https://doi.org/10.2352/issn.2169-2629.2020.28.41

Linear Spectral Estimate Refinement for Spectral Reconstruction from RGB

Tarek Stiebel and Dorit Merhof
RWTH Aachen University
tarek.stiebel@lfb.rwth-aachen.de

Abstract

Spectral signal recovery from RGB-images based on modern deep learning techniques demonstrated promising results in recent years and offers a feasible alternative to costly or otherwise more complex spectral imaging devices. The state-of-the-art deep learning is formed by approaches that learn a direct end-to-end mapping from RGB to spectral images from given RGB and spectral image pairs. Any prior knowledge, most importantly a known spectral responsivity of the imaging device, is not taken into account by the vast majority of deep learning based methods. Although attempts have been made to include prior knowledge with respect to the camera response functions, it remains unclear how to do so in a robust and constructive way.
In this work, we propose a hybrid processing method utilizing a handcrafted linear map to directly obtain a good estimate on the spectral signal. Deep learning is only used for a subsequent signal refinement. In contrast to previous work, our linear estimate on the spectral signal is not subject to any network optimization and relies on explicit knowledge on the camera response. It is finally demonstrated that the proposed hybrid processing strategy reduces spectral reconstruction errors.

Introduction

Spectral signal recovery from RGB images, also referred to as spectral super-resolution (SSR), has been subject to intense research efforts in recent years. For this purpose, convolutional neural networks (CNNs) were trained on corresponding image pairs of RGB- and spectral images to learn an end-to-end mapping from the RGB to the spectral domain. CNN based algorithms still form the state-of-the-art up until today. A first global benchmark was held in 2018 in form of the NTIRE challenge on spectral reconstruction from RGB [1] which may serve as a concise survey. The top performing methods were all based on established network structures such as DenseNets [2], ResNets [2, 3] or UNets [4, 5]. A new edition of the challenge was recently held as part of the NTIRE2020 workshop [6]. The complexity of the top performing methods increased with all top approaches implementing ensemble strategies. The best performing architecture was the adaptive weighted attention network proposed by Li [7]. While all the leading methods achieve great results, they mostly neglect prior knowledge, in particular on the camera response functions. Without explicitly incorporating the camera response functions, spectral predictions obtained through CNNs may actually be physically implausible, i.e. the reprojection into camera signal space does not match the input RGB signal. One approach that was partially considered is to incorporate the camera response functions within an additional term in the loss function [7, 8]. In such a case, the loss function for network training consists of both a spectral error metric as well as an RGB error metric obtained by reprojecting the recovered spectral image into RGB image space and comparison to the input RGB image. However, experiments have already shown that such a modified loss term has in the best case no impact at all on the reconstruction results. The more emphasis is placed on physically plausible spectra by a higher weighting of the respective loss subterm, the more will the spectral reconstruction metrics deteriorate [8]. It is instead common practice to solely rely on the network's capability to learn all necessary relationships on its own.
Recent work [8, 9] proposed potential solutions and demonstrated how a gain in performance can be achieved by explicitly separating the spectral domain into two subspaces with respect to a known camera response: particular solutions and metameric blacks. Modern deep learning approaches are subsequently applied to only predict the metameric blacks, effectively reducing the computational complexity while ensuring physically plausible reconstructions at the same time. The major downfall of such an approach is its high susceptibility to noise. Only applying signal quantization has already the potential to substantially harm the reconstruction accuracy [8].
So far, it is still unclear how to successfully utilize prior knowledge on the camera response functions for deep learning based methods in a robust and constructive way. However, we strongly believe that if explicit knowledge is available, it should be considered and not neglected. Within this work, we therefore propose a hybrid method consisting of classical signal processing (which is still common for multi-spectral imaging devices) in combination with modern deep learning. Based on an initial estimate on the spectral signal through a handcrafted algorithm, deep learning is only used for signal refinement. The approach is based on the common signal processing intuition that signal interpolation is more convenient when the underlying Laplacian energy is low. By incorporating a robust and accurate estimate of the spectral signal, we aim at effectively reducing the Laplacian energy of the signal which deep learning has to interpolate.
In summary, our proposed method consists of two processing steps:

- a direct estimation of the spectral stimulus based on known camera response functions and an appropriate spectral basis.
- a refinement of the initial estimate through modern deep learning.

