NTIRE 2020 Challenge on Spectral Reconstruction from an RGB Image

Boaz Arad Radu Timofte Ohad Ben-Shahar Yi-Tun Lin Graham Finlayson
Shai Givati Jiaojiao Li Chaoxiong Wu Rui Song Yunsong Li Fei Liu
Zhiqiang Lang Wei Wei Lei Zhang Jiangtao Nie Yuzhi Zhao Lai-Man Po
Qiong Yan Wei Liu Tingyu Lin Youngjung Kim Changyeop Shin
Kyeongha Rho Sungho Kim Zhiyu ZHU Junhui HOU He Sun Jinchang Ren
Zhenyu Fang Yijun Yan Hao Peng Xiaomei Chen Jie Zhao Tarek Stiebel
Simon Koppers Dorit Merhof Honey Gupta Kaushik Mitra Biebele Joslyn Fubara
Mohamed Sedky Dave Dyke Atmadeep Banerjee Akash Palrecha Sabarinathan
K Uma D Synthiya Vinothini B Sathya Bama S M Md Mansoor Roomi

Abstract

This paper reviews the second challenge on spectral reconstruction from RGB images, i.e.*, the recovery of whole-scene hyperspectral (HS) information from a 3-channel RGB image. As in the previous challenge, two tracks were provided: (i) a "Clean" track where HS images are estimated from noise-free RGBs, the RGB images are themselves calculated numerically using the ground-truth HS images and supplied spectral sensitivity functions (ii) a "Real World" track, simulating capture by an uncalibrated and unknown camera, where the HS images are recovered from noisy JPEG-compressed RGB images. A new, larger-than-ever, natural hyperspectral image data set is presented, containing a total of 510 HS images. The Clean and Real World tracks had 103 and 78 registered participants respectively, with 14 teams competing in the final testing phase. A description of the proposed methods, alongside their challenge scores and an extensive evaluation of top performing methods is also provided. They gauge the state-of-the-art in spectral reconstruction from an RGB image.*

1. Introduction

While conventional color cameras record scene spectral radiance integrated three spectral bands (red, green, and blue), hyperspectral imaging systems (HISs) can record the actual scene spectra over a large set of narrow spectral bands [14]. However, the rich, spectral, information provided by HISs comes with significant additional capture complexity: most common HISs rely on either spatial or spectral scanning (*e.g.* push-broom or variable-filter systems) and hence are unsuitable for real-time operation. Moreover, hyperspectral capture often requires a longer capture time and this means it is difficult to measure information from scenes with moving content. Although, recent advances in "Snapshot" HISs have continued to bridge the gap towards real-time spectral image acquisition - *e.g.* Mosaic [54, 21, 23] and light-field [11] based snapshot HISs can capture images at video-rates - these technologies record images with reduced spatial and spectral resolution. To date, both scanning and snapshot HISs remain prohibitively expensive for consumer grade use ("low-cost" HISs are often in the $10K-$100K range).

Due to these drawbacks of HISs, there has been a lot of research and industrial interest in developing methods for recovering spectra from the images of low cost and ubiquitous RGB cameras. Early work on RGB spectral recovery images leveraged sparse coding methods to recover HS data [6, 46, 2, 55]. In recent years, neural-net based methods have become more common [19, 13, 29, 7, 51, 50, 31, 49], with leading methods from the NTIRE 2018 spectral recovery challenge [7, 49] as well as more recent works [65, 37, 29, 16] adopting this approach. This transition to neural-net based methods highlights the need for larger data sets - both to facilitate improved training as well as improved evaluation. The latter consideration is crucial as neural-nets are prone to "overfitting" on small data sets and thus their test scores may not generalize well to real-world applications.

The inherent difficulty in evaluation neural-net based solutions was recently highlighted by Yi-Tun and Fin-

B. Arad (boazar@post.bgu.ac.il, Voyage81 & Ben-Gurion University of the Negev), R. Timofte, O. Ben-Shahar, Y.-T. Lin, G. Finlayson, S. Givati are the NTIRE 2020 challenge organizers, while the other authors participated in the challenge.
Appendix A contains the authors' teams and affiliations.
NTIRE 2020 webpage:
https://data.vision.ee.ethz.ch/cvl/ntire20/

layson [33] which evaluated top performing solutions from the NTIRE 2018 challenge under variable illumination conditions. Surprisingly, simply varying the brightness of input images (simulating longer/shorter camera exposures of the same scene) degraded the performance of neural net based methods to the point that they were outperformed by sparse-coding based methods (because the evaluated sparse-coding based methods were exposure invariant). Concomitantly, in Section 3 we present an extended evaluation of top performing methods - which includes the variable exposure test - to more thoroughly review the algorithms' performance.

Following NTIRE 2018, two potential experimental evaluation issues were identified and thus addressed here. First, the top performing methods in the NTIRE 2018 challenge obtained a percentage recovery error of about 1% mean relative absolute error (MRAE; *c.f*. Section 2.2 and Eq. 1), indicating that evaluation data may need better "dynamic range" beyond the one currently provided by the BGU HS Dataset [6] or that evaluation should extend beyond the spectral quantization levels of 31 bands currently in use. Second, it was found that the ranking of the algorithms in the previous challenge did not differ significantly between the clean and real world tracks, possibly indicating that the simulated "real world" camera did not add sufficient complexity relative to the clean track. To this end, the 2020 challenge presents a larger-than-ever data set nearly twice as large as the BGU HS data set (*c.f*. Section 2.1) as well as an improved real world track where camera noise is incorporated as well(*c.f*. Section 2.2).

2. NTIRE 2020 Challenge

The RGB to spectra recovery challenge [9] is one of the NTIRE 2020 challenges. The other challenges are: deblurring [40], nonhomogeneous dehazing [5], perceptual extreme super-resolution [63], video quality mapping [18], real image denoising [1], real-world super-resolution [35] and demoireing [60].

As in the NTIRE 2018 Spectral Recovery Challenge [7], the objectives of the NTIRE 2020 Challenge on Spectral Reconstruction are: (i) gauging and improving the state-of-the art in HS reconstruction from RGB images; (ii) comparing the different spectral recovery approaches; (iii) further expanding the amount of natural HS images available to the research community. Importantly, the 2020 challenge introduces not only a new and improved data set, but also an extended evaluation which attempts to gauge the expected performance of proposed methods beyond the scope of the challenge's test images.

2.1. ARAD HS Dataset

The NTIRE 2020 spectral reconstruction challenge provided a new, larger-than-ever, natural hyperspectral image data set. The data set included a total of 510 images: 450 training images, 30 validation images, and 30 test images. The training and test images were released during the challenge, while test images remain confidential to facilitate blind evaluation of future works. Figure 1 includes a set of sample images from the data set.

Figure 1: Sample images from the ARAD HS data set, note the variety of scene types (color and brightness have been manually adjusted for display purposes).

The ARAD data set was collected with a Specim IQ mobile hyperspectral camera. The Specim IQ camera is a stand-alone, battery-powered, push-broom spectral imaging system, the size of a conventional SLR camera (207 $\times$ 91 $\times$ 74 mm) which can operate independently without the need for an external power source or computer controller. The use of such a compact, mobile system facilitated collection of an extremely diverse data set with a large variety of scenes and subjects.

In addition to the ARAD data set, participants were invited to use the previously published BGU HS data set [6, 7] as well to obtain a total of 706 training images.

2.1.1 Radiometric Calibration

The Specim IQ camera provides RAW 512×512px images with 204 spectral bands in the 400-1000nm range. For the purpose of this challenge, manufacturer-supplied radiometric calibration has been applied to the RAW images, and the images have been resampled to 31 spectral bands in the visual range (400-700nm). Both RAW and radiometrically calibrated images have been made available to researchers.

The radiometric calibration corrects for measurement biases introduced by the camera systems CMOS sensor, con-

verting the recorded RAW per channel intensity data to accurate spectral measurements. "Lines" (image columns) with excessive interference are also removed by this process, resulting in a 482×512px image, resampled to 31 bands from 400nm to 700nm with a 10nm step.

