In Defense of Shallow Learned Spectral Reconstruction from RGB Images

Jonas Aeschbacher
Computer Vision Lab
D-ITET, ETH Zurich
aejonas@student.ethz.ch

Jiqing Wu
Computer Vision Lab
D-ITET, ETH Zurich
jwu@vision.ee.ethz.ch

Radu Timofte
CVL, D-ITET, ETH Zurich
Merantix GmbH
timofter@vision.ee.ethz.ch

Abstract

Very recent Galliani et al. *[13] proposed a method using a very deep CNN architecture for learned spectral reconstruction and showed large improvements over the recent sparse coding method of Arad* et al. *[6]. In this paper we defend the shallow learned spectral reconstruction methods by: (i) first, reimplementing Arad and showing that it can achieve significantly better results than those originally reported; (ii) second, introducing a novel shallow method based on A+ of Timofte* et al. *[33] from super-resolution that substantially improves over Arad and, moreover, provides comparable performance to Galliani's very deep CNN method on three standard benchmarks (ICVL, CAVE, and NUS); and (iii) finally, arguing that the train and runtime efficiency as well as the clear relation between its parameters and the achieved performance makes from our shallow A+ a strong baseline for further research in learned spectral reconstruction from RGB images. Moreover, our shallow A+ (as well as Arad) requires and uses significantly smaller train data than Galliani (and generally the CNN approaches), is robust to overfitting and is easily deployable by fast training to newer cameras.*

1. Introduction

Nowadays there is an ever-increasing variety of visual sensors used for image analysis. This enables devices to collect immense amounts of information from the environment. However, most cameras can only record a limited amount of information from the visible spectrum, often containing the standard RGB (red, green, blue) wavelength values matching the trichromaticity from the human visual system. This is caused mainly by the need to keep the costs of such sensors low and to achieve higher spatial resolution, in contrast to the hyperspectral cameras. For a given budget there is a trade-off between having high spectral and high spatial resolution for the camera captured imagery.

Capturing visual data with a camera with higher spectral resolution has been proven very useful in many areas, as for example in medical diagnosis [29, 12, 25], image segmentation [30, 10], modeling of computer generated imagery or general remote sensing tasks [16, 14, 8, 20, 7]. Unfortunately, there is often the decision whether to use a very high spectral or spatial resolution and this generally boils down to the (cheaper) latter one.
But since most of the resulting radiance of high spectral resolution images are a composition of the illumination and reflectance of the occurring materials in the image, it is reasonable to believe that only a small amount of different factors have an impact on a single pixel. This means that the RGB values and their corresponding hyperspectral radiance should be highly correlated [2, 11].

Only a reduced number of works (such as [21, 27, 39, 3, 1, 19, 22, 26, 35, 6, 13]) tried to infer a (full) hyperspectral image from its RGB image(s). Among them, Arad *et al.* [6] used high quality hyperspectral image priors to build a sparse dictionary of corresponding high (full spectrum) and low spectral (RGB) resolution pixels. In particular, the orthogonal matching pursuit [24] is used to decompose the input RGB pixels over the corresponding part of the dictionary to impose the decomposition coefficients on the corresponding high spectral resolution part and reconstruct the hyperspectral image. The assumed linearity between low and high spectral resolution signals is thus leveraged. Similar approaches can be found in prior super-resolution literature [32, 36, 38]. Very recently, Galliani *et al.* [13] proposed a convolutional neural network (CNN) which used best practices from the current super resolution and segmentation literature [18, 17, 15, 28, 31] and achieved better results than Arad *et al.* .

In this paper we reimplement and push the performance of Arad *et al.* method and, furthermore, we propose a new method to infer a hyperspectral image using a RGB image motivated by A+, the color single image super resolution method of Timofte *et al.* [33, 32, 34], and achieve top accuracies without the need of neural networks. The super-resolution problem usually works in the spatial domain of the RGB image and aims at restoring rich details/high frequencies. Clearly, it relates with our problem of reconstructing missing wavelengths from the spectrum based on the known RGB. Of course, non-trivial adaptation to reproduce a higher spectral resolution instead of recovering more pixels per image is required in our case. We compute a sparse dictionary containing the corresponding low and high spectral resolution atoms, which act as anchor points for the following computations. Then, for each anchor, we leverage the local Euclidean linearity of the spectral spaces to offline compute a projection matrix from RGB to hyperspectral image values, by using the nearest neighbors of the anchor. At runtime, for each RGB pixel there is only a nearest anchor search involved, followed by a projection to the hyperspectral values using the corresponding stored matrix. As shown in our experiments, in addition to the efficiency and fast runtime, we reach higher accuracies than Arad *et al.* [6], and comparable results to the very recent deep CNN approach of Galliani *et al.* [13].

Our main contributions are threefold:

(i) we efficiently improve the approach of Arad *et al.* [6] for better accuracy and runtime;

(ii) we propose a shallow A+ [33]-based method for spectral reconstruction;

(iii) we make a stand in defense of the efficient shallow models by achieving state-of-the-art performance.

This work sets strong shallow baselines for the future research in learned spectral reconstruction from RGB images. Deep end-to-end (CNN) models will likely benefit from the increased availability of train data to further push the performance, as happened in the super-resolution field [31].

The remainder of this paper is organized as follows. Section 2 reviews related works focusing on Arad *et al.* [6]. Section 3 proposes a novel method for spectral reconstruction based on A+ super-resolution method of Timofte *et al.* [33]. Section 4 describes the experiments, studies parameters and design choices, and compares in accuracy and runtime with the current state-of-the-art methods. Section 5 draws the conclusions.