The concept of further refining an initial estimate on spectral signals is not new, any ResNet might in fact be interpreted in such a way. Even explicit spectral estimate refinement approaches have been previously considered that are similar to ours, e.g. by Can [3] in form of two parallel processing paths for input RGB images: a single 7x7 convolutional layer to obtain an initial estimate on the spectral signal from the rgb input in conjunction with a CNN to learn the residual error of the initial prediction. The major difference of our approach is that the algorithm for obtaining the spectral estimate is handcrafted, explicitly depending on

the camera response and not subject to parameter optimization during network training.

Methods

Image formation for RGB imaging devices is modeled as

$$\vec{\varphi} = (\varphi_r, \varphi_g, \varphi_b)^T = \mathbf{R}\vec{s}, \tag{1}$$

where $\mathbf{R} \in \mathbb{R}^{3xn}$ denotes the spectral responsivity and $\vec{s} \in \mathbb{R}^n$ the incident spectral stimulus.

Linear Spectral Estimate

Spectral images are provided in a high-dimensional spectral space. However, it is well known that the intrinsic dimensionality of spectral data in natural scenes can be significantly lower [11]. In particular, for rather well behaved spectral curves without narrowband spikes caused for example by artificial illumination, as little as six dimensions can already be sufficient to hold the majority of information [12]. An intuitive way to reduce the dimensionality used to represent the spectral data is the principal component analysis (PCA) [13]. Given the orthonormal basis obtained through the application of the PCA on a set of representative spectral data, $\mathbf{T} = (\vec{t_1}, ..., \vec{t_n})^T \in \mathbb{R}^{nxn}$, where the individual basis vectors $\vec{t}$ are sorted in descending order according the signal variance they explain, the spectral subspace spanned by the major k principal components is denoted by $\mathbf{T_k} = (\vec{t_1}, ..., \vec{t_k})^T \in \mathbb{R}^{kxn}$ with $k \leq n$. Any spectral stimulus can then be approximately represented in the lower dimensional space by the associated principal components

$$\vec{s}_{pca} = \mathbf{T_k}\vec{s}. \tag{2}$$

The higher the dimensionality of the subspace k becomes, the better is the approximation. Combining Eq. 1 and Eq. 2 yields

$$\vec{\varphi} = \mathbf{R}\mathbf{T_k}^T\vec{s}_{pca}. \tag{3}$$

Of particular interest is the case where $k = 3$. If the matrix $\mathbf{S_{pca}} = \mathbf{R}\mathbf{T_3}^T \in \mathbb{R}^{3x3}$ is not singular, it contains a direct linear map from camera signals to spectral stimuli. However, spectral response functions are commonly chosen for the individual camera channels such that they capture distinct information, as well as to capture as much information on the spectral stimulus as possible. A matrix $\mathbf{S_{pca}}$ that is singular would correspond to response functions that are orthogonal to at least one of the three major principal components. This would correspond to a severe loss of captured information and directly contradict typical design decisions. In almost all practical applications, $\mathbf{S_{pca}}$ can therefore be assumed to be non-singular and thus invertible. As a result, we obtain a fast and direct way to approximate spectral stimuli from camera signals under the constraint that only the major three principal components are utilized to represent spectral stimuli,

$$\vec{s} = \mathbf{T_3}^T(\mathbf{R}\mathbf{T_3}^T)^{-1}\vec{\varphi}. \tag{4}$$

The direct linear map from 3-dimensional camera signal space to the 31-dimensional spectral signal space,

$$\mathbf{E} = \mathbf{T_3}^T(\mathbf{R}\mathbf{T_3}^T)^{-1}, \text{where } \mathbf{E} \in \mathbb{R}^{nx3}, \tag{5}$$

can be efficiently implemented as a convolutional layer.

