscience and Remote Sensing, 50*(8), 3066-3080. doi:10.1109/TGRS.2011.2178419

Nathan, D. S., Uma, K., Vinothini, D. S., Bama, B. S., & Roomi, S. (2020). Light weight residual dense attention net for spectral reconstruction from RGB images. *Sensors, 20*(8), 2426. doi:10.3390/s20082426

Nguyen, R. M., Prasad, D. K., & Brown, M. S. (2014). Training-based spectral reconstruction from a single RGB image. *European Conference on Computer Vision*, 186-201. doi:10.1007/978-3-319-10584-0_13

Nieves, J. L. (2020). Hyperspectral Imaging. *Encyclopedia of Color Science and Technology*. doi:10.1007/978-3-642-27851-8_425-1

Ojala, T., Pietikainen, M., & Harwood, D. (1994). Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. *12th International Conference on Pattern Recognition*, 1, 582-585. doi:10.1109/ICPR.1994.576366

Otsu, H., Yamamoto, M., & Hachisuka, T. (2018). Reproducing spectral reflectances from tristimulus colours. *Computer Graphics Forum*, 37(6), 370-381. doi:10.1111/cgf.13332

Park, J.-I., Lee, M.-H., Grossberg, M. D., & Nayar, S. K. (2007). Multispectral imaging using multiplexed illumination. *2007 IEEE 11th International Conference on Computer Vision*, 1-8. doi:10.1109/ICCV.2007.4409090

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. *IEEE Conference on Computer Vision and Pattern Recognition*, 2536-2544. doi:10.1109/CVPR.2016.278

Peng, H., Chen, X., & Zhao, J. (2020). Residual pixel attention network for spectral reconstruction from RGB images. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 486-487. doi:10.1109/CVPRW50498.2020.00251

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-assisted Intervention*, 9351, 234-241. doi:10.1007/978-3-319-24574-4_28

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290*(5500), 2323-2326. doi:10.1126/science.290.5500.2323

Shi, Z., Chen, C., Xiong, Z., Liu, D., & Wu, F. (2018). HSCNN+: Advanced cnn-based hyperspectral recovery from RGB images. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 939-947. doi:10.1109/CVPRW.2018.00139

Silva, R., & Melo-Pinto, P. (2021). A review of different dimensionality reduction methods for the prediction of sugar content from hyperspectral images of wine grape berries. *Applied Soft Computing, 113*, 107889. doi:10.1016/j.asoc.2021.107889

Smits, B. (1999). An RGB-to-spectrum conversion for reflectances. *Journal of Graphics Tools, 4*(4), 11-22. doi:10.1080/10867651.1999.10487511

Sowmya, V., Soman, K., & Hassaballah, M. (2019). *Hyperspectral image: Fundamentals and advances in Recent Advances in Computer Vision* (Vol. 804): Springer. doi:10.1007/978-3-030-03000-1_16

Stiebel, T., Koppers, S., Seltsam, P., & Merhof, D. (2018). Reconstructing spectral images from RGB-images using a convolutional neural network. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 948-953. doi:10.1109/CVPRW.2018.00140

Sun, Y., Fracchia, F. D., Calvert, T. W., & Drew, M. S. (1999). Deriving spectra from colors and rendering light interference. *IEEE Computer Graphics and Applications, 19*(4), 61-67. doi:10.1109/38.773965

Timofte, R., De Smet, V., & Van Gool, L. (2015). A+: Adjusted anchored neighborhood regression for fast super-resolution. *Asian Conference on Computer Vision*, 111-126. doi:10.1007/978-3-319-16817-3_8

Tsutsumi, S., Rosen, M., & Berns, R. (2007). Spectral gamut mapping using LabPQR. *Journal of Imaging Science and Technology, 51*(6), 473-485. doi:10.2352/J.ImagingSci.Technol.(2007)51:6(473)

Valero, E. M., Nieves, J. L., Nascimento, S. M., Amano, K., & Foster, D. H. (2007). Recovering spectral data from natural scenes with an RGB digital camera and colored filters. *Color Research & Application, 32*(5), 352-360. doi:10.1002/col.20339

Vapnik, V. N. (1999). *The Nature of Statistical Learning Theory*: Springer New York. doi:10.1007/978-1-4757-3264-1

Vermeulen, M., Smith, K., Eremin, K., Rayner, G., & Walton, M. (2021). Application of uniform manifold approximation and projection (UMAP) in spectral imaging of artworks. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 252*, 119547. doi:10.1016/j.saa.2021.119547

Willett, R. M., Duarte, M. F., Davenport, M. A., & Baraniuk, R. G. (2013). Sparsity and structure in hyperspectral imaging: Sensing, reconstruction, and target detection. *IEEE Signal Processing Magazine, 31*(1), 116-126. doi:10.1109/MSP.2013.2279507

Wu, G., Liu, Z., Yang, S., Zhu, M., & Liu, P. (2014). Weighted LabPQR interim connection space based on human color vision for spectral color reproduction. *Journal of Spectroscopy, 2014*. doi:10.1155/2014/595602

Wug Oh, S., Brown, M. S., Pollefeys, M., & Joo Kim, S. (2016). Do it yourself hyperspectral imaging with everyday digital cameras. *IEEE Conference on Computer Vision and Pattern Recognition*, 2461-2469. doi:10.1109/CVPR.2016.270

Xing, Z., Zhou, M., Castrodad, A., Sapiro, G., & Carin, L. (2012). Dictionary learning for noisy and incomplete hyperspectral images. *SIAM Journal on Imaging Sciences, 5*(1), 33-56. doi:10.1137/110837486

Xiong, Z., Shi, Z., Li, H., Wang, L., Liu, D., & Wu, F. (2017). HSCNN: cnn-based hyperspectral image recovery from spectrally undersampled projections. *IEEE International Conference on Computer Vision Workshops*, 518-525. doi:10.1109/ICCVW.2017.68

Yan, Y., Zhang, L., Li, J., Wei, W., & Zhang, Y. (2018a). Accurate spectral super-resolution from single RGB image using multi-scale CNN. *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 206-217. doi:10.1007/978-3-030-03335-4_18

Yasuma, F., Mitsunaga, T., Iso, D., & Nayar, S. K. (2010). Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. *IEEE Transactions on Image Processing, 19*(9), 2241-2253. doi:10.1109/TIP.2010.2046811

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H., & Shao, L. (2020). Learning enriched features for real image restoration and enhancement. *Computer Vision*, 492-511. doi:10.48550/arXiv.2003.06792

Zhang, J., Su, R., Ren, W., Fu, Q., & Nie, Y. (2021). A survey on computational spectral reconstruction methods from RGB to hyperspectral imaging. *Scientific reports*. doi:10.1038/s41598-022-16223-1

Zhang, L., Lang, Z., Wang, P., Wei, W., Liao, S., Shao, L., & Zhang, Y. (2020). Pixel-aware deep function-mixture network for spectral super-resolution. *AAAI Conference on Artificial Intelligence*, 34(07), 12821-12828. doi:10.1609/aaai.v34i07.6978

Zhao, Y., Po, L.-M., Yan, Q., Liu, W., & Lin, T. (2020). Hierarchical regression network for spectral reconstruction from RGB images. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 422-423. doi:10.1109/CVPRW50498.2020.00219

Appendix 1

Chapter 3.4.

Figure 1 compares the mean relative absolute error (MRAE) of the reconstructed spectrum in the 450 nm, 550 nm, and 650 nm channels.

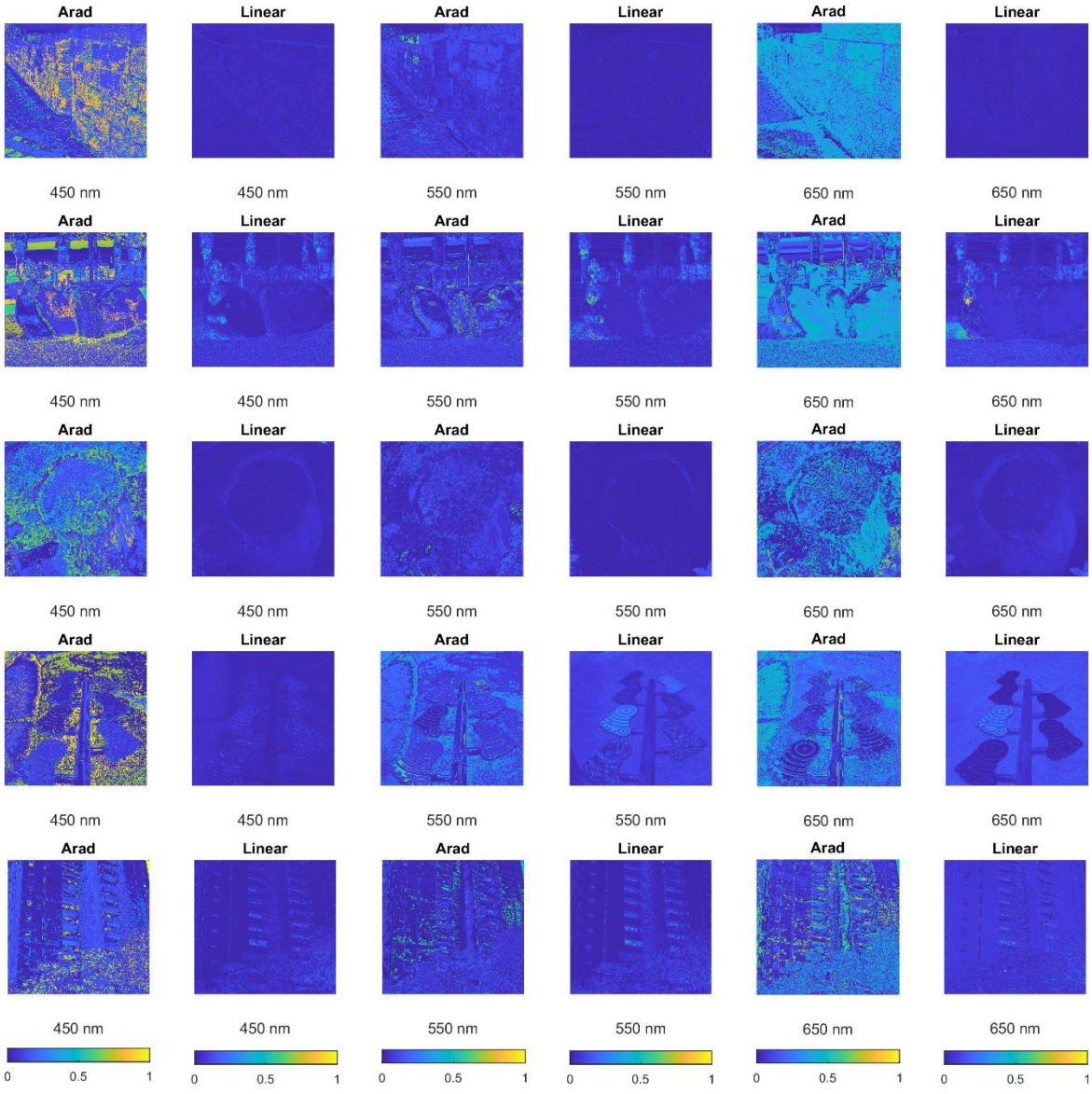


Figure 1. MRAE of the reconstructed spectrum in 450, 550 and 650nm of the NTIRE 2022, in most cases the linearly up-sample spectrum appears with less error.

Appendix 2

Chapter 4.1.3

Table 1 shows the conditional entropy and the additional information when involving simple statistical measurements of the image patch as additional context.

Table 1. Additional information when adding neighbour RGB pixel from 3×3 image patch.

	Variables	Random data 1	Random data 2	Edge data
	$H(N_p)$	7.17	7.19	7.26
RGB	$H(N_p, \mathrm{P})$	13.19	13.19	13.51
	$H(\mathrm{P})$	9.14	9.14	9.18
	$H(N_p \mid \mathrm{P})$	4.05	4.05	4.33
	I_{rgb}	3.12	3.14	2.93
Simple statistical measures				
Mean RGB from the patch	$H(N_p, \mathrm{P}A_{mean})$	14.80	14.78	15.25
	$H(\mathrm{P}A_{mean})$	11.77	11.76	12.09
	$H(N_p \mid \mathrm{P}A_{mean})$	3.03	3.02	3.16
	I_{mean}	1.02	1.02	1.17
Minimum RGB from the patch	$H(N_p \mid \mathrm{P}A_{min})$	15.00	14.99	15.70
	$H(\mathrm{P}A_{min})$	12.19	12.19	12.92
	$H(N_p \mid \mathrm{P}A_{min})$	2.81	2.80	2.78
	I_{min}	1.24	1.25	1.55
Maximum RGB from the patch	$H(N_p, \mathrm{P}A_{max})$	15.46	15.44	16.19
	$H(\mathrm{P}A_{max})$	12.81	12.81	13.70
	$H(N_p \mid \mathrm{P}A_{max})$	2.65	2.63	2.50
	I_{max}	1.40	1.41	1.84
Standard deviation of the patch	$H(N_p, \mathrm{P}A_{std})$	15.44	15.42	16.46
	$H(\mathrm{P}A_{std})$	12.87	12.87	14.15
	$H(N_p \mid \mathrm{P}A_{std})$	2.57	2.55	2.31
	I_{std}	1.48	1.49	2.02
Skewness of the patch	$H(N_p, \mathrm{P}A_{ske})$	19.25	19.24	18.99
	$H(\mathrm{P}A_{ske})$	18.41	18.40	17.82
	$H(N_p \mid \mathrm{P}A_{ske})$	0.84	0.83	1.17
	I_{ske}	3.21	3.21	3.16
Kurtosis deviation of the patch	$H(N_p, \mathrm{P}A_{tur})$	18.93	18.91	18.44
	$H(\mathrm{P}A_{tur})$	17.84	17.83	16.85
	$H(N_p \mid \mathrm{P}A_{tur})$	1.09	1.08	1.60
	I_{tur}	2.96	2.96	2.73

Appendix 3

Chapter 5.2

- **Rectangular area**

In this test, rather than incorporating additional information as random samples into the testing image patch, a rectangular area with a width of 20 and a length of 100 is added. This rectangular area has a mean RGB value of the painting patch, which is [21, 23, 16]. However, the reconstructed result, as depicted in Figure 1, did not recover any spectra within the painting shape. This test indicates that simply increasing the density of samples from the painting image does not influence the network's decision.

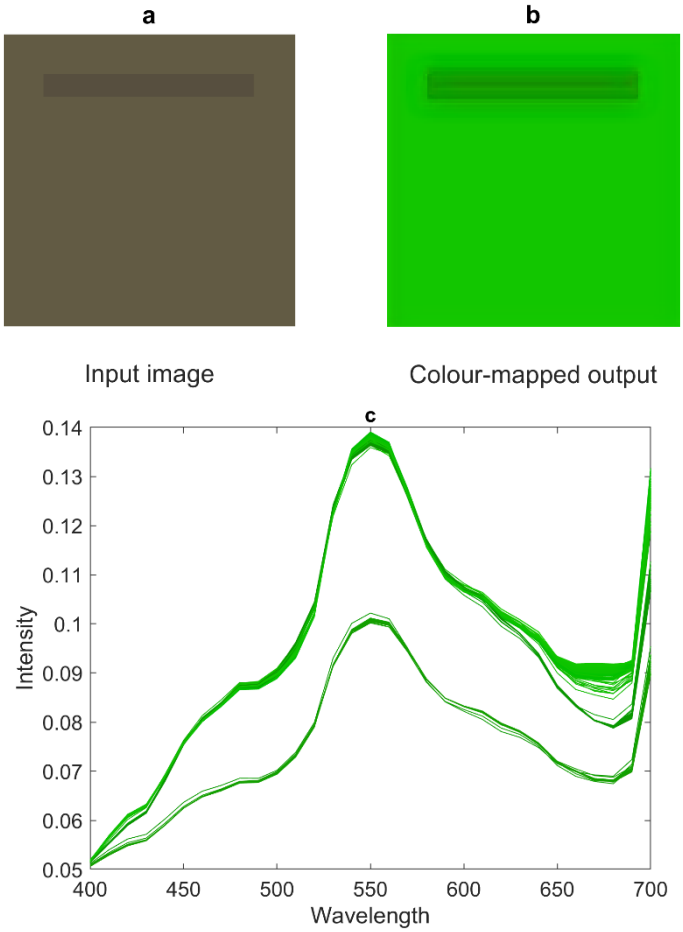


Figure 1. Reconstruction results when additional information in the form of a rectangular with RGB values from the painting image is added.

- **Lines**

In this test, instead of adding additional information as random samples to the testing image patch, lines with a width of 1 are used. These lines have the mean RGB value of the painting patch, which is [21, 23, 16]. However, the reconstructed result, as shown in Figure 2 & 3, did not recover any spectra within the painting shape. Interestingly, the network appears to treat horizontal and vertical lines differently, which may be related to the feature extraction filters learned by the network. However, it's worth noting that HSCNN+ does not rely on straight lines as additional information to reconstruct the spectrum into the painting shape.

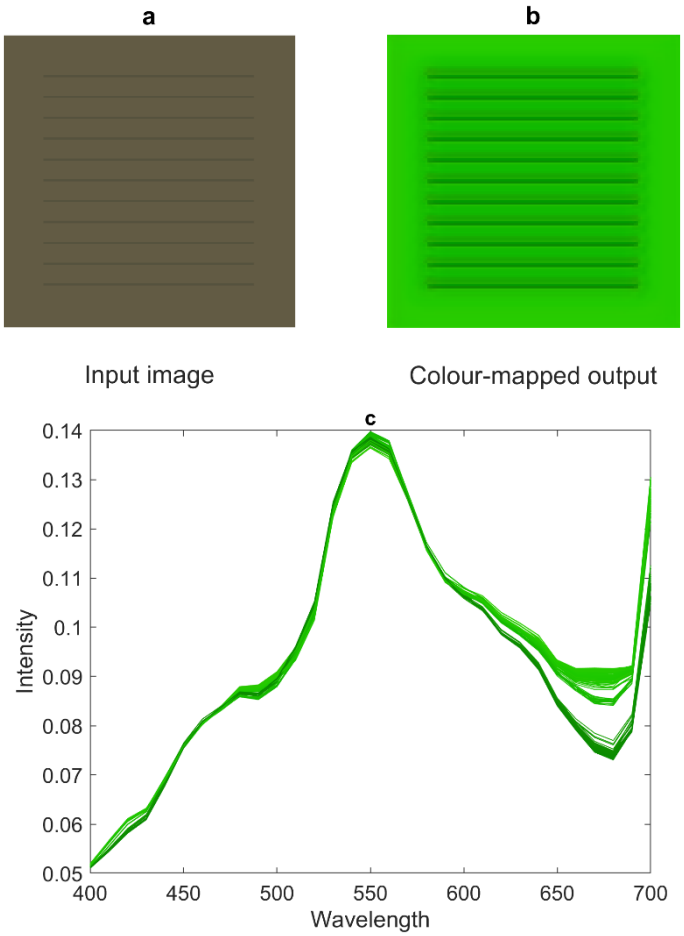


Figure 2. Reconstruction results when additional information in the form of horizontal lines with RGB values from the painting image is introduced.