2.2. Tracks

As in the previous iteration of this challenge [7], the NTIRE 2020 Spectral Recovery Challenge had two tracks, a "clean" and a "real world" track. While the clean track was similar to that of the previous challenge (NITRE 2018), the real world track was substantially updated to provide a more accurate simulation of physical camera systems.

Track 1: "Real World" simulates the recovery of spectral information from an unknown, uncalibrated camera. Participants were provided with 8-bit color images in compressed JPEG format created by applying the following procedure to spectral images:

1. Applying a real-world camera response function to a spectral image.
2. Subsampling the resulting 3 channel image to produce an RGGB Bayer mosaic image.
3. Adding simulated camera noise (Poisson shot noise and normally distributed dark noise) to the mosaic image.
4. Applying a demosaicing algorithm from the OpenCV [12] library to produce a three-channel RGB image.
5. Storing the image in compressed JPEG format.

The camera response and noise parameters used in the above procedure were kept confidential from challenge participants and shall remain confidential to facilitate equal ground comparisons of future works to the challenge results below.

Challenge participants were provided with code [8] (publicly available on the GitHub platform) used to generate both clean and real world track images.

Competitions competitions were hosted on the CodaLab platform [1], with a separate competition for each track. After registration, participants were able to access data and submit results for automatic evaluation on the competition test server. Due to constraints of the CodaLab platform, the validation and test set have been reduced to 10 images each track (for a total of 10 validation images and 20 test images).

Challenge phases The challenge had two phases:

1. **Development:** participants were provided with ground truth training hyperspectral/RGB image pairs for both tracks (450 image pairs for each track) as well as 10 RGB images for validation. A test server was made available to participant, allowing them to upload their results and receive an online evaluation score.
2. **Testing:** ground truth spectral validation images were released, along with final test images for each track. Participants were invited to upload their final solutions to the test server, and results were kept confidential until the challenge concluded.

Evaluation protocol . As in the 2018 competition [7], Mean Relative Absolute Error (MRAE) computed between the submitted reconstruction results and the ground truth images was selected as the quantitative measure for the competition. Root Mean Square Error (RMSE) was reported as well, but not used to rank results. MRAE and RMSE are computed as follows:

$$MRAE = \frac{\sum_{i,c} \frac{|P_{gt_{i_c}} - P_{rec_{i_c}}|}{P_{gt_{i_c}}}}{|P_{gt}|}, \tag{1}$$

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

where $P_{gt_{i_c}}$ and $P_{rec_{i_c}}$ denote the value of the c spectral channel of the i-th pixel in the ground truth and the reconstructed image, respectively, and $|P_{gt}|$ is the size of the ground truth image (pixel count $\times$ number of spectral channels).

3. Challenge Results

Submissions provided by challenge participants were evaluated against confidential ground-truth HS test-set images using the metrics described in Section 2.2 (*c.f.* Eq. (1), (2)). The results of the evaluations are shown in Table 1. Self-reported computational requirements and additional implementation details for submitted methods are reported in Table 2. The top performing method in the clean track (IPIC_SSR) achieved a MRAE of 0.0301 and a RMSE of 0.0129. The top performing method in the real world track (OrangeCat) achieved a MRAE of 0.0620 and a RMSE of 0.0192. For additional gains in accuracy top methods employed model ensemble and self-ensemble strategies [56]. All submitted solutions relied on recent-generation (and often state-of-the-art) GPUs for computation. Despite the use of powerful hardware, most solutions required at least 0.5 seconds to process a $\sim$ 0.25 megapixel (mp) image. The best placed solution that could recover the HSI in less than 0.5 seconds per-image (LFB) was ranked 8th in the clean track and 5th in the real world track. To achieve recovery in less than 0.1 seconds per-image (StaffsCVL) we needed to go down the ranked list

[1] https://codalab.org/

Figure 2: Sample "out-of-scope" images used to evaluate the proposed methods' generalization capabilities. Studio images (left, right) were recorded under halogen illumination. All images were images manually adjusted (color/brightness) for display.

respectively to 10th and 8th position for the clean and real world tracks.

In addition to the primary evaluation metrics, five additional auxiliary metrics were used to explore the stability and extrapolability of solutions proposed by participants. These metrics are described in the following sub-sections and were applied to the top-performing submissions. First, **"out-of-scope"** images which differ significantly from the training data were considered. Second, **"shuffled"** images where large-scale spatial features are broken down to randomly ordered 4×4 patches were used gauge the robustness of methods to unseen conditions and/or spatial features. Then, third, test image **brightness** is varied to assess methods' stability under varying illumination intensity. In a fourth test, **weighted** scoring is applied to accurately represent performance over spectral signatures which have lower abundance in the test data. Finally, recovered HS images are projected back to RGB space to examine the **physical consistency** of results - do the recovered HS images produce RGB projections which are similar to the query images? This last test is interesting because if a method does not meet this criterion, regardless of the MRAE, it must be recovering incorrect spectra.

3.1. Performance on "Out-of-Scope" Images

To study the generalizability of the proposed models, the top 3 models of each track are tested with 5 additionally images that were taken under drastically different settings, *e.g.* objects in a studio, halogen lighting, scenes with rare viewing perspective, etc. Example images are given in Figure 2.

The MRAE results of this study are given in the "Out-of-Scope" columns of Table 3. Perhaps unsurprisingly, the average MRAE error calculated for the out-of-scope images is more than *doubled* the MRAE score of all top performers, but perhaps more interestingly - solution ranking for top performers is varied significantly from their ranking on the challenge test set. Indicating, perhaps, that a slightly lower performance on data similar to the training set might be acceptable if the method is to generalise to spectral images that are quite different.

3.2. Dependence on Spatial Features

While most of the challenge participants exploit high-level information (*i.e.* image content) by mapping large image patches, many pixel-based spectral reconstruction methods in the prior art have already shown efficacy to a certain extent, *e.g.* [42, 6, 33, 2]. The purpose of this study is to examine: to what degree the proposed models can retain their efficacy if spatial information of the test images is, by construction, much more limited.

Models are tested with the "spatially shuffled" test images: each 4×4 patch in the original test images is randomly relocated. The MRAE results are given in the "Spatial" columns of Table 3. In the clean track, a significant degradation in performance can be seen and again - solution ranking for top performers is varied from their ranking on the challenge test set. In the real world track, degradation is dramatic, to the point where recovered data is unlikely to be usable (MRAEs of 0.22-0.45). It can be surmised that all top-performing solutions rely heavily on spatial information to overcome camera noise and compression artifacts in the real world track. In the noiseless clean track, dependence on spatial features remains significant, but much reduced relative to the real world track.

3.3. Dependence on Image Brightness

The RGB images can be brighter or dimmer depending on the exposure setting of the camera (*e.g.* shutter speed and aperture size) and/or the varying illumination intensity of the scene, which corresponds to linearly scaled ground-truth spectra. This means a linearly scaled hyperspectral image and its RGB counterpart is also a physically valid ground-truth pair. However, the best models in the 2018 competition [7] appear to perform poorly when the scene brightness changes [33].

In this year's challenge, the tests with two brightness modulations are included: half (HS images scaled down by a factor of 0.5) and double (scaled up by a factor of 2). The corresponding clean-track and real-world-track RGB images are simulated following the original methodology. The results are shown respectively in the "Brightness$\times 0.5$" and "Brightness$\times 2$" columns of Table 3. While varied exposure caused performance degradation in this years' top performers as well, the scale of this degradation is significantly reduced relative to the previous competition's top performers (MRAE degraded by 32% at most *vs.* 1245% at most for the 2018 top performer [33]).

