

Adversarial Networks for Spatial Context-Aware Spectral Image Reconstruction from RGB

Aitor Alvarez-Gila

TECNALIA / CVC - Universitat Autònoma de Barcelona
Derio, Spain

aitor.alvarez@tecnalia.com

Joost van de Weijer

CVC - Universitat Autònoma de Barcelona
Barcelona, Spain

joost@cvc.uab.es

Estibaliz Garrote

TECNALIA
Derio, Spain

estibaliz.garrote@tecnalia.com

Abstract

Hyperspectral signal reconstruction aims at recovering the original spectral input that produced a certain trichromatic (RGB) response from a capturing device or observer. Given the heavily underconstrained, non-linear nature of the problem, traditional techniques leverage different statistical properties of the spectral signal in order to build informative priors from real world object reflectances for constructing such RGB to spectral signal mapping. However, most of them treat each sample independently, and thus do not benefit from the contextual information that the spatial dimensions can provide. We pose hyperspectral natural image reconstruction as an image to image mapping learning problem, and apply a conditional generative adversarial framework to help capture spatial semantics. This is the first time Convolutional Neural Networks -and, particularly, Generative Adversarial Networks- are used to solve this task. Quantitative evaluation shows a Root Mean Squared Error (RMSE) drop of 44.7% and a Relative RMSE drop of 47.0% on the ICVL natural hyperspectral image dataset.

1. Introduction

Hyperspectral (HS) imaging has gained relevance over the last couple of years in the applied vision community. Remote sensing, UAV-based imaging, precision agriculture or autonomous driving are only some of the fields that are already benefiting from the use of imaging devices that provide a response that spans the spectral dimension with narrow-band channels to produce an image with higher spectral resolution than the standard RGB trichromatic one.

While the evolution of HS imaging devices has under-

gone major breakthroughs, it is also true that there is still a trade-off inherent to the fact that we are ultimately capturing three dimensional information with a two dimensional sensor, which limits the quality or resolution of the acquired signal in either of those dimensions: spatial, spectral or temporal. On top of that, the cost of such devices is orders of magnitude above that of conventional RGB cameras.

In this context, HS signal reconstruction from broadband or limited acquisition channels (typically, from RGB sensors) arises as a natural computational alternative, either to compete against native HS systems or to be included as part of their signal post-processing backends. The spectral reconstruction problem is a severely underconstrained, highly non-linear one, and the algorithms trying to solve this mapping should exploit the low dimensionality of the natural HS images [6] and learn informative priors of diverse forms from real world object reflectances, to be leveraged in the reconstruction phase. Note, however, that most of the existing solutions handle each pixel individually. By doing so, they are not taking advantage of the latent contextual information available in the spatially local neighborhood [6].

Generative adversarial Networks (GAN) are a class of neural networks which have shown to be able to successfully generate samples from the complex manifold of real images. In this work, we use this class of algorithms to learn a generative model of the joint spectro-spatial distribution of the data manifold of natural HS images and use it to optimally exploit spatial context information. To our knowledge, this is the first time Convolutional Neural Networks (CNN) are used in the task of spectral reconstruction of natural images. We quantitatively evaluate our approach on the largest HS natural image dataset available to date, i.e. ICVL, by comparing against [2], and show error drops

of 44.7% (RMSE) and 47.0% (relative RMSE) over their state of the art results.

1.1. Related work

A number of works are relevant to the proposed approach. This task was first addressed by isolating its spatial component and focusing on the reconstruction of homogeneous, well-established reflectances of real world surfaces such as Munsell chips, either from multispectral, RGB components [18] or from the tristimulus values [3, 1].

Initial attempts on the spectral reconstruction of natural images from full size RGB input required additional constraints or multiple input forms to help in their task: [22] and [5] use the aid of a low resolution HS measurement in addition to the RGB input, [28] restricts to the skylight samples domain, and [34, 35, 16], among others, rely on the aid of computational photography-like multiplexed narrow band lighting. The latter does, however, use spatial information for learning, as does [6], which focuses on the statistics for this class of images and defines a representation basis and computation method for the associated coefficients, but does not tackle reconstruction.

Solutions relying on a single RGB image input at test time are scarce, and almost none of them leverage the spatial context: [33] uses a Radial Basis Function network and produces an estimate of scene reflectance and global illuminant, but assumes a known camera color matching function, and directly depends on the performance of a white balancing stage as part of the workflow. [52] presents the *matrix R method* for spectral reflectance reconstruction, which additionally requires a calibration target to build a camera model. [2] learns a sparse dictionary of HS signatures as bases for the reconstruction. By treating each pixel independently, the ability to use the surround information is lost e.g. for producing distinct spectral outputs for metameric RGB pairs dependent on the context.

Remarkably, [40] exploits spatial material properties of the imaged objects by extracting not only spectral, but also convolutional features resulting from the application of the filter banks from [48], and adopting a constrained sparse coding-based reconstruction approach. In parallel to our development, we found a similar approach [15] which makes use of a CNN-based encoder-decoder to address this task.

Finally, there exists a certain relation between the HS reconstruction and the image colorization [7] tasks, which has been previously addressed in a similar fashion [51, 21], but under different evaluation requirements. We can think of the former being a generalization of the latter for an arbitrary number of input/output channels.

None of these methods would have been possible without the existence of publicly available HS natural image datasets. Until recently, the amount of images per set was the limiting factor for the development of HS reconstruc-

tion algorithms that learn on the basis of images or image patches [14, 50, 6, 33, 12, 13]. [2] changed this releasing a set of 201 high resolution images that we show is enough for the successful training of deep neural networks.

2. Adversarial spectral image reconstruction from RGB

This section describes the core functioning of our method, along with some of the mathematical developments that derived into the proposed models.