2. Related Work

We briefly review the method of Arad *et al.* [6] (see Fig. 1) based on the work of Zeyde *et al.* [38]. Arad *et al.* use a collection of hyperspectral signatures as a training prior to build an overcomplete dictionary with m hyperspectral signature atoms $\mathbf{h}_i$ via K-SVD [4]:

$$D_H = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_m\}, \quad \mathbf{h}_i \in \mathbb{R}^q \tag{1}$$

such that each training hyperspectral signature can be approximated by a sparse linear combination over the dictionary atoms as obtained via orthogonal matching pursuit (OMP) [24]. These atoms are then projected to the lower

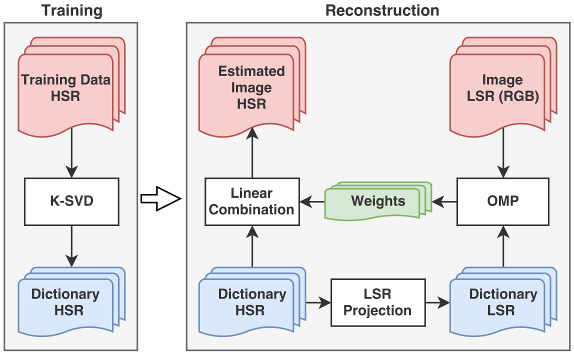


Figure 1. Training and reconstruction phases of the multichannel image restoration method by Arad *et al.* [6]

spectral resolution (LSR) space $\mathbf{l}_i \in \mathbb{R}^p$ (i.e., the RGB space), applying the appropriate camera sensitivity function $M(\mathbf{h}_i)$, such that it matches the sensors used to capture the sample for reconstruction. To later map the RGB sample to the higher spectral resolution (HSR) space, one needs to keep the correspondence between both dictionary atoms.

$$D_L = \mathbf{R} \cdot D_H = \{\mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_m\}, \quad \mathbf{l}_i \in \mathbb{R}^p \tag{2}$$

With the trained dictionaries it is possible to estimate the hyperspectral intensities from an RGB image, by first linearly decomposing each pixel $\mathbf{p}_L = (r_i, g_i, b_i)$ from RGB via OMP over D_L and then using the computed decomposition coefficients $\mathbf{w}$ to approximately reconstruct the corresponding HSR pixel $\mathbf{p}_H$:

$$D_L \cdot \mathbf{w} = \mathbf{p}_L \quad \Rightarrow \quad \mathbf{p}_H = D_H \cdot \mathbf{w} \tag{3}$$

The mapping of RGB values to the whole spectral space is severely underconstraint. Usually, there are a finite, low number of spectral wavelengths of interest. Arad *et al.* [6] consider 31 different wavelengths from the visible spectrum for reconstruction. The frequency of relative metameric pairs in this lower dimensional manifold needs to be low as well, which is the case in the visible spectrum used here, as further explained in [5, 23].

3. Proposed method

Our method (see Fig. 2) departs from the method of Arad *et al.* [6] and builds upon the A+ method of Timofte *et al.* [33, 32, 34] introduced for single image super-resolution. In our case, a hyperspectral, overcomplete sparse dictionary representation, trained with K-SVD and using OMP coefficients, is employed. The dictionary is then projected to a lower spectral resolution (RGB). In contrast to Arad's method, the color matching function, used for projecting the training HSR data to a LSR space, needs to be determined at training time, since the matrices are computed using both, RGB and hyperspectral atoms and not

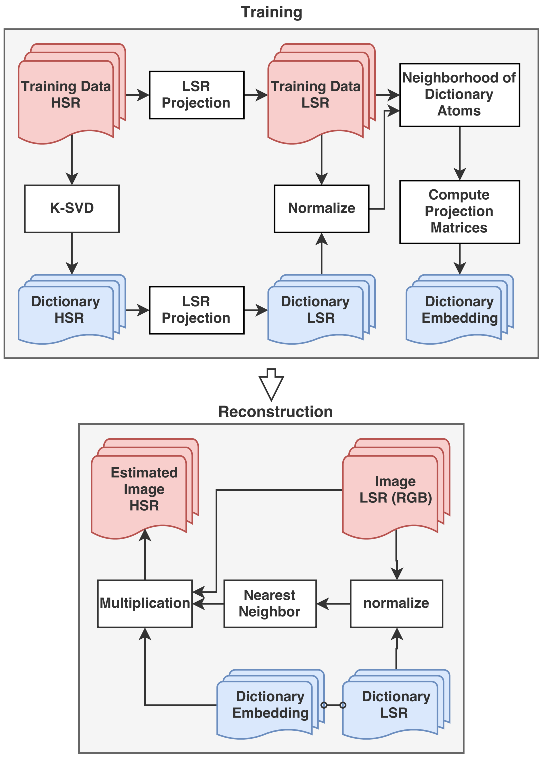


Figure 2. Training and reconstruction phases of our A+ multichannel image restoration method.

only the latter. Most neighbor embedding methods [6, 32] try to extract a linear combination of training samples to represent the low spectral resolution signature, followed by a reconstruction in the hyperspectral space, using the same coefficients. This computation can be moved into the training phase, as explained below. For each dictionary atom/anchor $\mathbf{l}_i$, we minimize the least squares error of the linear combination of its nearest neighbors ($\mathbf{N_L}$) from all available training data to $\mathbf{y}_L$:

$$\arg\min_{\boldsymbol{\alpha}} ||\mathbf{y}_L - \mathbf{N}_L\boldsymbol{\alpha}||_2^2 + \lambda||\boldsymbol{\alpha}||_2^2 \tag{4}$$

where λ regularization stabilizes the closed form solution:

$$\boldsymbol{\alpha} = (\mathbf{N}_L^T\mathbf{N}_L + \lambda\mathbf{I})^{-1}\mathbf{N}_L^T \cdot \mathbf{y}_L \tag{5}$$

As a result of applying the same assumptions as described in Section 2, it is possible to get a linearized projection matrix for the neighborhood samples of each corresponding dictionary atom from the LSR space (RGB) to a HSR with the same coefficients $\boldsymbol{\alpha}$.

$$\mathbf{y}_H = \mathbf{N}_H\boldsymbol{\alpha} \tag{6}$$

$$\mathbf{P}_i = \mathbf{N}_H(\mathbf{N}_L^T\mathbf{N}_L + \lambda\mathbf{I})^{-1}\mathbf{N}_L^T \tag{7}$$

$$\mathbf{y}_H = \mathbf{P}_i \cdot \mathbf{y}_L \tag{8}$$

where $\mathbf{N}_H$ are the corresponding HSR neighborhood samples to $\mathbf{N}_L$ for an anchor $\mathbf{l}_i$. After offline computing and storing all the projection matrices $\mathbf{P}_i$, at runtime the RGB samples can be embedded into the hyperspectral space using the projection of the nearest dictionary atom. For a more in-depth explanation we refer the interested reader to [33, 32].

