Multispectral illumination estimation using deep unrolling network

Yuqi Li Qiang Fu Wolfgang Heidrich
King Abdullah University of Science & Technology, Saudi Arabia
{yuqi.li, qiang.fu, wolfgang.heidrich}@kaust.edu.sa

Abstract

This paper examines the problem of illumination spectra estimation in multispectral images. We cast the problem into a constrained matrix factorization problem and present a method for both single-global and multiple illumination estimation in which a deep unrolling network is constructed from the alternating direction method of multipliers(ADMM) optimization for solving the matrix factorization problem. To alleviate the lack of multispectral training data, we build a large multispectral reflectance image dataset for generating synthesized data and use them for training and evaluating our model. The results of simulations and real experiments demonstrate that the proposed method is able to outperform state-of-the-art spectral illumination estimation methods, and that it generalizes well to a wide variety of scenes and spectra.

1. Introduction

As an intrinsic physical property of materials, spectral reflectance is a rich information source for a wide range of vision tasks, including object recognition and material reproduction, as well as man technical and scientific imaging problems. However, the acquisition of accurate spectral reflectance images requires an extra per-image calibration to compensate for the illumination conditions in the scene, for example with a known reference [21] or a dedicated measurement device [2]. Unfortunately this calibration is generally cumbersome and frequently fails in complex lighting situations with multiple, different illumination sources.

A practical solution is estimating the illumination spectra from captured multispectral images and separating the reflectance and illumination spectral for further purposes. This illumination estimation problem is highly underdetermined; thus, regularizations are required to constrain the solution to satisfy the image priors of multispectral reflectances and illumination. Recently, many approaches have been studied to estimate illumination spectra. Most of them either utilize the statistics of multispectral images or extract the specular reflection component to estimate illumination spectra, but don't sufficiently consider the image priors of multispectral reflectance and illumination. There are also some attempts applying convolutional neural networks (CNN) to estimate illumination spectra. However, in our opinion, such methods are not flexible to adapt to various illumination spectra, and they are limited in the lack of large image datasets. Moreover, to the best of our knowledge, none of the existing methods can handle multiple illuminants in multispectral images. These limitations motivate us to build a large image dataset of multispectral reflectance and put forward a flexible illumination estimation method based on deep neural network (DNN).

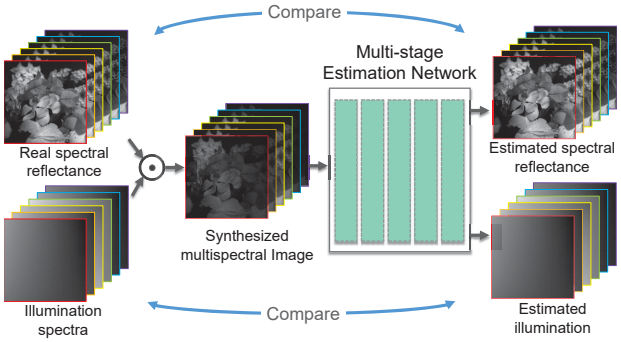


Figure 1. The framework of our spectral reflectance and illumination estimation method.

Our method leverages a multispectral image dataset synthesized by using real spectral reflectance images and illumination spectra to train the network, as shown in Fig. 1. We cast the illumination estimation problem as a constrained matrix factorization problem, and propose an optimization-inspired unrolling multi-stage network to solve the problem. Our network can achieve more accurate estimation than previous methods by training on the synthesized dataset and utilizing the non-local prior of reflectance images, and the low-rank and total-variation (TV) prior of illumination spectra. The contributions of this paper can be summarized as follows:

- We present a *new take* on constrained matrix factorization using a multi-stage loop-unrolled neural net-

work to solve the multispectral illumination estimation problem. Our data-driven method explicitly learns the non-local prior from the real captured reflectance images and improves the estimation accuracy with a low-rank and a TV regularization on illumination spectra. The proposed method significantly outperforms existing approaches.

- We present the *first work* to handle both single-global illumination and multiple illumination estimation for multispectral images. Both simulations on synthesized multispectral images and experiments on real images show the high flexibility, effectiveness, and generalization ability of our method.
- We build *the largest existing spectral reflectance image dataset* consisting of 400 high-quality multispectral images, providing a general-purpose benchmark for the illumination estimation model's training and evaluation.

2. Related Works

Recently, many efforts have been made to study color constancy of a trichromatic RGB image [6, 19, 30, 26, 13, 1, 31] and illumination spectral estimation [17, 18, 16, 15, 28, 35, 24]. As for the task of illumination spectral estimation, most of the previous methods are inspired by color constancy methods, and can be divided into statistics-based and learning-based approaches.

Statistics-based methods A number of proposed methods are based on statistical properties of captured scenes, and estimate illuminantion spectra by exploiting the priors from such assumptions. In the color constancy problem, some classical methods assume that illuminantion color can be directly obtained from the average response of the whole captured images (Gray-World) [6], from the pixel which has the maximum spectral response (Max-RGB) [19], or from the edge pixels in the captured images (GrayEdge) [30]. These assumptions also hold and perform well in spectral illumination estimation problem [17, 18]. With these assumptions, the intensity of each wavelength channel of the spectral power distribution (SPD) function is estimated separately. The limitation of such methods is they usually require high spectral diversity in the scene, and tend to fail in scenes with less diversity. Some others assume that the specular reflection in the captured images can provide information of scene illumination [16, 28], but they rely on non-Lambertian surfaces for extracting the specular component. In addition, the low-rank property of spectral reflectance is also exploited to constrain the separation of illumination and reflectance spectra [35]. However, exclusively relying on a single assumption is limited and insufficient for accurate spectral illumination estimation in general scenes. Moreover, most of these methods are applied by using fixed set of hyperparameters, and the estimation accuracy is sensitive to the choice of hyperparameters and the specific imaging scenes.

Learning-based methods Illumination color is estimated via deep learning in the great majority of recent color constancy methods [26, 13, 1, 31]. Taking advantage of large-scale image data with ground truth illumination color, these methods can generally achieve higher accuracy with respect to statistics-based methods.

To extend the use of learning-based model to multispectral domain, a CNN-based method [24] exploits multi-scale retinex model as the front-end network and estimate the illumination spectrum using a CNN. However, the main difficulty in applying the above learning-based methods to the multispectral domain is the lack of large-scale image data sets of multispectral images with ground truth illumination spectra. Moreover, since the dimensionality of spectral illumination is much higher than that of 3-channel color illumination, previous CNN-based methods tend to have difficulties generalizing to illumination spectra that differ from the training set. Therefore, building large-scale data sets and designing more flexible network architectures are the two critical challenges in estimating illumination spectra using deep learning. Recently, optimization-based loop-unrolled network have been used with excellent results on various sensing matrix in image restoration tasks [33, 20]. In this paper, considering the high dimensionality of illumination spectra, we will adopt a unrolling network architecture to deal with the illumination estimation problem.