2.1. Adversarial learning

Generative Adversarial Networks (GANs) GAN-s [17] are generative statistical models that learn to produce realistic samples y that lay in the data manifold by relying on a setup consisting on two competing agents: the generator G takes noise z as input as a source of randomness, and creates *fake* data samples $G(z)$. It is trained to make the generated samples as realistic as possible. On the other end, the aim of the discriminator, D , which randomly takes as input both samples from the training data set and those generated by G , is to learn to tell if the received input samples are real or fake. Typically, both G and D are neural nets, and they are trained iteratively to progressively become better in their respective tasks. The objective function associated to such a setting is:

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_{y \sim p_{data}(y)} [\log D(y)] + \mathbb{E}_{z \sim p_{noise}(z)} [\log(1 - D(G(z)))] \quad (1)$$

where G tries to minimize this loss and D attempts to maximize it, yielding the objective function:

$$G^* = \underset{G}{\operatorname{argmin}} \max_D \mathcal{L}_{GAN} \quad (2)$$

This adversarial framework has successfully been applied to the unsupervised generation of data of different modalities, including natural images [11], and empirical architecture guidelines for G and D have been derived [37] for such cases, along with common tricks to stabilize the training process [43].

Conditional Generative Adversarial Networks (cGANs) cGANs [31] extend this framework by feeding both G and D with additional information x to be used to condition on the output of the generator. Such conditioning input could adopt different modalities, and range from simple categorical labels [31] to more sophisticated content, such as text [38] or images [27], either alone or as a combination of multiple input modalities [39, 53]. This has been proved useful for a number of tasks and output types [49, 30]. Eq. 3 shows the updated loss function for conditional GANs. In this case, G attempts to generate images that look realistic

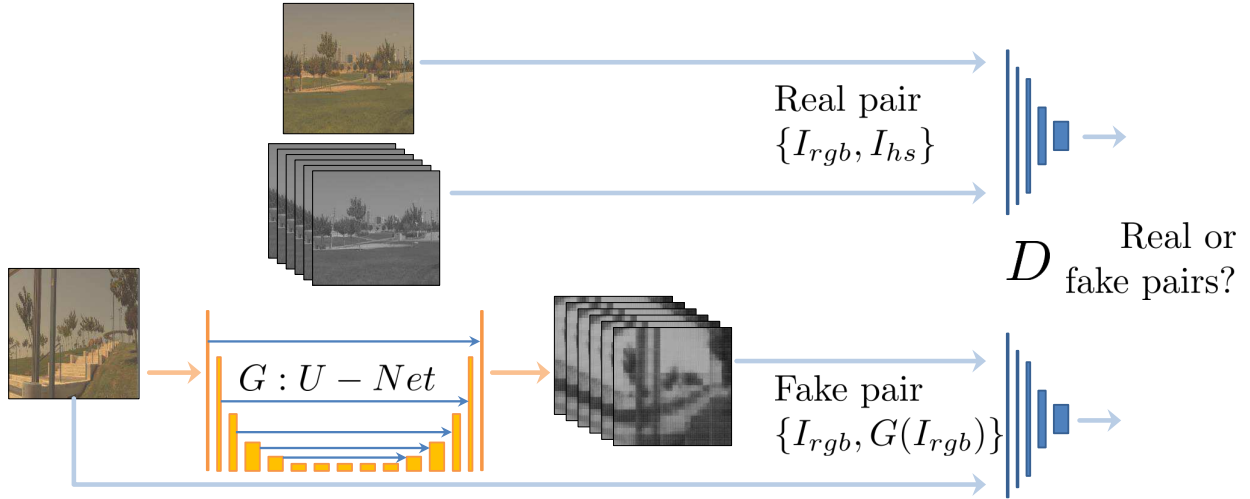


Figure 1. Adversarial spatial context-aware spectral image reconstruction model

given the additional provided input x (be it the class of y , a descriptive text, or an additional image), and D tries to determine whether the given (x, y) pair makes sense or not as a mapping.

$$\begin{aligned} \mathcal{L}_{cGAN}(G, D) = & \\ = \mathbb{E}_{x, y \sim p_{data}(x, y)} [\log D(x, y)] + & \\ + \mathbb{E}_{x \sim p_{data}(x), z \sim p_{noise}(z)} [\log(1 - D(x, G(x, z)))] \quad (3) \end{aligned}$$

As a result, cGANs open the door to using generative statistical modeling for our HS reconstruction problem by conditioning the generation of an HS outcome on a given input RGB image.

Adversarial image to image mapping Many modern computer vision tasks can better be regarded under the common reference framework of image to image mapping learning, in which a generator model G is learned that translates an input image x into the most probable representation y of such image in the output domain. This is the case *e.g.* for semantic segmentation [45], instance segmentation [10], or depth and surface normal estimation from single image [4], among others. Most of these tasks have been recently addressed making use of Convolutional Neural Networks that yield deterministic results as generators, and which are specifically tailored, in terms of architecture design, objective function or other specific training details, for their respective tasks.

There are, in addition, some tasks for which this mapping is not unique, and one same input image could have multiple equally correct representations in the output domain. Realistic image rendering from semantically labeled images (inverse of the semantic segmentation problem) or from hand-drawn sketches, or image colorization [7], are

just a few examples of this. The choice of the objective functions to use in each of these cases is a particularly challenging design aspect; applying an otherwise useful ℓ_2 loss to x, y image pairs is known to be problematic and yield blurry results [26], as the generator tends to average over the space of valid image representations.

For all of the above, [21] proposes a common image to image mapping learning framework based on the cGAN adversarial setting, which, provided that one can feed it with co-registered image pairs of input and output domains, is able to learn the most suitable loss function for each of the tackled tasks in a data-driven approach. This is done implicitly using the adversarial objective from eq. 3, enforced by the discriminator trying to identify the fake images and, this way, encouraging the generator to become better at trying to deceive it.