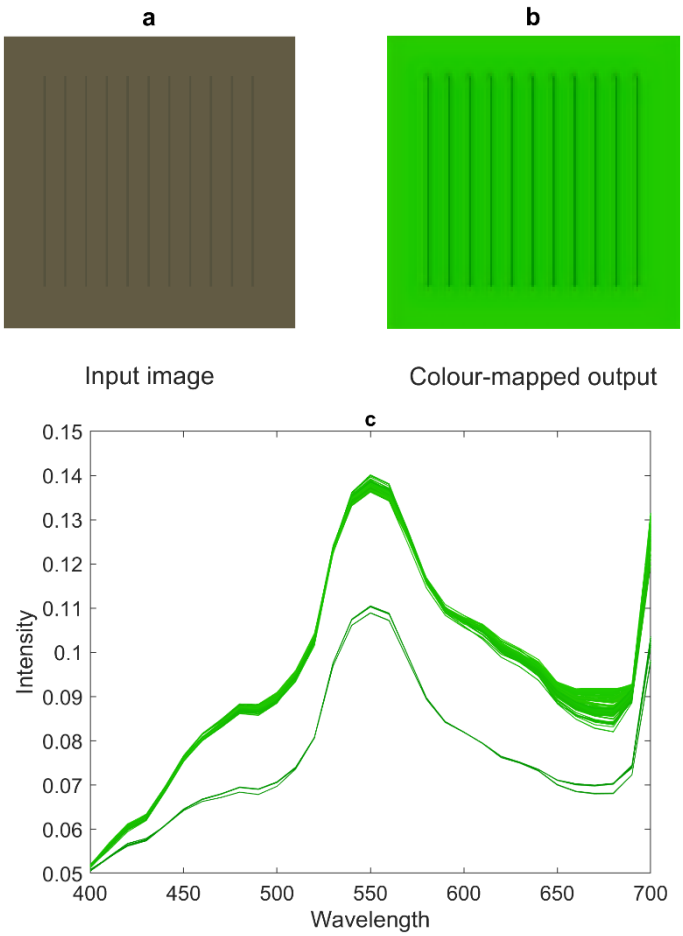


Figure 3. Reconstruction results when additional information in the form of vertical lines with RGB values from the painting image is introduced.

- **Edge with RGB values purely from the grass**

In this specific test, additional information was added to the identified edge of the painting patch. However, the RGB value used was [21, 21, 16], which is exclusively found in grass images and not in the painting image. The reconstructed result is shown in Figure 4, and it reveals that no samples were recovered in the form of the painting. The added RGB value lacks representation within the painting dataset, which means incorporating a sample with this RGB value could increase the likelihood of HSCNN+ reconstructing the spectrum as a grass shape.

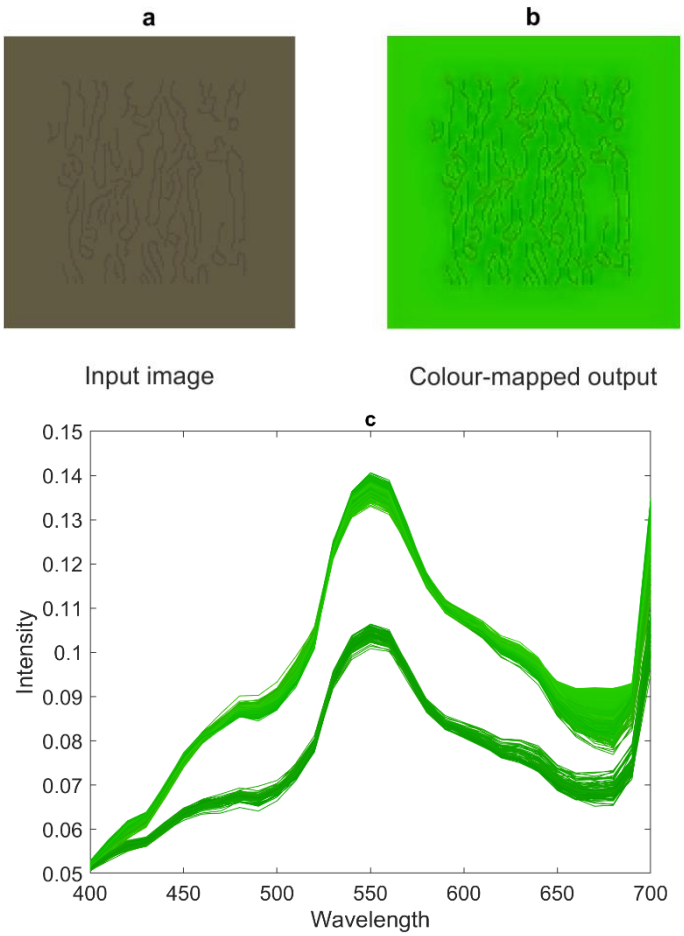


Figure 4. Reconstruction results when additional information in the form of edge with RGB values purely from the plant is introduced.

- **Edges detected in grass image.**

In this test, the additional information is added as samples at the detected edge of the grass instead of the painting. Figure 5 shows the detected edge, since this patch is from the grass dataset, it contains the texture information of the natural plant. The edges are detected with a Canny filter with the default threshold. The background of the patch is in [27 29 19]. The added samples are in [21 23 16] which is the average RGB value of the image patch cropped from the painting patch.

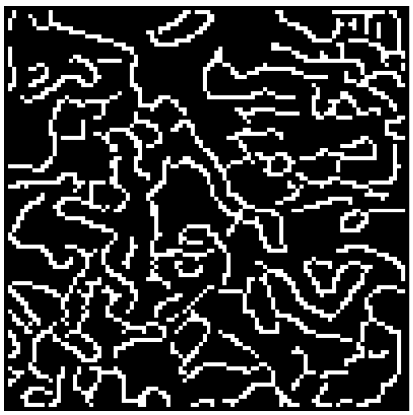


Figure 5. the detected edge of the image patch from the manmade dataset.

The reconstructed result is presented in Figure 6, where some of the recovered spectra appear with the shape of the painting, as indicated by the red spectrum in Figure 6 (b). The corresponding pixels exhibiting the painting shape are illustrated in Figure 6 (c). One possible explanation for this observation is that the texture of the painting and the grass are quite similar, given that they are both present in a plant scene. Consequently, the HSCNN+ is able to extract the similar spatial information from the edge feature.

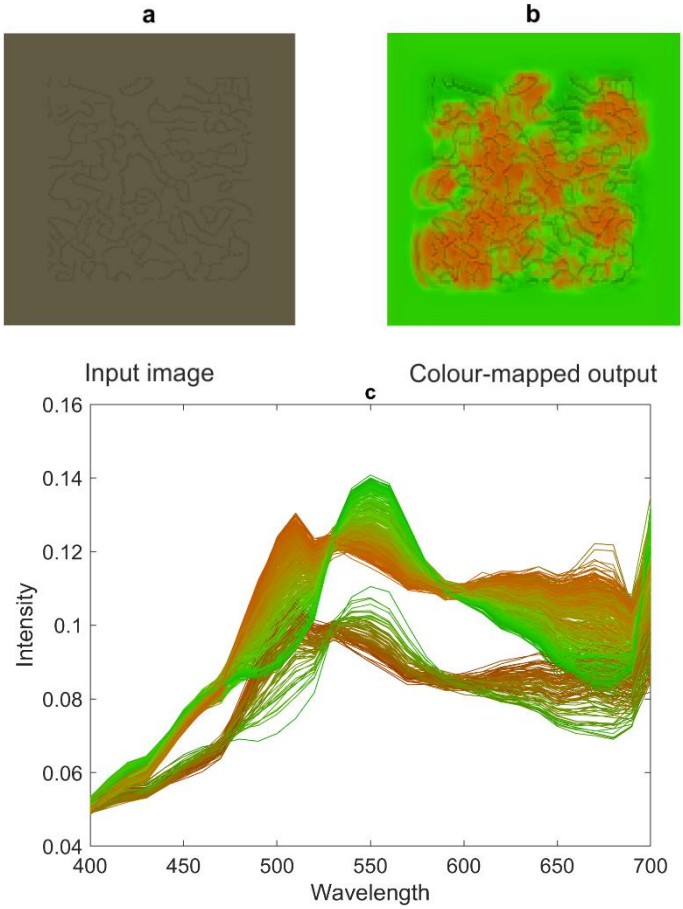


Figure 6. Reconstruction results when additional information in the form of edges detected in plant image.

The results of this experiment clearly demonstrate that in order to resolve metamerism, the HSCNN+ network requires spatial information such as texture as contextual information. However, it is currently unclear whether both colour and spatial information are necessary or if the network relies solely on texture information. Therefore, in the next section, we will investigate how the network responds when the colour information is removed. This will allow us to determine whether the network can still effectively resolve metamerism without the presence of colour information.

- **Edge with RGB values in grey**

In this test, the additional information is added as samples at the detected edge of the 'painting' patch. Unlike test 4, the RGB value of the edge pixels in this test is set to grey [21 21 21]. The objective is to eliminate colour information and evaluate whether the texture (edge) of the painting dataset alone is sufficient for HSCNN+ to identify the input as belonging to the painting category.

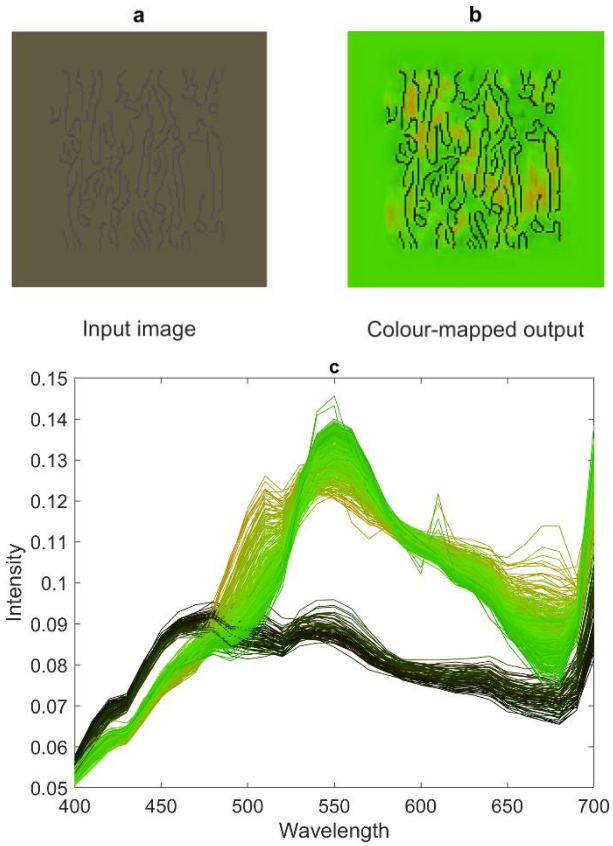


Figure 7. Reconstruction results upon adding additional information in the form of edges with a grey RGB values. Few samples with edge feature have been recovered into the painting shape.

In this case, the painting shape is only reconstructed for a few spectra, as shown in Figure 7. None of the samples on the edge have been recovered into the painting shape, indicating that the network needs specific RGB requirements to determine the shape of its output. When texture information is present in the scene, the network is more likely to identify the target pixels as belonging to the

painting rather than the plant because the RGB of the background can be found in both materials. Therefore, in this case, the HSCNN+ requires both spatial information and RGB values as context to resolve metamerism.

- **'X' shape with RGB values from the 'Painting' image**

In the last test, the HSCNN+ appears to be sensitive to the diagonal edges, therefore, in this test, additional information is added as diagonal lines. In this test, the setting of the background and additional samples are same to the last test.

Figure 10 shows the reconstruction result, there is no sample has been reconstructed into the shape of the 'painting'. It seems the HSCNN+ have not interested in straight lines. The texture requires some curvature to be recognised as a 'painting'.

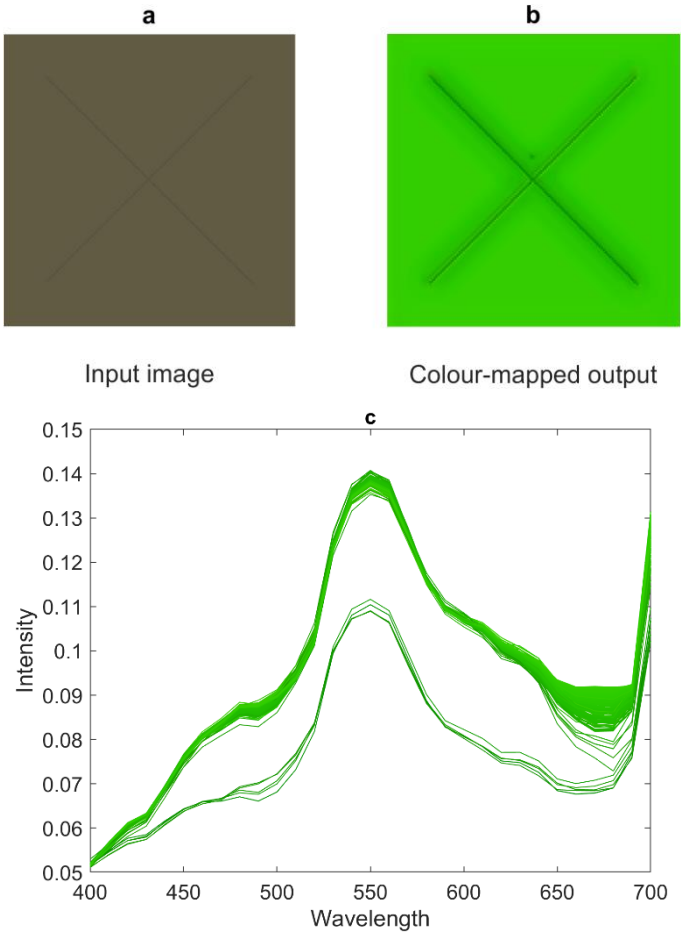


Figure 8. Reconstruction results when additional information in the form of a 'x' shape with RGB values from the painting image is added.

Appendix 4

Chapter 5.3.1. Remove all information expect the neighbour pixels

Figure 1 illustrates the RMSE as a function of the radius for HRNET.

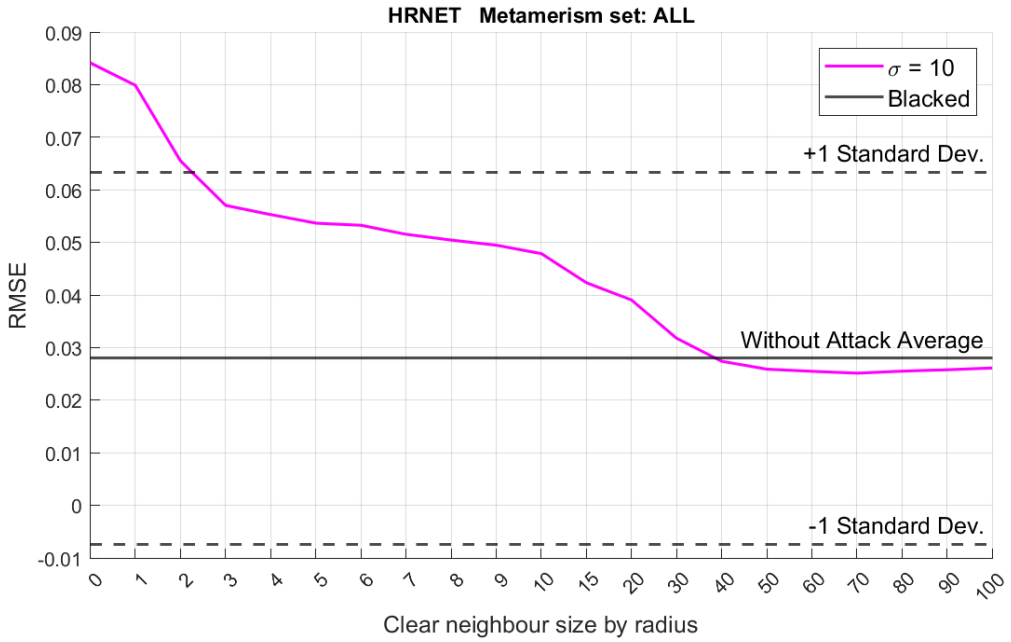


Figure 1. The mean RMSE as a function of the untouched neighbour size when all information has been removed except the neighbour area. HRNET are sensitive to close neighbour pixels.

Figure 2-5 illustrates the SAM as a function of the radius for other networks.

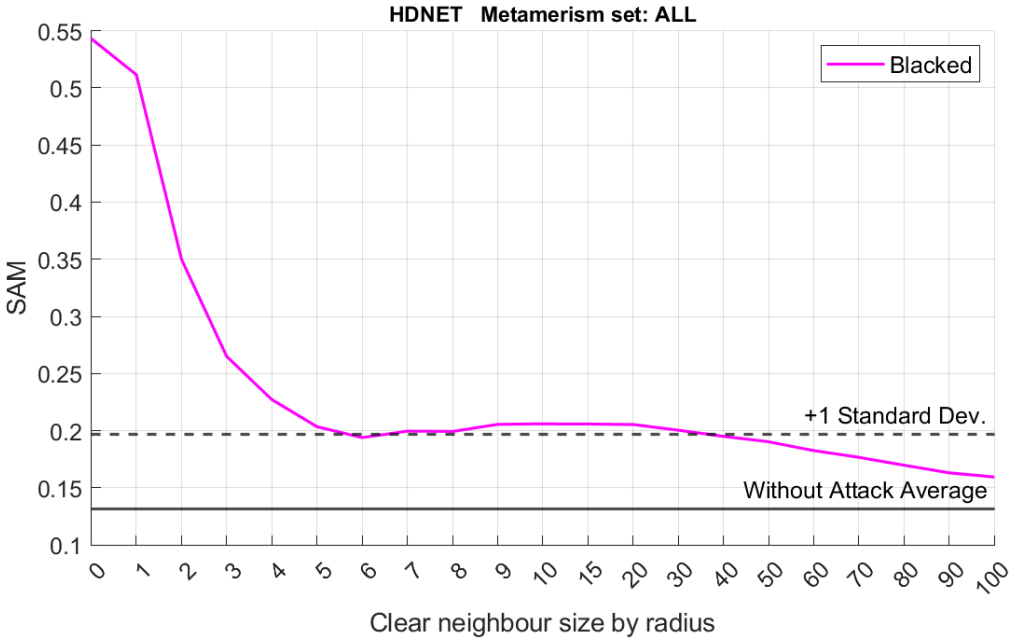


Figure 2. The SAM as a function of the untouched neighbour size when all information has been removed except the neighbour area of HDNET.

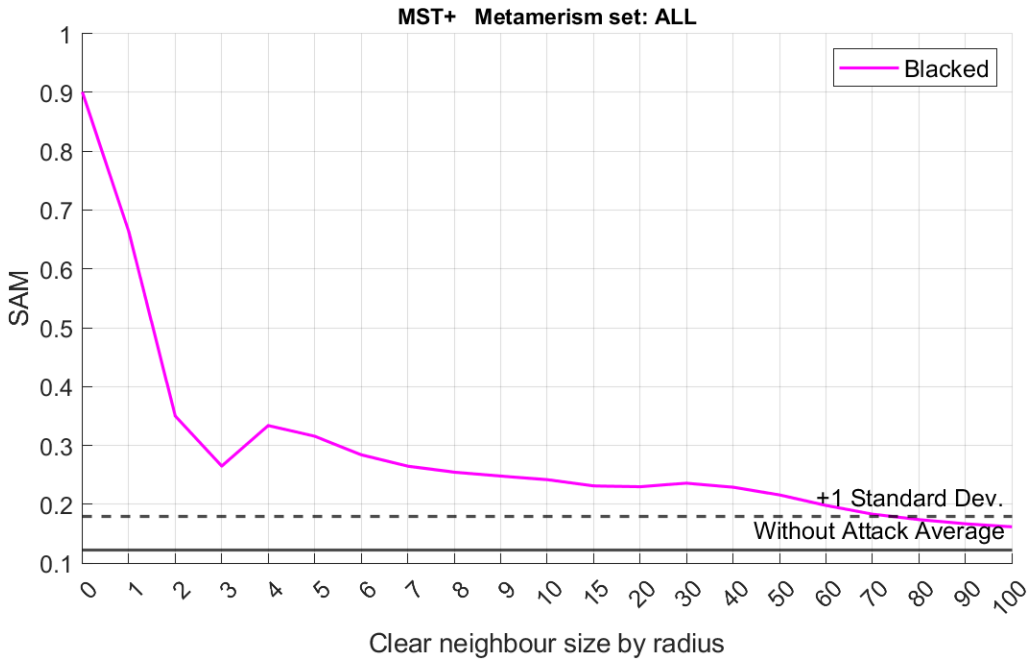


Figure 3. The SAM as a function of the untouched neighbour size when all information has been removed except the neighbour area of MST++.