We keep the name A+ since our proposed approach can be seen as the A+ method of Timofte *et al.* [33] adapted to the spectral reconstruction domain. In contrast with Timofte *et al.* [33] we work directly with pixel values and not with gradient responses (patch features) and residuals. Moreover, we regress from RGB values to multiple spectral values instead of regressing from low frequencies to high frequencies patches. In contrast with Arad *et al.* [6], instead of online computing OMP coefficients over the dictionary to impose in the hyperspectral space reconstruction, we learn offline anchored regressors from the low to the high spectral spaces using the pool of train samples.

3.1. Implementation

For the proposed method, as shown in Fig. 2, we first train an overcomplete dictionary with K-SVD (using OMP decomposition coefficients) and the hyperspectral priors. Then we obtain a normalized dictionary, containing the atoms to represent the hyperspectral signal with a linear combination of those anchor points. In the next step, the collection of atoms and the training data are projected to the desired lower dimensional spectral manifold. We use the CIE 1964 color matching function, which embeds the HSR sample into the RGB space. In Fig. 7 one can see the qualitative plot for the weights used to imitate the camera's sensitivity function, to project the pixels to LSR space.

Having these projections, they are normalized and used to extract the c nearest neighbors of the training signatures for each dictionary atom. The next step is to take the unnormalized neighbors for the atoms and use equation (7) to compute the embedding matrices.

After that, the nearest dictionary atom is selected for each RGB image pixel and we multiply it with the corresponding stored projection matrix. The simplicity of the reconstruction phase leads to a relatively efficient reconstruction. As shown in Table 2, even with an unoptimized Matlab implementation and using only an average processor core and no parallel implementation, the runtime is very efficient compared to other methods.

4. Experiments

In this section we describe the experimental benchmark, then analyze the major parameters of our methods [1] and

[1] Our codes, trained models, and results are available at: http://www.vision.ee.ethz.ch/~timofter/

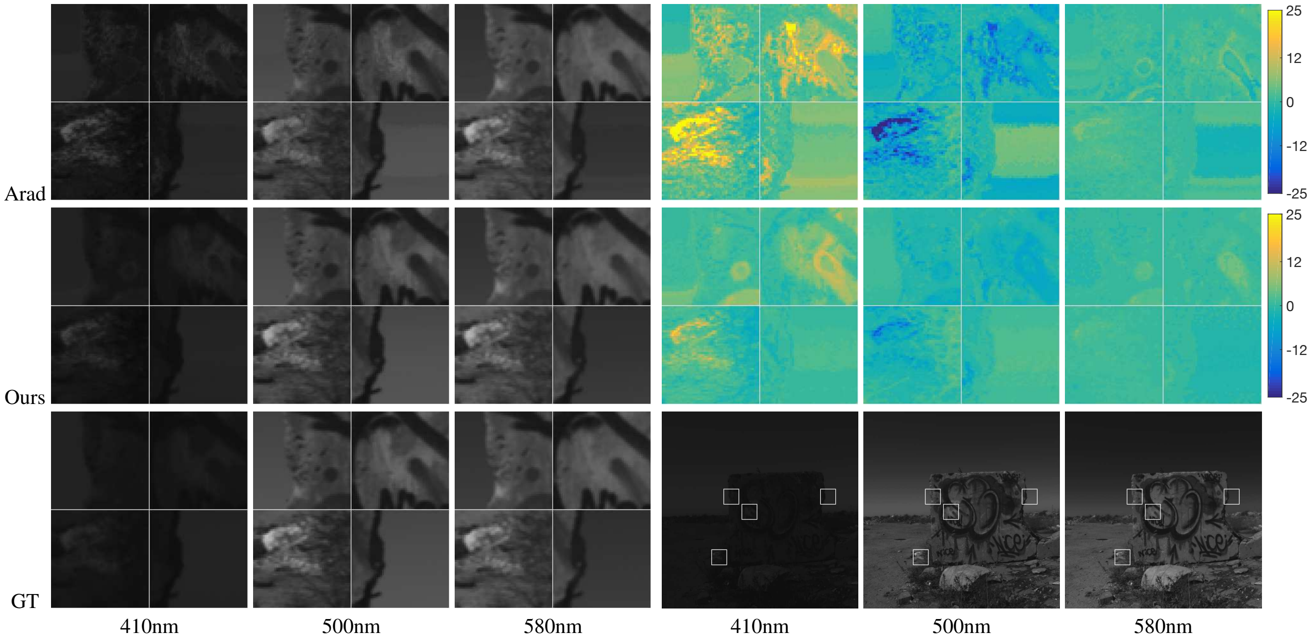


Figure 3. Visual comparison of Arad (our implementation) vs. our method based on A+ for three wavelengths on crops from an ICVL image. For reference we show also the ground truth and vizualize the pixel errors. Best zoomed on screen.

their impact on the achieved performance. In the end, we directly compare with current state-of-the-art methods.

4.1. Hyperspectral Image Databases

We evaluate our methods on three public benchmarks.

ICVL dataset [6] was recently released by Arad *et al.* , together with their method for sparse recovery of hyperspectral signal from natural RGB images. The dataset consists of 200 images collected by line scanner camera (Specim PS Kappa DX4 hyperspectral), which captures pictures at a spatial resolution of 1392×1300 over 519 spectral bands/wavelengths between 400 and 1000 nm. To facilitate the use of these pictures they were downsampled to a spectral resolution of 31 channels from 400 to 700nm at 10nm increments.