Spectral Refinement

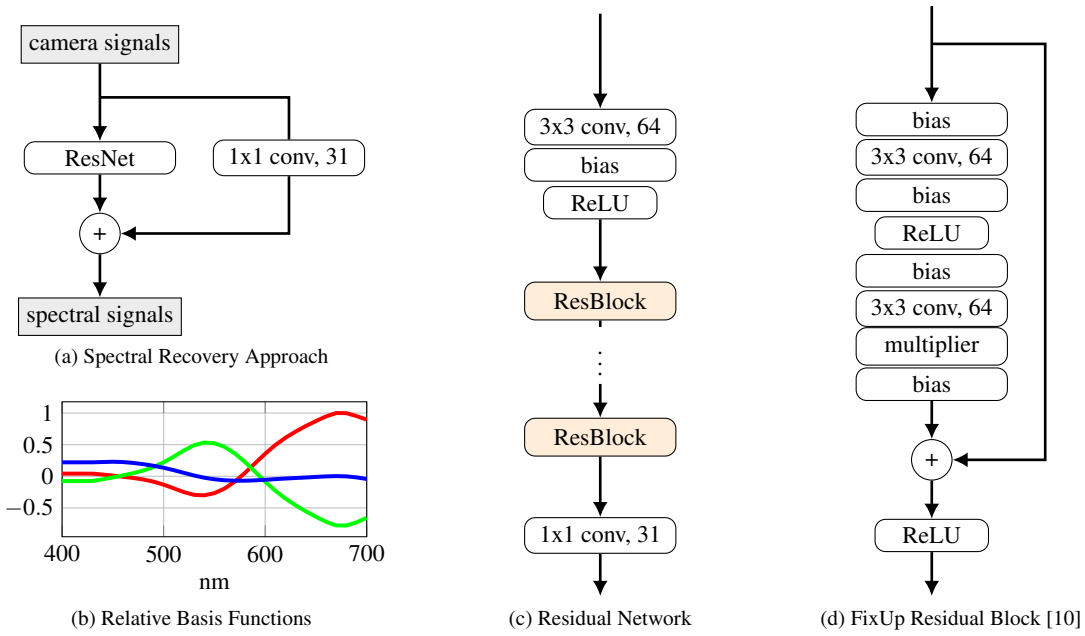
Based on the spectral estimate, a neural network is required for signal refinement. Since the neural network itself is not the primary focus of this work, we utilize a ResNet based architecture as it was one of top performing architectures at the NTIRE2018 challenge in form of HSCNN+ [2] while still being comparably simple. In particular, a ResNet-18 with fixed-update initialization [10] was implemented within this work and evaluated. The respective network architecture is shown in Fig. 1c. Input RGB-signals are upsampled to 64 features by an initial convolutional layer, following an additive bias and a rectified linear unit (ReLU). The resulting feature map is subsequently refined by eight FixUp residual blocks [10]. The precise structure of each residual block is visualized in Fig. 1d. One last convolutional layer maps the final 64 feature channels back to the spectral domain. All convolutional layers rely on replication-padding to hold the spatial image dimension constant. In total, the network has a receptive field of 34x34 pixels. The network's architecture is rather lightweight in comparison to the top performing methods of the NTIRE2020 challenge.
The network's output is finally added to the previous linear estimate on the spectral signal as it is shown in Fig. 1a, resulting in the complete processing approach for spectral signal recovery.

Training Details

The network for spectral refinement is trained in conjunction with the convolutional layer representing the direct spectral estimate as described in Eq. 4. However, the layer yielding the spectral estimate remains constant throughout the training process and is excluded from any update steps. All the training data is processed utilizing a batch size and a patch size of both 50. In conjunction with our patch extraction, a single epoch consists of 2754 iterations. Training begins with an initial learning rate of 0.0001 and Adam optimization [14]. The learning rate is automatically reduced by a factor of 10 if within 20 epochs no training progress can be observed on the validation data. Automatic termination of the training process occurs if the learning rate has been reduced three times.
Pytorch was used as deep learning framework. The training itself was run on an NVidia GTX 2080TI.