3.4. Physical Consistency of Results

The hyperspectral and RGB images are physically related. Indeed, following a specified pipeline, RGB images can be accurately simulated from hyperspectral images (refer to section 2.2). The so-called *physical consistency* asks the question: if the reconstructed hyperspectral images are

Team	Username	Track 1: Clean MRAE	Track 1: Clean RMSE	Track 2: Real World MRAE	Track 2: Real World RMSE
IPIC_SSR [32]	Deep-imagelab	0.03010 (1)	0.01293	0.06216 (3)	0.01991
MDISL-lab	ppplang	0.03075 (2)	0.01268	0.06212 (2)	0.01946
OrangeCat [68]	zyz987	0.03231 (3)	0.01389	0.06200 (1)	0.01923
AIDAR	PARASITE	-	-	0.06514 (4)	0.02065
VIPLab[1]	ZHU_zy	0.03475 (4)	0.01475	-	-
TIC-RC	sunnyvick	0.03516 (5)	0.01567	0.07032 (7)	0.02191
VIPLab[2]	ninaqian	0.03518 (6)	0.01511	-	-
GD322	Hpeng	0.03601 (7)	0.01695	0.06780 (6)	0.02071
LFB	Tasti	0.03633 (8)	0.01690	0.06732 (5)	0.02124
CI Lab	honeygupta	0.03769 (9)	0.01677	0.07581 (9)	0.02253
StaffsCVL [17]	fubarabjs	0.04401 (10)	0.01978	0.07141 (8)	0.02173
Pixxel AI	akashpalrecha	0.04441 (11)	0.01645	0.09322 (10)	0.02255
Image Lab	sabarinathan	0.04577 (12)	0.01595	-	-
disqualified	Achiever11	0.17382 (13)	0.04573	0.16459 (11)	0.04743

Table 1: NTIRE 2020 Spectral Reconstruction Challenge results and final rankings on the ARAD HS test data.

Team	Reported runtime per image (sec) Clean	Reported runtime per image (sec) Real World	Platform	CPU	GPU	Training Time	Notes	Ensemble/Fusion
IPIC_SSR [32]	0.56	0.56	Pytorch	E5-2678	2x NVIDIA 2080Ti 11G	36 hours	Self-ensemble used only for "Real World" track.	self-ensemble, model-ensemble
MDISL-lab	16	16	PyTorch		NVIDIA 1080 Ti 12G	48 hours		10 model ensemble
OrangeCat [68]	3.74	3.74	PyTorch		2x NVIDIA Titan Xp 12G	7 days	8-setting ensemble strategy for both tracks	self-ensemble, model-ensemble
AIDAR	-	30	PyTorch		NVIDIA Titan Xp 12G	16 hours		self-ensemble, model-ensemble
VIPLab[1]	¯1	-	PyTorch		4x UNKNOWN	12 hours		
TIC-RC	0.7	0.7	PyTorch	Intel(R) Xeon(R) CPU	Tesla K80 12GB	13.9 hours		
VIPLab[1]	¯1	-	PyTorch		4x UNKNOWN	12 hours		
GD322	1.35	1.35	PyTorch	E5-2680	NVIDIA Titan Xp	13 hours		
LFB	0.31	0.30	PyTorch	Intel Xeon W-2133	NVIDIA GTX 2080Ti	8 hrs		
CI Lab	0.4	0.4	TensorFlow	Intel i9-9900X	NVIDIA GTX 2080Ti	36 hours		
StaffsCVL [17]	0.034	0.034	PyTorch	Intel Core i5	NVIDIA RTX 2080Ti	2.7 hours		
Pixxel AI	0.154	0.154	FastAI	Intel(R) Xeon(R)	NVIDIA V100	8.6 hours		
Image Lab	0.69	-	Keras	Intel Core i7	NVIDIA GTX 1080			
disqualified	3.75	3.75	Keras	Intel Core i7	NVIDIA GTX 1080	33 hours		

Table 2: Self-reported runtimes per image on the ARAD HS test data and additional implementation details.

Team	Track 1: Clean (MRAE) Out-of-Scope	Spatial	Brightness×0.5	Brightness×2	Physical	Weighted	Track 2: Real World (MRAE) Out-of-Scope	Spatial	Brightness×0.5	Brightness×2	Physical	Weighted
IPIC_SSR [32]	0.08511	0.09580	0.03273	0.03969	0.00117	0.03746	0.12556	0.45796	0.08000	0.06598	0.03290	0.07944
MDISL-lab	0.08076	0.07948	0.03562	0.03390	0.00053	0.03901	0.14005	0.21058	0.08203	0.06832	0.03563	0.07981
OrangeCat [68]	0.09233	0.07670	0.04052	0.04419	0.00103	0.04169	0.13019	0.22689	0.08097	0.06784	0.03346	0.07945

Table 3: Auxiliary test results of the top 3 models in each track. **Out-of-scope:** performance on images which differ significantly from the training data. **Spatial:** performance on images which were broken down to randomly ordered 4×4 patches. **Brightness ($\times 0.5$, $\times 2$):** performance over images where intensity was decreased/increased by $\times 0.5 / \times 2$. **Physical:** correlation between RGB projection of recovered spectra and input RGB images. **Weighted:** accuracy over representative spectra samples without accounting for their abundance.

applied with the original pipeline and re-generate the RGB images, how far off are these re-generated RGB images from the original ones?

The results are presented as the MRAE between the ground-truth and re-generated RGB images in the "Physical" columns of Table 3. Top performers presented relatively high consistency with images in the clean track, and slightly reduced consistency with images in the real world track. Reduced consistency in the latter is likely attributable to simulated camera noise and compression artifacts. However, although the RGB MRAE numbers are small we make two additional comments. First, assuming (approximately the following assumption holds) that a 1% MRAE error correlates roughly with Just Noticeable Differences (1 JND is a concept from psychophysics where an observer can just see the difference between stimuli) an MRAE of 3% correlates with a color difference of 3 which in turn correlates with perceived colors in images than can be seen to be different. Second, the MRAE hides the fact that the, for example, 95% quantile error can be large (> 10). This kind of error means that the recovered spectrum, when projected back to the RGB, results in a color which is instantly noticeable as different.

Curiously, because the recovered spectra do not reproject

to the same RGB, these spectra cannot be the correct answer (irrespective of any MRAE).

3.5. Weighted Accuracy

The spectral properties of the pixels representing the same material are expected to be similar. However, the abundance of one material in the scene does not indicate its importance. This study aims to provide a fair assessment across different materials in each scene. First, similar spectra are grouped into 1000 clusters. Then, the mean MRAE of each groups are calculated individually. Finally, the weighted MRAE is the mean of the groups' performances.

The results are provided in the "Weighted" columns of Table 3.

4. Conclusions

The NTIRE 2020 Challenge on Spectral Reconstruction from an RGB Image provides the most extensive evaluation to date of methods for spectral recovery from RGB images in terms of both participation and evaluation scope. Participants were provided with a larger-than-ever natural hyperspectral image data set and presented a wide variety of neural net based solutions to the task of spectral recovery from RGB images. Analysis of the proposed solutions revealed several intriguing areas for future exploration, namely: high-performance spectral recovery for video and/or edge devices, reducing dependence on spatial features, and increased robustness to unseen scenes.

Top performing methods required *at least* 0.5 seconds to process a ~ 0.25mp image on two state-of-the-art GPUs, the fastest method required ~ 34ms on a single state-of-the-art GPU. While the latter could claim processing at "video rates" (30fps), this would only hold true for 0.25mp video on a GPU based platform. Extrapolating from this information, processing a single frame of 4K video (8.5mp) would require approximately 34 and 1.15 seconds on a single GPU for the most accurate and fastest method respectively. Processing on an edge device (*e.g.* cellular phone) without a discrete GPU can be expected to take an order-of-magnitude longer. Future challenges may include an "edge device" track where solutions are scored on their computation requirements as well as their recovery performance.

All top performers were found to have a nontrivial dependency on spatial features when recovering spectral information from RGB images. The impact of this dependence becomes clear when one considers possible uses of recovered spectral information, for example: differentiating between similar objects based on their spectral reflectance (*e.g.* real fruit vs. fake plastic fruit). For this reason future challenges may emphasize dependence on spatial features when scoring proposed methods and possibly include an application-based test metric as well.