The amount of ICVL data is relatively large in comparison with other databases [22, 37, 11, 9] and is therefore well suited to analyze different parameters and techniques for our method before cross validating the final approach on other data sets.

To train a dictionary, a global training and test split was prepared. For that, the images of all the different environmental settings were evenly distributed into two sets to then randomly sample a certain amount of pixels from every picture in the database. With these pixels one can train two distinct dictionaries and later evaluate the performance by using all images of the remaining set not used for dictionary training. This is different to the original method of Arad *et al.* as they took for each testing image 1000 samples from each of the remaining images for training.

CAVE [37] database consists of 32 different images, with a spatial resolution of 512×512 pixels, also at 31 different spectral bands between 400 and 700nm, shot with a cooled CCD camera (Apogee Alta U260). It is a diverse collection of objects, containing faces, fake and real fruits, candy, paint, textiles and a lot more.

Because of the small number of pictures, we used a 4 fold crossvalidation, dividing the set into four groups. 24 images are then used to train the model, while the remaining 8 can be fed into the model to evaluate it. The different scenes were distributed as evenly as possible.

NUS [22] dataset contains 66 spectral images and their corresponding illuminations between 400 to 700nm, and 10nm increments. The pictures were taken with a Specim's PFD-CL-65-V10E spectral camera. Different illumination conditions were used, considering natural sunlight, artificial wide band lights using metal halide lamps and a commercial off-the-shelf LED [22].

We used the default training-test split of the database, with 25 images for testing and 41 for training.

4.2. Compared methods

Arad [6] method described in Section 2 uses a collection of hyperspectral signatures as a training prior to build an overcomplete dictionary with K-SVD [4] and OMP [24]. Having computed the atoms for the collection, it is projected to the RGB space and used to represent low spectral resolution samples with a linear combination of atoms. The sparse set of OMP weights can be applied to the hyperspectral dictionary atoms to reconstruct the final spectral signature.

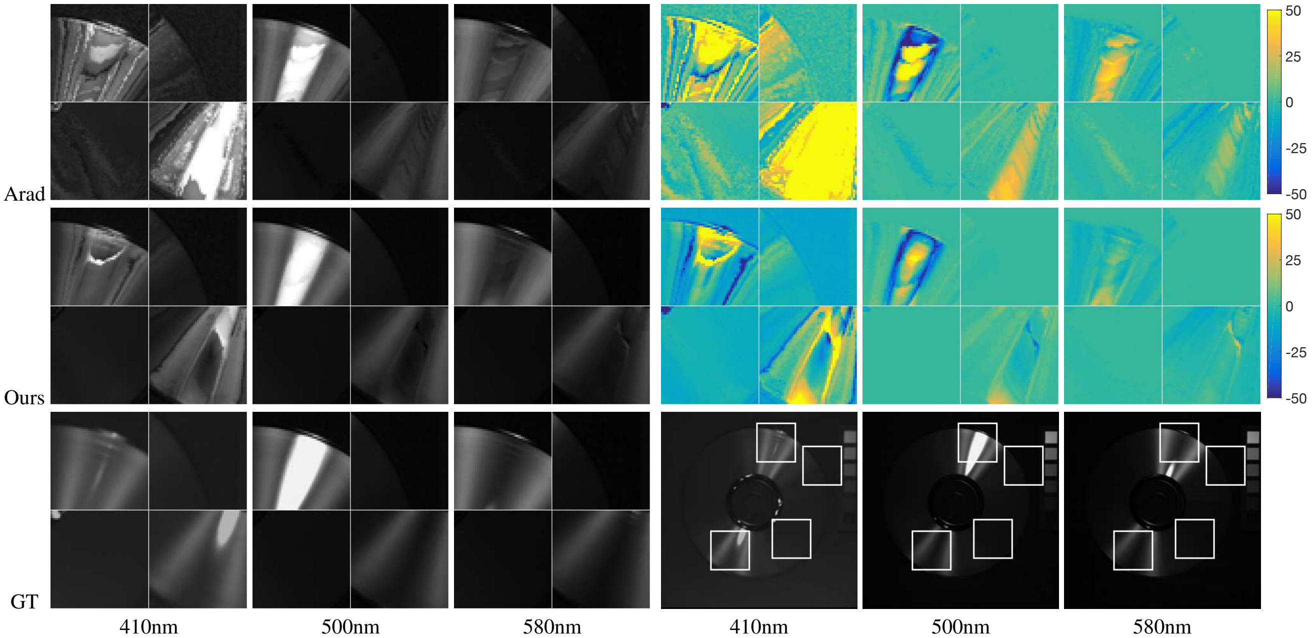


Figure 4. Visual comparison of Arad (our implementation) vs. our method based on A+ for three wavelengths on crops from a CAVE image. For reference we show also the ground truth(GT) and vizualize the pixel errors. Best zoomed on screen.

Arad *et al.* proposed to use a dictionary size of $m = 512$ and a sparsity of $k = 28$. For each test image a new model is trained using 1000 samples from each remaining image of the dataset.

Galliani [13] approach used a deep convolutional neural network as a model inspired by current semantic segmentation architecture Tiramisu of Jegou *et al.* [18]. An end-to-end mapping of the complete image from RGB to hyperspectral values is trained. The network first uses several densely connected convolutional layers with max pooling to subsequently scale the input down and extract important information. To recover the whole HSR image from learned features, subpixel upsampling is then used as proposed by Shi *et al.* [28]. The layers are interconnected to speed up the learning and reduce the vanishing gradient problem.

Nguyen [22] model is based on a learned mapping between hyperspectral response and their corresponding RGB values for a given camera sensitivity mapping. It uses a non-linear mapping, applying a radial basis function network. Additionally, the input data is processed with a white balancing step to reduce the effect of different illumination conditions on the mapping from RGB to hyperspectral reflectance.