By doing this, [21] manages to get rid of the blur inherent to ℓ_2 distance-based models and produce sharp results. Nevertheless, it has been previously shown [36, 46] that combining one of the traditional loss functions with the adversarial objective \mathcal{L}_{cGAN} can help produce more spatially consistent results and make the generator less prone to artifacts inherent to the adversarial scheme. They thus place an additional ℓ_1 term (eq. 4) on the generator, which is known to yield less blur:

$$\mathcal{L}_{\ell_1}(G) = \mathbb{E}_{x, y \sim p_{data}(x, y), z \sim p_{noise}(z)} [\|y - G(x, z)\|_1] \quad (4)$$

and produce the following combined objective function:

$$G^* = \underset{G}{\operatorname{argmin}} \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{\ell_1}(G) \quad (5)$$

where λ is a weighting factor for the ℓ_1 term, which is set to 100 in [21]. In essence, $\mathcal{L}_{cGAN}(G, D)$ would be in charge

of producing sharp, realistic looking results, while ℓ_1 takes care of the global image structure.

Interestingly, the stochastic output pursued by the noise input to cGAN-like models does not manifest itself under this design (see details in section 2.2), and the resulting mapping is a fundamentally deterministic one. A probable interpretation is G learning to ignore the effect of the noise. As a result, [21] gets rid of the noise input and leaves test-time dropout as unique source of randomness.

Adversarial spectral reconstruction networks The forward correspondence learning between the RGB and hyperspectral signals is a heavily under-constrained one, which could benefit from an approach that aims at exploiting the underlying priors present in both the spectral and spatial dimensions and learn a model that specifically produces realistic outcomes as a target. It not only requires mapping a 3-dimensional image to a much higher dimensional one (typically 31 spectral channels and the two spatial dimensions), but such mapping can be context-dependent as well, as is in the case of metameric colors. The inverse mapping, however, i.e. the rendition of RGB images from their spectral counterparts, is well defined, and deterministic under the only assumption of the color matching functions defining the observer, or the spectral sensitivity functions that characterize specific sensors. This makes it immediate to generate perfectly aligned (RGB, hyperspectral) image pairs (see section 3) to be used under the described solution.

Hyperspectral image reconstruction from RGB can then be posed as one of the aforementioned image to image mapping learning problems and thus be solved under the conditional adversarial network-based image to image translation framework proposed by [21].

The resulting adversarial and combined objectives would then become:

$$\mathcal{L}_{adv} = \mathbb{E}_{I_{rgb}, I_{hs} \sim p_{data}(I_{rgb}, I_{hs})} [\log D(I_{rgb}, I_{hs})] + \mathbb{E}_{I_{rgb} \sim p_{data}(I_{rgb})} [\log(1 - D(I_{rgb}, G(I_{rgb})))] \quad (6)$$

$$\mathcal{L}_{rgb2hs}(G, D) = \mathcal{L}_{adv} + \lambda \mathcal{L}_{\ell_1} = \mathcal{L}_{adv} + \lambda \mathbb{E}_{I_{rgb}, I_{hs} \sim p_{data}(I_{rgb}, I_{hs})} [\|I_{hs} - G(I_{rgb})\|_1] \quad (7)$$

where I_{hs} represents the original hyperspectral image, I_{rgb} is the corresponding input image in the RGB domain and λ is scalar weight used to balance both loss terms (and is set to 100 in all our experiments, unless otherwise stated). Note that we have explicitly removed any reference to the input noise, and the RGB image remains as the only input to G .

Figure 1 shows an overview of the whole adversarial spatial context-aware spectral image reconstruction process. We depart from a database of perfectly aligned RGB and hyperspectral image pairs, which are extracted one pair at a time. In a first iteration, a first pair of real images of size

$H \times W$ is taken: $\{I_{RGB}, I_{HS}\}$. The generator G takes I_{RGB} as input, and yields the corresponding hyperspectral reconstruction of size $H \times W$, \hat{I}_{HS} . The discriminator D is now fed with two pairs of images, $\{I_{RGB}, I_{HS}\}$ and $\{I_{RGB}, \hat{I}_{HS}\}$ and uses the associated labels indicating if they are real or fake $\{1, 0\}$ to compute the adversarial loss and update its gradients. G 's weights are also updated, and both D and G continue to become better at their respective tasks iteratively.

2.2. Architecture design and training

As for the specific implementation of the models, since G needs to yield full-size detailed images, a *U-Net*-like architecture [42] is used. Regular autoencoder networks [25] exhibit a progressively reduced representation size until a bottleneck layer and there is no way for the last layers of accessing the original data, which negatively affects the results when we aim at detailed outcomes. Unlike these, the *U-Net* incorporates skip connections between layers of equal representation size, and concatenates local activations from the upscaling phase with those coming from the downscaling stages, which has shown to achieve superior performance on tasks where the details are relevant. It was first proposed with semantic segmentation tasks in mind, but original spectral signal reconstruction falls within the kind of tasks that can clearly benefit from accessing the original input levels at each sample (i.e. pixel).

The discriminator D , defined as *PatchGAN*, is simpler in terms of convolutional layer count, and is focused solely on modeling high-frequency structure. Each of the $M \times M$ output neurons is restricted to see only a limited $N \times N$ receptive field from the input image, which can be significantly smaller than the input image size. Consequently, only the adversarial loss term is placed over D (eq 5).