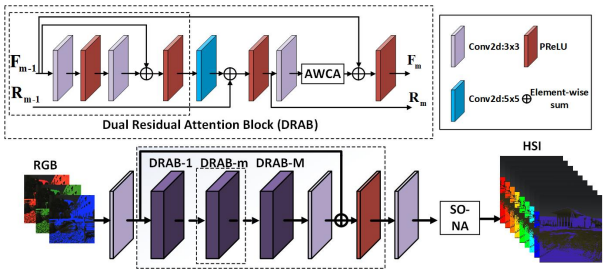


Figure 3: Network architecture of the IPIC_SSR adaptive weighted attention network (AWAN).

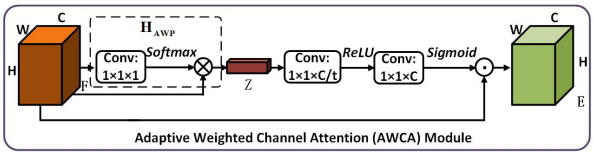


Figure 4: Diagram of adaptive weighted channel attention (AWCA) module. $\odot$ denotes element-wise multiplication.

The tests on image brightness and physical consistency are interesting. For the same scene, exposure - how well the same physical object is lit - varies across the scene. But, despite this we would expect to recover the spectrum (up to a scaling factor) and this was found not to be the case for existing methods. The physical consistency test is interesting and surprising. All challenge methods do not find spectra consistent with the original RGB. Even though their MRAE may be small, these methods *must* recover the wrong answer.

Finally, "out-of-scoope" image tests reveal that none of the top performers were able to robustly extrapolate to new settings. This indicates that while the training data set provided to participants is the largest of its kind, it can be further extended to cover additional settings. Namely indoor scenes and scenes under a larger variety of illumination conditions should be added to future data sets. The constantly increasing portability and ease-of-use of modern HISs is expected to facilitate the collection of larger and more varied data sets.

5. Challenge Methods and Teams

5.1. IPIC_SSR - Adaptive Weighted Attention Network with Camera Spectral Sensitivity Prior for Spectral Reconstruction from RGB Images[32]

As shown in Figure 3, a novel deep adaptive weighted attention network (AWAN) is presented for spectral reconstruction from RGB images. Specifically, the backbone architecture of the AWAN network is constituted of 8 dual residual attention blocks (DRAB). Each DRAB consists of

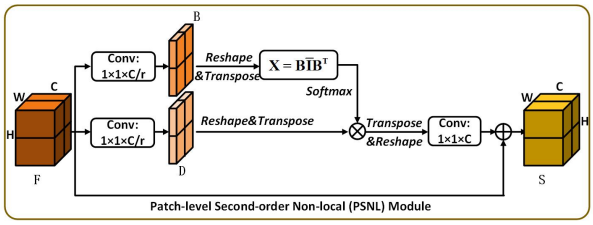


Figure 5: Diagram of patch-level second-order non-local (PSNL) module. $\otimes$ denotes matrix multiplication.

a traditional residual module and additional paired convolutional operations with a large (5×5) and small size (3×3) kernels, where the long and short skip connections to form the dual residual learning in the block. Typically, the output channel of each convolutional layer is set to 200. The adaptive weighted channel attention (AWCA) module (see in Figure 4) embedded in the DRAB adaptively integrates channel-wise interdependencies. At the tail of the AWAN network, a patch-level second-order non-local (PSNL) module is employed to capture long-range spatial contextual information via second-order non-local operations. The diagram of the PSNL module is illustrated in Figure 5).

Since the "Clean" track aims to recover hyperspectral images (HSIs) from the noise-free RGB images created by applying a known spectral response function to ground truth hyperspectral information, the camera spectral sensitivity (*i.e.* spectral response function) prior is introduced to improve the quality of spectral reconstruction. Considering the fact that the reconstructed RGB can be calculated naturally through the super-resolved HSI, the final loss is a linear combination of the discrepancies of RGB images and the differences of HSIs

$$l = l_h + \tau l_r \tag{3}$$

where τ denotes the tradeoff parameter and is set to 10 empirically. Given the ground truth $\mathbf{I}_{HSI}$ and the spectral super-resolved HSI $\mathbf{I}_{SSR}$, the two loss functions are specifically defined as

$$l_h = \frac{1}{N}\sum_{n=1}^{N}(|\mathbf{I}_{HSI}^{(n)} - \mathbf{I}_{SSR}^{(n)}|/\mathbf{I}_{HSI}^{(n)}) \tag{4}$$

$$l_r = \frac{1}{N}\sum_{n=1}^{N}(|\mathbf{\Phi}(\mathbf{I}_{HSI}^{(n)}) - \mathbf{\Phi}(\mathbf{I}_{SSR}^{(n)})|) \tag{5}$$

where $\mathbf{I}_{HSI}^{(n)}$ and $\mathbf{I}_{SSR}^{(n)}$ denote the n-th pixel value and $\mathbf{\Phi}$ is camera spectral sensitivity function. N is the total number of pixels. However, the camera spectral sensitivity is unknown in the "Real World" track, thus the AWAN network is optimized by stochastic gradient descent algorithm with individual constraint l_h.

5.1.1 Global Method Description

Training During the training, 64×64 RGB and HSI sample pairs are cropped with a stride of 32 from the original dataset. The batch size of our model is 32 and the parameter optimization algorithm chooses Adam modification with $\beta_1 = 0.9$, $\beta_2 = 0.99$ and $\epsilon = 10^{-8}$. The reduction ratio t value of the AWCA module is 16. The learning rate is initialized as 0.0001 the polynomial function is set as the decay policy with $power = 1.5$. The network training is stopped at 100 epochs. The proposed AWAN network has been implemented on the Pytorch framework and approximately 36 hours are required for training a network with 8 DRABs and output channel $= 200$ on 2 NVIDIA 2080Ti GPUs .

Testing In our experiments, different spectral recovery ways are tried and compared with their scores in the validation sets of the two tracks. One way is to split the input images into small overlapping patches, then average and stitch their outputs together on the GPU. The other is to feed the entire image to the AWAN network for inference on the CPU. Finally, the whole image is inputted into the network to fulfill the spectral recovery on the "Clean" track and at least 64G CPU is required for inference. The inference-time per image (CPU time) is 57.05s for both validation and test data. For the "Real World" track, the entire image is split into 128×128 overlapping patches with a stride of 64 and perform spectral reconstruction on an NVIDIA 2080Ti GPU with 11G memory. The AWAN network takes 0.56s per image (GPU time) for both validation and test data. By the way, in the "Clean" track, we can also achieve fast spectral reconstruction in the same way as the "Real World" track on the GPU, but the results will be slightly worse.

5.1.2 Ensembles and fusion strategies

For the "Clean" track, four models are trained for model-ensemble strategy, including two models with 8 DRABs and 200 channels and two models with 20 DRABs and 128 channels. Different from the "Clean" track, for the "Real World" track, the self-ensemble method [56] is firstly adopted for single AWAN network. Concretely, the RGB input is flipped up/down to acquire a mirrored output. Then the mirrored output and the original output are averaged into the target result. Also, three models with 8 DRABs and 200 channels and one model with 10 DRABs and 180 channels are trained for model-ensemble of AWAN network. Please refer to [32] for specific details.

5.2. MDISL-lab - Improved Pixel-aware Deep Function-Mixture Network

One fact is that the spectral of different pixels in a image vary widely. However, most existing Deep Convolutional Neural Networks (DCNNs) based Spectral Reconstruction (SR) methods treat all pixels in Hyper-Spectral Images

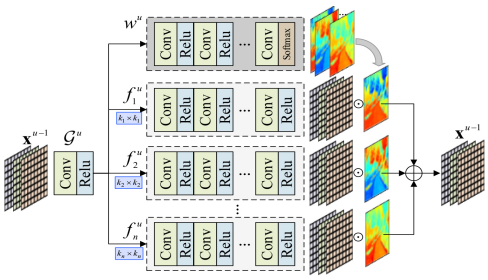


Figure 6: Architecture of the MDISL-lab function-mixture block.

equally and learn a universal mapping function, as shown in Figure. Based on the observation, we present a pixel-aware deep function-mixture network for SR, which is flexible to pixel-wisely determine the receptive field size and the mapping function.