Experimental Results

The evaluation is conducted on the spectral dataset, as it was published during the NTIRE2020 [6] challenge on spectral reconstruction from RGB. This new dataset consists of 450 training images and, respectively, 10 validation and test images. Each spectral image contains a sampled version of the visible wavelength from 400nm to 700nm in 10nm steps, effectively yielding 31 channels. It should be noted that each spectral image is individually scaled such that its maximum signal value equals one. All spectral images therefore provide a radiometric quantity up to arbitrary scale. The spatial resolution per image is 512x482 pixels. All images were captured outdoors under natural illumination, mostly bright sunlight. The characteristics of all captured spectral stimuli can therefore be considered as well behaved.
A corresponding RGB image is computed for each spectral image by applying a camera response function according to Eq. 1. The challenge comprised two tracks. Within the 'clean' track, the response function is known and was chosen as the CIE 1964 human standard observer. All computed RGB images are finally subject to 8bit quantization. In contrast, the camera response is not known within the 'real world' track. Additionally, a real world camera processing pipeline is simulated, including demo-



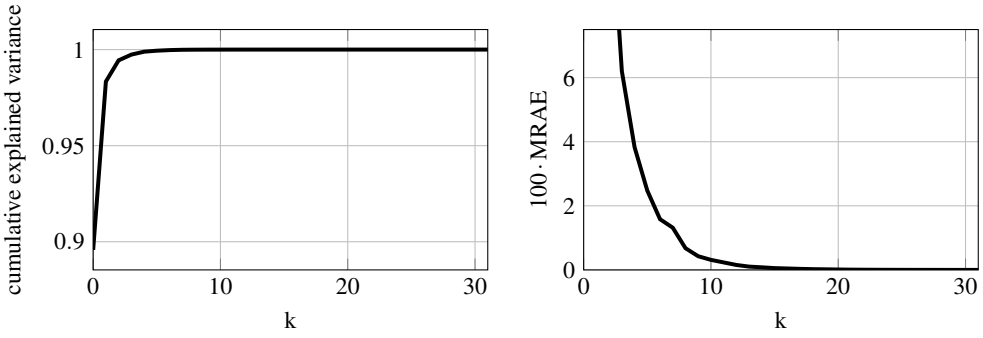
(a) Spectral Recovery Approach

(b) Relative Basis Functions

(c) Residual Network

(d) FixUp Residual Block [10]

Figure 1. *Visualization of the complete process for spectral signal recovery from RGB.*



(a) Explained variation in our training data.

(b) Resulting spectral error in our test data.

Figure 2. *Spectral dimensionality reduction of the NTIRE 2020 [6] spectral dataset through principal components.*

saicking and JPEG-compression. Since our method relies on explicit knowledge of the camera response, we only consider the 'clean' track. All spectral predictions are compared to the ground truth in terms of the mean relative absolute error (MRAE)

$$MRAE = \frac{1}{31xS} \sum_{i,q} \frac{|s^q_{gt}[i] - s^q_{rec}[i]|}{s^q_{gt}[i]}, \tag{6}$$

where $\vec{s^q}$ denotes the spectral stimulus at the q'th pixel within a spectral image and S the image size, i.e. total amount of pixels.
There are official data splits provided. However, the spectral images of the official test split are not disclosed to allow reporting unbiased results of future works. For each dataset, we therefore randomly split the official training images into two subsets, training and validation, and we consider the official validation images as our test set. This allows us to easily perform and report a variety of different evaluations. In summary, our training set comprises 440 images, while our validation and test set both comprise 10 images.
Finally, objective results on the official NTIRE2020 test images are additionally reported for the proposed hybrid spectral recovery approach, offering a fair comparison to other state-of-the-art methods.