The use of this design solution for D is consistent with our initial hypothesis that local spatial context can help better reconstruct the spectral signal. Specifically, we hypothesize that the proposed approach could help disentangling the illuminant and object body reflectance components of a pixel's trichromatic response, as defined by the dichromatic reflection model [44]. The design of D , with its attached ℓ_1 objective, helps capture the high frequencies that characterize the textures in the image. These are, together with the body color component, one of the main features characteristic of the different materials which, ultimately, produce distinct spectral responses. Therefore, convolutionally integrating the trichromatic response of adjacent pixels should yield a better estimate of the central spectral response. To this respect, the *PatchGAN* design isolates D 's response associated to pixels separated by more than one input patch. For small enough patch sizes, this effectively implies that the discriminator is learning a loss function tailored for texture or material recognition, making sure that the recon-



Figure 2. Random RGB samples from the ICVL dataset [2].

structed spectra falling within the patch are not only plausible in the spectral domain, but also spatially consistent in the close proximities.

The illuminant-specific component of [44], on the other hand, is typically largely constant or slowly varying across big portions of the image (especially in terms of chromaticity and conversely, spectral shape), and the ℓ_1 norm does a good job taking care of its global image-wide consistency, along with that of the low-mid frequency spatial structures.

Avoiding Batch Normalization Given the intrinsically exact nature of our task (some of the described design choices help leverage spatial structure consistency for our task, but we ultimately want the reconstructed spectra to be accurate), we choose to remove all the Batch Normalization [20] layers present in the generator architectures proposed in [21]. While this technique has shown to be useful to help accelerate and regularize the training process for a wide variety of tasks by reducing the internal covariate shift, the fact that it makes the signal lose track of its original value, along with the deterministic nature of the desired output, makes it non-advisable for reconstruction tasks. We experimentally found that including Batch Normalization produced inferior results.

2.3. Implementation details

We now provide some details on the configurations used for our implementation. We use Keras with Theano backend and take the implementation of [21] made by [9] as starting point, modifying it for our purposes. We use Adam optimizer [24] for both G and D , with a learning rate of $2 \cdot 10^{-4}$ and $\beta_1 = 0.5$. We use a minibatch size of 1 in order to benefit from the regularization provided by the gradient estimation noise [23], and following common practice [21]. The training is performed iteratively and alternates between the two models: at each step, the discriminator is first trained for 50 iterations and then the generator gets trained for 25 more minibatches.

We crop the original 1392×1300 images during the training phase by extracting one random crop of size 256×256 (the H, W values from section 2) per image and epoch. The models are fed with these crops during training, while, for the testing phase, each full size RGB image is divided in tiles of 256×256 with no overlap, which effectively yields image sizes of 1280×1280 pixels. Each tile gets processed by the generator independently and we reconstruct the full image back before evaluating it.

The generator G accepts input images of size 256×256 . Its encoding stage is composed by eight successive 3×3 convolutions with stride 2 and a leaky ReLU after each of them, thus yielding a 1×1 activation in the most narrow point of the main branch. The initial number of filters is 64, which gets doubled at each convolutional layer up to 512, keeping it constant after that. On the decoding part, eight transposed convolution blocks successively double the activation size up until the original 256×256 size, while progressively reducing the number of filters in a symmetric way with respect to the encoding stage. Each block comprises the transposed convolution itself, followed by a train-time-only Dropout layer (with a drop rate of 10%) and a leaky ReLU activation. After each Dropout, the correspondent activations from the encoding stage are concatenated, thus producing eight skip connections between levels of equivalent activation size. Finally, two 1×1 convolutions are added at the end before the output \tanh activation, with a leaky ReLU in between, in order to get the direct input images adequately combined with the upstream features.

The discriminator D is a simple single-branch net composed of four 3×3 convolutional layers with stride 2, each of them followed by a leaky ReLU, with filter numbers doubling at each step. A fifth 3×3 convolution with a sigmoid yields the output 8×8 prediction.

3. Experimental evaluation

This section contains an overview of the experiments performed to quantitatively assess our algorithm's performance as compared to previous methods.

3.1. Dataset

Given the amount of images, diversity and resolution, we evaluate our approach on the dataset presented in [2]. At the time of writing, it comprised 201 hyperspectral images (see Figure 2 for RGB renditions of a few random samples) of 1392×1300 spatial resolution and 519 spectral bands in the $400nm - 1000nm$ range, with a spectral resolution of $1.25nm$. As for the acquisition, a *Specim PS Kappa DX4* hyperspectral camera was used, together with a rotary stage for spatial scanning. This aspect is noticeable in some of the samples, in which common objects such as cars exhibit aspect ratios that do not match those we find in real life. There is also a spectrally downsampled version of 31 bands in the $400nm - 700nm$ range. Following practice from [2], we use the latter for our reconstruction experiments. There is no illuminant information available for each of the images, which would allow for object reflectance recovery; therefore, our task consists on the estimation of the radiance correlate represented by the captured hyperspectral images.

3.2. Preparation

In order to get the aligned image pairs dataset required by our method, and given the deterministic correspondence between spectral and RGB samples once the observer (or sensor sensitivity functions) and the output color space are specified, we render wide band trichromatic RGB versions of the spectral images in the sRGB color space as follows: we first obtain the CIE XYZ tristimulus values for each spectral image pixel location x , making use of the color matching functions corresponding to the CIE 1964 10° standard observer:

$$\mathbf{X}(x) = K(x) \sum_{\lambda=400nm}^{700nm} S(\lambda, x) \bar{\mathbf{x}}(\lambda) \Delta\lambda \quad (8)$$

where $S(\lambda, x)$ is the relative spectral power distribution of pixel x , $\mathbf{X} = \{X, Y, Z\}$, $\bar{\mathbf{x}}(\lambda) = \{\bar{x}(\lambda), \bar{y}(\lambda), \bar{z}(\lambda)\}$ are the color matching functions, $\Delta\lambda = 10nm$ and $K(x)$ is the normalization factor, defined, for illuminant $L(\lambda, x)$, as:

$$K(x) = \frac{100}{\sum_{\lambda=400nm}^{700nm} L(\lambda, x) \bar{y}(\lambda) \Delta\lambda} \quad (9)$$

Note that, before going through this computation, the original spectral power distribution captured by the camera for each image $S'(\lambda, x)$ is preprocessed with min value subtraction and max value scaling. The final step is producing the sRGB renders. We do so by applying the associated

3×3 transformation matrix and unlinearizing (i.e. *gamma-correcting*) the result with a $1/2.4$ power law gamma with a linear segment in low luminance values.