Arad (ours) Our implementation of Arad's method has improved results as shown in Table 1. This can be explained by the fact that we used more training samples for the training of the dictionary and because we adopted a global train and test split, instead of a split for every single image. This resulted in 300000 training samples instead of only 200000. Besides these changes, the remaining parameters stayed the same.

A+ (ours) For our approach we use a pretrained overcomplete dictionary, not as proposed by Arad for a superposition but as anchor points to perform a nearest neighbor search. A projection matrix is (offline) computed for each anchor, using a collection of neighboring samples from the complete training set to approximate a local mapping from RGB to hyperspectral values. The method is described in Section 3.

4.3. Quantitative measures

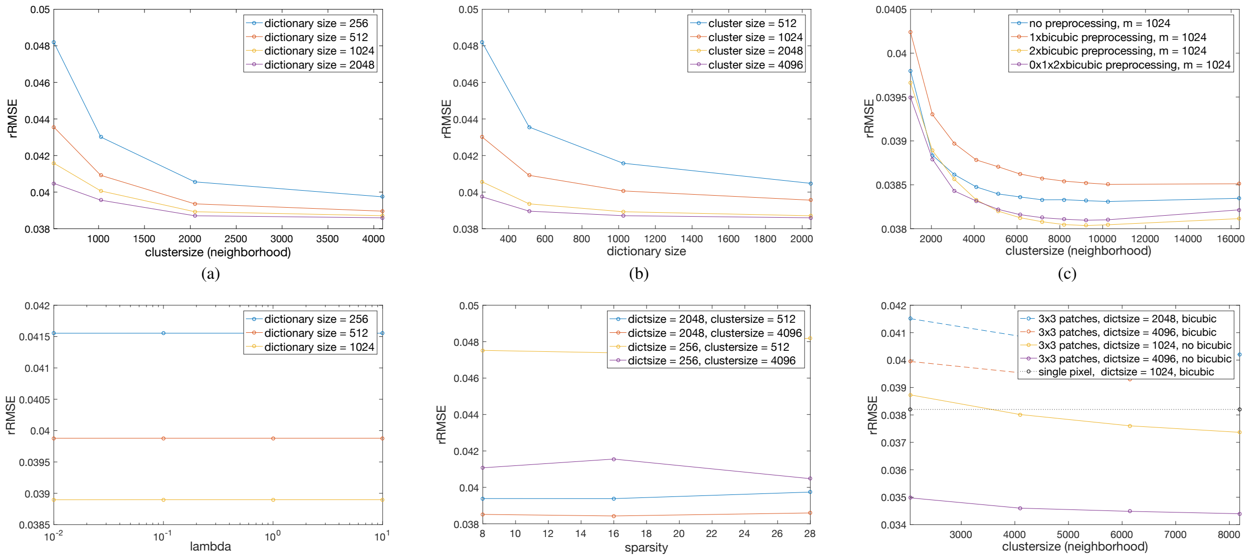
Root-mean-square error (RMSE) is a standard quantitative measure for accuracy. Arad *et al.* [6] and Galliani *et al.* [13] used the absolute and relative RMSE and in order to facilitate direct comparison we use them both.

RMSE The absolute RMSE is computed over 8 bit intensity pixel values. Equation (9) shows the formula used by Arad *et al.* , while equation (10) by Galliani *et al.* . $I_E^{(i)}$ and $I_G^{(i)}$ represents the ith element of the estimated or ground truth image and n are the number of pixels.

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(I_E^{(i)} - I_G^{(i)})^2} \tag{9}$$

$$RMSE_G = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(I_E^{(i)} - I_G^{(i)})^2} \tag{10}$$

rRMSE Arad *et al.* compute the relative RMSE by dividing the luminance error by the ground truth luminance, thus preventing a bias towards low errors in low luminance pix-



(a) (b) (c)

(d) (e) (f)

Figure 5. (a) Influence of the cluster size c (used for the computation of the projection matrices) on the accuracy, using $\lambda = 0.1, k = 28$ (At the top left). (b) Influence of the dictionary size c (number of atoms) on the accuracy, using $\lambda = 0.1, k = 28$ (At the top right). (c) Influence of the cluster size c and preprocessing on training image (no preprocessing or using bicubic interpolation) on the accuracy. ($\lambda = 0.1, k = 8$). (d) Influence of λ on the accuracy, using $k = 16, c = 4096$, (At the bottom left). (e) Influence of the dictionary sparsity k on the accuracy, using $\lambda = 0.1$, (At the bottom right). (f) Performance of a bigger patch size. Dashed lines represent trained models with bicubic downsampled training samples, while the continuous line uses unprocessed input data for training.

els as shown in (11). Galliani *et al.* take the average of the ground truth ($\bar{I}_G$) to get the relative RMSE, in (12).

$$rRMSE = \frac{1}{n}\sum_{i=1}^{n}(\sqrt{(I_E^{(i)} - I_G^{(i)})^2}/I_G^{(i)}) \qquad (11)$$

$$rRMSE_G = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((I_E^{(i)} - I_G^{(i)})/\bar{I}_G)^2} \qquad (12)$$

4.4. Parameters

The main parameters of our method are: dictionary size (m), cluster size (c), sparsity of the dictionary (k) and λ, which regularizes the least squares solution for the A+ computation.

To analyze the parameters, we were training our method on one half of the ICVL images and testing it with the remaining ones and vise versa. While evaluating the parameters, the training set was built of 2000 randomly selected samples (pixels) from each image, resulting in 200000 feature vectors, to train the dictionaries. To compute the projection matrices the same process was used, but with an increased amount of sample pixels per image. This leads to 2000000 signatures. More samples does not increase the matrix computation time as much as for the dictionary training, while significantly increasing the accuracy.