One fact is that the spectral of different pixels in a image vary widely. However, most existing Deep Convolutional Neural Networks (DCNNs) based Spectral Reconstruction (SR) methods treat all pixels in Hyper-Spectral Images equally and learn a universal mapping function, as shown in Figure. Based on the observation and inspired by [64], a pixel-aware deep function-mixture network is presented for SR, which is flexible to pixel-wisely determine the receptive field size and the mapping function.

It is worth noting that, in order to reduce the computational complexity, different receptive fields of different sizes are achieved by stacking multiple 3x3 convolution layers. To further improve the learning ability of the network, a SE module [26] is placed after each branch and at the end of each module.

Specifically, a new module, termed the function-mixture (FM) block, is firstly developed. Each FM block consists of some parallel DCNN based subnets, among which one is termed the *mixing function* and the remaining are termed *basis functions*. The basis functions take different-sized receptive fields and learn distinct mapping schemes; while the mixture function generates pixel-wise weights to linearly mix the outputs of the basis functions, as shown in Figure 6. In this way, the pixel-wise weights can determine a specific information flow for each pixel and consequently benefit the network to choose appropriate RGB context as well as the mapping function for spectrum recovery. Then, several such FM blocks are stacked to further improve the flexibility of the network in learning the pixel-wise mapping. Furthermore, to encourage feature reuse, the intermediate features generated by the FM blocks are fused in late stage. The overall architecture of proposed network is shown in figure 7.

Training The paired spectral and RGB patches with a spatial size of 64x64 are cropped from the original images with a stride of 64. For data augmentation, horizontal flip and 90 degree rotation were randomly performed. Teh model was trained by ADAM optimizer and the mini-batch size is set to 64. The initial learning rate was set to 3e-4 and halved at every 60,000 iterations for three times. The model was implemented through PyTorch framework and trained on a NIVIDIA 1080 Ti GPU and Intel(R) Xeon(R) CPU E5-2640.

5.3. OrangeCat - Hierarchical Regression Network for Spectral Reconstruction from RGB Images [68]

Generally, we propose a 4-level hierarchical regression network (HRNet) [68] architecture for high-quality spectral reconstruction from RGB images, as shown in Figure 8. The PixelShuffle layers [48] are utilized to downsample the input to each level without adding parameters. Thus, the number of pixels of input is fixed while the spatial resolution decreases. Since PixelShuffle only reshapes feature maps and does not introduces interpolation operation, it allows HRNet to learn upsampling operation.

For each level, the process is decomposed to inter-level integration, artifacts reduction, and global feature extraction. The top level uses the most blocks to effectively integrate features and reduce artifacts thus produce high-quality spectral images. For inter-level learning, the output features of subordinate level are pixel shuffled, then concatenated to superior level, which uses an additional convolutional layer to unify the channels. In order to effectively reduce artifacts, we adopt a series of dense connection blocks [27, 24], containing 5 convolutional layers and a residual. The residual global block [26, 24] with short-cut connection of input is used to extract different scales of features. In this block, each remote pixel is connected with other pixels to model the global attention due to MLP layers. The illustration of these blocks are in Figure 2.

Since the features are most compact in bottom level, there is a 1×1 convolutional layer attached to the last of bottom level in order to enhance tone mapping by weighting all channels. The two mid levels process features at different scales. Moreover, the top level uses the most blocks to effectively integrate features and reduce artifacts thus produce high-quality spectral images. The illustration of these blocks are in Figure 9.

Training OnlyL1 loss is used duriung the training process, which is a PSNR-oriented optimization for the system. The L1 loss is defined as:

$$L_1 = \mathbb{E}[||G(x) - y||_1], \tag{6}$$

where x and y are input and output, respectively. The $G(*)$ is the proposed HRNet.

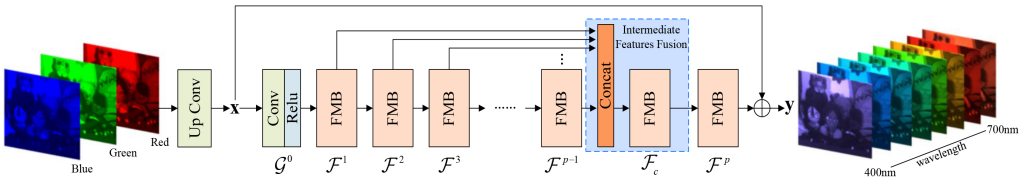


Figure 7: Architecture of the MDISL-lab pixel-aware deep function-mixture network. UpConv denotes a 3x3 convolution layer, which is used to increase the channels of input RGB image to the same as the output. FMB denotes the function-mixture block.

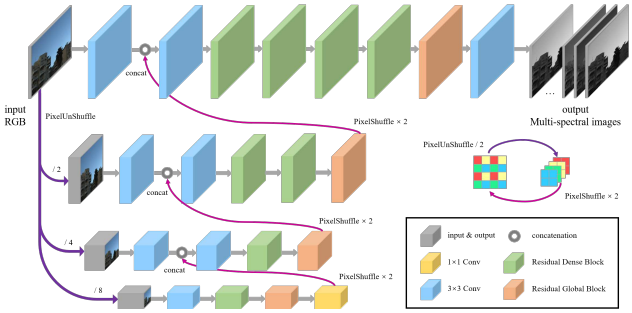


Figure 8: Illustration of the HRNet architecture.

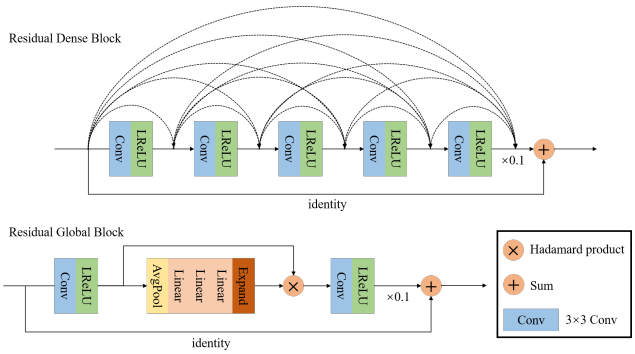


Figure 9: Illustration of the residual dense block (RDB) and residual global block (RGB) architectures.

The input RGB image and output spectral images were randomly cropped to a 256×256 region, then rescaled to [0, 1]. The parameters of network are Xavier initialized [22]. The whole system was trained for 10000 epochs in total. The initial learning rate was 1×10^{-4} and halved every 3000 epochs. For optimization, the Adam optimizer was used with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and a batch size of 8. Reflection padding was used in the system to avoid border artifacts. The LeakyReLU activation [36] function was attached to each convolution layer. No normalization were used in the proposed architecture. All the experiments were performed on 2 NVIDIA Titan Xp GPUs.

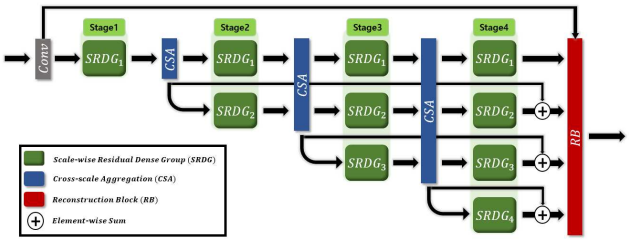


Figure 10: Overall framework of the AIDAR cross-scale aggregation network (CSAN).