While not suffering from the same lack of an adequate performance evaluation method that affects typical generative modeling tasks [47], spectral signal reconstruction algorithms assessment is an active research field that lacks consensus on what is the most adequate metric to measure spectral match of signals [19]. When the signals comprise the visual spectrum, the task can be tackled from a variety of perspectives, ranging from the pure signal processing point of view of spectral curve difference metrics, to a full spectrum of metric families that place different levels of perceptual load on their computation: metameric indexes, CIE ΔE color difference equations, or weighted spectral metrics.

If we widen the scope onto full reference image difference metrics, little work has been done on the spectral extension of these families [32]. We here focus on four of the most widely used metrics, namely RMSE (Root Mean Squared Error, computed across the spectral dimension for each pixel and then averaging for whatever number of pixels present in the image or the dataset), RMSERel (i.e. RMSE relative to the value of the real signal), GFC (Goodness of Fit Coefficient [41]) and ΔE_{00} (CIEDE2000) perceptual color difference formula [8] computed over the reconstructed tristimulus values.

3.3. Experiments and discussion

We compare our method with the only other one having reported on the ICVL dataset, i.e. [2]. In their general experiment over the whole set (we do not have enough samples in each of the domain-specific ones that they define to be able to learn), which was back then composed of 100 images, they perform a leave-one-out procedure, and learn from pixels sampled along the whole set except for the unique image being tested at a time. We choose to split the dataset in two equal partitions of 100 images each, training on one and reporting on the other, running two full train-test cycles and averaging the results across folds and runs. Table 1 compares the obtained values for the aforementioned metrics over each of the testing sets, showing an average per-pixel error drop of 44.7% in terms of RMSE and 47.0% in terms of RMSERel with respect to [2]. While [2] does not provide any further evaluation metric, note that our average GFC values are all above the GFC threshold which [41] considers a *very good reconstruction*, and one which implies missing only 0.2% of the signal energy in the process. Also, the average per-pixel color difference (which does not account for spatial perceptual effects) is constrained around as low as $2\Delta E_{00}$ units.

Figure 3 shows the sRGB rendition of original and reconstructed hyperspectral images for some randomly chosen test image samples. In addition, for each image, we

| Method | RMSE | RMSERel | GFC | ΔE_{00} |
|-----------------------------|-------------------------------------|---------------------------------------|---|-------------------------------------|
| Arad <i>et al.</i> [2] | 2.633 | 0.0756 | - | - |
| Ours (weighted avg.) | 1.457 ± 0.040 | 0.0401 ± 0.0024 | 0.99921 ± 0.00012 | 2.044 ± 0.341 |
| Ours (fold 0) | 1.452 ± 0.101 | 0.0383 ± 0.0024 | 0.99906 ± 0.00001 | 1.861 ± 0.324 |
| Ours (fold 1) | 1.463 ± 0.022 | 0.0420 ± 0.0024 | 0.99936 ± 0.00023 | 2.228 ± 0.358 |

Table 1. Summary results of the conducted experiments over ICVL dataset. Black pixels contained in the original hyperspectral images (derived from the variable image width) are not taken into account for evaluation purposes in any of the experiments, and folds are weighted accordingly. RMSE values are in the $[0 - 255]$ range. Two train-test cycles were run and the results averaged.