Neighborhood/cluster size (c) As shown in Fig. 5a the cluster size c has a big influence on the performance of our method. One can reduce the needed dictionary size for optimal results by using more neighbors for projection matrix computation, which will finally shorten the reconstruction time which involves a search over the dictionary atoms. Larger neighborhoods only increase training time.

regularization (λ) stabilizes the solution and our method is relatively robust to the selection of λ value as seen in Fig. 5d. We set λ to 0.1.

Dictionary size (m) rRMSE improves with the number of atoms m in the dictionary up to a saturation as shown in Fig. 5b,c. m affects both training and reconstruction, while the cluster size c only the training time. Therefore, it is beneficial to use large neighborhoods (c) and medium to small dictionaries (m).

Dictionary sparsity (k) The sparsity has low impact on the rRMSE of our method (see Fig. 5e). However, it heavily impacts the dictionary training as both K-SVD and OMP employed techniques depend on sparsity.

For later tests, the values for $\lambda = 0.1$ and $k = 8$ are fixed.

Mining of samples The techniques of Arad *et al.* and Timofte *et al.* work mainly due to the (assumed) linearity of the two spectral spaces in correspondence to each other. By representing the space as linear combinations of atoms (Arad) and as a linear projection in the case of A+ (Timofte), there already occurs smoothing by a certain degree. Often when linearizing a function, this can have a stabilizing effect on the results. Thus, it is important to verify whether it is possible to reduce the impact of noise on the training data even

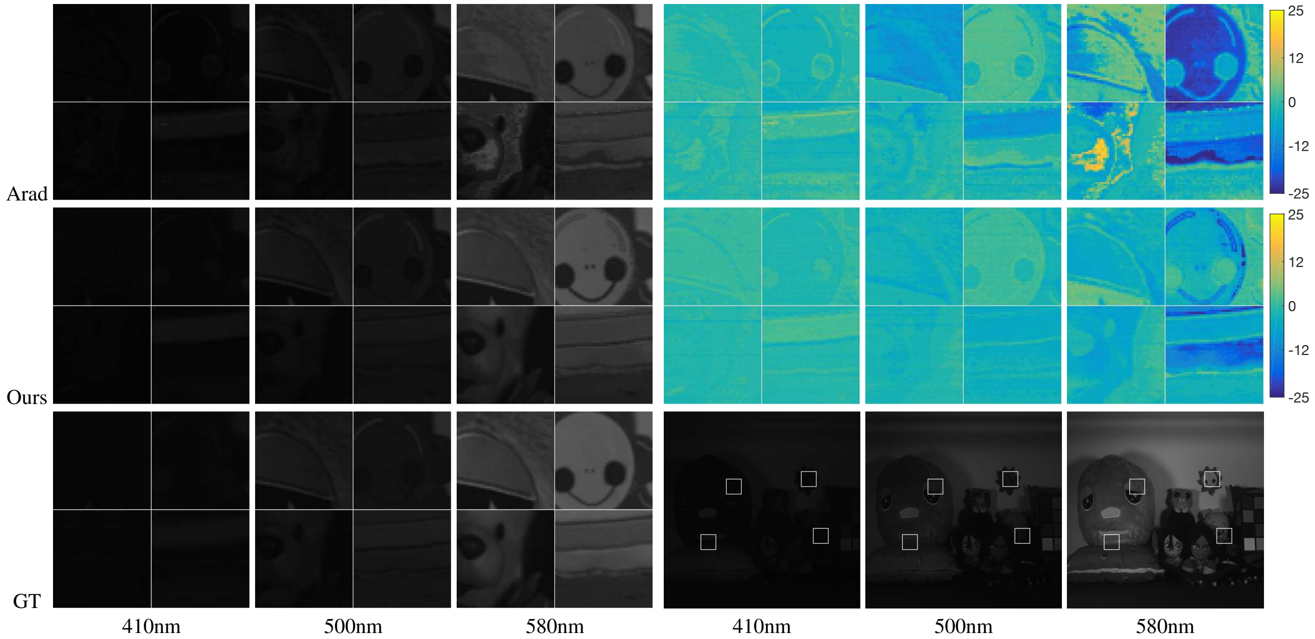


Figure 6. Visual comparison of Arad (our implementation) vs. our method based on A+ for three wavelengths on crops from a NUS image. For reference we show also the ground truth and vizualize the pixel errors. Best zoomed on screen.

further, by preprocessing the images used for training. This is done by scaling down the images using bicubic interpolation and, thus, removing some of the noise.

As shown in Fig. 5c, using training data which was scaled down by different factors (0x - no downscaling, 1x bicubic downscaling by a factor of 2^1, 2x - bicubic downscaling by a factor 2^2) , or even combinations of different downscaled samples, reached higher accuracy scores than the unprocessed one. A cluster size larger than 10000 leads to an increased error for all the settings considered. This is due to an overly linearized projection from LSR to HSR signals in relation with the ratio between total number of train samples and number of dictionary atoms/anchors. We conclude that $c = 8192$, $m = 1024$, and 2x maximize the accuracy and efficiency.

Patch vs. pixel support As seen in previous experiments bicubic downscaling of train images helped to reduce pixel noise and to improve the results. Therefore, we further investigate on using not only a pixel but a surrounding patch as description of that pixel. We used patches of size 3×3 and larger. This way one can use adjacent pixel intensities directly to infer additional information.

Fig. 5f shows that using patch support of size 3×3 does not benefit from training on downscaled (thus, less noisy) train images. This is because at runtime the model can not handle well the noise unseen in the training. The method with 3×3 patch support achieves better performance than the reference 1×1 for 1024 dictionary size and further improves with larger dictionaries (here 4096) and neighborhoods (here above 4000). For the final evaluations both settings are tested to have a good comparison, but for efficiency reasons, one would prefer the model trained with only single pixel feature vectors to keep the time for training and testing as low as possible.