5.4. AIDAR - Cross-scale Aggregation Network for Spectral Reconstruction

A cross-scale aggregation network (CSAN) is proposed with a novel feature fusion mechanism across multiple resolution branches. In CSAN, the scale-wise residual dense groups (SRDGs) exploit hierarchical feature information over different spatial resolutions in parallel. The SRDG is a series of residual dense blocks to fully achieve all the hierarchical representation capability. Here, a novel multi-scale feature fusion module is designed, which are named as cross-scale aggregation module (CSA), for compounding and aggregating the multi-resolution feature information from the prior SRDGs. It generates refined features at each resolution-level by fusing useful information across all the scale-levels. Such a function combines complementary characteristics from dynamic cross-scale representations in a stage-by-stage fashion. Also, the hierarchical levels are extended to explore strong contextual information from the low-resolution representations. It further has inner shortcut connections at each spatial level to improve gradient flow throughout the network. In addition, a global skip connection routes data between two ends of the network, improving further the ability of the network to accurately recover fine details. Finally, the reconstruction block select useful set of features from each branch representations with step-wise refinement. The CSAN can generate high-quality spectral images without noticeable artifacts, as will be confirmed by our results.

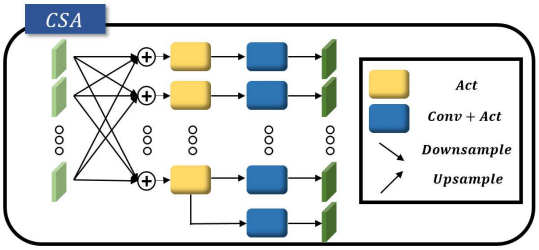


Figure 11: Cross-scale aggregation (CSA).

5.4.1 Global Method Description

Total method complexity: a pre-processing conv-block, 10 SRDGs(4 stages), a reconstruction block.
Training training a CSAN roughly takes 16 hours with a single NVIDIA Titan XP GPU for 1000 epochs. Training input was Self-ensemble and model-ensemble, learning rate was 3e-5, batch size was 8, and the Adam optimizer was used.
Testing testing a CSAN roughly takes about 30 seconds with a single Titan XP GPU.

5.4.2 Ensembles and fusion strategies

Self-ensemble and model-ensemble were used. Quantitatively, MRAE of baseline model with the ensemble is about 0.015 higher than that without the ensemble. Moreover, it shows qualitative improvement with some great recovery of words in the hyperspectral images.

The baseline model is described in Figure 10. For model ensemble, the model is modified by applying different width(the number of channels) and height(the number of stages) of the model, maintaining the volume of the models.

5.5. VIPLab

We separately learn the mean and corresponding residual of each image using DNNs for the res learning net we normalize output feature to zero-mean.

5.5.1 Ensembles and fusion strategies

we train 3 network one uses RELU activation the other uses swish activation finally a network trained with fine-tune result.

5.6. TIC-RC - Hyperspectral Spectral Super-resolution via an Improved HSCNN+

In the NTIRE2018 Spectral Reconstruction Challenge [7], the HSCNN+ [49] has achieved the best performance with a ResNet-based and a DenseNet-based approaches. Therefore, the ResNet-based HSCNN+ has been utilized as the baseline , which has also shown a good performance on this challenge. However, there are two problems for the baseline method. The first one is the huge computational burden for us, one training epoch costs about 150s with our computational resource (Google Colab with K40 GPU), which is not acceptable. The second one is the reconstruction performance can be improved. According to that, an efficient framework with less ResNet blocks is desired whilst improving or maintaining the reconstruction performance.

In the image dehazing field [45], multiple input RGB images after pre-processing have proved to be useful for recovering more information. Therefore, more RGB inputs are generated from the provided RGB images, including the white balanced (WB) image and the gamma correction (gamma) image. An example is shown in Figure 12. Figure 13 depicts the suggested architecture.

(a) (b) (c) (d) (e) (f)

Figure 12: Comparison between original RGB images, generated WB images, and gamma images: (a) Clean image (b) WB of clean image (c) gamma of clean image, (d) real world image (e) WB of real world image, (f) gamma of real world image.

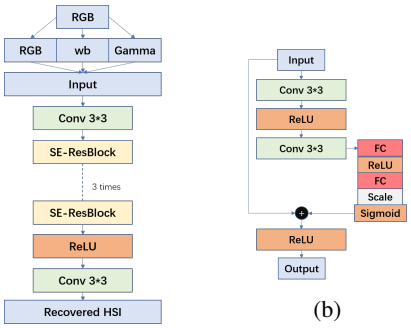
Methods	MRAE	Time(per epoch)
ResNet-based HSCNN+	0.0718251689	152s
Ours with raw input	0.0714754468	49s
Ours	0.0687300712	50s

Table 4: Track 2 reconstruction performance for baseline HSCNN+ and the proposed Improved HSCNN+ network.

5.7. GD322 - Residual pixel attention network for spectral reconstruction from RGB Images[44]

A residual pixel attention network (RPAN) is designed for spectral reconstruction from RGB images as shown in Figure 14.The RPAN we proposed was inspired by the RCAN[66], so it should be noted that the proposed RPAN and RCAN have some similarities in composition.

The proposed network adopts a global residual architecture, then the main architecture of the RPAN network is constituted of 4 residual attention group blocks (RAG) with 64 filters. The RAGs also are stacked to form a Concat layer followed by 1×1 convolutions layer. By skip connection, they add 64 feature maps from the first 3×3 convolution layer by global skip connection. Then output 31 channel



(a) (b)

Figure 13: (a) Improved HSCNN+ network framework (b) SE-ResBlock.

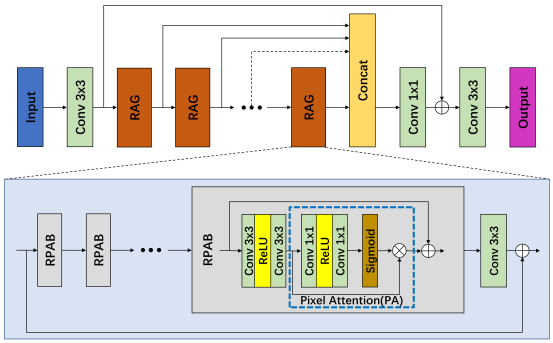


Figure 14: RPAN (up) and a RAG block (down).

spectral image is through the last 3×3 convolution layer. Each RAG is composed of 8 RPABs (residual pixel attention block) with 64 filters, and the RPAB is composed of residual blocks and a novel module is called pixel attention (PA) inside the residual block.

The features of different positions and different channels should not be the same in importance, some positions and channels are more helpful for the spectral reconstruction. In order to better treat the features of different channels and different positions, the PA is firstly developed in our RPAN network, which can rescale the pixel-wise features in each channel adaptively to improve the quality of hyperspectral image reconstruction from RGB images, as shown in Figure 14.

Training During the training, the input RGB image and hyperspectral images are cropped into small pieces of 64×64 from the training dataset and the batchsize is set to 16. In the "Clean" track, all biases are removed from each convolution layer, because the "Clean" track image contains no noise. In the "Real World" track, biases in all convolutional layers are reserved to compensate for the noise and JPG compression.

Adam optimizer is used for optimizing the proposed RPAN network with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and the weight decay was set to 10^{-6}. The initial learning rate is set to 8×10^{-5}, the learning rate decays by 0.8 after every 5 epochs, the network is ended the optimization at the 50-th epoch. MRAE loss function is used for training the RPAN network, zero-padding is used in all 3×3 convolutional layers to keep the feature map size unchanged. The proposed RPAN is trained by Pytorch platform on a single NVIDIA Titan Xp GPU. It takes about 13 hours to train the RPAN network for each track.

Testing Instead of cropping image into small blocks, the complete RGB image is used to get a complete spectral image on an NVIDIA Titan Xp. The RPAN network takes 1.35s(including inference time and spectral image reconstruction time) per image.

5.8. LFB - Linear Spectral Estimate Refinement for Spectral Reconstruction from RGB

The basic idea of the method used is that signal interpolation is more convenient when the underlying Laplacian signal energy is low. To reduce the Laplacian signal energy, a linear estimation of the spectral stimulus is considered. The proposed method can thus be described as two processing steps:

- a direct linear estimation on the spectral stimulus based on known camera response functions and an appropriate spectral basis.
- a refinement of the initial estimate through a convolutional neural network.