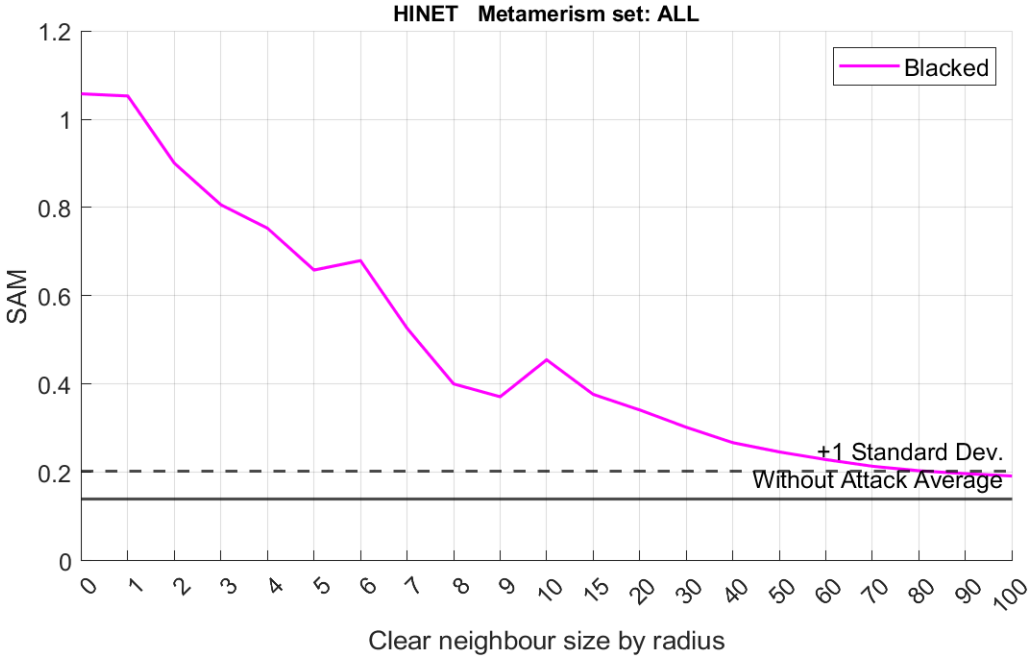


Figure 4. The SAM as a function of the untouched neighbour size when all information has been removed except the neighbour area of HINET.

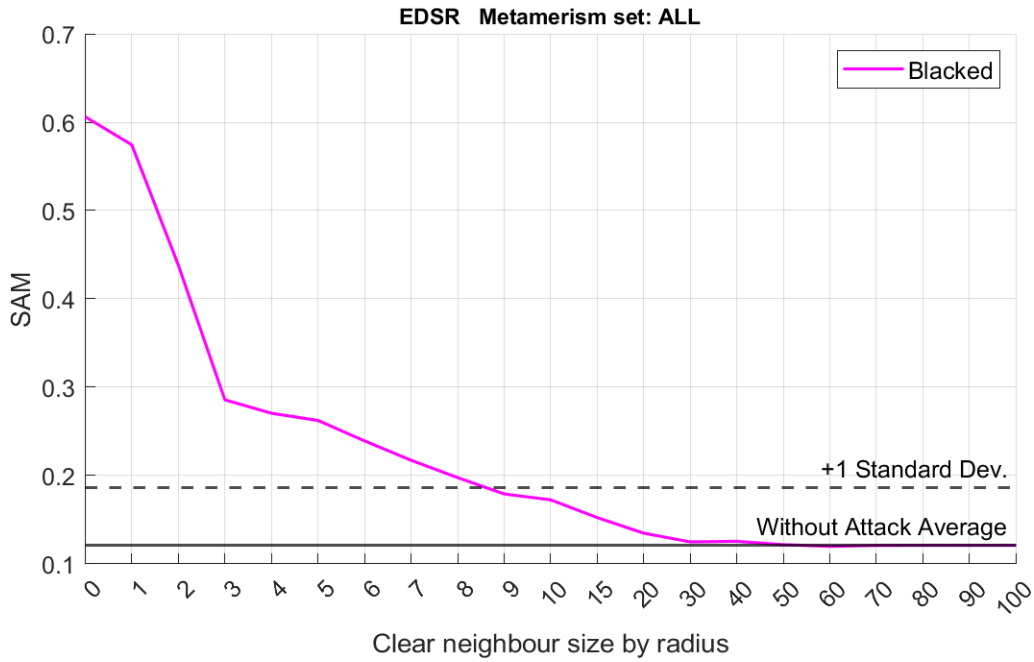


Figure 5. The SAM as a function of the untouched neighbour size when all information has been removed except the neighbour area of EDSR.

Chapter 5.3.2. Blurring the image outside the neighbour area

Figure 6 illustrates the RMSE as a function of the radius for HRNET.

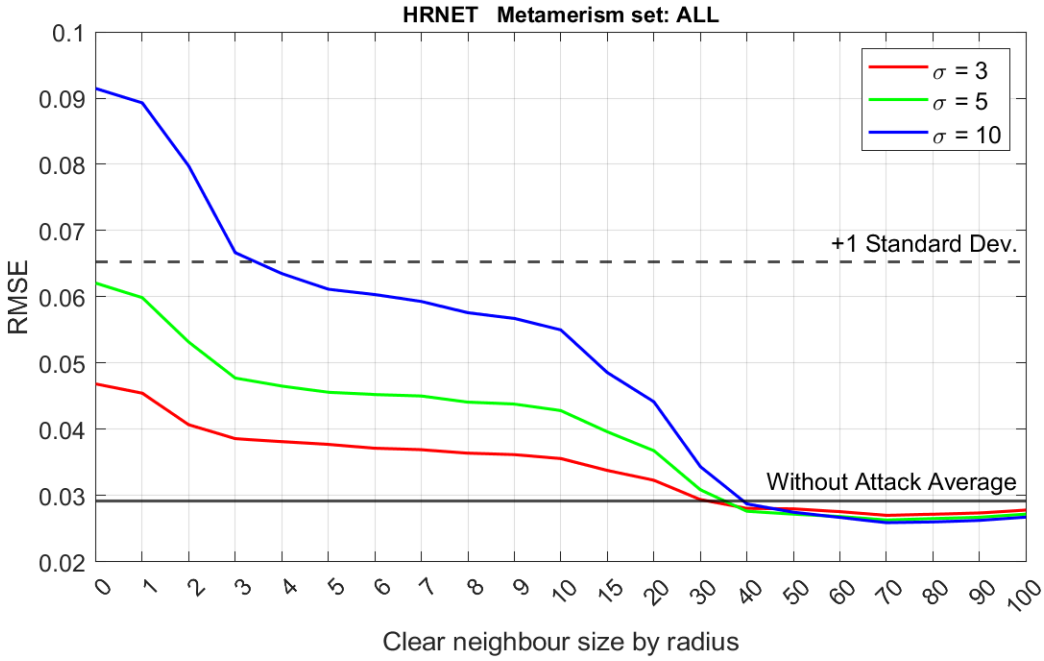


Figure 6. The RMSE as a function of the untouched neighbour size when the rest of image is blurred. When the size of clear neighbour is larger than 30 pixels, HRNET have sufficient information for reconstruction.

Figure 7-10 illustrates the SAM as a function of the radius for other networks.

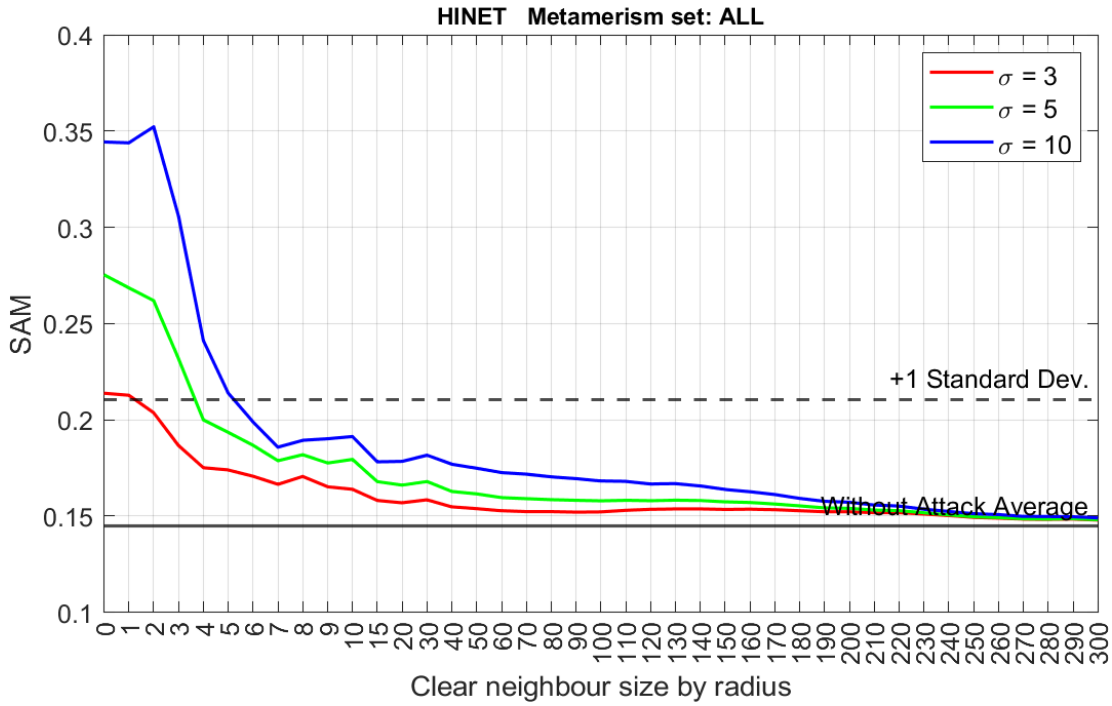


Figure 7. Responses of networks that extract global information when the image is blurred. The reconstruction gets close to the best performance when almost all image is untouched.

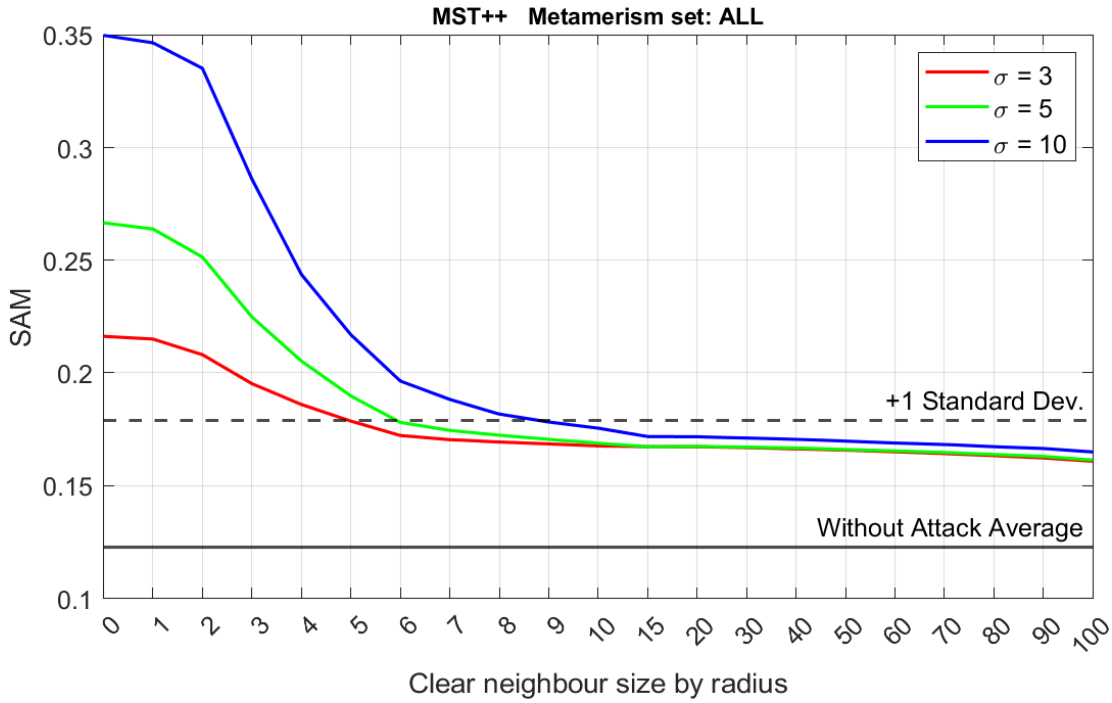


Figure 8. MST++ relies on global spatial information to achieve its best performance, but it also displays higher sensitivity to local spatial information.

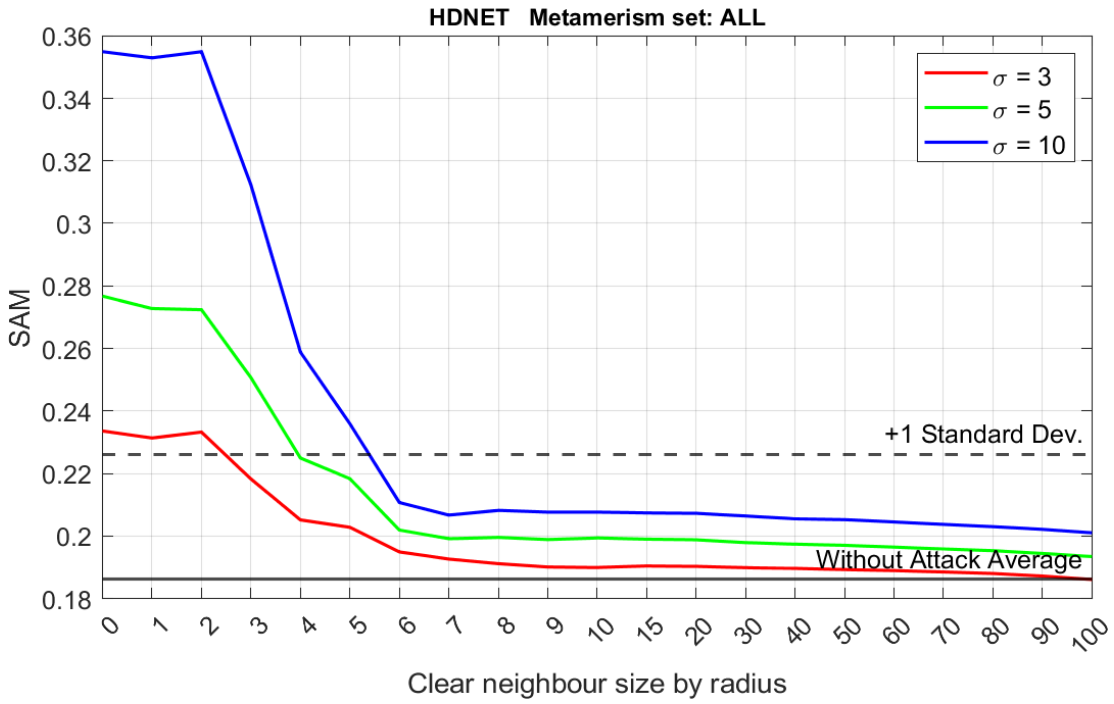


Figure 9. HDNET relies on global spatial information to achieve its best performance, but it also displays higher sensitivity to local spatial information.

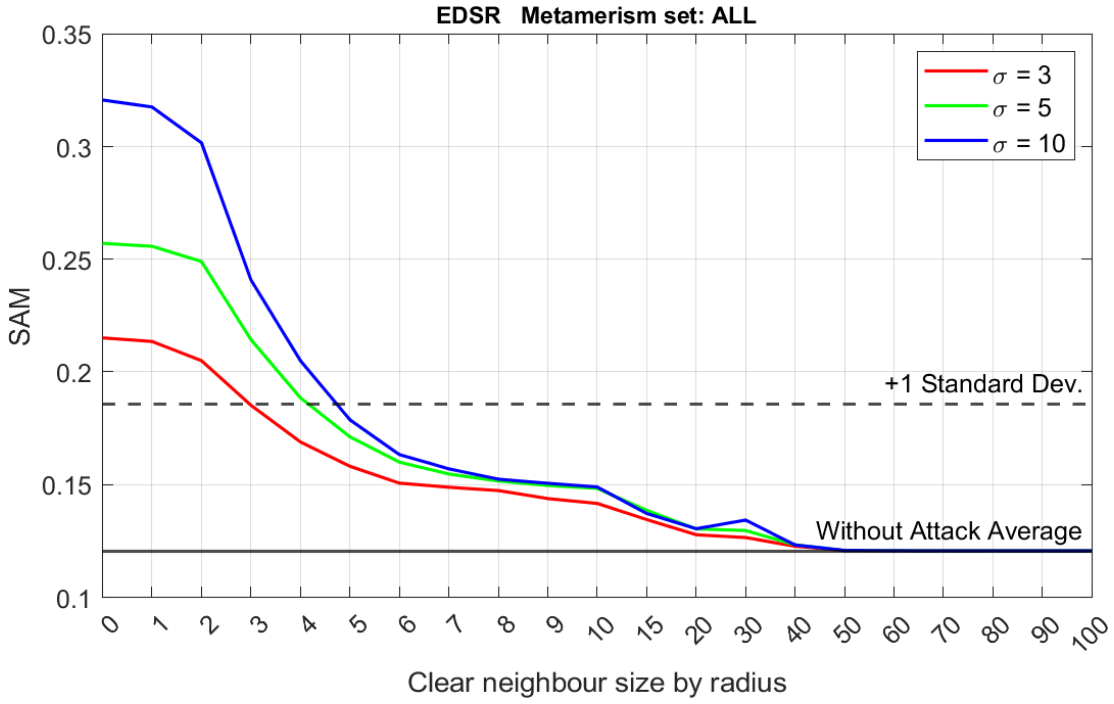


Figure 10. The SAM as a function of the untouched neighbour size when the rest of image is blurred. When the size of clear neighbour is larger than 30 pixels, EDSR have sufficient information for reconstruction.

Chapter 5.3.3. Adding noise to the image except the neighbour area

Figure 11-15 illustrates the SAM as a function of the radius for other networks.