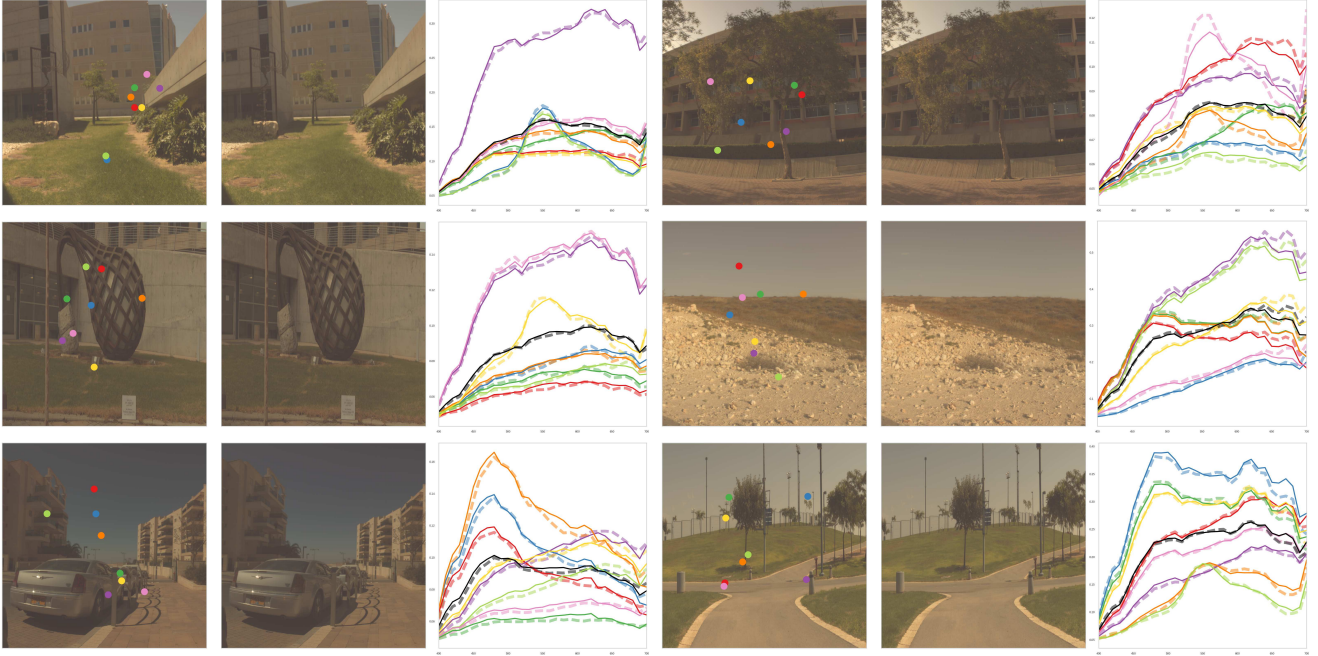


Figure 3. Sample results for our method. For each triplet, left, center: sRGB rendition of original and reconstructed hyperspectral signals, respectively. Right: Original (dashed) and reconstructed (solid) spectra of eight random pixels identified by the colored dots.

show the original and estimated spectra for eight randomly selected pixels from the image.

3.3.1 Does the spatial information actually help?

In an attempt to empirically validate our main hypothesis of contextual spatial information on a local neighborhood being relevant for the correct spectral reconstruction of any given central pixel, we conduct a branch pruning experiment. We depart from a minimal version of our net, in which both the main branch and all the skip connections have been removed, except for the one connecting the 256×256 input with the last pair of 1×1 convolutions (such model predicts each output pixel independently by design, without incorporating any spatial contribution), and keep adding skip connections at successively deeper levels (extending the receptive field of the model and thus increasing the spatial contribution at each step) until we end up with the full net, after the addition of the

main 1×1 stream branch. Figure 4 shows the results of running at least two train-test cycles on each of these nets, and testing over the 1280×1280 versions of the images in fold 1. All the four metrics show a closely correlated outcome, with a very significant average performance improvement (-20.8% RMSE, -23.5% RMSERel, -47.1% ΔE_{00}) when transitioning from the model with a single skip connection and a 1×1 receptive field to that with 2 skip connections and a 3×3 receptive field. Further increases of the model’s theoretical receptive field (by adding new branches) yield only marginally better results (models labeled as $3/7 \times 7$, $4/15 \times 15$) and, from there on, additional deeper skip connections produce increasing test error rates. We hypothesize that this is due to the influence of overfitting for experiments $5/31 \times 31$ and onwards. Given this, a straightforward way of improving the reported results could be that of increasing the regularization associated to the deepest branches by, e.g., increasing their dropout rate.