Most of the above methods assume uniform illuminations across the scene and only focus on estimating a single-global illumination. Multiple illumination estimation can handle spatially varying illumination in more general scenes.

Multiple illumination estimation A straightforward strategy to handle multiple illuminations is to cluster pixels into superpixel regions according to the possible illumination and the associated reflectance [10], and then either apply single-illuminantion estimation methods [9] or seek for references in the dataset [14] to estimate pixel-wise illumination locally. Other clustering-based methods estimate local illumination first and exploit conditional random field [4] and factor graph [23] to optimize their global distribution. The success of these methods demonstrates that clustering can help to improve multi-illumination estimation. In addition, some methods [12, 7] impose constraints on the number of lights that illuminate the scene or on the smoothness of the varying illuminations across the scene, which can boost the performance of illumination and reflectance separation since the reflectance usually has a much higher diversity than the illumination in both the spatial and the color domain. Deep learning-based methods [5, 27] for

the problem are also proposed recently. Our method is inspired by the thought of clustering and the used constraints in the above multiple illumination estimation methods to deal with multiple illuminations. We will start with the problem formulation to introduce our multispectral illumination estimation model.

3. Methodology

3.1. Problem formulation

Let us consider a multispectral image $\mathbf{I}$ taken from a multispectral camera with a resolution of $m \times n \times c$, where $m \times n$ denote the spatial resolution of the image, and c denotes the number of spectral channels in the image. Our method represents the multispectral image as a two-dimensional $mn \times c$ matrix and aims to decompose it into a per-pixel illumination spectrum $\mathbf{L}$ and a per-pixel spectral reflectance $\mathbf{R}$, both of which are also represented as matrices of dimension $mn \times c$. That is, $\mathbf{I} = \mathbf{L} \odot \mathbf{R}$, where $\odot$ represents the Hadamard product.

It is worth noting that our goal is to handle both single-global illumination and multiple illumination estimation. In the case of single illumination estimation, the illumination spectra vector $\mathbf{L}$ is treated as a rank-one matrix that all pixels share the same spectrum. While in the case of multiple illumination estimation, we need to estimate pixel-wise illumination spectra for the image.

Our method regards the illumination estimation as a constrained matrix factorization problem:

$$\begin{aligned} \min_{\mathbf{L},\mathbf{R}} \tfrac{1}{2}||\mathbf{I} - \mathbf{L} \odot \mathbf{R}||_F^2 + \eta_r||\mathbf{L}||_* + \eta_t||\mathbf{L}||_{TV} + Q(\mathbf{R}), \\ s.t.\ \mathbf{L} \geq \mathbf{0}, \mathbf{0} \leq \mathbf{R} \leq \mathbf{1}, \end{aligned} \tag{1}$$

where $Q(\mathbf{R})$ denotes a non-local prior of reflectance $\mathbf{R}$, $||\mathbf{L}||_*$ denotes the nuclear norm of $\mathbf{L}$ as a low-rank denoising regularizer of illumination in the spectral dimension, $||\mathbf{L}||_{TV}$ denotes the total variation (TV) regularizer of illumination in the spatial dimension which is defined as the integral of the absolute gradient, η_r and η_t denote the weighted parameters. The use of a non-local regularizer for reflectance takes into account non-local data redundancy and is motivated by existing RGB illumination estimation methods that cluster image regions for more robust and accurate estimation as mentioned in Sec. 2. In the image processing field, the total variation regularizer penalizes the spurious detail while preserving the edges in images [25]; the low-rank regularizer aims at reconstructing images with only a few basis[29]. The low rank regularization for illumination is added because scenes are usually illuminated by only a few distinct illumination sources thus the illumination matrix can be presented as a low dimensional linear combination of spectra; and the TV regularization is for estimating smoothly varying illumination in spatial dimension. Note the estimated illumination $\mathbf{L}$ is constrained to be non-negative, and the estimated reflectance $\mathbf{R}$ is constrained to be within the range $[0, 1]$ for physical plausibility.

3.2. Optimization

The ADMM algorithm is adapted to efficiently solve the above optimization problem. Our goal is to unroll the ADMM to a multistage network. By introducing auxiliary variables $\mathbf{M}$, $\mathbf{L}_r$, $\mathbf{L}_t$, and $\mathbf{R}_q$, Eq. (1) is equivalent to:

$$\begin{aligned} \min_{\mathbf{L},\mathbf{R}} \tfrac{1}{2}||\mathbf{I} - \mathbf{M}||_F^2 + Q(\mathbf{R}_q) + \eta_r||\mathbf{L}_r||_* + \eta_t||\mathbf{L}_t||_{TV}, \\ s.t.\ \mathbf{M} = \mathbf{L} \odot \mathbf{R},\ \mathbf{L}_t, \mathbf{L}_r = \mathbf{L},\ \mathbf{R}_q = \mathbf{R}, \\ \mathbf{L}_r, \mathbf{L}_t \geq \mathbf{0}, \mathbf{0} \leq \mathbf{R}_q \leq \mathbf{1}. \end{aligned} \tag{2}$$

By converting the constrained optimization problem into an unconstrained one, the augmented Lagrangian for the above problem is given by:

$$\begin{aligned} &\mathcal{L}_{\alpha_0,\alpha_1,\alpha_2,\alpha_3}(\mathbf{L}, \mathbf{R}, \mathbf{M}, \mathbf{L}_r, \mathbf{L}_t, \mathbf{R}_q;\ \mathbf{U}_\mathbf{M}, \mathbf{U}_{\mathbf{L}_r}, \mathbf{U}_{\mathbf{L}_t}, \mathbf{U}_{\mathbf{R}_q}) \\ &= \tfrac{1}{2}||\mathbf{I}-\mathbf{M}||_F^2 + \langle \mathbf{U}_\mathbf{M}, \mathbf{M}-\mathbf{L} \odot \mathbf{R} \rangle + \tfrac{\alpha_0}{2}||\mathbf{M}-\mathbf{L}\odot\mathbf{R}||_F^2 + Q(\mathbf{R}_q) \\ &\quad + \langle \mathbf{U}_{\mathbf{R}_q}, \mathbf{R}_q - \mathbf{R} \rangle + \tfrac{\alpha_1}{2}||\mathbf{R}_q - \mathbf{R}||_F^2 + \eta_r||\mathbf{L}_r||_* + \eta_t||\mathbf{L}_t||_{TV} \\ &\quad + \langle \mathbf{U}_{\mathbf{L}_r}, \mathbf{L}_r - \mathbf{L} \rangle + \tfrac{\alpha_2}{2}||\mathbf{L}_r - \mathbf{L}||_F^2 + \langle \mathbf{U}_{\mathbf{L}_t}, \mathbf{L}_t - \mathbf{L} \rangle + \tfrac{\alpha_3}{2}||\mathbf{L}_t - \mathbf{L}||_F^2, \end{aligned} \tag{3}$$

where $\mathbf{U}_\mathbf{M}$, $\mathbf{U}_{\mathbf{L}_r}$, $\mathbf{U}_{\mathbf{L}_t}$, and $\mathbf{U}_{\mathbf{R}_q}$ are Lagrangian multipliers representing dual variable, α_0, α_1, α_2, and α_3 are weighted parameters.