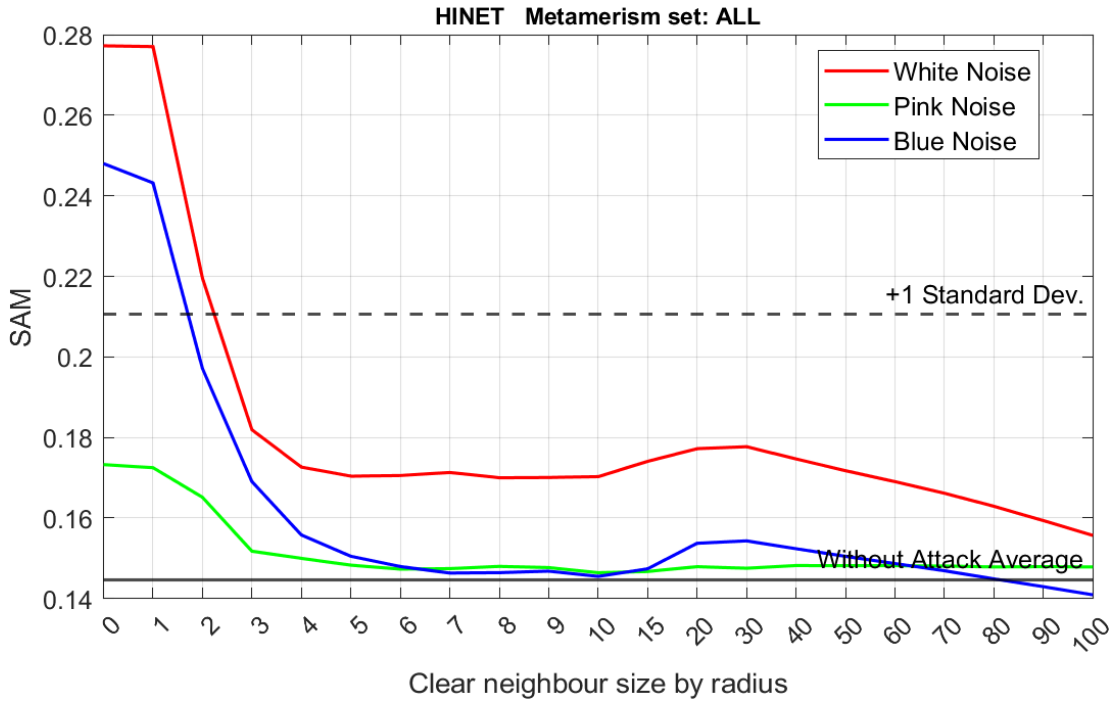


Figure 11. The SAM as a function of the untouched neighbour size when the rest of image is noised. HINET are more sensitive to local spatial information.

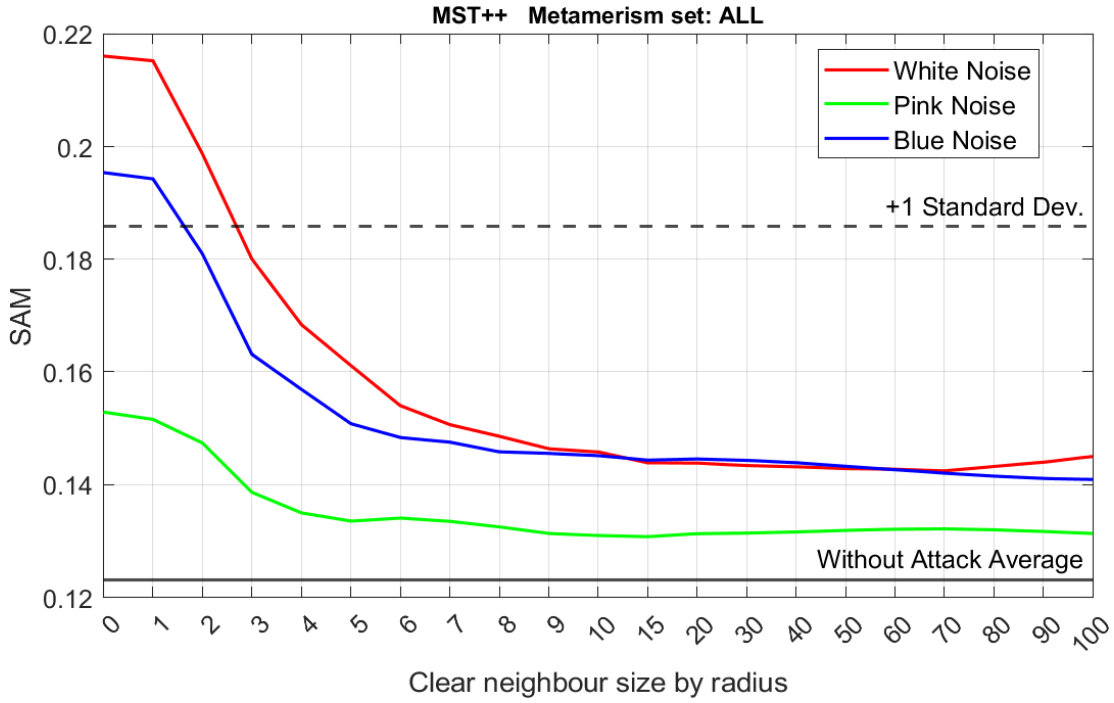


Figure 12. The SAM as a function of the untouched neighbour size when the rest of image is noised. MST++ are more sensitive to local spatial information.

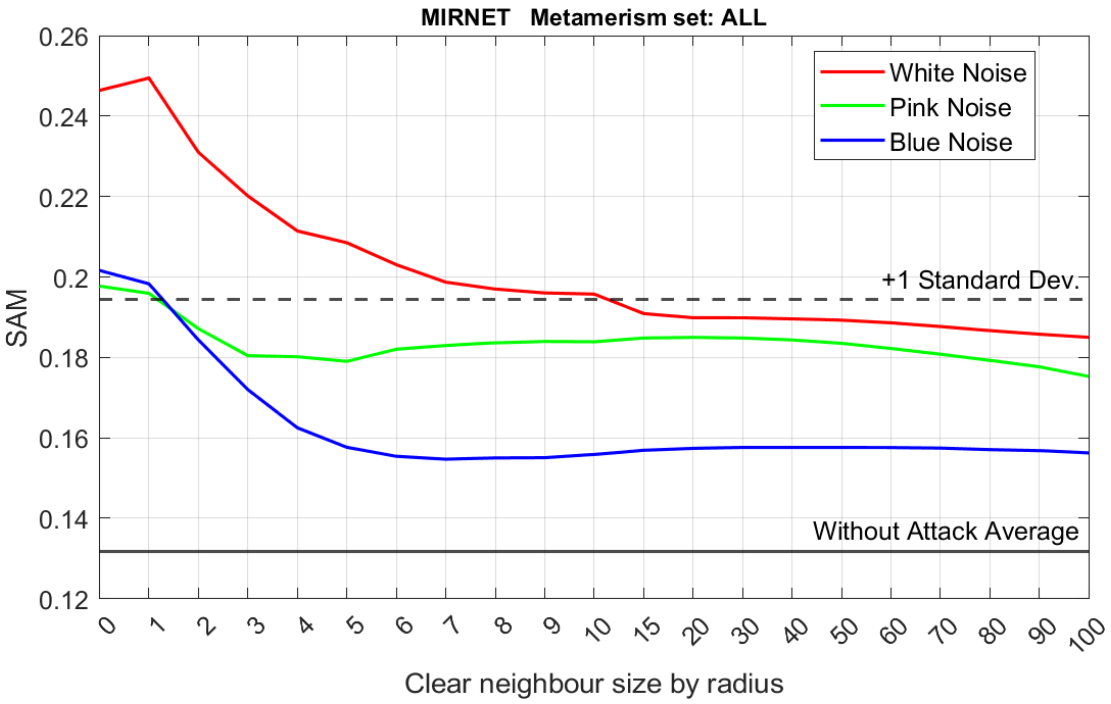


Figure 13. The SAM as a function of the untouched neighbour size when the rest of the image is noised. MIRNET uses global information to resolve metamerism but still exhibits sensitivity to local information.

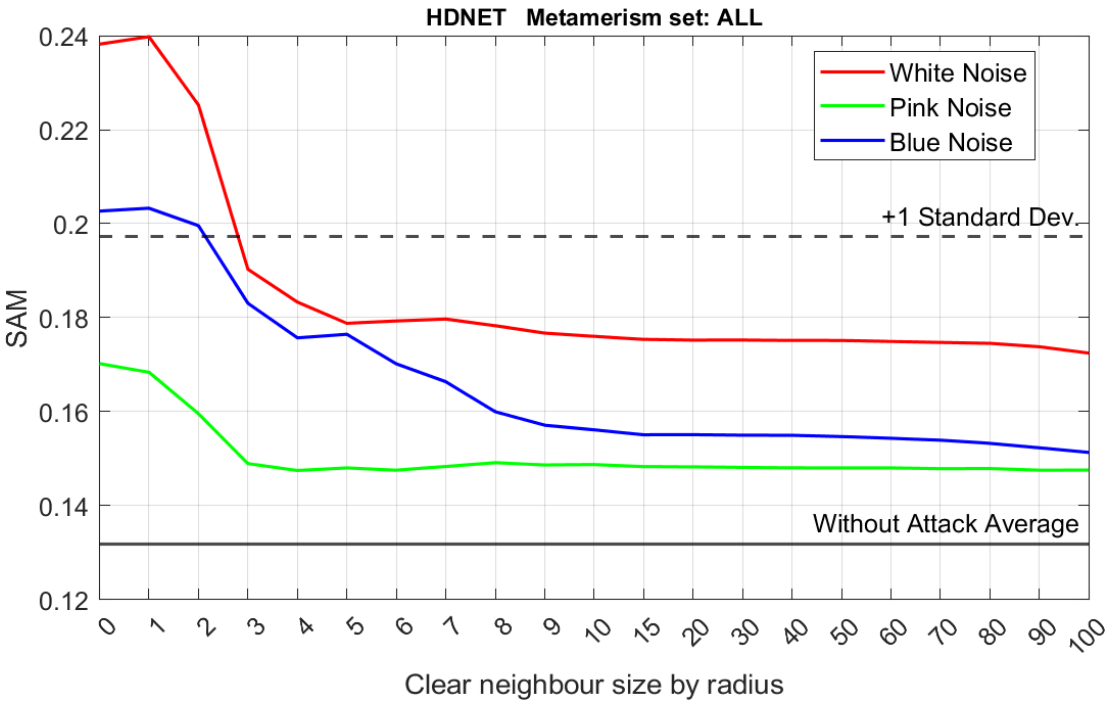


Figure 14. The SAM as a function of the untouched neighbour size when the rest of the image is noised. HDNET uses global information to resolve metamerism but still exhibits sensitivity to local information.

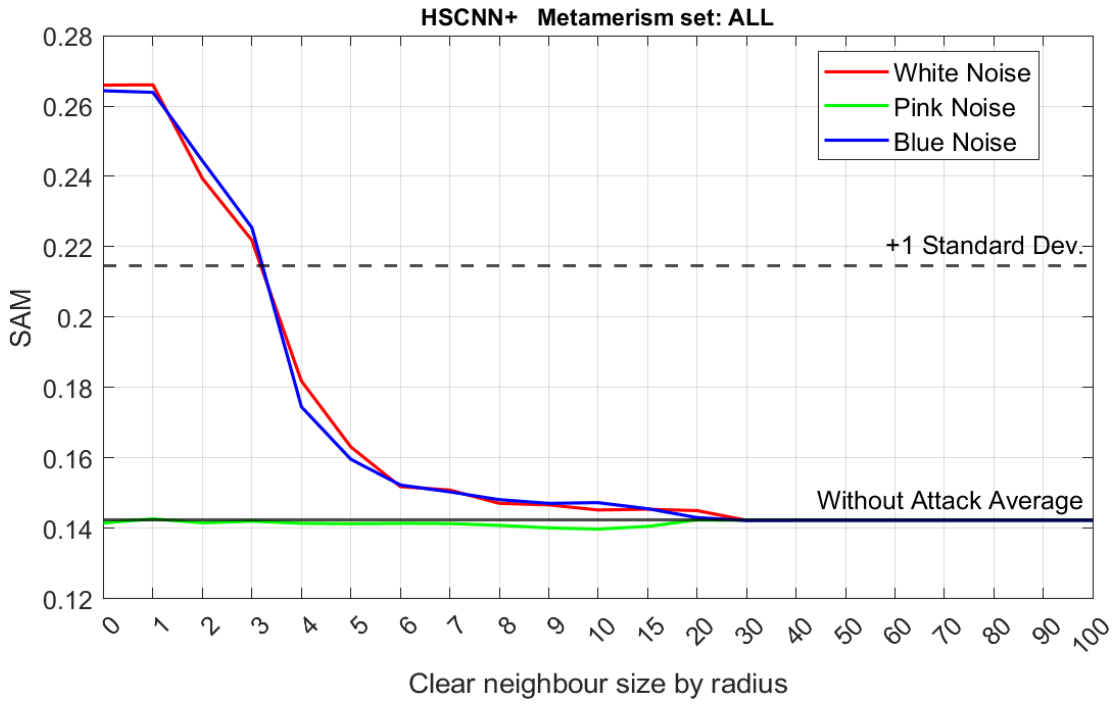


Figure 15. The SAM as a function of the untouched neighbour size when the rest of the image is noised. When the size of clear neighbour is larger than 30 pixels, HSCNN+ have sufficient information for reconstruction.

Chapter 5.3.4. Attacks in the colour of the input image

Figure 16-20 illustrates the SAM as a function of the radius for other networks.

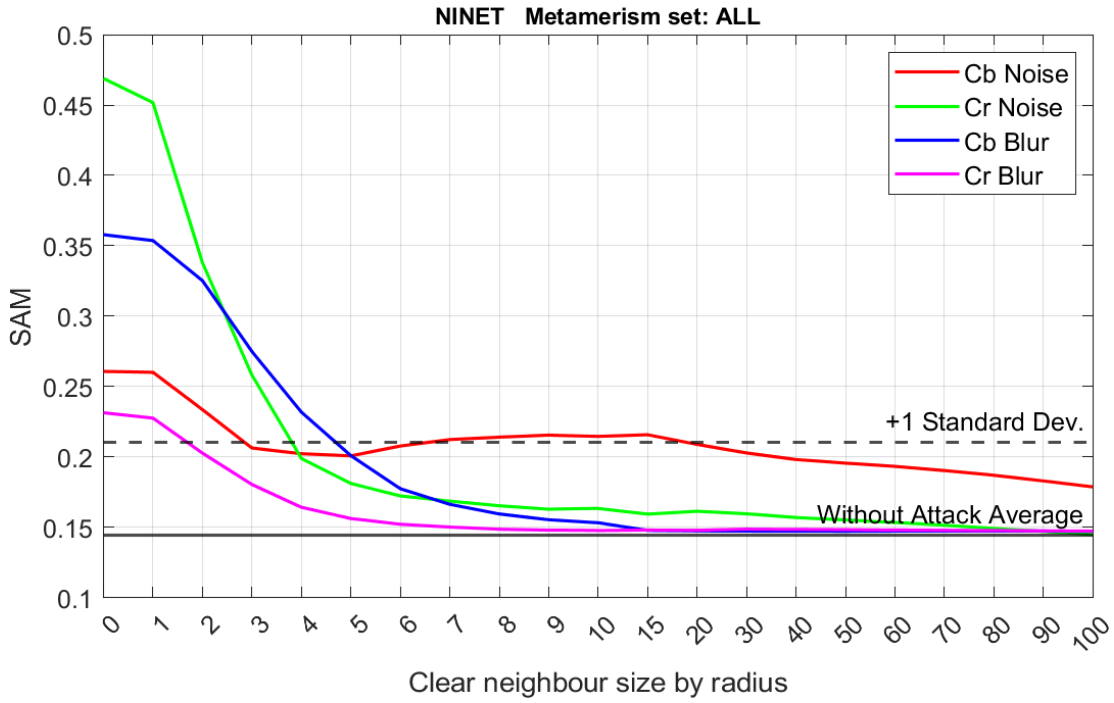


Figure 16. The average SAM as a function of the untouched neighbour size when the rest of the image is attacked in YCbCr colour space.

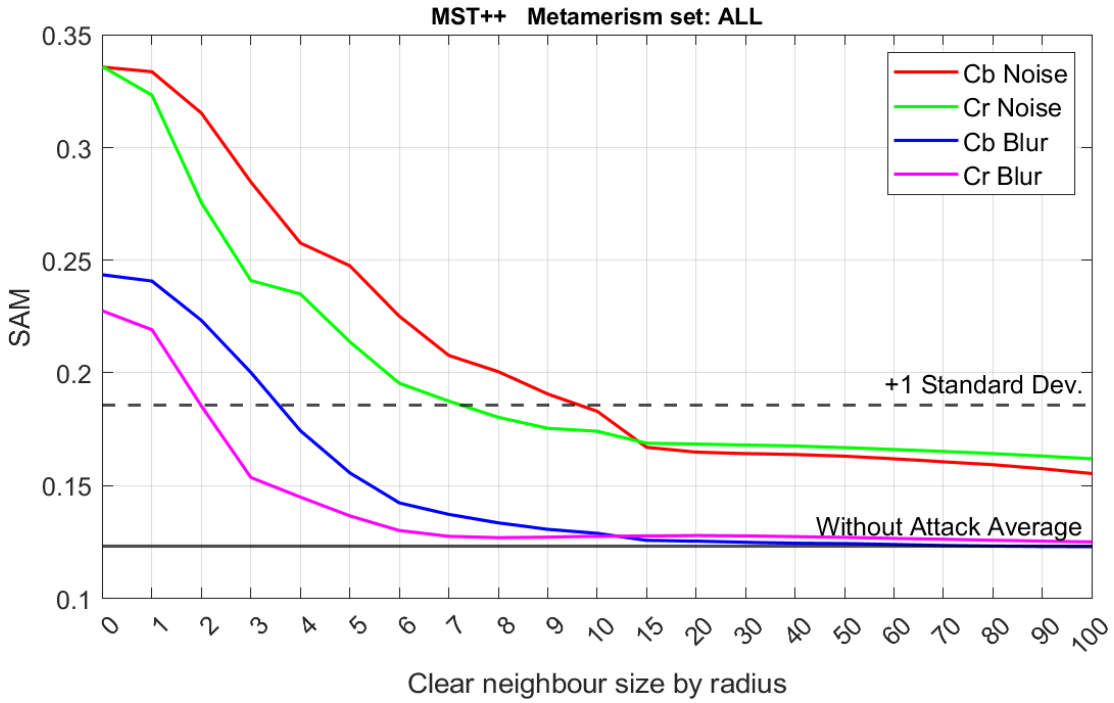


Figure 17. The average SAM as a function of the untouched neighbour size when the rest of the image is attacked in YCbCr colour space.

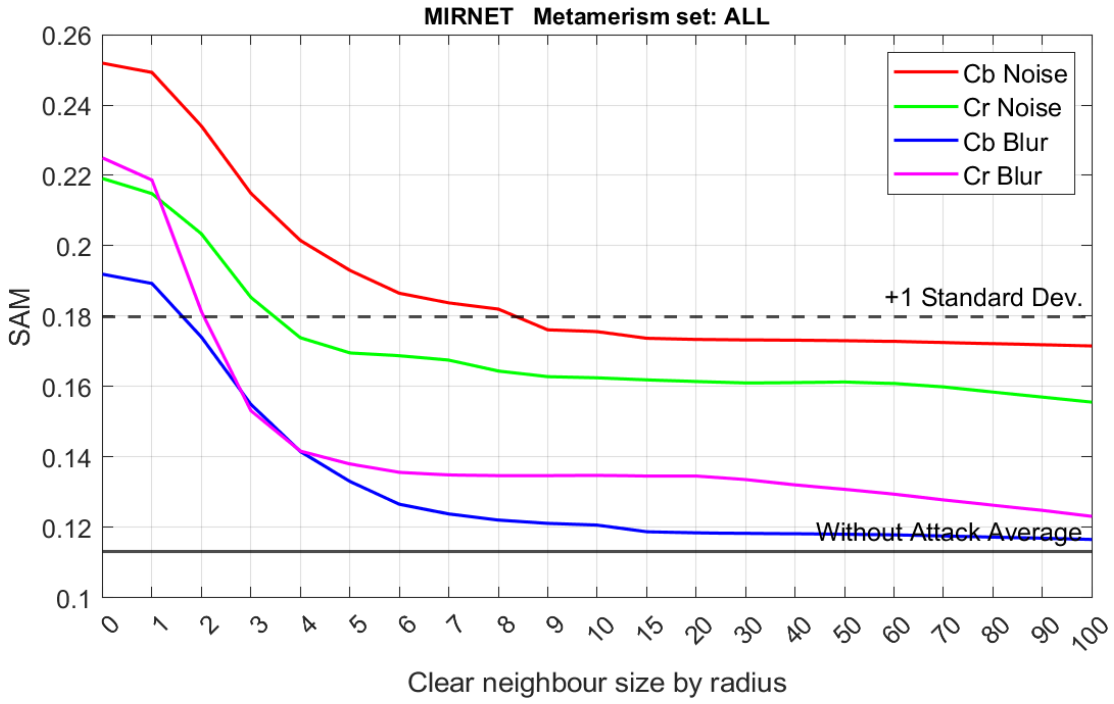


Figure 18. The average SAM as a function of the untouched neighbour size when the rest of the image is attacked in YCbCr colour space.