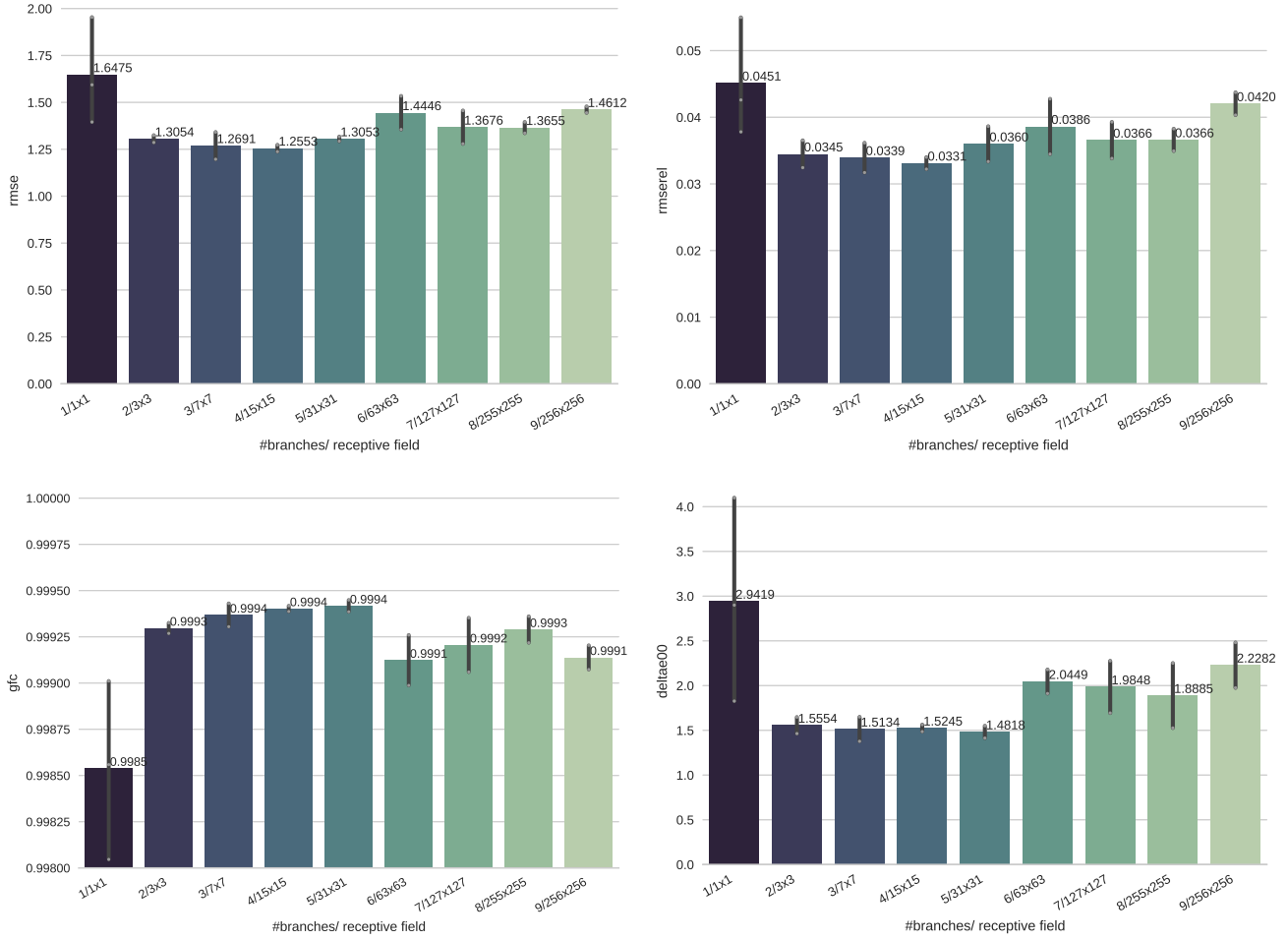


Figure 4. Branch pruning experiment results. Top-left: RMSE. Top-right: RMSERel. Bottom-left: GFC. Bottom-right: ΔE_{00} . Leftmost bar is the model with a single skip connection at 256×256 activation size level and 1×1 receptive field (RF). Each additional bar adds one skip connection at increasingly deeper levels of the U-Net. The rightmost bar is the full net, resulting from the addition of the main branch, and its RF (which would be 512×512 in an unconstrained scenario), is here limited by the 256×256 patch size. The addition of this last layer is justified by the notion of effective RF presented in [29], which may be significantly smaller than its theoretical counterpart.

It is also noticeable the substantially higher variance that the results seem to suggest for the model with a 1×1 receptive field. This would mean that the addition of local spatial information does not only improve the overall prediction accuracy, but it does so in a more robust manner as well.

4. Conclusion

We propose a convolutional neural network architecture that successfully learns an end-to-end mapping between pairs of input RGB images and their hyperspectral counterparts. We adopt an adversarial framework-based generative model that shows itself effective in capturing the structure of the data manifold, and takes into account the spatial contextual information present in RGB images for the spectral

reconstruction process. State of the art results in the ICVL dataset suggest that individual pixel-based approaches suffer from the fundamental limitation of not being able to effectively exploit the local context when applied to spectral image data in their attempt to build informative priors. The observed performance in terms of both reconstruction error and speed open the door to a full range of potential higher level applications in sectors of increasing demand for spectral footage at a lower cost.

Acknowledgements This work has been partially funded by the Ministerio de Economía y Competitividad of Spain under grant number DPI2015-64571-R. We also acknowledge the Spanish project TIN2016-79717-R, and the CHIST-ERA project M2CR (PCIN-2015-251).

References

- [1] 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, Oct. 2008. 2
- [2] B. Arad and O. Ben-Shahar. Sparse Recovery of Hyperspectral Signal from Natural RGB Images. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII*, pages 19–34. Springer International Publishing, Cham, 2016. 1, 2, 5, 6, 7
- [3] F. Ayala, J. F. Echávarri, P. Renet, and A. I. Negueruela. Use of three tristimulus values from surface reflectance spectra to calculate the principal components for reconstructing these spectra by using only three eigenvectors. *JOSA A*, 23(8):2020–2026, Aug. 2006. 2
- [4] A. Bansal, X. Chen, B. Russell, A. Gupta, and D. Ramanan. PixelNet: Representation of the pixels, by the pixels, and for the pixels. *arXiv:1702.06506 [cs]*, Feb. 2017. 3
- [5] X. Cao, X. Tong, Q. Dai, and S. Lin. High resolution multi-spectral video capture with a hybrid camera system. In *CVPR 2011*, pages 297–304, June 2011. 2
- [6] A. Chakrabarti and T. Zickler. Statistics of Real-World Hyperspectral Images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 193–200, 2011. 1, 2
- [7] Z. Cheng, Q. Yang, and B. Sheng. Deep Colorization. In *The IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015. 2, 3
- [8] CIE. CIE 142-2001 Improvement to Industrial Colour-Difference Evaluation. Technical Report CIE 142-2001, Commission Internationale de L’éclairage, Vienna, 2001. ISBN 978 3 901906 08 4. 6
- [9] P. Costa, A. Galdran, M. I. Meyer, M. D. Abràmoff, M. Niemeijer, A. M. Mendonça, and A. Campilho. Towards Adversarial Retinal Image Synthesis. *arXiv:1701.08974 [cs, stat]*, Jan. 2017. 5
- [10] J. Dai, K. He, and J. Sun. Instance-Aware Semantic Segmentation via Multi-Task Network Cascades. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3
- [11] E. L. Denton, S. Chintala, a. szlam, and R. Fergus. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1486–1494. Curran Associates, Inc., 2015. 2
- [12] J. Eckhard, T. Eckhard, E. M. Valero, J. L. Nieves, and E. G. Contreras. Outdoor scene reflectance measurements using a Bragg-grating-based hyperspectral imager. *Applied Optics*, 54(13):D15–D24, May 2015. 2
- [13] D. H. Foster, K. Amano, and S. M. Nascimento. Time-lapse ratios of cone excitations in natural scenes. *Vision Research*, 120:45–60, Mar. 2016. 2
- [14] D. H. Foster, K. Amano, S. M. C. Nascimento, and M. J. Foster. Frequency of metamerism in natural scenes. *Journal of the Optical Society of America A*, 23(10):2359, Oct. 2006. 2
- [15] S. Galliani, C. Lanaras, D. Marmanis, E. Baltsavias, and K. Schindler. Learned Spectral Super-Resolution. *arXiv:1703.09470 [cs]*, Mar. 2017. 2
- [16] M. Goel, E. Whitmire, A. Mariakakis, T. S. Saponas, N. Joshi, D. Morris, B. Guenter, M. Gavrilu, G. Borriello, and S. N. Patel. HyperCam: Hyperspectral Imaging for Ubiquitous Computing Applications. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp ’15*, pages 145–156, New York, NY, USA, 2015. ACM. 2
- [17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 2
- [18] V. Heikkinen, R. Lenz, T. Jetsu, J. Parkkinen, M. Hauta-Kasari, and T. Jäskeläinen. Evaluation and unification of some methods for estimating reflectance spectra from RGB images. *JOSA A*, 25(10):2444–2458, Oct. 2008. 2
- [19] F. H. Imai, M. R. Rosen, and R. S. Berns. Comparative study of metrics for spectral match quality. In *Conference on Colour in Graphics, Imaging, and Vision*, volume 2002, pages 492–496. Society for Imaging Science and Technology, 2002. 6
- [20] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 448–456, 2015. 5
- [21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-To-Image Translation With Conditional Adversarial Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 3, 4, 5
- [22] R. Kawakami, Y. Matsushita, J. Wright, M. Ben-Ezra, Y. W. Tai, and K. Ikeuchi. High-resolution hyperspectral imaging via matrix factorization. In *CVPR 2011*, pages 2329–2336, June 2011. 2
- [23] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *arXiv:1609.04836 [Cs, Math]*, Toulon, FR, Apr. 2017. 5
- [24] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, USA, 2015. 5
- [25] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2013. 4
- [26] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1558–1566, New York, New York, USA, 20–22 Jun 2016. PMLR. 3