To minimize Eq.(3) with respect to these variables, ADMM divides the problem of Eq.(3) into subproblems for each variable and alternatively optimizes the variables iteratively. The updates of each variable in the k-th iteration are given by Eq.(4) and Eq.(6).

$$\begin{cases} \mathbf{R}^{(k+1)} \leftarrow \frac{\mathbf{L}^{(k)} \odot (\alpha_0 \mathbf{M}^{(k)} + \mathbf{U}_\mathbf{M}^{(k)}) + \alpha_1 \mathbf{R}_q^{(k)} + \mathbf{U}_{\mathbf{R}_q}^{(k)}}{\alpha_0 \mathbf{L}^{(k)} \odot \mathbf{L}^{(k)} + \alpha_1}, \\ \mathbf{R}_q^{(k+1)} \leftarrow \arg\min_{\mathbf{0} \leq \mathbf{R}_q \leq \mathbf{1}} \frac{\alpha_1}{2}||\mathbf{R}_q - \mathbf{R}^{(k+1)} + \frac{\mathbf{U}_{\mathbf{R}_q}^{(k)}}{\alpha_1}||_F^2 + Q(\mathbf{R}_q), \\ \mathbf{M}^{(k+1)} \leftarrow \frac{\mathbf{I} + \alpha_0 \mathbf{L}^{(k)} \odot \mathbf{R}^{(k+1)} - \mathbf{U}_\mathbf{M}^{(k)}}{\alpha_0 + 1}, \\ \mathbf{U}_{\mathbf{R}_q}^{(k+1)} \leftarrow \mathbf{U}_{\mathbf{R}_q}^{(k)} + \alpha_1 \left(\mathbf{R}^{(k+1)} - \mathbf{R}_q^{(k+1)}\right). \end{cases} \tag{4}$$

As can be seen, variables $\mathbf{R}^{(k+1)}$, $\mathbf{M}^{(k+1)}$, and $\mathbf{U}_{\mathbf{R}_q}^{(k+1)}$ have closed-form update rules. Note that there is an inequality constraint $\mathbf{0} \leq \mathbf{R}_q \leq \mathbf{1}$ exists in the update step of variable $\mathbf{R}_q$, an auxiliary variable $\mathbf{R}_{q+}$, a dual variable $\mathbf{U}_{\mathbf{R}_{q+}}$, and the corresponding weighted parameter α_4 are introduced for the nonnegative constraint. The variables are updated as follows:

$$\begin{cases} \mathbf{R}'^{(k+1)}_q \leftarrow \frac{\alpha_1 \mathbf{R}^{(k+1)} + \alpha_4 \mathbf{R}_{q+}^{(k)} + \mathbf{U}_{\mathbf{R}_{q+}}^{(k)} - \mathbf{U}_{\mathbf{R}_q}^{(k)}}{\alpha_1 + \alpha_4}, \\ \mathbf{R}_q^{(k+1)} \leftarrow \underset{\mathbf{R}_q}{\operatorname{argmin}} ||\mathbf{R}_q - \mathbf{R}'^{(k+1)}_q||_F^2 + \frac{2}{\alpha_1+\alpha_4} Q(\mathbf{R}_q), \\ \mathbf{R}_{q+}^{(k+1)} \leftarrow clip\left(\mathbf{R}_q^{(k+1)} - \mathbf{U}_{\mathbf{R}_{q+}}^{(k)}/\alpha_4, [\mathbf{0}, \mathbf{1}]\right), \\ \mathbf{U}_{\mathbf{R}_{q+}}^{(k+1)} \leftarrow \mathbf{U}_{\mathbf{R}_{q+}}^{(k)} + \alpha_4 (\mathbf{R}_q^{(k+1)} - \mathbf{R}_{q+}^{(k+1)}). \end{cases} \tag{5}$$

Instead of explicitly giving the regularization model $Q(\cdot)$ and the proximal operator to optimize $\mathbf{R}_q$, we directly learn

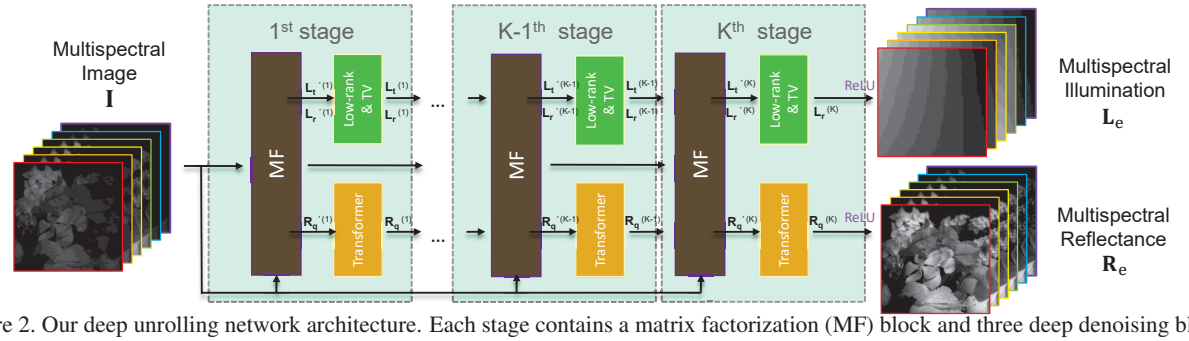


Figure 2. Our deep unrolling network architecture. Each stage contains a matrix factorization (MF) block and three deep denoising blocks (Low-Rank, Total Variation, and multi-head transformer).

a solver for the proximal operator with a self-attention neural network block. In this manner, the spatial-spectral reflectance image prior is not explicitly modeled but learned from the network block. The network block will be introduced in the Sec.3.3.

The update rules for the remaining variables are:

$$\begin{cases} \mathbf{L}^{(k+1)} \leftarrow \frac{(\alpha_0 \mathbf{M}^{(k)} + \mathbf{U}_{\mathbf{M}}^{(k)}) \odot \mathbf{R}^{(k+1)} + \alpha_2 \mathbf{L}_r^{(k)} + \alpha_3 \mathbf{L}_t^{(k)} + \mathbf{U}_{\mathbf{L}_r}^{(k)} + \mathbf{U}_{\mathbf{L}_t}^{(k)}}{\alpha_0 \mathbf{R}^{(k+1)} \odot \mathbf{R}^{(k+1)} + \alpha_2 + \alpha_3}, \\ \mathbf{L}_r^{(k+1)} \leftarrow \arg\min_{\mathbf{L}_r \geq \mathbf{0}} \frac{\alpha_2}{2} ||\mathbf{L}_r - \mathbf{L}^{(k+1)} + \frac{\mathbf{U}_{\mathbf{L}_r}^{(k)}}{\alpha_2}||_F^2 + \eta_r ||\mathbf{L}_r||_*, \\ \mathbf{L}_t^{(k+1)} \leftarrow \arg\min_{\mathbf{L}_t \geq \mathbf{0}} \frac{\alpha_2}{2} ||\mathbf{L}_t - \mathbf{L}^{(k+1)} + \frac{\mathbf{U}_{\mathbf{L}_t}^{(k)}}{\alpha_2}||_F^2 + \eta_t ||\mathbf{L}_t||_{TV}, \\ \mathbf{U}_{\mathbf{L}_r}^{(k+1)} \leftarrow \mathbf{U}_{\mathbf{L}_r}^{(k)} + \alpha_2 \ (\mathbf{L}^{(k+1)} - \mathbf{L}_r^{(k+1)}), \\ \mathbf{U}_{\mathbf{L}_t}^{(k+1)} \leftarrow \mathbf{U}_{\mathbf{L}_t}^{(k)} + \alpha_3 \ (\mathbf{L}^{(k+1)} - \mathbf{L}_t^{(k+1)}), \\ \mathbf{U}_{\mathbf{M}}^{(k+1)} \leftarrow \mathbf{U}_{\mathbf{M}}^{(k)} + \alpha_0 \ (\mathbf{L}^{(k+1)} \odot \mathbf{R}^{(k+1)} - \mathbf{M}^{(k+1)}), \end{cases} \tag{6}$$

Again, variables $\mathbf{L}^{(k+1)}$, $\mathbf{U}_{\mathbf{L}_r}^{(k+1)}$, $\mathbf{U}_{\mathbf{L}_t}^{(k+1)}$, and $\mathbf{U}_{\mathbf{M}}^{(k+1)}$ have closed-form solutions. Note that both the objective functions of the $\mathbf{L}_r^{(k+1)}$ and $\mathbf{L}_t^{(k+1)}$ constrain them to be nonnegative, thus other two sets of auxiliary variable $\mathbf{L}_{r^+}$, $\mathbf{L}_{t^+}$, dual variable $\mathbf{U}_{\mathbf{L}_{r+}}$, $\mathbf{U}_{\mathbf{L}_{t+}}$, and the corresponding weighted parameter α_5, α_6 are introduced to deal with the nonnegative constraint.

$$\begin{cases} \mathbf{L}'^{(k+1)}_r \leftarrow \frac{\alpha_2 \mathbf{L}^{(k+1)} + \alpha_5 \mathbf{L}_{r+}^{(k)} + \mathbf{U}_{\mathbf{L}_{r+}}^{(k)} - \mathbf{U}_{\mathbf{L}_r}^{(k)}}{\alpha_2 + \alpha_5}, \\ \mathbf{L}_r^{(k+1)} \leftarrow \underset{\mathbf{L}_r}{\text{argmin}} ||\mathbf{L}_r - \mathbf{L}'^{(k+1)}_r||_F^2 + \frac{2\eta_r}{\alpha_2 + \alpha_5} ||\mathbf{L}_r||_*, \\ \mathbf{L}'^{(k+1)}_t \leftarrow \frac{\alpha_2 \mathbf{L}^{(k+1)} + \alpha_6 \mathbf{L}_{t+}^{(k)} + \mathbf{U}_{\mathbf{L}_{t+}}^{(k)} - \mathbf{U}_{\mathbf{L}_t}^{(k)}}{\alpha_2 + \alpha_6}, \\ \mathbf{L}_t^{(k+1)} \leftarrow \underset{\mathbf{L}_t}{\text{argmin}} ||\mathbf{L}_t - \mathbf{L}'^{(k+1)}_t||_F^2 + \frac{2\eta_t}{\alpha_2 + \alpha_6} ||\mathbf{L}_t||_{TV}, \\ \mathbf{L}_{r+}^{(k+1)} \leftarrow max\left(\mathbf{L}_r^{(k+1)} - \mathbf{U}_{\mathbf{L}_{r+}}^{(k)}/\alpha_5, \ \mathbf{0}\right), \\ \mathbf{L}_{t+}^{(k+1)} \leftarrow max\left(\mathbf{L}_t^{(k+1)} - \mathbf{U}_{\mathbf{L}_{t+}}^{(k)}/\alpha_6, \ \mathbf{0}\right), \\ \mathbf{U}_{\mathbf{L}_{r+}}^{(k+1)} \leftarrow \mathbf{U}_{\mathbf{L}_{r+}}^{(k)} + \alpha_5 \ (\mathbf{L}_r^{(k+1)} - \mathbf{L}_{r+}^{(k+1)}), \\ \mathbf{U}_{\mathbf{L}_{t+}}^{(k+1)} \leftarrow \mathbf{U}_{\mathbf{L}_{t+}}^{(k)} + \alpha_6 \ (\mathbf{L}_t^{(k+1)} - \mathbf{L}_{t+}^{(k+1)}). \end{cases} \tag{7}$$

To optimize the associated components $\mathbf{L}_r^{(k+1)}$ and $\mathbf{L}_t^{(k+1)}$, we simply apply the soft thresholding operation $\mathcal{T}(\mathbf{x}, \tau) = max(\mathbf{x} - \tau, \mathbf{0}) \frac{\mathbf{x}}{|\mathbf{x}|}$ to each update step respectively. A soft thresholding function is applied to optimize variable $\mathbf{L}_r$:

$$\mathbf{L}_r{}^{(k+1)} \leftarrow \sum_{i=1}^{c} \mathcal{T}(\Sigma_i, \tau_{rank}^{(k)}) \mathcal{U}_i \mathcal{V}_i^*, \tag{8}$$

where $\mathcal{U} \Sigma \mathcal{V}^*$ is the SVD factorization of matrix $\mathbf{L}'^{(k+1)}_r$, and $\tau_{rank}^{(k)}$ denotes the threshold parameter. Similarly, the $\mathbf{L}_t$ is updated as:

$$\begin{aligned} \mathbf{L}_t{}^{(k+1)} = & (1 - \gamma^{(k)}) \mathbf{L}_t^{(k)} + \gamma^{(k)} \mathbf{L}'^{(k+1)}_t \\ & + \rho^{(k)} \mathbf{D}^T \left(\mathcal{T}(\mathbf{D}\mathbf{L}_t^{(k)}, \tau_{tv}^{(k)}) - \mathbf{D}\mathbf{L}_t^{(k)} \right), \end{aligned} \tag{9}$$

where $\mathbf{D}$ is the matrix to calculate the image gradient, $\rho^{(k)}$ denotes a weighting parameter, and $\tau_{tv}^{(k)}$ denotes the threshold parameter.

3.3. Unrolled network

Based on the mathematical formulation of the estimation procedure of the problem, we propose unrolling the optimization to construct a K-stages neural network, as shown in Fig. 2. The network is trained end-to-end to obey the basic multiplicative model and exploit the proposed priors simultaneously. Each network stage, consisting of one linear matrix factorization block and three denoising blocks, represents one iteration in the optimization. Compared with using the parameters with fixed values, relaxing the parameters can accelerate the convergence of the optimization thus can reduce the amount of iterations. Besides, the optimal parameters can also help understanding the real contribution of the blocks in each iteration. We will show the benefit brought by optimized parameters in the ablation study.