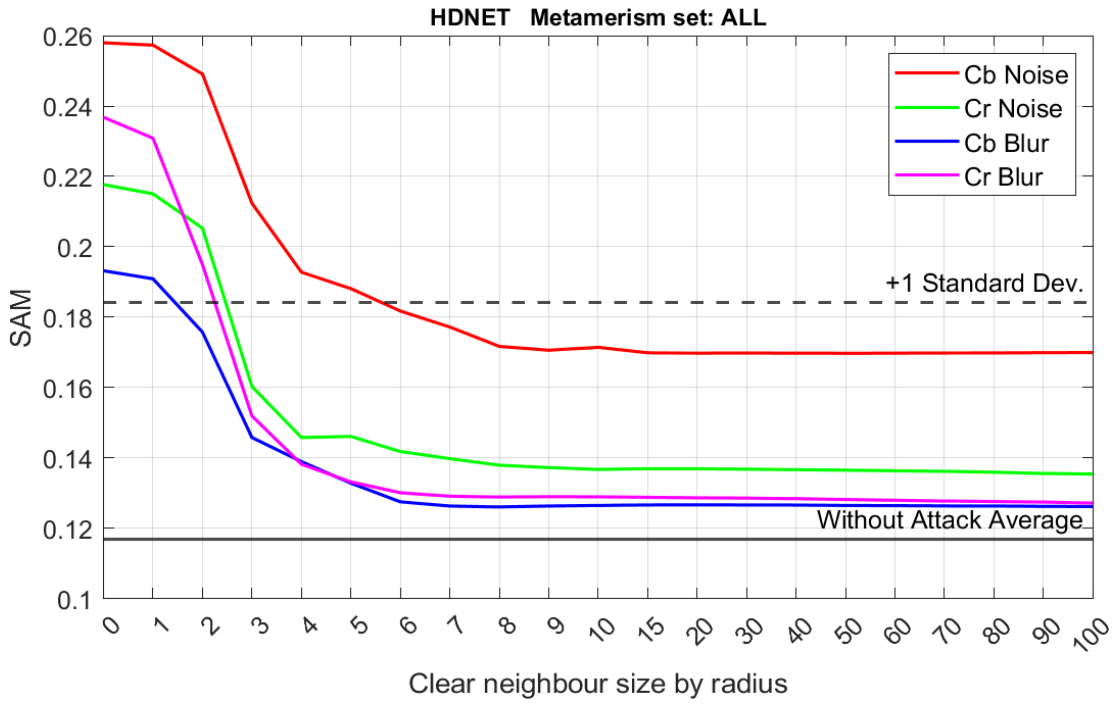


Figure 19. The average SAM as a function of the untouched neighbour size when the rest of the image is attacked in YCbCr colour space.

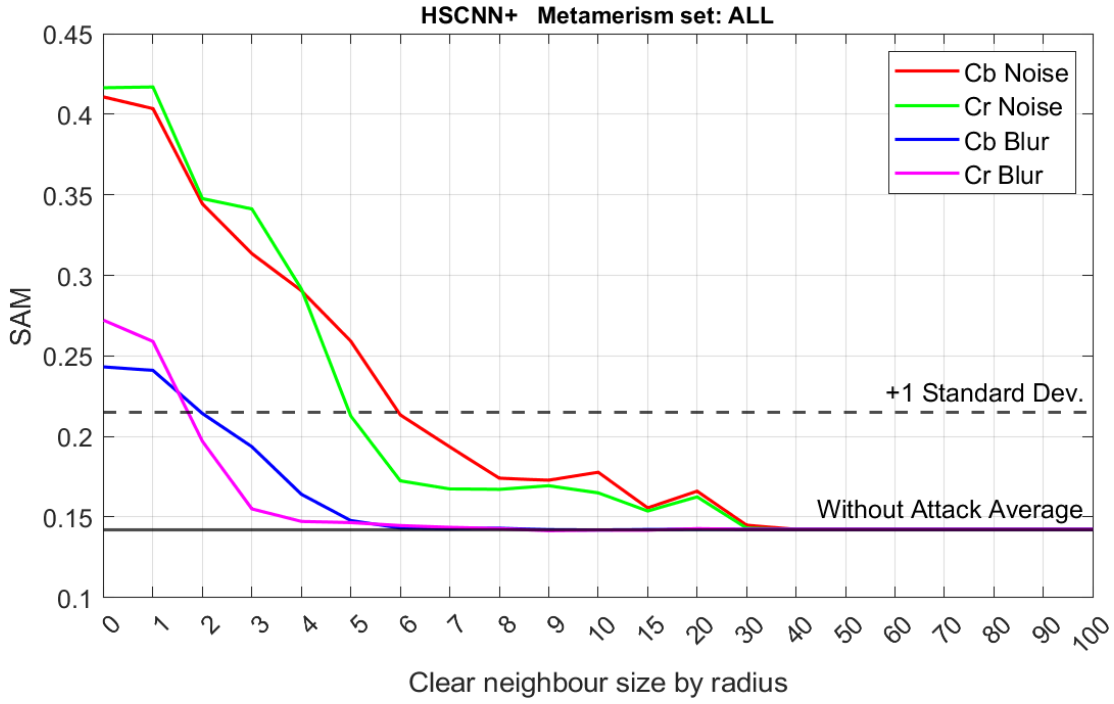
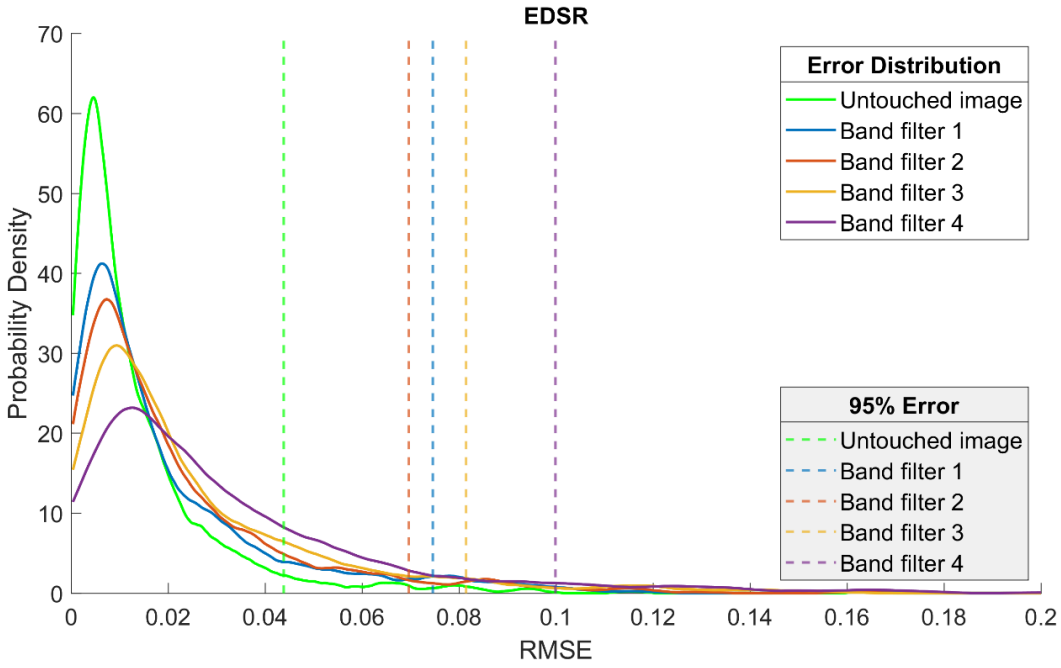


Figure 20. The average SAM as a function of the untouched neighbour size when the rest of the image is attacked in YCbCr colour space.

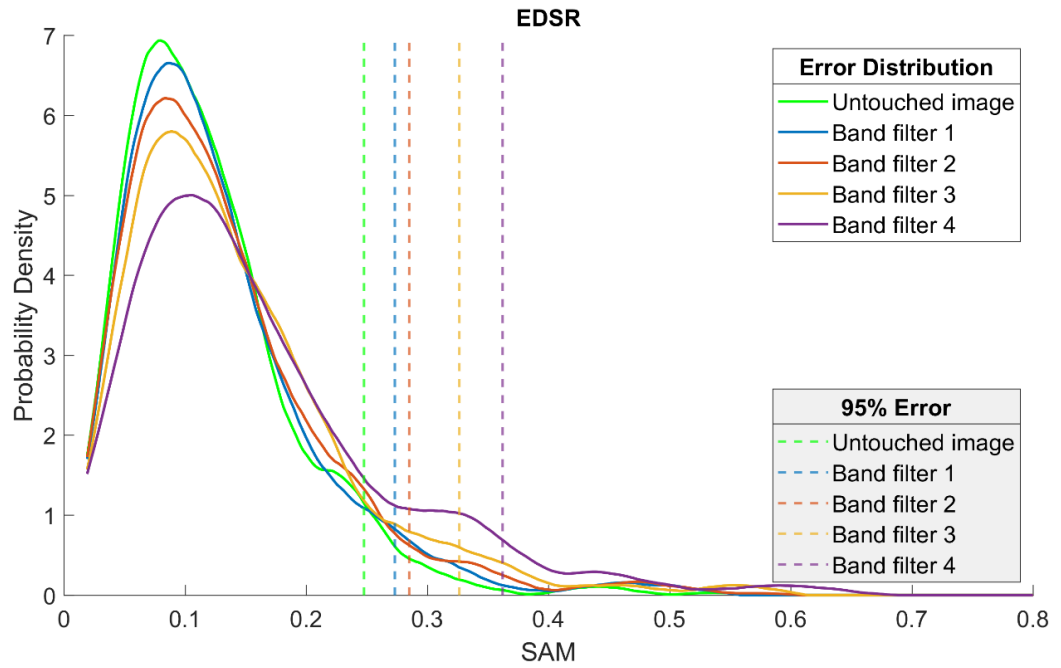
Appendix 5

Chapter 5.4.1.

Figure 1-4 demonstrate the reconstruction accuracy for other networks, after the neighbouring area is attacked by band-stop filters.

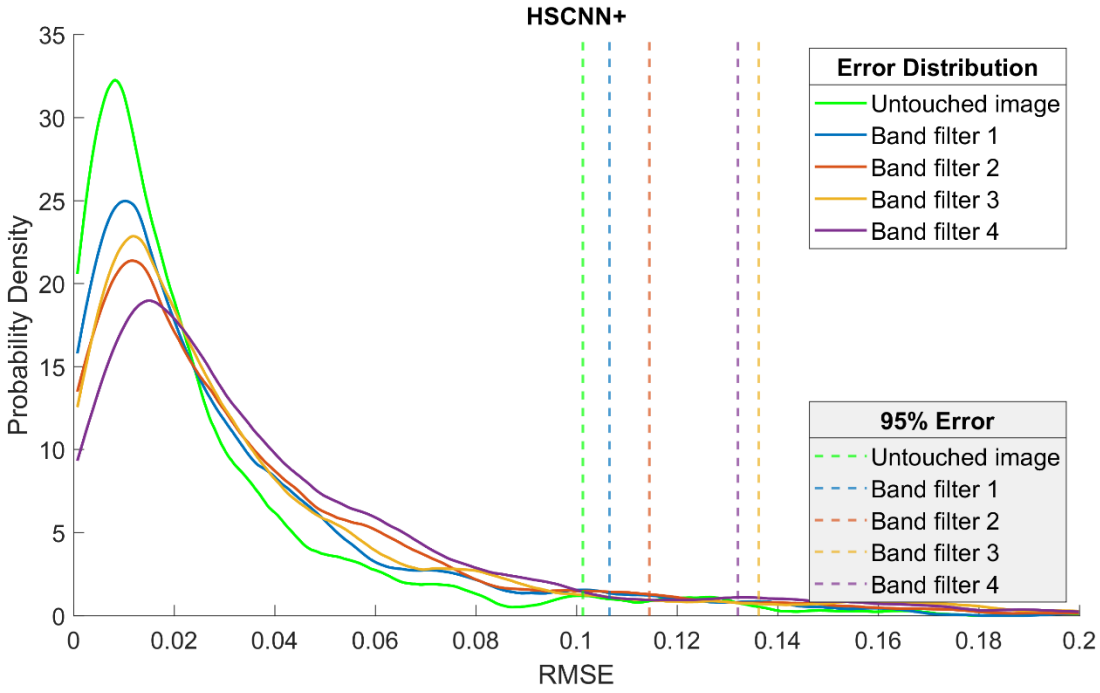


(a)

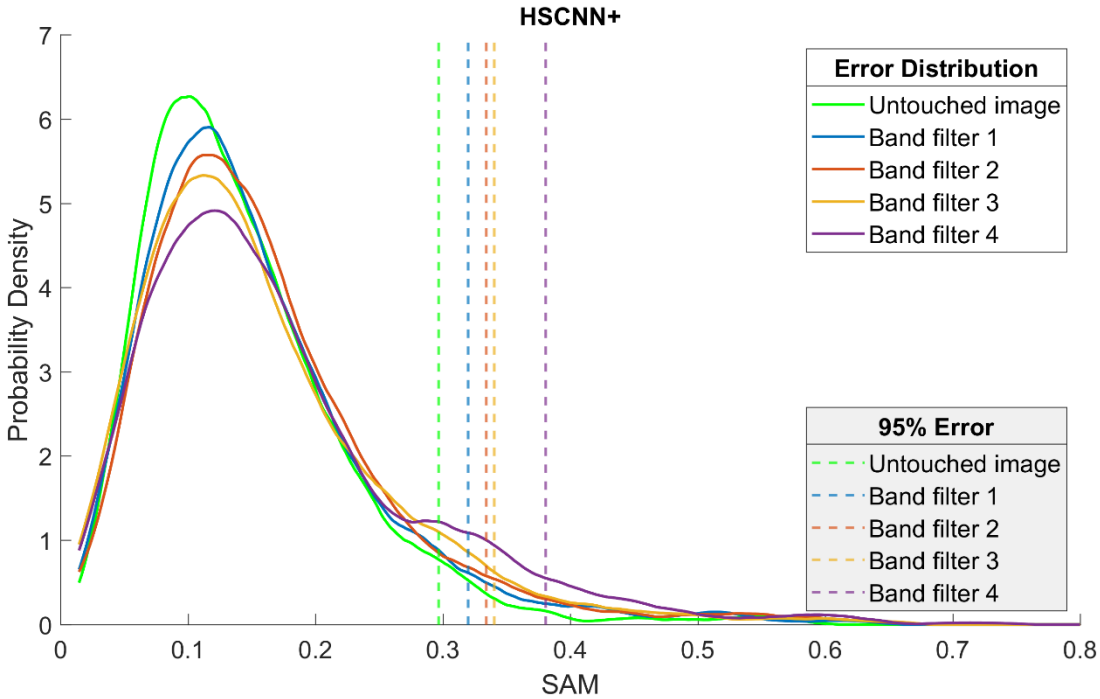


(b)

Figure 1. The error distribution and 95% error when the image is attacked by band stop filter of EDSR.

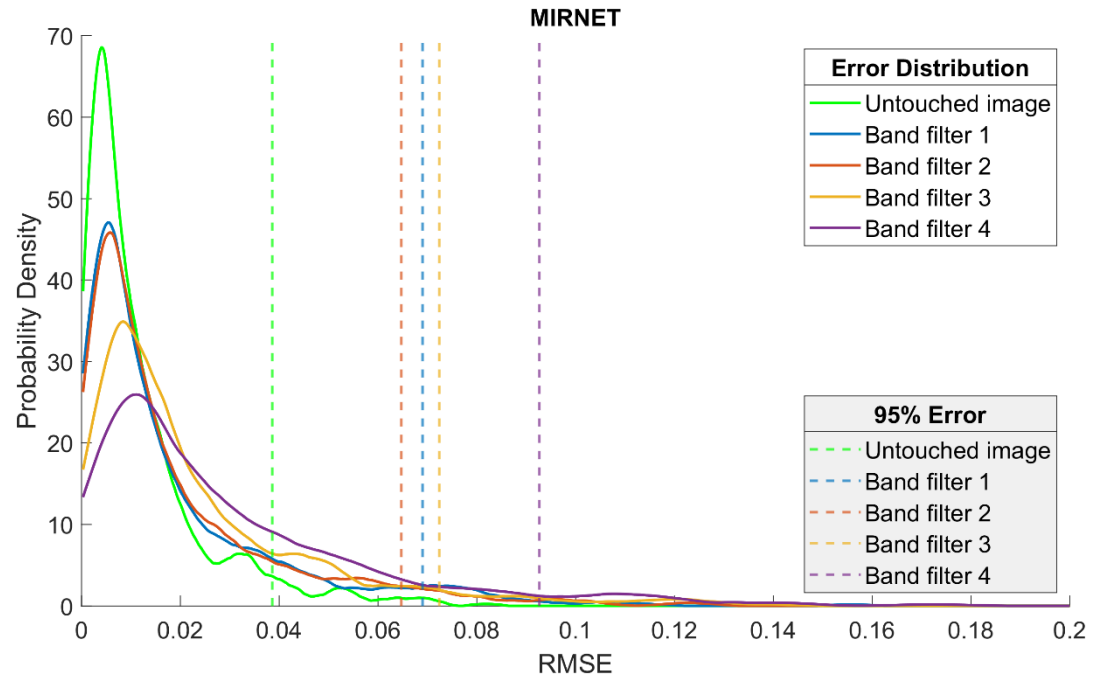


(a)

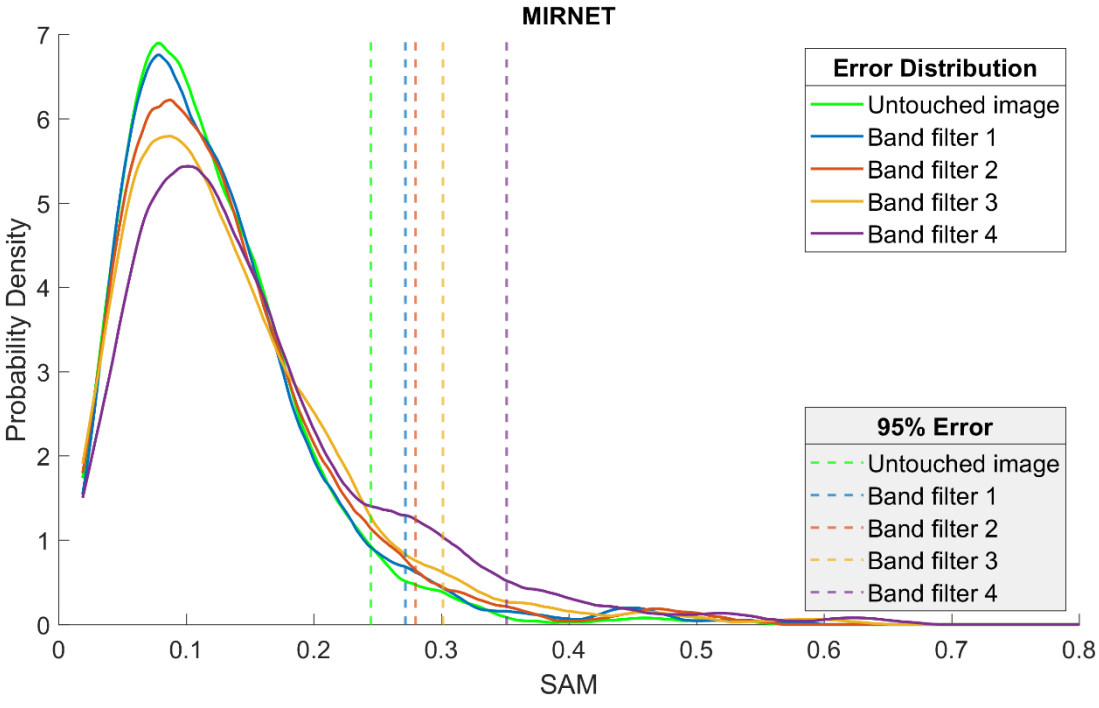


(b)

Figure 2. The error distribution and 95% error when the image is attacked by band stop filter of HSCNN+.