- [27] C. Li and M. Wand. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III*, pages 702–716. Springer International Publishing, Cham, 2016. 2
- [28] M. A. López-Álvarez, J. Hernández-Andrés, J. Romero, F. J. Olmo, A. Cazorla, and L. Alados-Arboledas. Using a trichromatic CCD camera for spectral skylight estimation. *Applied Optics*, 47(34):H31–H38, Dec. 2008. 2
- [29] W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In D. D. Lee, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances In Neural Information Processing Systems 29*, pages 4898–4906. Curran Associates, Inc., 2016. 8
- [30] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In *International Conference on Learning Representations (ICLR 2016)*. arXiv preprint arXiv:1511.05440, 2015. 2
- [31] M. Mirza and S. Osindero. Conditional generative adversarial nets. In *arXiv Preprint arXiv:1411.1784*, 2014. 2
- [32] S. L. Moan and P. Urban. Image-Difference Prediction: From Color to Spectral. *IEEE Transactions on Image Processing*, 23(5):2058–2068, May 2014. 6
- [33] R. M. H. Nguyen, D. K. Prasad, and M. S. Brown. Training-Based Spectral Reconstruction from a Single RGB Image. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision ECCV 2014*, number 8695 in Lecture Notes in Computer Science, pages 186–201. Springer International Publishing, Sept. 2014. 2
- [34] J. I. Park, M. H. Lee, M. D. Grossberg, and S. K. Nayar. Multispectral Imaging Using Multiplexed Illumination. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, Oct. 2007. 2
- [35] M. Parmar, S. Lancel, and B. A. Wandell. Spatio-spectral reconstruction of the multispectral datacube using sparse recovery. In *2008 15th IEEE International Conference on Image Processing*, pages 473–476, Oct. 2008. 2
- [36] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context Encoders: Feature Learning by Inpainting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3
- [37] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, 2016. 2
- [38] S. Reed, Z. Akata, X. Yan, L. Logeswaran, H. Lee, and B. Schiele. Generative Adversarial Text to Image Synthesis. In *International Conference on Machine Learning (ICML)*, 2016. 2
- [39] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning What and Where to Draw. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 217–225. Curran Associates, Inc., 2016. 2
- [40] A. Robles-Kelly. Single Image Spectral Reconstruction for Multimedia Applications. In *ACM Multimedia*, 2015. 2
- [41] J. Romero, A. García-Beltrán, and J. Hernández-Andrés. Linear bases for representation of natural and artificial illuminants. *JOSA A*, 14(5):1007–1014, 1997. 6
- [42] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer, Cham, Oct. 2015. 4
- [43] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen. Improved Techniques for Training GANs. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2226–2234. Curran Associates, Inc., 2016. 2
- [44] S. A. Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, Dec. 1985. 4, 5
- [45] E. Shelhamer, J. Long, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2016. 3
- [46] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning From Simulated and Unsupervised Images Through Adversarial Training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3
- [47] L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, 2016. 6
- [48] M. Varma and A. Zisserman. Classifying Images of Materials: Achieving Viewpoint and Illumination Independence. In *Computer Vision – ECCV 2002*, Lecture Notes in Computer Science, pages 255–271. Springer, Berlin, Heidelberg, May 2002. 2
- [49] X. Wang and A. Gupta. Generative Image Modeling Using Style and Structure Adversarial Networks. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV*, pages 318–335. Springer International Publishing, Cham, 2016. 2
- [50] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar. Generalized Assorted Pixel Camera: Postcapture Control of Resolution, Dynamic Range, and Spectrum. *IEEE Transactions on Image Processing*, 19(9):2241–2253, Sept. 2010. 2
- [51] R. Zhang, P. Isola, and A. A. Efros. Colorful Image Colorization. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III*, pages 649–666. Springer International Publishing, Cham, 2016. 2
- [52] Y. Zhao and R. S. Berns. Image-based spectral reflectance reconstruction using the matrix R method. *Color Research & Application*, 32(5):343–351, Oct. 2007. 2

- [53] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative Visual Manipulation on the Natural Image Manifold. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, pages 597–613. Springer International Publishing, Cham, 2016. [2](#)