In the matrix factorization block, the variables ($\mathbf{L}$, $\mathbf{L}_t$, $\mathbf{L}_r$, $\mathbf{L}_{t+}$, $\mathbf{L}_{r+}$, $\mathbf{R}$, $\mathbf{R}_q$, $\mathbf{R}_{q+}$, $\mathbf{M}$, $\mathbf{U}_{\mathbf{L}_t}$, $\mathbf{U}_{\mathbf{L}_r}$, $\mathbf{U}_{\mathbf{L}_{t+}}$, $\mathbf{U}_{\mathbf{L}_{r+}}$, $\mathbf{U}_{\mathbf{R}_q}$, $\mathbf{U}_{\mathbf{R}_{q+}}$, $\mathbf{U}_{\mathbf{M}}$) are updated with seven trainable parameters α_i ($i = 0, \cdots, 6$), according to Eq.(4)(5)(6)(7). The illumination variable $\mathbf{L}_r$ and $\mathbf{L}_t$ are updated through the low-rank and TV denoising blocks respectively, with trainable parameter τ_{rank}, τ_{tv}, and ρ according with Eq. (8) and Eq. (9). Therefore, each stages contains ten trainable weight or threshold parameters.

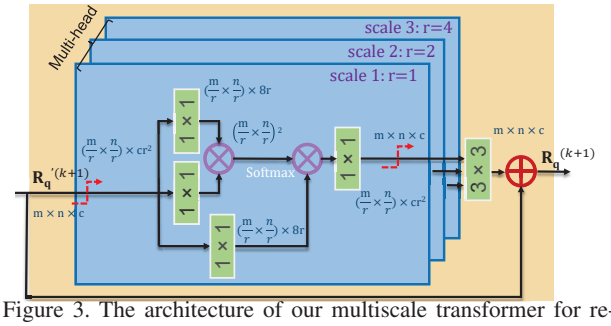


Figure 3. The architecture of our multiscale transformer for reflectance image denoising. Red dashed line denotes patches extraction and tensor reshaping. $m \times n \times c$ means c channels with height m and width n, $\left(\frac{m}{r} \times \frac{n}{r}\right) \times cr^2$ means $\left(\frac{m}{r} \times \frac{n}{r}\right)$ extracted patches with c channels and the spatial resolution $r \times r$. $\otimes$ denotes matrix multiplication. $\oplus$ denotes element-wise addition.

The reflectance variable $\mathbf{R}_q$ is updated through the transformer-based denoising block as shown in Fig. 3. Transformer is a widely used architecture which adopts nonlocal self-attention mechanism for various vision tasks [11], and the superiority of the nonlocal attention on image restoration have been shown in previous literatures [22, 34]. In our network, the multi-head-transformer-based blocks efficiently search coherent patches from the input reflectance image to extract deep features, cluster them according to their similarity in a soft manner, and finally improve each patch across different scale to reduce the noise caused in the optimization. The self-attention mechanism is both effective and flexible in reflectance image denoising since it can benefit from the redundancy among patches and have no constraint on the spectral contents; while other denoising architectures, such as residual convolutional blocks, despite having hundreds of parameters of the transformer, show very little improvement in reflectance image denoising when embedded in our network.

One difficulty of the factorization method is the existence of degenerate trivial solutions. To avoid such solutions, most of the previous methods use a cosine similarity loss of illumination to train or evaluate their models. However, this loss is sensitive to image noise in low-intensity pixels. To ensure the robustness of the proposed method and force the estimated reflectance to satisfy the prior of real reflectance, we choose the weighted combination of mean square error (MSE) of the scaled estimated illumination and reflectance as our loss function, expressed as:

$$\mathcal{L} = ||s\mathbf{L}_e - \mathbf{L}_{gt}||_2^2 + \phi_1 \, ||\mathbf{R}_e/s - \mathbf{R}_{gt}||_2^2 + \phi_2 |\log(s)| \quad \text{with} \quad s = \frac{\langle \mathbf{L}_{gt}, \mathbf{L}_e \rangle}{\langle \mathbf{L}_e, \mathbf{L}_e \rangle},$$

where s denotes the scale of illumination, ϕ_1 and ϕ_2 denote the weighting parameters (we set $\phi_1 = 0.2$ and $\phi_2 = 0.001$ empirically), $\mathbf{L}_{gt}$ and $\mathbf{L}_e$ denotes the ground truth and estimated illumination, $\mathbf{R}_{gt}$ and $\mathbf{R}_e$ denotes the ground truth and estimated reflectance. Note that the third term forces the scale to be 1 as close as possible to enhance the robustness of the network and avoid trivial solutions.

We construct a five-stage unrolling network, which ran on an NVIDIA GeForce GTX 1080 Ti GPU. We built our model using Pytorch and trained it up to 40 epochs. For optimization, Adam is employed with a batch size of 2 and a learning rate of 3×10^{-4}. The computation of a single image can be accomplished in three seconds.

4. Simulations and Experiments

4.1. Spectral dataset

We captured 400 high-quality multispectral images as our training reflectance data (see Fig. 4(a)). The images, consisting of a mixture of indoor and outdoor scenes, are acquired using a compact scanning-based hyperspectral camera: Specim IQ. The captured images have a spatial resolution of 512×512 pixels and 204 spectral bands ranging from 400nm to 1000nm. From these images we synthesized a large training set with a range of different simulated illumination conditions.

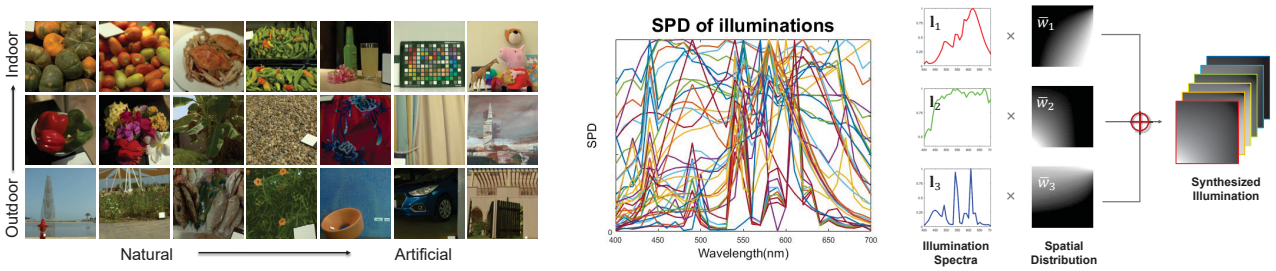
To prepare the data for training, the reflectance spectrum at each pixel is computed from the measured multispectral image using a white reference chart. We uniformly sampled the spectra in the visible range from 400nm to 700 nm into 10 nm intervals, resulting in a total of 31 channels. We then cropped the reflectance images into 256×256 subimages to remove the white reference surface. We normalized the multispectral reflectance in the range of [0,1], and finally randomly selected 320 reflectance images for generating the training set and tested our method's performance on the remaining 80 images. Each set has a balanced number of indoor/outdoor scenes.