The concept of refining an estimate on spectral signals is not new. The major difference of our approach is that the algorithm for obtaining the spectral estimate is handcrafted, explicitly dependant on the camera response and not subject to parameter optimization during network training. The hybrid approach is therefore limited to the "clean" track.

Based on the spectral estimate, the neural network is tasked with signal refinement. Since the architecture of the convolutional neural network was not the focus of this work, a ResNet-18 was utilized. All the weights were initialized using fixed-update initialization [61]. The precise network architecture is summarized in Fig. 15c.

Only the new NTIRE2020 data was used for training and evaluation. The network was trained using Adam optimization with an initial learning rate of 10^{-4} and both a patch size and batch size of 50. The training itself was executed on a NVIDIA 2080TI graphics card. Pytorch was utilized as a framework.

It was found that the proposed hybrid approach outperforms the stand-alone ResNet and it was concluded that the hybrid approach is superior to the stand-alone ResNet. Figure 15 illustrates the suggested architecture and relative basis functions.

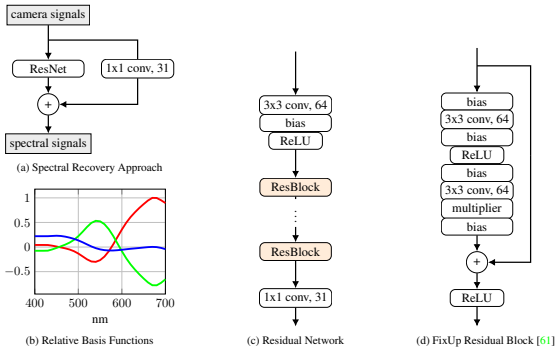


Figure 15: Visualization of the LFB hybrid approach for spectral signal recovery from RGB.

5.9. CI Lab - HSCNND++: Dense-connection based CNN network for Hyperspectral reconstruction

The backbone of this network is the DenseNet based model proposed in HSCNN+ [49] that has shown superior performance for hyperspectral reconstruction during the NTIRE2018 Spectral Reconstruction Challenge.

The backbone CNN model contains 2 parallel branches, each containing 30 dense-blocks with 4 convolutional layers and a fusion block, which also has 30 dense-blocks and merges the features from the two-parallel blocks and provids the predicted hyperspectral image. It was observed that the original HSCNN-D based model often got stuck in a local minima when trained with the mean relative absolute error(MRAE) or mean squared error(MSE), hence a compound loss function was proposed that achieved better convergence on the modified version of HSCNN-D model. This modified HSCNN-D model with the compound loss function is termed as HSCNND++.

The compound loss function consists of three components: an absolute difference or $L1$ loss, a structural loss and a gradient loss. The $L1$ loss is defined as $L_1 = ||G(x_{RGB}) - y_{HS}||_1$, which is a global loss that aims to reduce pixel-wise difference between the output, $G(x_{RGB})$, and the ground-truth, y_{HS}, images. To enforce further constraints, the structural and gradients losses were used. The structural loss is defined as $L_{SSIM} = 1 - SSIM(G(x_{RGB}), y_{HS})$ and the gradient loss is defined as $L_\nabla = ||\nabla(G(x_{RGB})) - \nabla(y_{HS})||_1$, where ∇ is the Laplacian operator. The weights for the individual losses were $10, 1, 1$ for L_1, L_{SSIM} and L_∇, respectively.

Training The "Clean" model was trained for $20k$ epochs and the "Real" model for $10k$ epochs with a varying learning rate. The dataset was augmented by taking random crops of size 48×48. Random flips and rotation were also performed on the images at each iteration. Adam optimizer was used with $\beta_1 = 0.9$, $\beta_2 = 0.99$ for training the models.

Testing guy While testing, the model takes around $40ms$ to reconstruct the hyperspectral image from an input RGB image of dimensions 482×512, on a single NVIDIA 2080Ti GPU.

5.10. StaffsCVL - RGB to Spectral Reconstruction via Learned Basis Functions and Weights [17]

.

Instead of predicting a 31-channel spectral image, the model predicts weights for a set of basis functions which are learned at the same time as the weights. In classical spectral reconstruction literature, the spectrum was recovered by weighted combination of basis functions[3] or sparse coding[43]. This method combines the simplicity of weighted basis functions and the performance and robustness of deep learning. The network predicts 10 weights for each pixel as well as learns a set of 10 basis functions which is then combined to form the final spectral image cube.
A modified UNet[47] network is used with skip connections to allow lower level features to flow to deeper layers. The 2x2 pooling layers are replaced with linear downsampling layers and four contracting steps are done. The cropping step before concatenation in the expansive path is replaced with a direct concatenation as cropping might dispose of edge information which could be useful for robust prediction, especially around the edges of the image. The same network architecture and training policy is used for both the clean and real world tracks. The proposed method is depicted in figure 16.

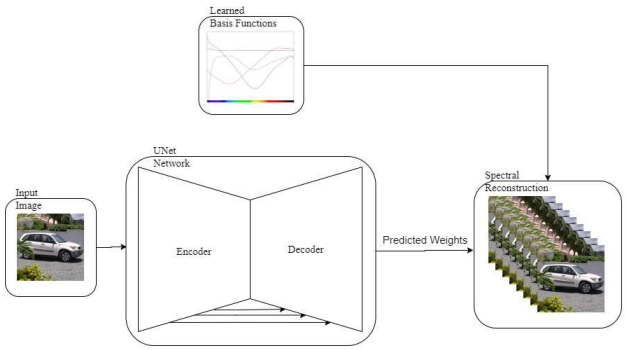


Figure 16: Architecture of the StaffsCVL method for spectral recovery.

The models from both tracks were trained on only the NTIRE 2020 challenge dataset without the use of any pretrained models. They were trained on patches from the RGB image and spectral cubes. The RGB image and spectral cubes were resized to 512×512, and 64×64 patches were extracted deterministically, which were used for training. The training batch size was 128, learning rate was 1e-4 and

the Adam optimiser [30] was used during training. Random horizontal and vertical flips were used for data augmentation, with a weight decay of 1e-5. The use of weight decay proved helpful to avoid overfitting. The basis functions are learned as a 10×31 matrix variable during training without going through any neural network layer. During inference, the saved trained matrix is simply loaded into memory and used. At test time, the full RGB image is passed through the CNN. The spectral cube is then generated as a weighted combination of the basis functions, using the predicted weights. The proposed method is advantageous because it is able to reconstruct the spectral cube using fewer parameters than would normally be required (*i.e.* predicting 10 weights per pixel instead of 31, a 67.74% reduction in predicted output). This becomes even more significant when predicting 301 spectral bands (96.68% reduction in this case). More detailed information on the method can be found in [17].

5.11. Pixxel AI - MXR-U-Nets for Real Time Hyperspectral Reconstruction [10]

The approach combines some of the very recent advancements in image classification, segmentation and Generative Adversarial Networks. At a high level, a model based on the U-Net [47] architecture with self-attention [62] to project RGB images to their Hyperspectral counterparts was used. The loss function for the model is a slightly modified version of perceptual losses (See section 3.2 in [10]) [28] [20].

The encoder backbone is an XResnet [25] with Mish [39] activation function (Replacing ReLU). The XResnet encoder, as proposed by He *et al*. (referred to as Resnet-D in [25]), improves upon the original Resnet [24] in classification performance. The solution combines both of these improvements and the model will be referred to as *mxresnet*. *mxresnet* implementation used in the implementation: [58]. Sub-pixel convolution layers [48] layers are used as upsampling layers in decoder blocks with ICNR [4] initialization scheme and weight normalization [59]. The sub-pixel convolution layers serve to conserve information during the upsampling part of a decoder. To reduce the checkerboard artifacts introduced by the sub-pixel convolution layers, ICNR initialization with weight normalization is used. Each sub-pixel convolution layer is followed by a blur [53] layer consisting of an average pooling layer with a 2×2 filter and a stride of 1. This improvement adds to the previous solution for dealing with checkerboard artifacts in the outputs of pixel shuffle layers. The decoder has a Self Attention block as proposed by Zhang *et al*. in [62] to help the network focus more on the relevant parts of the image. Figure 17 and Figure 18 depict the proposed architecture and U-Net block respectively.