Intrinsic Spectral Dimensionality

We will begin by inspecting the intrinsic spectral dimensionality of the dataset. A PCA was thus respectively performed based on the training images. The cumulative explained variance is plotted in Fig. 2. All variance within the spectral data can almost be expressed after only six components and little improvement is achieved by further increasing the component count. It is of interest to also evaluate the PCA in terms spectral error metrics. All spectral signals are thus projected onto the lower dimensional subspace spanned by the major k components. The resulting approximated spectral images are compared to the ground truth images in terms of the MRAE metric. Additionally, our test set is considered to remove potential bias. The average spectral error depending on the amount of principal components is plotted in Fig. 2b. It can be observed that in contrast to before, more components are required to decrease the spectral error. The approximation error completely vanishes when $k >= 15$, which roughly halves the dimensionality of the spectral space.
By including knowledge of the camera response functions, as well as considering the major three principal components, we can establish the direct linear map, $\mathbf{E}$, from camera signals to the spectral domain according to Eq. 5. Since the CIE human stan-

Table 1: average Laplacian energy of the ground truth spectra as well as the difference signals due to our linear estimate

Image Index	0451	0453	0455	0456	0457	0459	0462	0463	0464	0465	avg
Lapl. En. Ground truth	6.05	6.03	4.20	2.85	1.70	2.48	1.87	5.98	4.10	2.13	3.74
Lapl. En. Difference Signal	2.44	5.28	2.93	0.80	0.75	0.69	0.80	2.39	2.03	1.41	1.95

dard observer was considered within the 'clean' track, it was utilized in conjunction with the previously computed principal components to calculate the linear map **E**. The individual columns of the matrix can be seen as a spectral basis, that is only weighted by the measured camera signals. For illustration, the resulting basis functions are plotted in Fig. 1b in relative scale.

Laplacian Signal Energy

Spectral reconstruction from RGB can be viewed as interpolating a heavily subsampled spectral signal in form of RGB back to its original spectral resolution. The proposed hybrid approach for spectral signal recovery is built on the premise that signal interpolation becomes easier the lower the Laplacian energy of the underlying signal is. In the borderline case where the underlying signal takes on the form of a line - associated with a second order derivative of zero - signal interpolation is obviously trivial.
When considering SSR, the signal interpolation is performed by a CNN. By utilizing a proper estimation on the spectral signals, the Laplacian energy of the signal to be recovered by the CNN is effectively reduced. When training any CNN as a stand-alone network in an end-to-end fashion, it has to interpolate the original ground truth spectra. For the proposed workflow, it is only required to interpolate the difference signal of the ground truth to the linear estimate. As a result, signal interpolation becomes easier.
For clarification, we consider the signal energy of the spectral stimulus $\vec{s}$ as

$$E = \sum_{i=0}^{31} s[i]^2 = \vec{s}^T \vec{s}, \tag{7}$$

and, consequently, the signal energy of the second order derivative as

$$E_l = \vec{s}^T \mathbf{D} \vec{s} \tag{8}$$

with

$$\mathbf{D} = \begin{pmatrix} 1 & -1 & 0 & \dots & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \ddots & \ddots & 0 \\ 0 & -1 & 2 & -1 & 0 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & -1 & 2 & -1 & 0 \\ \vdots & \ddots & \ddots & 0 & -1 & 2 & -1 \\ 0 & \dots & \dots & \dots & 0 & -1 & 1 \end{pmatrix}. \tag{9}$$

Our hypothesis is evaluated by comparing the average energy of all the ground truth spectra, $E_l^{gt}(\vec{s}_{gt})$, to the energy of the remaining estimation error when utilizing our proposed linear estimation, $E_l^{dif}(\vec{s}_{gt} - \vec{s}_{est})$. The result is shown in Tab. 1. Indeed, the average Laplacian energy of the signal that the neural network has to reconstruct could be successfully reduced. This result is consistent for all images where on average an improvement of a factor of approximately two could be achieved.