To synthesize the multispectral images under various simulated illuminations, we constructed our spectral illumination dataset by collecting the SPDs of standard illuminants, some artificial light sources from public datasets[1], and solar lights from public datasets[2]. Some samples are shown in Fig. 4(b). To avoid over-fitting, we excluded some similar SPDs, selected 40 representative spectra illuminations to generate our training set, and selected the other ten spectral illuminations to generate the test set.

We synthesized the multispectral images dataset by random multiplication of the reflectance images and the SPDs of illuminations. Both single-global illumination and multiple illumination estimation are simulated. In the case of single illumination estimation, we directly multiplied the SPD of a single illumination by a reflectance image to simulate a captured multispectral image. While in the case of multiple illumination estimation, we first randomly chose

[1] http://galileo.graphycs.cegepsherbrooke.qc.ca/app/fr/lamps

[2] http://www.nrel.gov/grid/solar-resource/assets



(a) Multispectral reflectance images (b) Illumination spectra (c) Spatially varying illumination synthesis

Figure 4. (a-b) Representative samples of our dataset. The RGB images are rendered under CIE standard illuminant D65 with the CIE 1964 10° standard observer. (c) The operation of synthesizing spatially varying illumination.

three illumination spectra $\mathbf{l}_i(i = 1, 2, 3)$ from the illumination dataset, and generated three random two-dimensional sine-based functions $w_i(x, y)$ for the spatial distribution of each illumination. The random spatial distribution function represents the contribution of each illumination, which is defined as:

$$w_i(x, y) = \frac{A_i sin(\omega_{i1}(x+p_{i1}))sin(\omega_{i2}(y+p_{i2}))}{(\omega_{i1}(x+p_{i1}))(\omega_{i2}(y+p_{i2}))} \tag{10}$$

where A, ω, p is the random intensity, frequency, phase respectively. We normalized the spatial distribution functions w_i to $\bar{w}_i$ and ensured the sum of the each contribution be a constant, here we let the sum of the normalized function $\sum_{i=1,2,3} \bar{w}_i = 1$. Our spatially varying illumination is generated by the linear combination of the three illumination spectra as shown in Fig. 4(c). Finally we synthesized a large amount of multispectral images by multiplying the reflectance image with the generated illumination for training and evaluation.

To quantify the estimation accuracy of the methods, we respectively calculated absolute error ΔS and angular error ΔA between the ground truth illuminant $\mathbf{l}_g$ and the estimated illuminant $\mathbf{l}_e$. The two errors are commonly used in previous work:

$$\Delta S = ||\mathbf{l}_g - \mathbf{l}_e \cdot \frac{\langle \mathbf{l}_g, \mathbf{l}_e \rangle}{\langle \mathbf{l}_e, \mathbf{l}_e \rangle}||_1, \quad \Delta A = \arccos\left(\frac{\langle \mathbf{l}_g, \mathbf{l}_e \rangle}{||\mathbf{l}_g|| \cdot ||\mathbf{l}_e||}\right).$$

4.2. Ablation study

To sufficiently investigate the contribution of trainable weight parameters and each denoising block (LR: low-rank, TV: total-variation, NL: nonlocal self-attention transformer) in our unrolling network, we carried out six ablation simulations for estimating multiple illumination on our dataset. The quantitative results are shown in Table 1.

The results show that the trainable parameters can bring 15 percent absolute error reduction compared with using fixed values of parameters. It is also evident that the TV and LR are the two most crucial regularizers in our simulation. The possible reason is that the two regularizers explicitly constrains the illumination spectra estimation. Combining all the three regularizers, our method leads to a significant improvement in reconstruction quality.

In the ablation study, we attempt to replace the transformer in each stage by using a residual block (resblock) which consists of ten 3×3 convolution layers and two skip connections. However, the resblock is largely ineffective in our optimization framework, which indicates that the blocks only considering local information are not suitable to deal with the denoising of reflectance images in our model.

Table 1. Ablation Studies for multiple illumination estimation.

MF	TV	LR	NL	ΔS		ΔA	
				mean	std	mean	std
fixed	fixed	fixed	×	4.23	2.36	0.29	0.23
✓	×	✓	✓	4.35	2.02	0.34	0.27
✓	✓	×	✓	4.21	2.44	0.35	0.26
✓	✓	✓	×	3.63	1.90	0.25	0.19
✓	✓	✓	resblock	3.50	1.85	0.24	0.18
✓	✓	✓	✓	**2.84**	**1.37**	**0.21**	**0.15**

4.3. Single global illumination estimation

Estimating a single, global illumination spectrum is a traditional problem that assuming the illumination spectrum distributes uniformly on the captured image. Our approach can be directly applied to estimate global illumination by performing an average pooling on the estimated illumination spectral cube, without resorting to additional learning.

We compared our method with five existing methods, including GrayEdge [17], LRMF [35], ISNL [28], and P-WIR [24]. Specifically, GrayEdge is a variant of a classical color constancy method; LRMF and ISNL either ultilize the low rank prior of spectral reflectance or extract the specular reflections to estimate global illumination; PWIR is a CNN-based method finding the illuminant-invariant features in images. Although PWIR can predict full-resolution images of illumination, it actually uses bicubic interpolation to achieve the goal, so we classify PWIR as a single global illumination estimation method.

In the simulation, we compared the five methods with ours on the two synthesized datasets generated using our multispectral reflectance images and the CAVE multispec-

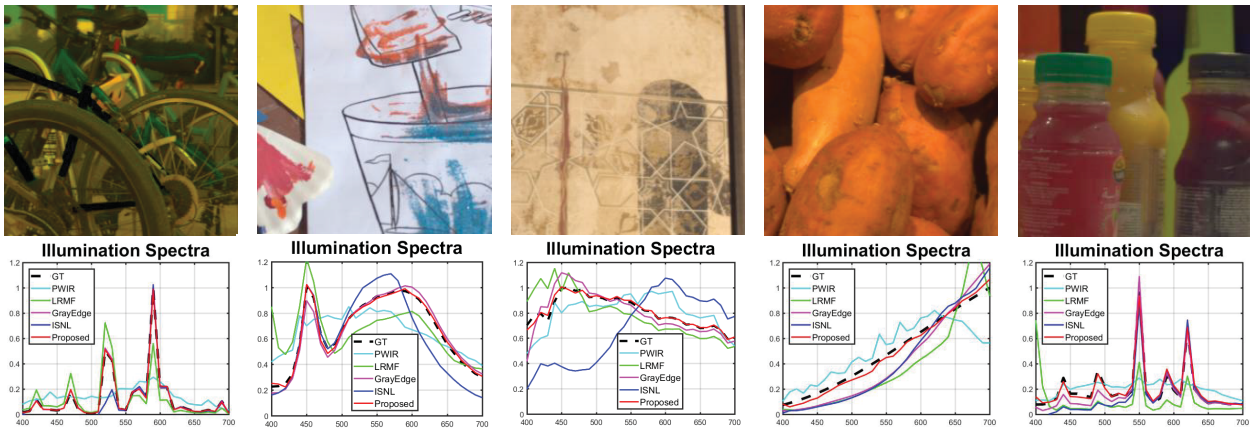


Figure 5. Comparison of the five single-global illumination estimation methods. Top: the rendered RGB images of captured multispectral images. (Left to Right: bicycles, paintings, wall, potatoes, drinks) Bottom: the SPDs of estimated illumination spectra and ground truth.