4.5. Performance evaluation

As reported in Fig. 5 and in the previous sections our shallow method is well-behaved with respect to its main parameters, i.e. the accuracy improves with the increase of dictionary and/or neighborhood / train pool of samples. In the next, we report results for our method with two slightly different settings. For the first '1×1', the parameters are set to $k = 8$, $m = 1024$, $c = 8196$, $\lambda = 0.1$ and the values of a single pixel were used as input. The second '3×3' had a bigger dictionary of size $m = 4096$ and a patch support of 3×3 pixels. For both we used a total of $300'000$ samples to learn the dictionary and $3'000'000$ to compute the projection matrices. At training for the single pixel setting we used bicubic downscaled images (factor 2^2).

Quantitative evaluation The rRMSE of our A+ method at different wavelengths is shown in Figure 7 on ICVL data. Our method is very accurate at wavelengths corresponding to the RGB values, whereas the average and variance of the RMSE start to deviate more in between and at the edges of the visual spectrum.

Table 1 reports the results of our methods in comparison with those of Galliani, Arad, and Nguyen methods on three benchmarks: ICVL, CAVE and NUS. Since two different ways of calculating the relative and absolute RMSE were proposed by Galliani *et al.* and Arad *et al.* , all of them

Table 1. Quantitative comparison on ICVL, CAVE and NUS datasets. Best results are in bold.

	ICVL dataset [6]					**CAVE dataset [37]**					**NUS dataset [22]**				
	Galliani	Arad	**Arad**	**A+**	**A+**	Galliani	Arad	Arad	**A+**	**A+**	Nguyen	Galliani	Arad	**A+**	**A+**
	[13]	[6]	(ours)	1x1	3x3	[13]	[6]	(ours)	1x1	3x3	[22]	[13]	(ours)	1x1	3x3
rRMSE	-	0.0756	0.0507	0.0382	**0.0344**	-	-	0.4998	**0.4265**	0.4443	0.2145	-	0.1904	**0.1420**	0.1466
rRMSE$_G$	0.0587	-	0.0873	0.0599	**0.0584**	**0.2804**	-	0.7755	0.3034	0.3401	0.3026	0.234	0.3633	**0.2242**	0.2303
RMSE	-	2.6	1.70	1.12	**1.04**	-	5.4	5.61	**2.74**	2.90	12.44	-	4.44	2.92	**2.90**
RMSE$_G$	1.98	-	3.24	2.00	**1.96**	**4.76**	-	20.13	6.70	7.60	8.06	5.27	9.56	**5.17**	**5.17**

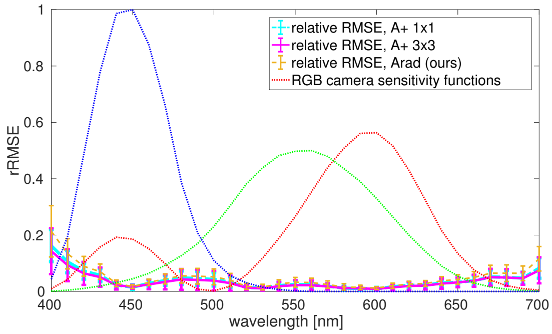


Figure 7. rRMSE at different wavelengths for our A+ and Arad methods on ICVL. RGB sensitivity is provided for reference.

Table 2. Runtime measurement for spectral reconstruction of one ICVL image of size 1300×1392.

	Arad[6]	Arad (ours)	**A+ (ours)**	**A+ (ours)**
			1x1	3x3
training (offline)	-	2.8h	**1.5h**	5.7h
runtime	1.5h+100s	130s	**30s**	110s

were computed.

On ICVL dataset our Arad reimplementation clearly improves over the original method and reported results by Arad *et al.* [6], moreover our shallow A+ methods significantly improve over Arad and are on par with or slightly better than the deep CNN method of Galliani *et al.*

On CAVE data collection again our A+ results are better than Arad and comparable to Galliani. Interestingly this time the single pixel approach is better than the one using patches as input. Of course it could be that one has to slightly adjust the parameters k, m, c, λ to adapt for different images to reach an optimal performance. Another problem might be the fact, that the images contained a lot of dark areas, which probably reduced the useful amount of patches for the second approach, which usually needs more samples to avoid over- and underfitting.

On the NUS collection, the anchored learned spectral reconstruction methods performed better than all other methods.

From the results in Table 1 we conclude that our shallow method perform comparable or better than the very recent deep CNN approach of Galliani *et al.* and significantly better than the recent sparse coding method of Arad *et al.* and the relatively older approach of Nguyen *et al.*

Visual assessment In Figs. 3, 4 & 6 we show reconstructed intensity values by Arad (our implementation) and our A+ method (single pixel setting) in comparison with the ground truth at different wavelengths of hyperspectral images from ICVL, CAVE, and NUS dataset, respectively. The displayed crops of interesting areas show that our A+ method provides smoother and more accurate results handling noisy images better than Arad.

Runtime In table 2 one can see the time needed for training and also how long it takes to reconstruct an image from the ICVL database. Our A+ has a runtime ×4.3 lower than our efficient reimplementation of Arad's method. Considering that these measurements were done with a CPU, implemented on Matlab without any vast parallelism, the method could easily be accelerated. The same goes for the training time, which can be reduced by at least half with the single pixel approach. Simply reducing the sparsity for the dictionary to $k = 1$ would ensure a large speedup.

5. Conclusions

In this work we defended the shallow methods in comparison with the very recent Galliani *et al.* [13] proposed very deep CNN-based method for learned spectral reconstruction. For this, (i) first, we improved the recent sparse coding shallow method of Arad *et al.* [6] and achieved significantly better results than those originally reported; (ii) second, we introduced a novel shallow method based on A+ of Timofte *et al.* [33] from super-resolution that substantially improves over Arad and, moreover, provides comparable performance to Galliani on three standard benchmarks (ICVL, CAVE, and NUS); and (iii) finally, we argued that the efficiency in training and at runtime as well as the clear relation between its parameters and the achieved performance makes from our shallow A+ a strong baseline for further research in learned spectral reconstruction from RGB images. Our A+ shallow method requires significantly smaller train data than Galliani (and generally the CNN approaches), is robust to overfitting and can be easily deployed to newer cameras by fast training.