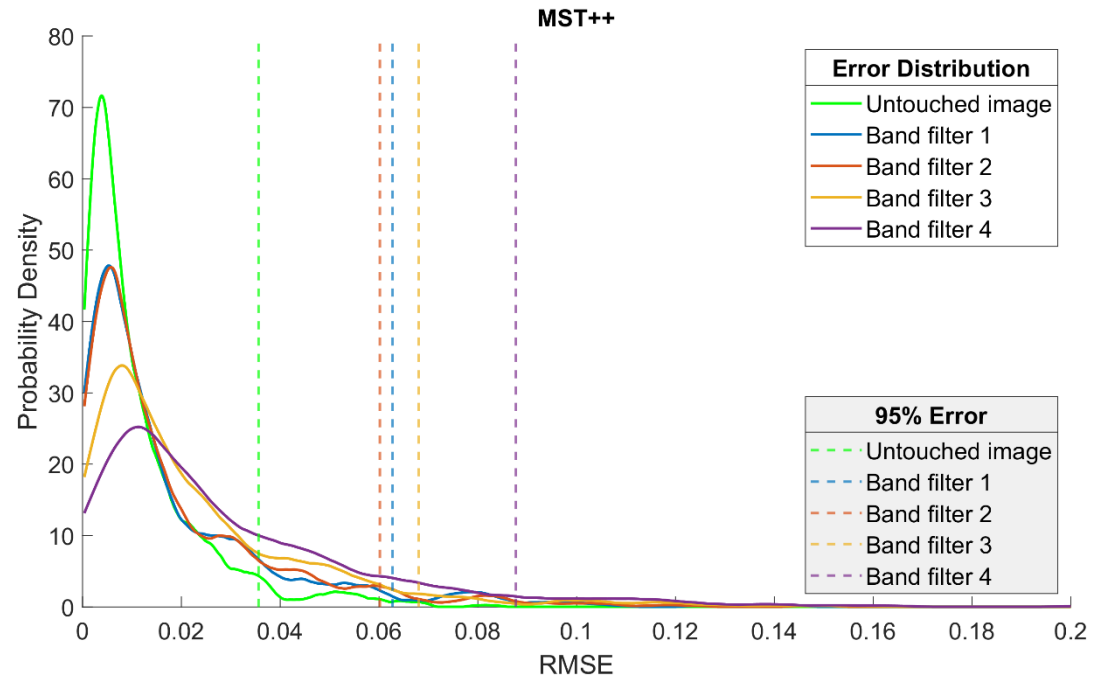


(a)

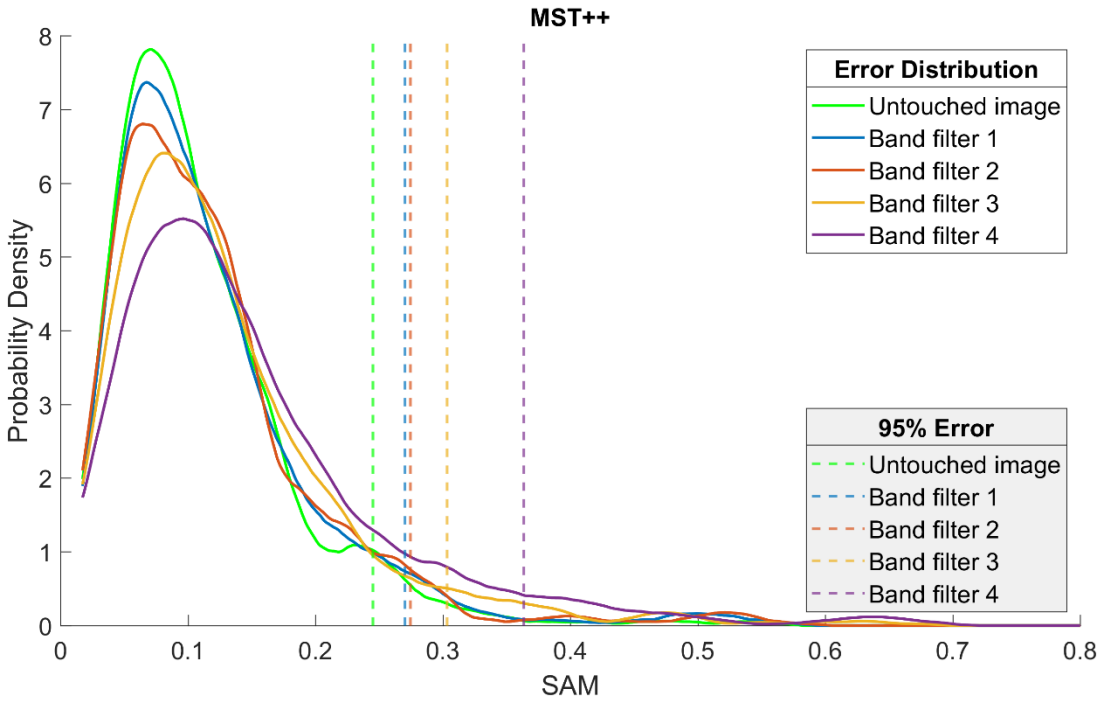


(b)

Figure 3. The error distribution and 95% error when the image is attacked by band stop filter of MIRNET.



(a)

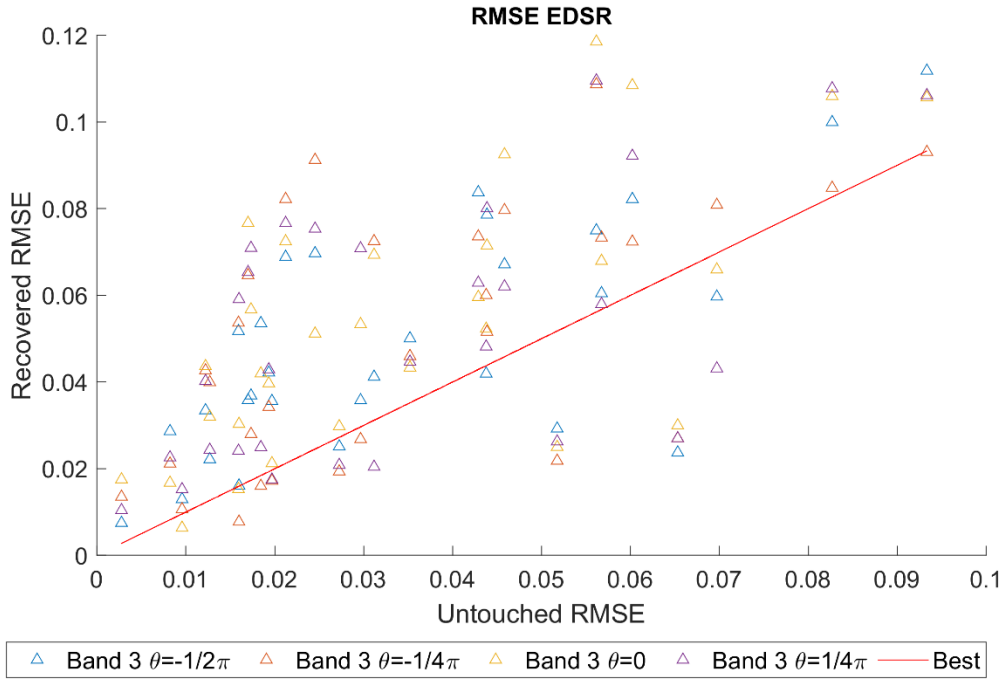


(b)

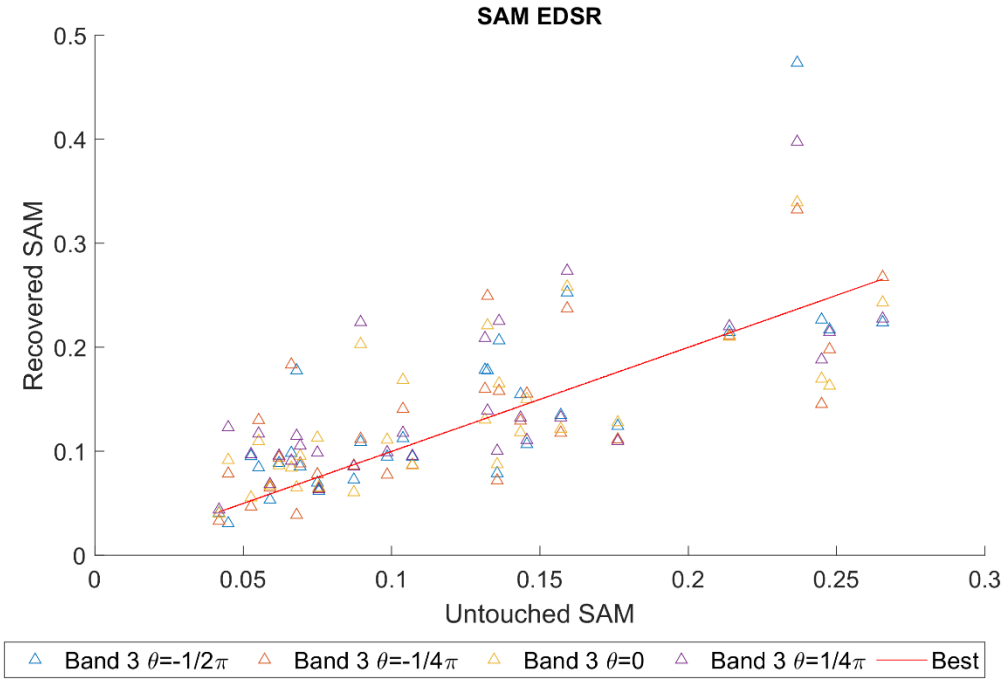
Figure 4. The error distribution and 95% error when the image is attacked by band stop filter of MST++.

Chapter 5.4.2.

Figure 5-7 (a) and (b) illustrates the RMSE and SAM of the recovered samples from a metamerism set when local textural features are removed across different orientations.

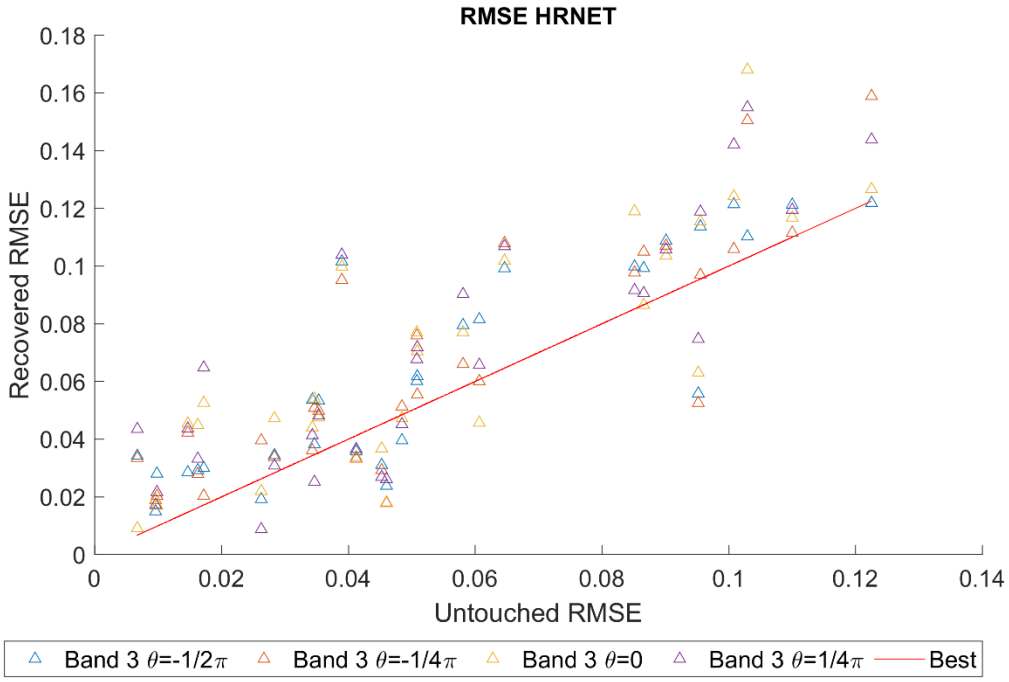


(a)

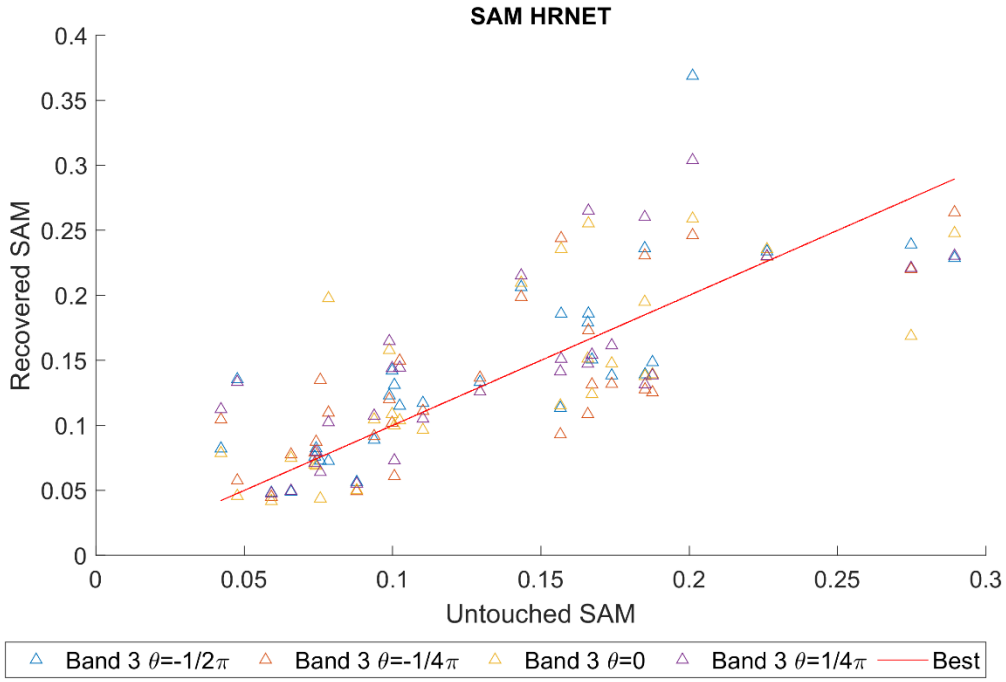


(b)

Figure 5. Reconstruction accuracy affected by removing local texture features in varying orientations but same scale. The EDSR are sensitive to local textural feature in different orientations, appearing with the change of reconstruction accuracy of each spectral sample.

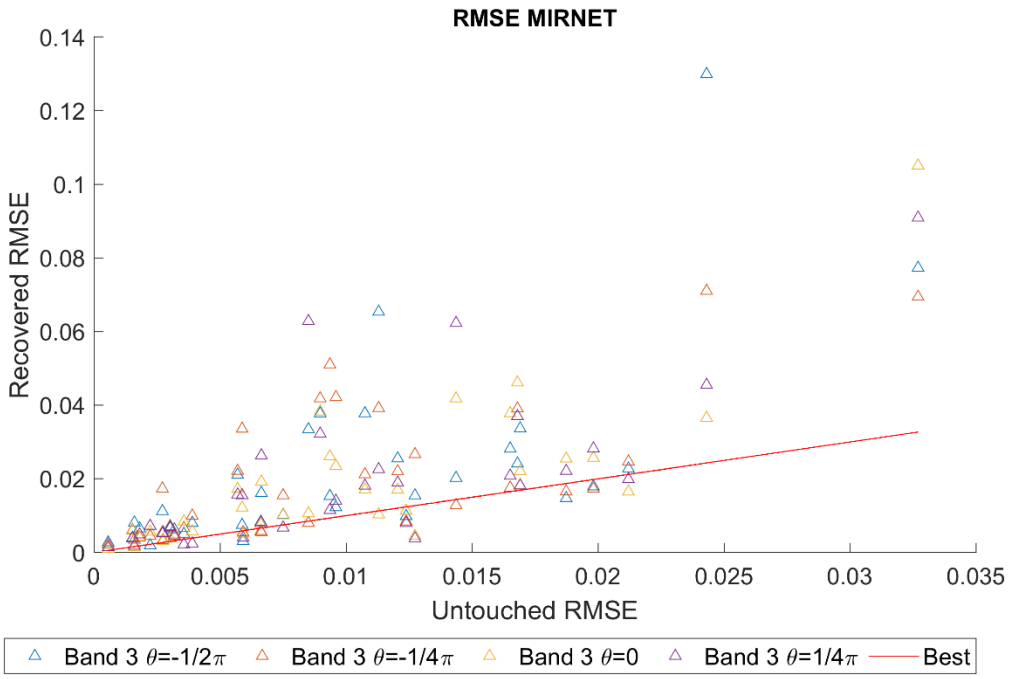


(a)

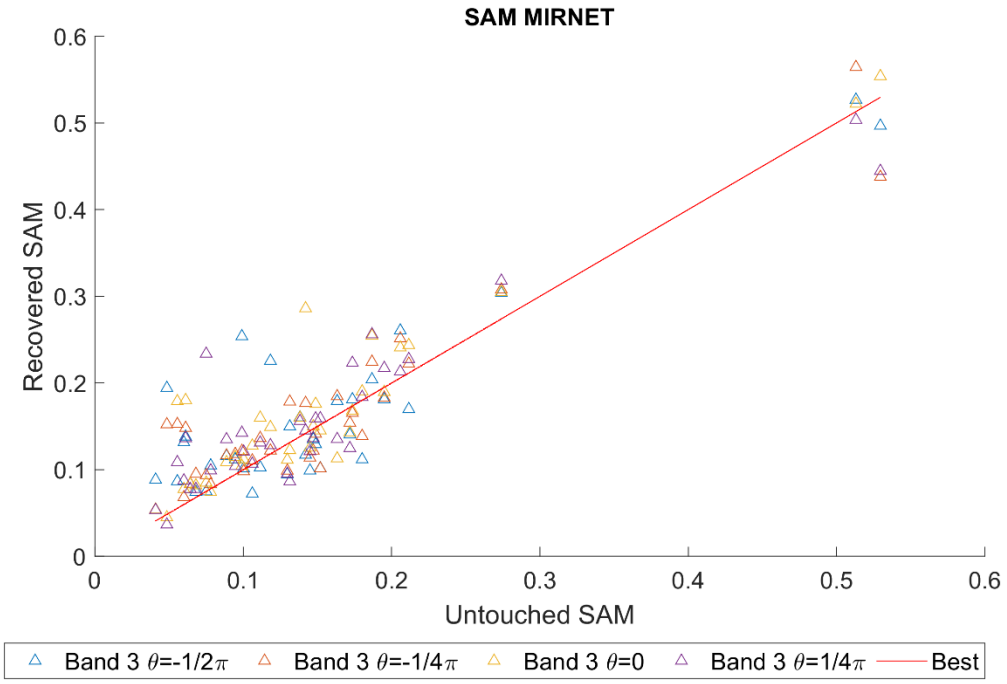


(b)

Figure 6. Reconstruction accuracy affected by removing local texture features in varying orientations but same scale. The HRNET are sensitive to local textural feature in different orientations, appearing with the change of reconstruction accuracy of each spectral sample.