tral dataset [32] respectively. Fig. 5 shows the comparisons of the estimated SPDs of the five methods on our dataset. Our method performs best, regardless of whether the spectra are smooth or have sharp spikes. To our surprise, the classical GrayEdge method outperforms many of the more recent methods, but it is unable to match the performance of our method, especially when estimating smooth illumination spectra. Although the CNN-based method PWIR is trained with more parameters than our method, it fails in most of the cases. We believe this is because the CNN has difficulties in generalizing to different spectra other than the ones it has been trained on. We can also observe that ISNL performs poorly in purely diffuse scenes (e.g. third column in Fig. 5), and the performance of LRMF is not stable and highly dependent on the choice of hyperparameters.

Table 2. Comparison of single-illumination estimation methods.

Datasets	Methods	ΔS		ΔA	
		mean	std	mean	std
Ours	GrayEdge	2.29	2.42	0.20	0.14
	LRMF	4.34	5.40	0.42	0.41
	ISNL	3.96	3.06	0.35	0.20
	PWIR	3.52	**0.62**	0.49	0.37
	Proposed	**1.80**	1.33	**0.16**	**0.11**
CAVE	GrayEdge	2.48	2.40	0.32	0.26
	LRMF	2.81	2.60	0.35	0.27
	ISNL	3.72	2.47	0.40	0.29
	PWIR	4.62	1.62	0.61	0.37
	Proposed	**1.87**	**1.40**	**0.28**	**0.24**

Table 2 shows the estimation error of the five methods along with the error statistics (mean and standard deviation). The PWIR and our method are trained on our dataset and are tested on both datasets to evaluate the generalization ability. The estimation accuracy of our method shows significantly superior results to others irrespective of the dataset used. PWIR performs worse on the CAVE dataset than on our dataset. Together with the low variance of the error on our dataset this again suggests overfitting of the PWIR CNN, and difficulties with generalization to other scenes.

4.4. Multiple illumination estimation

Since there is no existing multiple illumination estimation method for multispectral images besides PWIR, we compared our method with two state-of-the-art multiple illumination estimation methods for RGB images, by transforming their method from RGB domain to spectral domain. The two methods are BUTD [8] and AngularGAN [27], in which BUTD is a method extracting color-invariant structure in RGB images and mapping colors based on statistics, and AngularGAN is a pixel-to-pixel GAN-based method with a U-net generator. We trained and compared the three methods on our dataset and CAVE dataset.

Table 3. Comparison of multi-illumination estimation methods.

Datasets	Methods	ΔS		ΔA	
		mean	std	mean	std
Ours	BUTD	4.25	1.21	0.41	0.37
	angularGAN	3.77	**0.96**	0.38	0.30
	Proposed	**2.84**	1.37	**0.21**	**0.15**
CAVE	BUTD	4.06	1.54	0.48	0.35
	AngularGAN	3.89	**1.16**	0.41	0.29
	Proposed	**2.65**	1.22	**0.31**	**0.16**

As shown in Table 3, the estimated error of our method is significantly less than those of others. It is worth noticing that our method even shows competitive results compared to the results of other methods in the single-illumination estimation case. The amount of parameters of our method is only one-tenth of that of AngularGAN. Acting like another ordinary CNN-based method (PWIR) in the single-illumination estimation, AngularGAN performs more stable than others but also easier to get overfitting.

A comparison of illumination estimation angular error is visualized in Fig. 6. We show the angular error by rendering the estimated reflectance to RGB images with a s-

Figure 6. Top: RGB images rendered from synthesized multispectral image, and the estimation angular error of the three methods. Bottom: RGB images rendered from spectral reflectance image of ground truth and the estimated reflectance images estimated by the three methods.

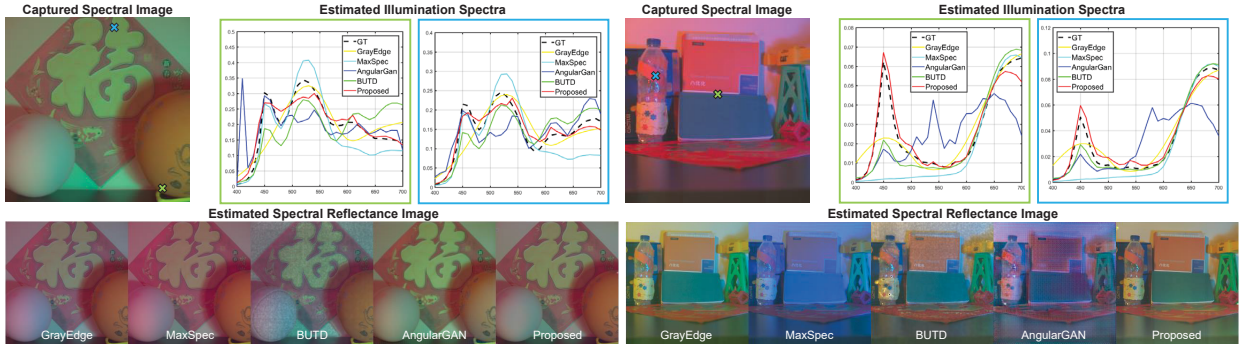


Figure 7. The comparisons of the estimated illumination spectra and reflectance images. Both scenes were illuminated by two full-spectrum lamps with filters. We show the estimated spectra of the five illumination estimation methods at the positions of two crosses in each scene. The reflectance images are rendered to RGB images for visualization.

tandard D65 illumination. Our reflectance images show a more uniform appearance that indicates a more accurate multiple-illumination estimation. We can also observe that the estimation accuracy of BUTD is highly dependent on the choice of parameters, and it may introduce deviated color artifacts; while AngularGAN cannot handle spatially varying illumination well.

4.5. Experiments on Real Images

We compared our method with two single-illumination methods(GrayEdge and MaxSpec[17]) and two multiple-illumination estimation methods (BUTD and AngularGAN) on real multispectral images captured under more than one illumination. The used multispectral camera is also Specim IQ with spatial resolution 512×512. The used light sources include sunlight, halogen lamp, and LED lamp. We added color filters in front of these light sources to increase the diversity of the illumination spectra. To obtain the ground truth of illumination spectra, we attached a few white references on the objects of each captured scene.