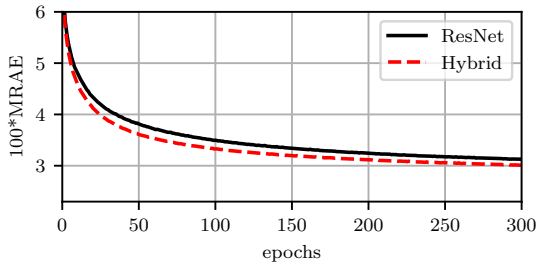


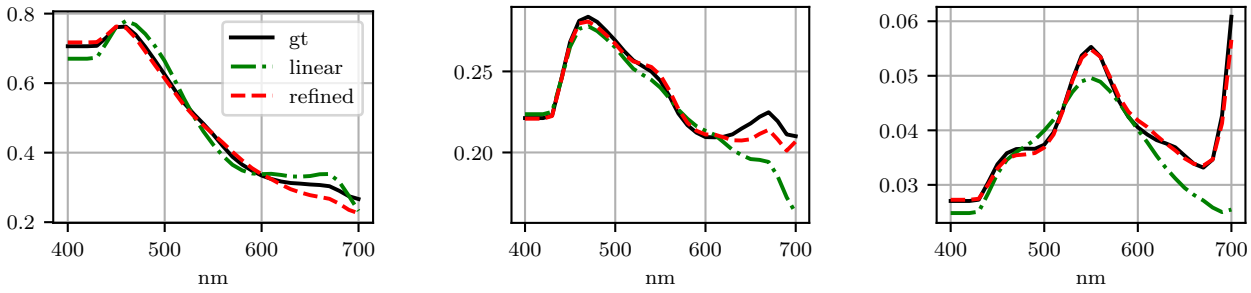
Figure 3. *Training loss for the stand-alone ResNet and our hybrid approach.*

Evaluation

The proposed hybrid approach consisting of a linear estimation in conjunction with deep spectral refinement is trained on the training data. In order to assess the influence of the linear estimate, we also train the utilized ResNet alone. Fig. 3 displays the training loss for both scenarios. The gain in performance is directly observable. When our stopping criteria have been reached, the hybrid approach achieves a reconstruction error of 4.02% in terms of MRAE on our test data while the ResNet converges at 4.15%.
We attribute the gain in performance to the high quality of our initial linear estimate. To strengthen this claim, we recovered all spectral images of our test set by only applying the linear estimation on its own. The detailed results are listed in Table 2. An average reconstruction error of 9.82% in terms MRAE was found. This already is a rather good approximation considering the computational simplicity of the approach. In comparison to high-end and very complex neural networks, it is only worse by factor of roughly 2.5. For a better understanding, Fig. 4 displays some qualitative reconstruction results for both the linear estimate as well as the refined version. The linear estimate generally approximates the ground truth spectrum well. Most of the differences arise at the lower- and higher-end wavelengths. This can directly be explained by the shape of the underlying spectral responsivity. In the wavelength bands where the response functions also show their peak responsibility, the reconstruction quality is already very accurate. Thus, the prediction task for the neural network is heavily simplified and it can focus on its capability to add more complex information in wavelength bands where the camera is unsensitive.
Finally, the hybrid method was used to recover the spectral images of the official test set of the NTIRE2020 data. In summary, an MRAE of 3.633% was achieved on average. At the moment, the best method reached an MRAE of 3.010%. However, all top methods featured more complex network architectures, some sort of ensemble strategies and non-trivial pixel awareness. For a better comparison, alternative architectures based on the HSCNN+ network respectively achieved reconstruction errors of 3.516% and 3.769% [6], which is comparable to our result. However, our proposed network consists of significantly fewer parameters. Due to the linear estimate, a network with less parameters is sufficient. In summary, the proposed hybrid approach does not