(a)



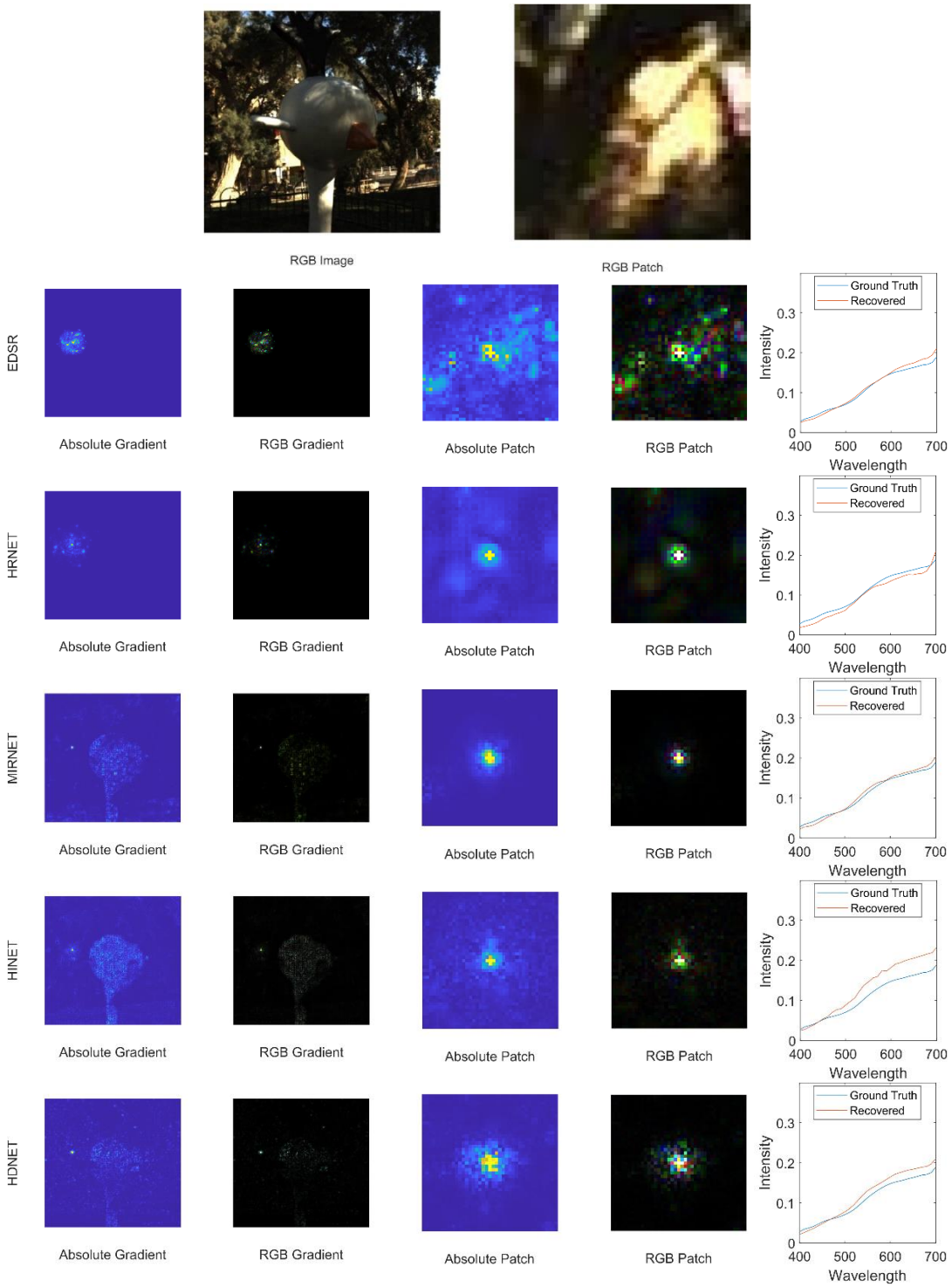
(b)

Figure 7. Reconstruction accuracy affected by removing local texture features in varying orientations but same scale. The HRNET are sensitive to local textural feature in different orientations, appearing with the change of reconstruction accuracy of each spectral sample.

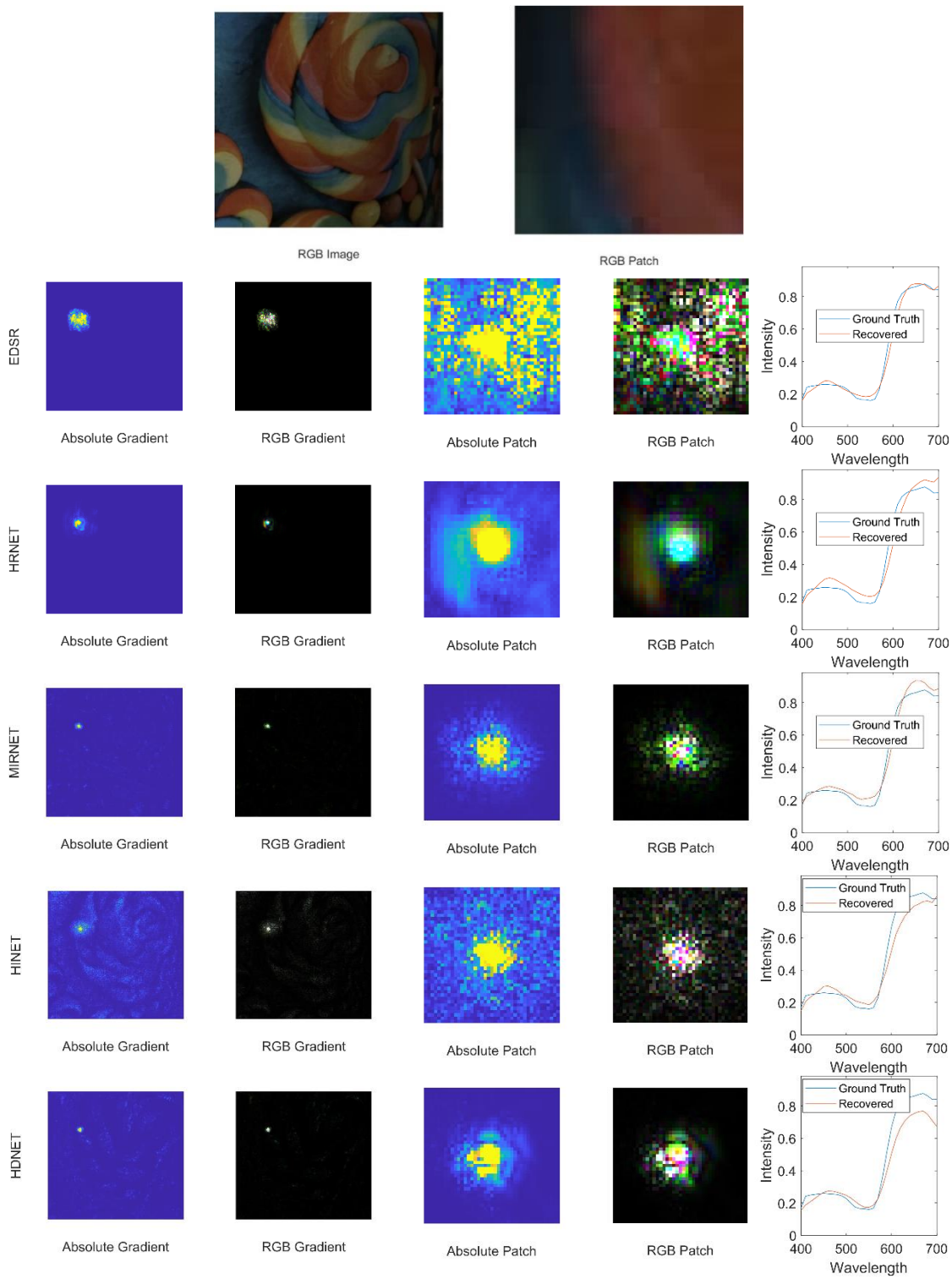
Appendix 6

Chapter 5.5.4.

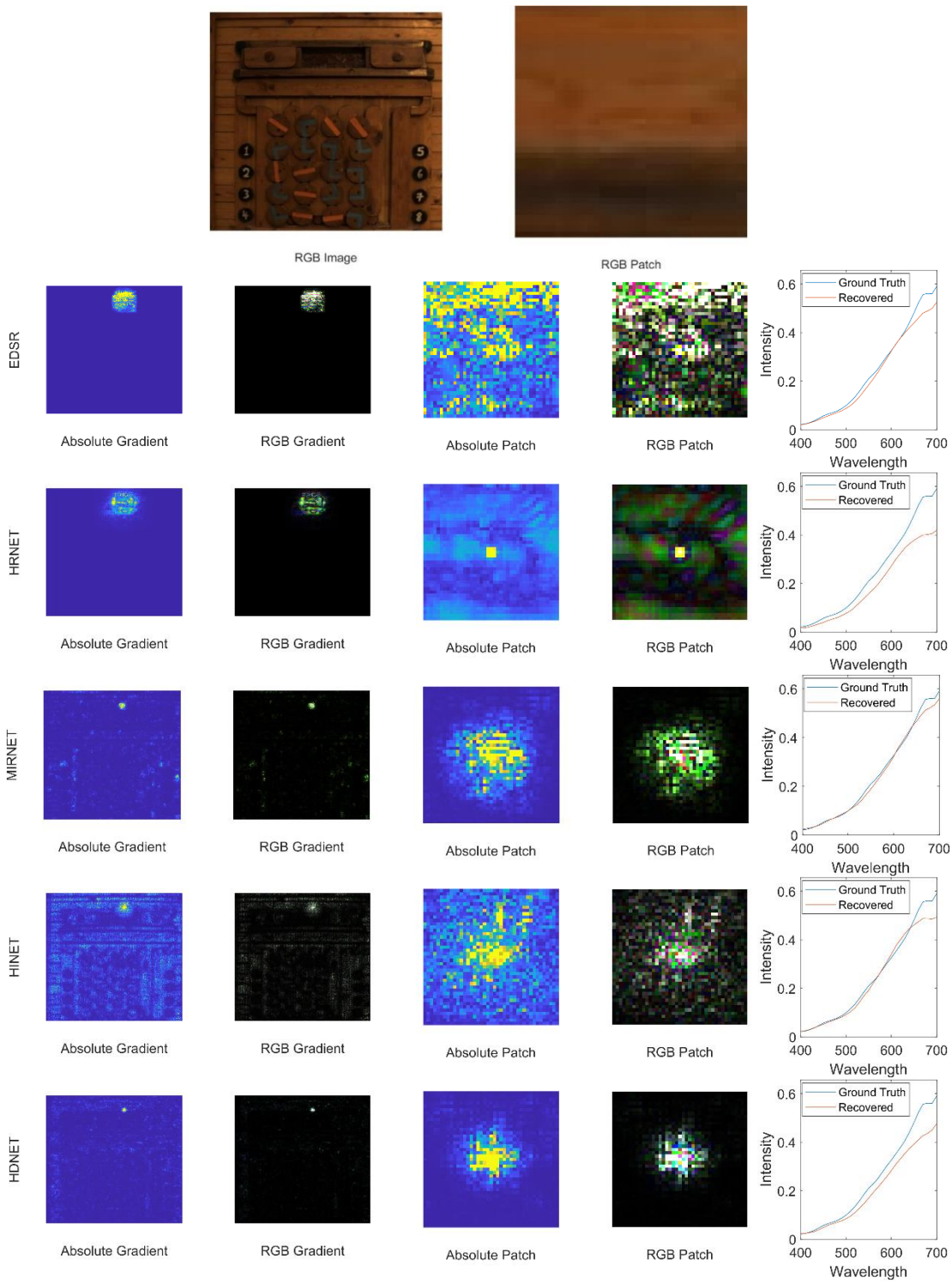
Figure 1 illustrates how networks appear to be sensitive to the global information within the input RGB image.



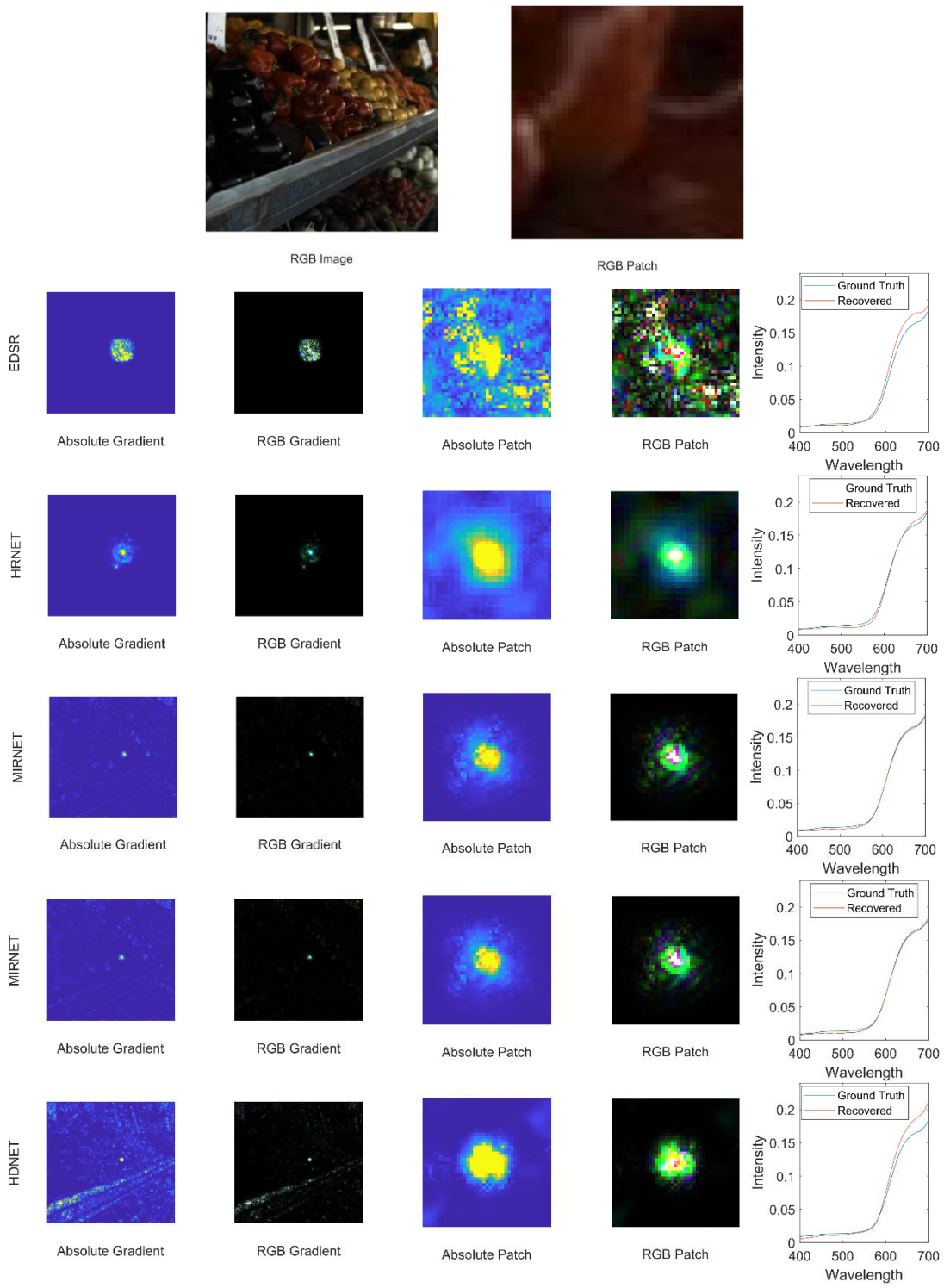
(a)


RGB Image
RGB Patch
EDSR
HRNET
MIRNET
HINET
HDNET
Absolute Gradient
RGB Gradient
Absolute Patch
RGB Patch
Ground Truth
Recovered
Intensity
Wavelength
0
0.2
0.4
0.6
0.8
400
500
600
700

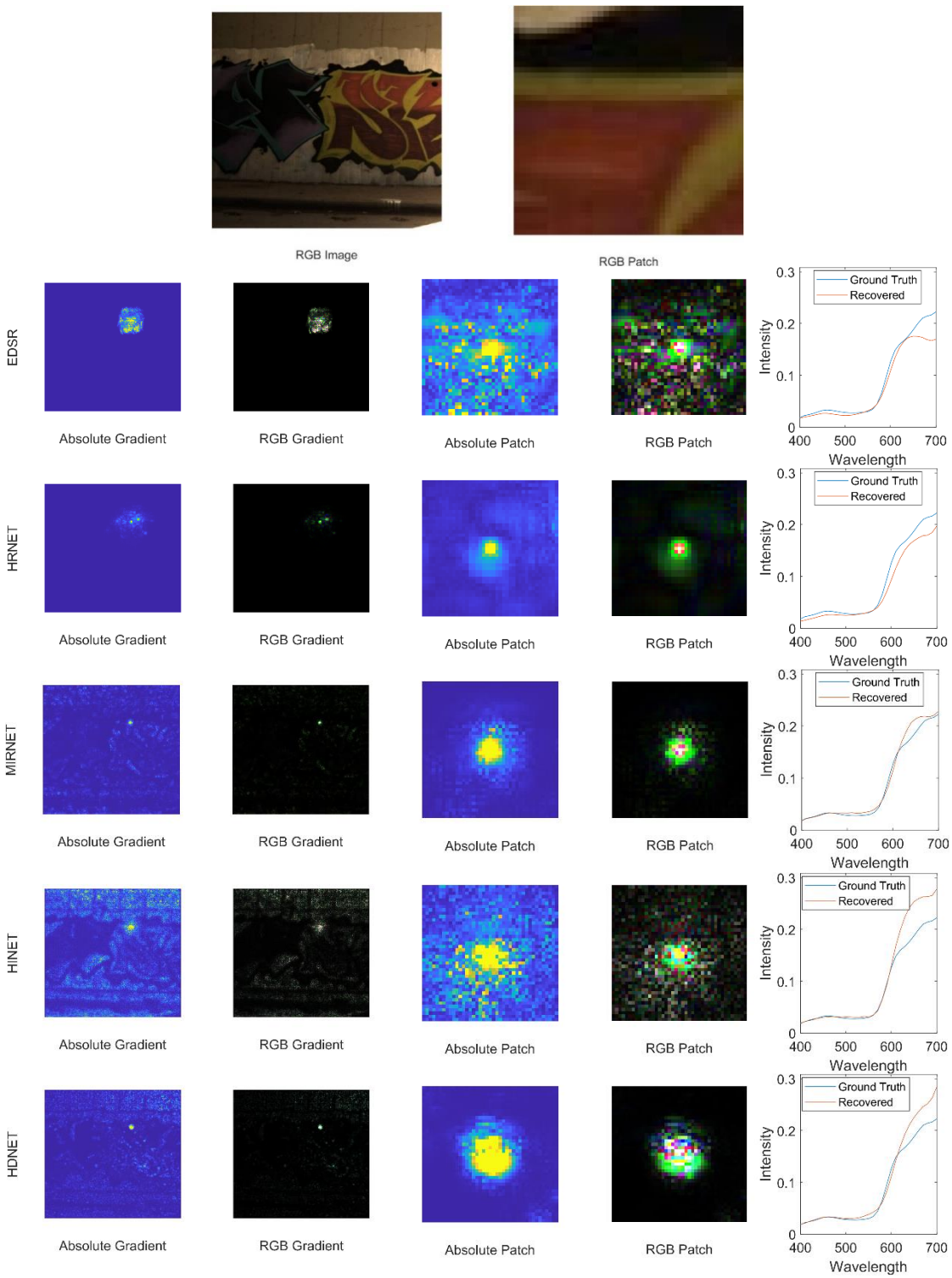
(b)