Acknowledgments

This work was supported by ETH Zurich General Fund (OK).

References

[1] F. M. Abed, S. H. Amirshahi, and M. R. M. Abed. Reconstruction of reflectance data using an interpolation technique. *JOSA A*, 26(3):613–624, 2009.

[2] J. Adams, M. Smith, and A. Gillespie. *Simple models for complex natural surfaces: a strategy for the hyperspectral era of remote sensing*. IGARSS, 1989.

[3] F. Agahian, S. A. Amirshahi, and S. H. Amirshahi. Reconstruction of reflectance spectra using weighted principal component analysis. *Color Research & Application*, 33(5):360–371, 2008.

[4] M. Aharon, M. Elad, and A. Bruckstein. k-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[5] K. Amano, S. Nascimento, M. Foster, et al. Frequency of metamerism in natural scenes. *Journal of the Optical Society of America A*, 2006.

[6] B. Arad and O. Ben-Shahar. Sparse recovery of hyperspectral signal from natural rgb images. In *European Conference on Computer Vision*, pages 19–34. Springer, 2016.

[7] E. Belluco, M. Camuffo, S. Ferrari, L. Modenese, S. Silvestri, A. Marani, and M. Marani. Mapping salt-marsh vegetation by multispectral and hyperspectral remote sensing. *Remote sensing of environment*, 105(1):54–67, 2006.

[8] M. Borengasser, W. S. Hungate, and R. Watkins. *Hyperspectral remote sensing: principles and applications*. Crc Press, 2007.

[9] G. J. Brelstaff, A. Párraga, T. Troscianko, and D. Carr. Hyperspectral camera system: acquisition and analysis. In *Satellite Remote Sensing II*, pages 150–159. International Society for Optics and Photonics, 1995.

[10] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson. Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. *IEEE Signal Processing Magazine*, 31(1):45–54, 2014.

[11] A. Chakrabarti and T. Zickler. Statistics of real-world hyperspectral images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 193–200. IEEE, 2011.

[12] D. T. Dicker, J. Lerner, P. Van Belle, D. Guerry, 4th, M. Herlyn, D. E. Elder, and W. S. El-Deiry. Differentiation of normal skin and melanoma using high resolution hyperspectral imaging. *Cancer biology & therapy*, 5(8):1033–1038, 2006.

[13] S. Galliani, C. Lanaras, D. Marmanis, E. Baltsavias, and K. Schindler. Learned spectral super-resolution. *arXiv preprint arXiv:1703.09470*, 2017.

[14] D. Haboudane, J. R. Miller, E. Pattey, P. J. Zarco-Tejada, and I. B. Strachan. Hyperspectral vegetation indices and novel algorithms for predicting green lai of crop canopies: Modeling and validation in the context of precision agriculture. *Remote sensing of environment*, 90(3):337–352, 2004.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[16] E. K. Hege, D. O'Connell, W. Johnson, S. Basty, and E. L. Dereniak. Hyperspectral imaging for astronomy and space surviellance. In *Optical Science and Technology, SPIE's 48th Annual Meeting*, pages 380–391. International Society for Optics and Photonics, 2004.

[17] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.

[18] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. *arXiv preprint arXiv:1611.09326*, 2016.

[19] R. Kawakami, H. Zhao, R. T. Tan, and K. Ikeuchi. Camera spectral sensitivity and white balance estimation from sky images. *International Journal of Computer Vision*, 105(3):187–204, 2013.

[20] T. Lillesand, R. W. Kiefer, and J. Chipman. *Remote sensing and image interpretation*. John Wiley & Sons, 2014.

[21] L. T. Maloney. Evaluation of linear models of surface spectral reflectance with small numbers of parameters. *JOSA A*, 3(10):1673–1683, 1986.

[22] R. M. Nguyen, D. K. Prasad, and M. S. Brown. Training-based spectral reconstruction from a single rgb image. In *European Conference on Computer Vision*, pages 186–201. Springer, 2014.

[23] S. E. Palmer. *Vision science: Photons to phenomenology*. MIT press, 1999.

[24] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40–44. IEEE, 1993.

[25] L. L. Randeberg, I. Baarstad, T. Løke, P. Kaspersen, and L. O. Svaasand. Hyperspectral imaging of bruised skin. In *Biomedical Optics 2006*, pages 60780O–60780O. International Society for Optics and Photonics, 2006.

[26] A. Robles-Kelly. Single image spectral reconstruction for multimedia applications. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 251–260. ACM, 2015.

[27] J. Romero, A. Garcıa-Beltrán, and J. Hernández-Andrés. Linear bases for representation of natural and artificial illuminants. *JOSA A*, 14(5):1007–1014, 1997.

[28] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016.

[29] G. N. Stamatas, C. J. Balas, and N. Kollias. Hyperspectral image acquisition and analysis of skin. In *Biomedical Optics 2003*, pages 77–82. International Society for Optics and Photonics, 2003.

[30] Y. Tarabalka, J. Chanussot, and J. A. Benediktsson. Segmentation and classification of hyperspectral images using watershed transformation. *Pattern Recognition*, 43(7):2367–2379, 2010.

[31] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July.

[32] R. Timofte, V. De Smet, and L. Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[33] R. Timofte, V. De Smet, and L. Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conference on Computer Vision*, pages 111–126. Springer, 2014.

[34] R. Timofte, R. Rothe, and L. Van Gool. Seven ways to improve example-based single image super resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[35] S. Wug Oh, M. S. Brown, M. Pollefeys, and S. Joo Kim. Do it yourself hyperspectral imaging with everyday digital cameras. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[36] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

[37] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar. Generalized assorted pixel camera. Technical report, Tech. Report, Department of Computer Science, Columbia University, 2008.

[38] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010.

[39] Y. Zhao and R. S. Berns. Image-based spectral reflectance reconstruction using the matrix r method. *Color Research & Application*, 32(5):343–351, 2007.