Two comparisons of the estimated spectral plots and the RGB images rendered from estimated reflectance images are given in Fig. 7. Exploiting non-local denoising prior, our method provides more accurate and clean results than AngularGAN and BUTD, and performs much better on large patches with the same spectral reflectance which is challenging in illumination estimation problem. It is also evident that our method provides the most stable estimation results as the color(spectral reflectance) of the background wall are consistent in the two scenes.

5. Conclusion

We present an end-to-end unrolling network architecture for solving multispectral illumination estimation problem. Unlike previous methods, our method can handle both single-global illumination and multiple illumination estimation, and significantly outperforms previous methods due to the utilized denoising prior. The promising performance of the proposed method on synthesized and real images showed its effectiveness, flexibility, and generalization ability. We constructed a large spectral reflectance image dataset for training and evaluating, and the community can use it for future training and analysis work. In the future, we will enlarge the spectral reflectance image dataset, and explore the use of hardware encoding [3] to apply spectral illumination estimation to general imaging systems.

References

[1] Mahmoud Afifi and Michael S Brown. Deep white-balance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1397–1406, 2020.

[2] Sara Alvarez-Cortes, Timo Kunkel, and Belen Masia. Practical low-cost recovery of spectral power distributions. In *Computer Graphics Forum*, volume 35, pages 166–178. Wiley Online Library, 2016.

[3] Seung-Hwan Baek, Hayato Ikoma, Daniel S Jeon, Yuqi Li, Wolfgang Heidrich, Gordon Wetzstein, and Min H Kim. End-to-end hyperspectral-depth imaging with learned diffractive optics. *arXiv preprint arXiv:2009.00463*, 2020.

[4] Shida Beigpour, Christian Riess, Joost Van De Weijer, and Elli Angelopoulou. Multi-illuminant estimation with conditional random fields. *IEEE Transactions on Image Processing*, 23(1):83–96, 2013.

[5] Simone Bianco, Claudio Cusano, and Raimondo Schettini. Single and multiple illuminant estimation using convolutional neural networks. *IEEE Transactions on Image Processing*, 26(9):4347–4362, 2017.

[6] Gershon Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin institute*, 310(1):1–26, 1980.

[7] Dongliang Cheng, Abdelrahman Abdelhamed, Brian Price, Scott Cohen, and Michael S Brown. Two illuminant estimation and user correction preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 469–477, 2016.

[8] Shao-Bing Gao, Yan-Ze Ren, Ming Zhang, and Yong-Jie Li. Combining bottom-up and top-down visual mechanisms for color constancy under varying illumination. *IEEE Transactions on Image Processing*, 28(9):4387–4400, 2019.

[9] Arjan Gijsenij, Rui Lu, and Theo Gevers. Color constancy for multiple light sources. *IEEE Transactions on image processing*, 21(2):697–707, 2011.

[10] Lin Gu, Cong Phuoc Huynh, and Antonio Robles-Kelly. Segmentation and estimation of spatially varying illumination. *IEEE transactions on image processing*, 23(8):3478–3489, 2014.

[11] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020.

[12] Eugene Hsu, Tom Mertens, Sylvain Paris, Shai Avidan, and Frédo Durand. Light mixture estimation for spatially varying white balance. In *ACM SIGGRAPH 2008 papers*, pages 1–7. 2008.

[13] Yuanming Hu, Baoyuan Wang, and Stephen Lin. Fc4: Fully convolutional color constancy with confidence-weighted pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4085–4094, 2017.

[14] Hamid Reza Vaezi Joze and Mark S Drew. Exemplar-based color constancy and multiple illumination. *IEEE transactions on pattern analysis and machine intelligence*, 36(5):860–873, 2013.

[15] Haris Ahmad Khan. *Multispectral constancy for illuminant invariant representation of multispectral images*. PhD thesis, Bourgogne Franche-Comté, 2018.

[16] Haris Ahmad Khan, Jean-Baptiste Thomas, and Jon Yngve Hardeberg. Towards highlight based illuminant estimation in multispectral images. In *International Conference on Image and Signal Processing*, pages 517–525. Springer, 2018.

[17] Haris Ahmad Khan, Jean-Baptiste Thomas, Jon Yngve Hardeberg, and Olivier Laligant. Illuminant estimation in multispectral imaging. *JOSA A*, 34(7):1085–1098, 2017.

[18] Haris Ahmad Khan, Jean-Baptiste Thomas, Jon Yngve Hardeberg, and Olivier Laligant. Spectral adaptation transform for multispectral constancy. *Journal of Imaging Science and Technology*, 62(2):20504–1, 2018.

[19] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 61(1):1–11, 1971.

[20] Yuqi Li, Miao Qi, Rahul Gulve, Mian Wei, Roman Genov, Kiriakos N Kutulakos, and Wolfgang Heidrich. End-to-end video compressive sensing using anderson-accelerated unrolled networks. In *2020 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2020.

[21] Yuqi Li, Chong Wang, Jieyu Zhao, and Qingshu Yuan. Efficient spectral reconstruction using a trichromatic camera via sample optimization. *The Visual Computer*, 34(12):1773–1783, 2018.

[22] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1680–1689, 2018.

[23] Lawrence Mutimbu and Antonio Robles-Kelly. Multiple illuminant color estimation via statistical inference on factor graphs. *IEEE Transactions on Image Processing*, 25(11):5383–5396, 2016.

[24] Antonio Robles-Kelly and Ran Wei. A convolutional neural network for pixelwise illuminant recovery in colour and spectral images. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 109–114. IEEE, 2018.

[25] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[26] Wu Shi, Chen Change Loy, and Xiaoou Tang. Deep specialized network for illuminant estimation. In *European conference on computer vision*, pages 371–387. Springer, 2016.

[27] Oleksii Sidorov. Conditional gans for multi-illuminant color constancy: Revolution or yet another approach? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[28] Tong Su, Yu Zhou, Yao Yu, Xun Cao, and Sidan Du. Illumination separation of non-lambertian scenes from a single hyperspectral image. *Optics express*, 26(20):26167–26178, 2018.

[29] Cheng Tai, Tong Xiao, Yi Zhang, Xiaogang Wang, and E Weinan. Convolutional neural networks with low-rank regularization. In *ICLR 2016*, 2016.

[30] Joost Van De Weijer, Theo Gevers, and Arjan Gijsenij. Edge-based color constancy. *IEEE Transactions on image processing*, 16(9):2207–2214, 2007.

[31] Bolei Xu, Jingxin Liu, Xianxu Hou, Bozhi Liu, and Guoping Qiu. End-to-end illuminant estimation based on deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3616–3625, 2020.

[32] Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K Nayar. Generalized assorted pixel camera: post-capture control of resolution, dynamic range, and spectrum. *IEEE transactions on image processing*, 19(9):2241–2253, 2010.

[33] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938, 2017.

[34] Y Zhang, K Li, K Li, B Zhong, and Y Fu. Residual non-local attention networks for image restoration. In *International Conference on Learning Representations*, 2019.

[35] Yinqiang Zheng, Imari Sato, and Yoichi Sato. Illumination and reflectance spectra separation of a hyperspectral image meets low-rank matrix factorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1779–1787, 2015.