Table 2: average reconstruction quality of the linear estimation per image

Image Index	0451	0453	0455	0456	0457	0459	0462	0463	0464	0465	mean
MRAE in percent	6.04	29.27	14.45	2.98	8.51	2.59	8.42	6.78	5.28	13.90	9.82



(a) (b) (c)

Figure 4. *Achieved spectral reconstructions by only the initial linear estimate (linear) as well as in conjunction with parallel refinement (refined). Figure 5 displays the corresponding pixel positions.*

Figure 5. *Exemplary image of the validation set rendered in sRGB assuming CIE D50 illumination.*

achieve top ranking results, but the reconstruction quality is still competitive while offering computational simplicity. This can for example proven by considering the runtime per image. A single image consists of roughly 0.25 mega-pixel. All the top performing methods require at least 0.5s to process such an image, most methods require more than 1s and the slowest algorithm takes 30s. In contrast, the proposed method requires an average processing time of 0.3s per image. In fact, the proposed method offers the best reconstruction results for all methods that perform the spectral recovery in less than 0.5s per image [6].
For future research, it might interesting to act contrary to the current trend of making neural networks more complex and as a result more powerful, but instead reduce the network's architecture as much as possible while maintaining the general performance. Such a lightweight and thus fast method would be crucial for mass market adoption, e.g. smartphones.

Conclusion

Within this work, we proposed to merge traditional, non-deep learning based signal processing with a state-of-the-art neural network for the task of spectral signal recovery from RGB images. It was shown that our hybrid approach outperforms the stand-alone neural network. As a prerequisite, explicit knowledge on the camera response functions is essential. It was shown how a linear map from camera signals to the spectral domain can be constructed from the known camera response in combination with an appropriate spectral basis. The spectral basis was obtained by employing a principal component analysis. It was found that the linear map already provides an acceptable approximation to the ground truth spectra. A refinement of the initial estimate was subsequently performed by a FixUp-ResNet, resulting in good reconstruction results proven on the NTIRE2020 [6] data while offering a short processing time in comparison to other state-of-the-art methods. This is due to the computational simplicity which could be achieved with the help of the proposed linear estimate.

References

[1] B. Arad *et al.*, "Ntire 2018 challenge on spectral reconstruction from rgb images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[2] Z. Shi, C. Chen, Z. Xiong, and D. Liu, "Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images," 06 2018, pp. 1052–10 528.

[3] Y. B. Can and R. Timofte, "An efficient CNN for spectral reconstruction from RGB images," 2018.

[4] T. Stiebel, S. Koppers, P. Seltsam, and D. Merhof, "Reconstructing spectral images from rgb-images using a convolutional neural network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[5] A. Alvarez-Gila, J. Weijer, and E. Garrote, "Adversarial networks for spatial context-aware spectral image reconstruction from rgb," 09 2017.

[6] B. Arad *et al.*, "Ntire 2020 challenge on spectral reconstruction from rgb images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[7] J. Li, C. Wu, R. Song, Y. Li, and F. Liu, "Adaptive weighted attention network with camera spectral sensitivity prior for spectral reconstruction from rgb image," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[8] T. Stiebel, P. Seltsam, and D. Merhof, "Enhancing deep

spectral super-resolution from rgb images by enforcing the metameric constraint," 01 2020, pp. 57–66.

[9] Y.-T. Lin and G. D. Finlayson, "Physically plausible spectral reconstruction from rgb images," 2020.

[10] H. Zhang, Y. N. Dauphin, and T. Ma, "Residual learning without normalization via better initialization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=H1gsz30cKX

[11] F. J. Sanchez-Marin, "Principal wavelengths in the formation of spectral images of natural scenes," *Journal of Biomedical Optics*, vol. 18, no. 4, pp. 1 – 9, 2013.

[12] H. Fairman and M. Brill, "The principal components of reflectances," *Color Research & Application*, vol. 29, pp. 104 – 110, 04 2004.

[13] I. Jolliffe, *Principal Component Analysis*. Springer, 2002.

[14] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.