RGB Image
RGB Patch
EDSR
HRNET
MIRNET
HINET
HDNET
Absolute Gradient
RGB Gradient
Absolute Patch
RGB Patch
Ground Truth
Recovered
Intensity
Wavelength
0
0.2
0.4
0.6
400
500
600
700

(c)


RGB Image
RGB Patch
EDSR
HRNET
MIRNET
MIRNET
HDNET
Absolute Gradient
RGB Gradient
Absolute Patch
RGB Patch
Ground Truth
Recovered
Intensity
Wavelength
0
0.05
0.1
0.15
0.2
400
500
600
700

(d)

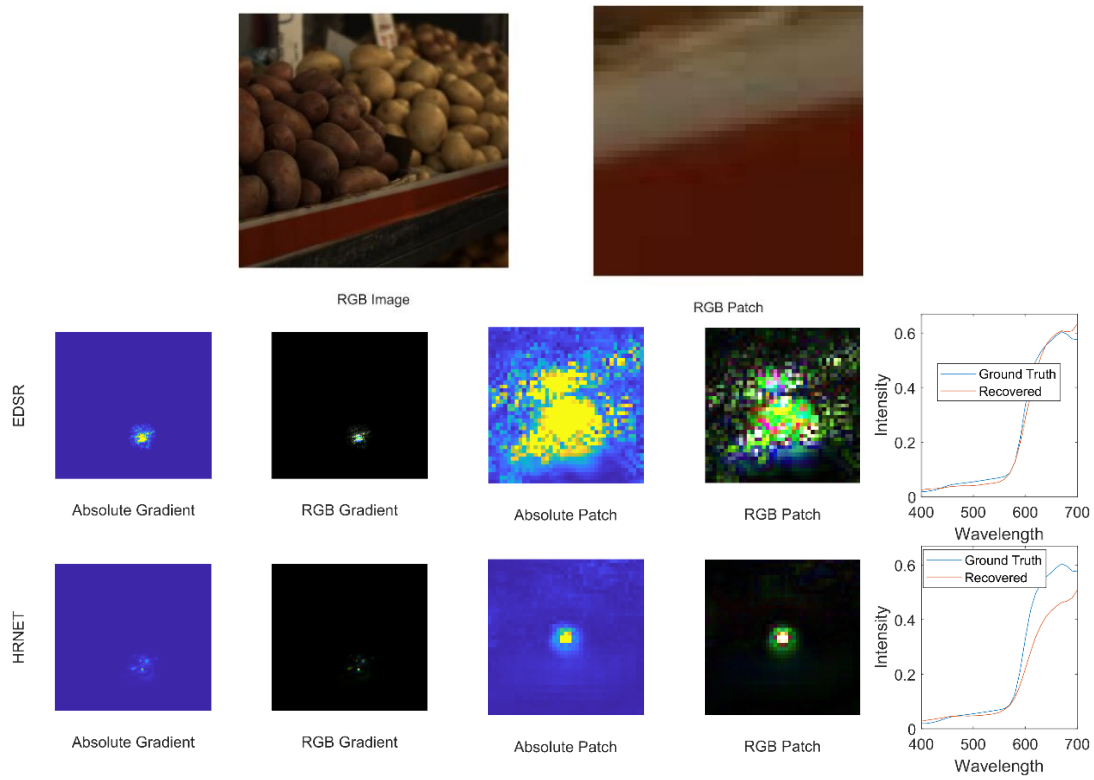


(e)

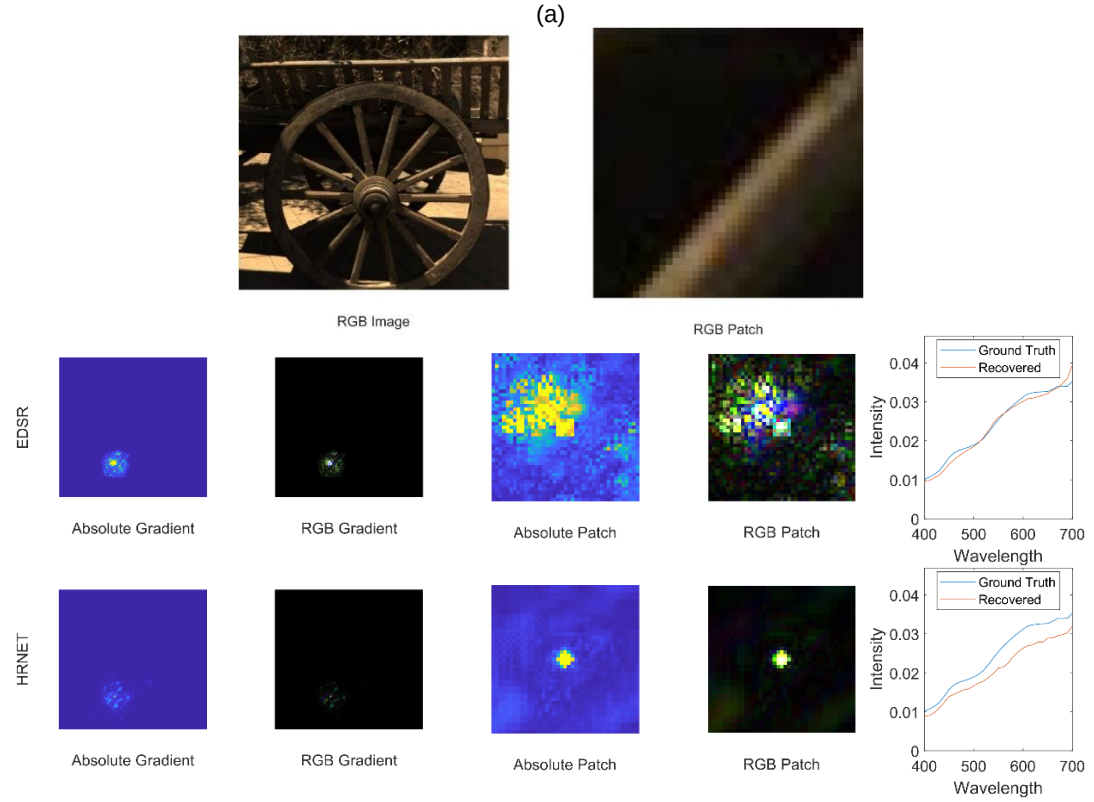
Figure 1. Gradient of input image that the listed networks are sensitive to global information.

Chapter 5.5.5.

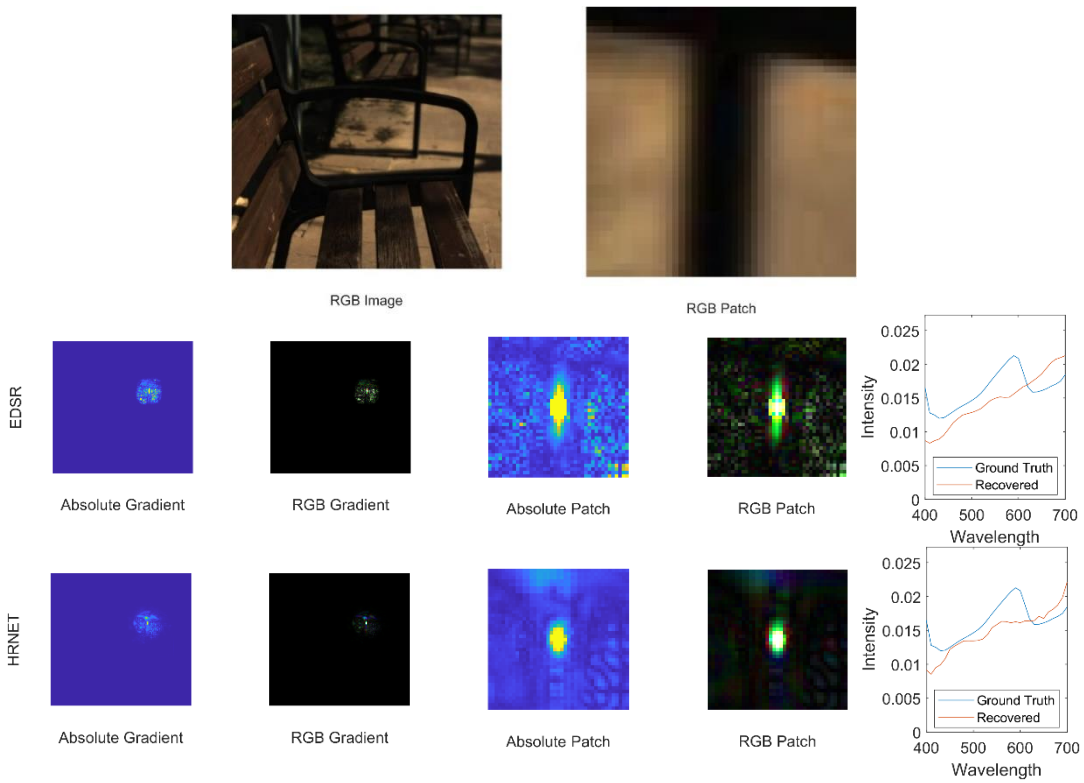
Figure 2 illustrates how networks appear to be sensitive to the local information within the input RGB image.



(a)



(b)



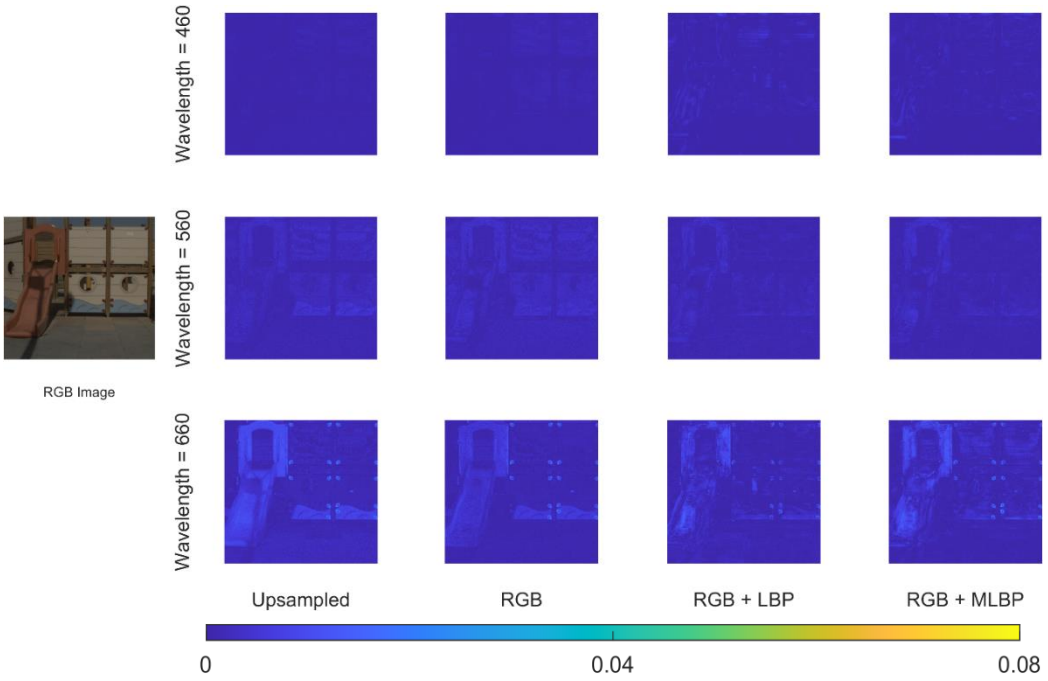
(c)

Figure 2. Gradient of input image where networks are sensitive to local information.

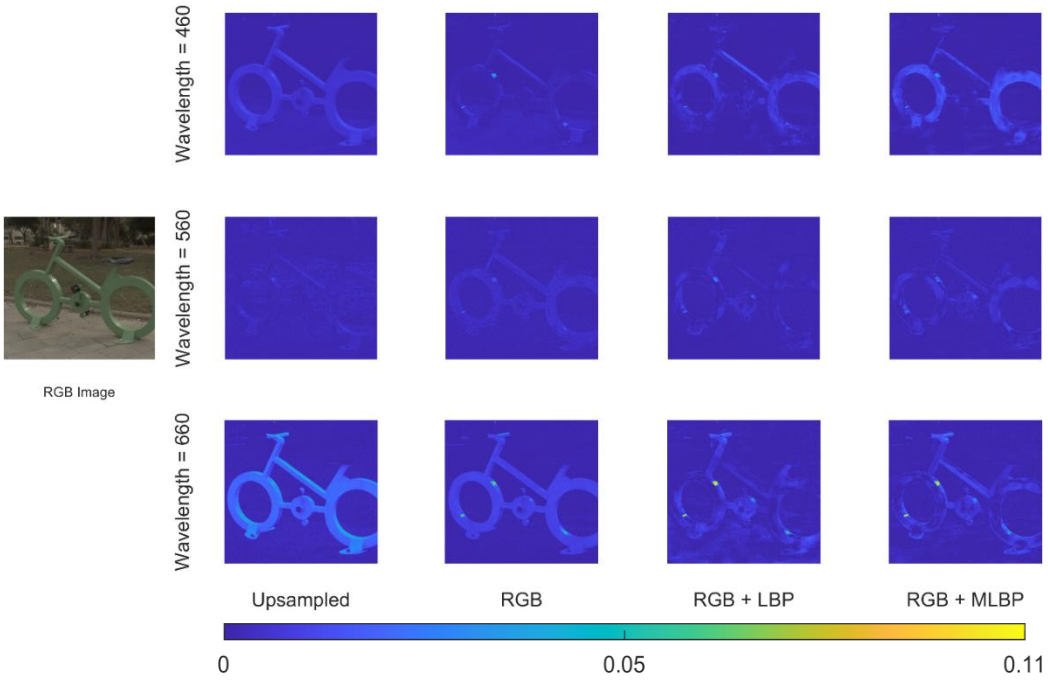
Appendix 7

Chapter 6.2.2.

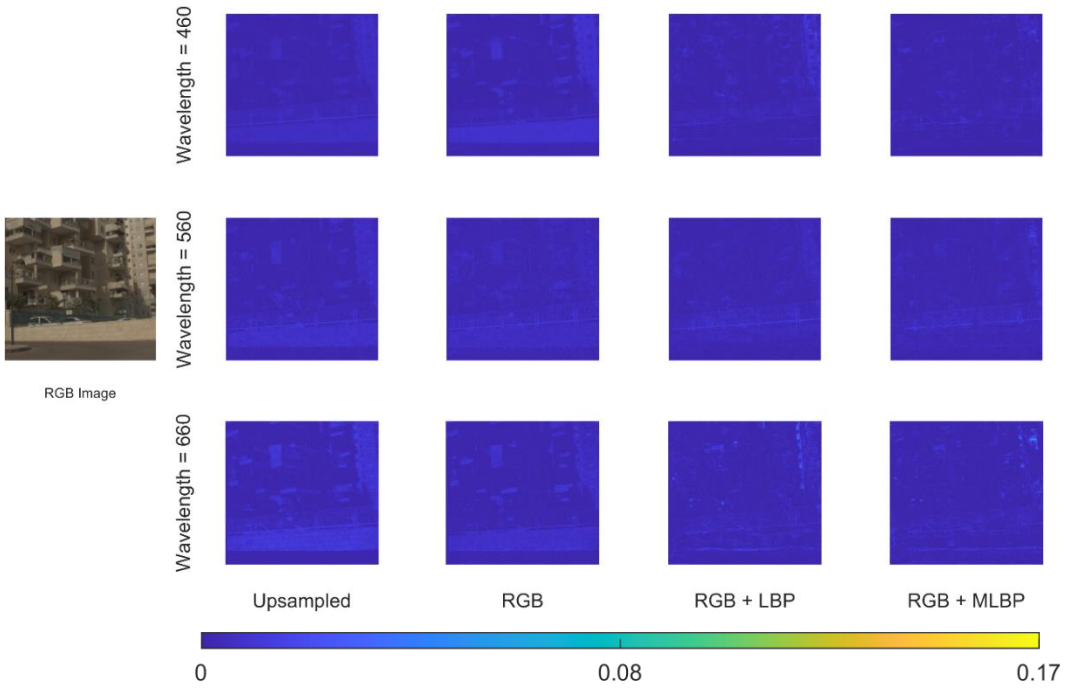
Figure 1 compares the reconstruction accuracy at three selected wavelengths corresponding to blue, green, and red wavelengths.



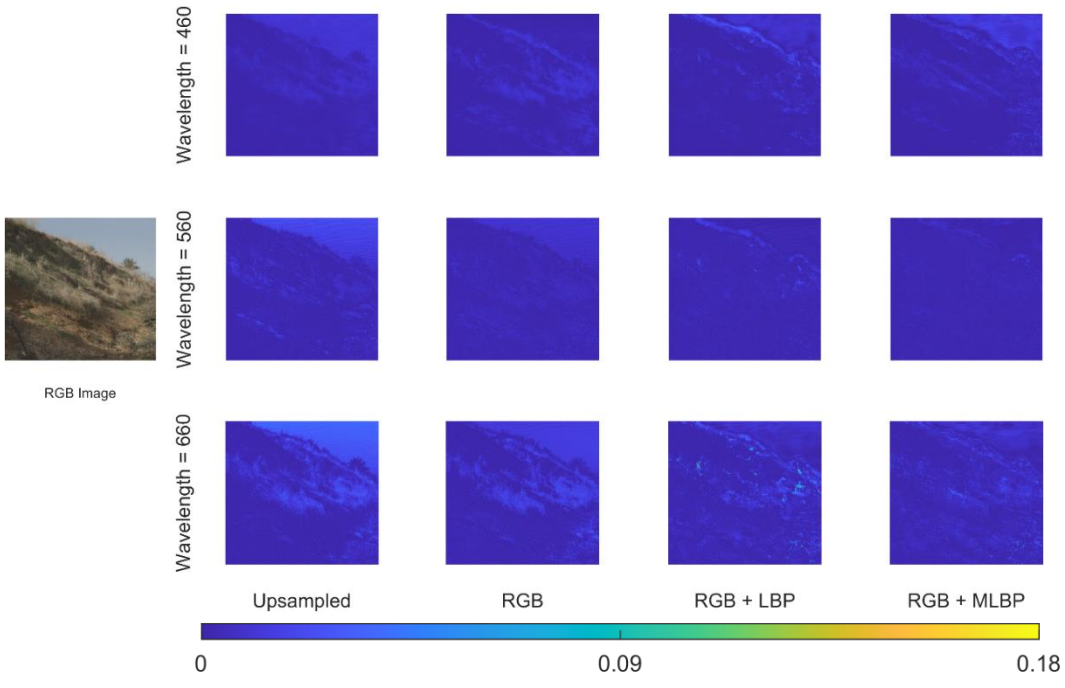
(a)



(b)



(c)



(d)

Figure 1. Reconstructed image from the left to right: without estimating the residual; estimating residual from RGB value; estimating residual from RGB value and single scale LBP context and estimating residual from RGB value and multi-scale LBP context.