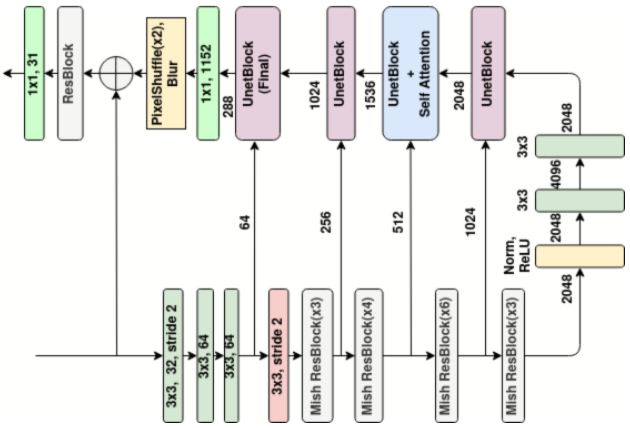


Figure 17: Architecture of the proposed mxresnet50 model.

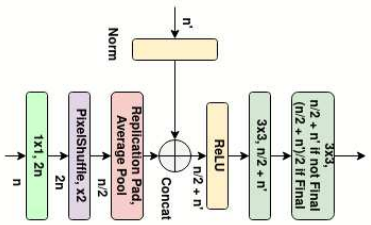


Figure 18: U-Net block inside the proposed mxresnet50 model.

Training An *mxresnet50* encoder based model is trained for 200 epochs with the AdamW [34] optimizer with a weight decay of $1e-3$. The learning rate follows the OneCycle schedule [52]. Under this schedule, the learning rate is started at 1e-5 and increased to $1e-3$ over 60 epochs following a half cosine curve. After the learning rate peaks, it is reduced to $1e-9$ over another 140 epochs following a similar half cosine curve. The model was trained using mixed-precision training [38] to lower training time and memory requirements. A single V100 GPU was used for all training runs.

5.12. Image Lab - Light Weight Residual Dense Attention Net for Spectral Reconstruction

An ensemble of convolution layer with Residual Dense Attention block (RDAB) connected at multi-scale level are used for spectral reconstruction. Specifically, in each block, certain significant features are given more importance spatially and spectrally by its dedicated attention mechanism [57], henceforth multi-scale hierarchical features are extracted at multi-level to widen the forward paths for higher capacity. The proposed network [41] for Spectral Reconstruction is shown in Figure 19.

In this Network, the RGB image serves as an input, the coordinal features are extracted to improve its spatial in-

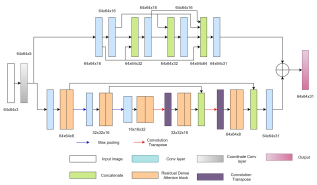


Figure 19: Architecture of the Light Weight Residual Dense Attention Net for Spectral Reconstruction.

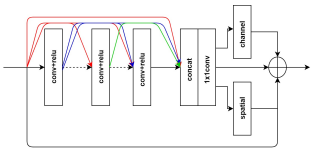


Figure 20: Residual Dense Attention Block.

formation [15]. The weights from the Coordinate convolution block are shared by two independent feature extraction mechanisms, one by dense feature extraction and the other by the multiscale hierarchical feature extraction. The Dense features are extracted by a dedicated dense block [67] connection whereas the multiscale hierarchical features are extracted by the Residual Dense Attention Block. The block diagram of Residual Dense Attention Block is shown in Figure 20. The Residual Dense Block (RDB) generates the local hierarchical feature. RDAB blocks are connected at multi-scale level in a U-net fashion, where the encoding phase consists of Maxpooling layer in between the RDAB blocks meanwhile the decoding phase consists of Transpose Convolution between them. This Transpose Convolution helps to reconstruct the image to the same spatial resolution as that of the input. Finally, the features from both the feature extraction mechanisms are globally fused to produce the 31 spectral bands.

$$LossFunction = L2 + (1 - SSIM)$$

Acknowledgments

We thank the NTIRE 2020 sponsors: HUAWEI Technologies Co. Ltd., OPPO Mobile Corp., Ltd., Voyage81, MediaTek Inc., DisneyResearch|Studios, and ETH Zurich (Computer Vision Lab). Graham Finlayson is grateful for the support of EPSRC grant EP S028730. Ohad Ben-Shahar gratefully acknowledges the support of the ISF-FIRST program grant 555/19.

A. Teams and affiliations

NTIRE2018 team

Title: NTIRE 2018 Challenge on Spectral Reconstruction from RGB Images

Members: *Boaz Arad*[1,4]*(boazar@post.bgu.ac.il)*, Radu Timofte [2], Ohad Ben-Shahar [1], Yi-Tun Lin [3], Graham Finlayson [3], Shai Givati [1]
Affiliations:
[1] Ben-Gurion University of the Negev, Israel
[2] Computer Vision Lab, ETH Zurich, Switzerland
[3] University of East Anglia, UK
[4] Voyage81

IPIC_SSR

Title: Adaptive Weighted Attention Network with Camera Spectral Sensitivity Prior for Spectral Reconstruction from RGB Images

Members: *Jiaojiao Li*[1] *(jjli@xidian.edu.cn)*, Chaoxiong Wu[1], Rui Song[1], Yunsong Li[1], Fei Liu[1]
Affiliations:
[1] Xidian University, Xi'an, China

MDISL-lab

Title: Improved Pixel-wise Deep Function-Mixture Network

Members: *Zhiqiang Lang*[1] *(2015303107lang@mail.nwpu.edu.cn)*, Wei Wei[1], Lei Zhang[1], Jiangtao Nie[1]
Affiliations:
[1] Chang'an campus of Northwestern Polytechnical University

OrangeCat

Title: Hierarchical Regression Network for Spectral Reconstruction from RGB Images

Members: *Yuzhi Zhao*[1] *(yzzhao2-c@my.cityu.edu.hk)*, Lai-Man Po[1], Qiong Yan[2], Wei Liu[2,3], Tingyu Lin[1]
Affiliations:
[1] City University of Hong Kong
[2] SenseTime Research
[3] Harbin Institute of Technology

AIDAR

Title: Cross-scale Aggregation Network for Spectral Reconstruction

Members: *Youngjung Kim*[1] *(read12300@add.re.kr)*, Changyeop Shin[1], Kyeongha Rho[1], Sungho Kim[1]
Affiliations:
[1] Agency for Defense Development

VIPLab

Title: VIPLab

Members: *Junhui HOU (zbzhzhy@gmail.com)*, Zhiyu ZHU, Yue QIAN

Affiliation: City University of Hong Kong

TIC-RC

Title: Hyperspectral Spectral Super-resolution via an Improved HSCNN+
Members: *He Sun*[1] *(h.sun@strath.ac.uk)*, Jinchang Ren[1], Zhenyu Fang[1], Yijun Yan[1]
Affiliations:
[1] University of Strathclyde, Glasgow, U.K.

GD322

Title: Residual pixel attention network for spectral reconstruction from RGB Images
Members: *Hao Peng*[1] *(Hao_Peng@outlook.com)*, Xiaomei Chen[1], Jie Zhao[1]
Affiliations:
[1] Beijing Institute of Technology

LFB

Title: Linear Spectral Estimate Refinement for Spectral Reconstruction from RGB
Members: *Tarek Stiebel*[1] *(tarek.stiebel@lfb.rwth-aachen.de)*, Simon Koppers[1], Dorit Merhof[1]
Affiliations:
[1] Institute of Imaging & Computer Vision, RWTH Aachen University

CI Lab

Title: HSCNND++: Dense-connection based CNN network for Hyperspectral reconstruction
Members: *Honey Gupta*[1] *(hn.gpt1@gmail.com)*, Kaushik Mitra[1]
Affiliations:
[1] Indian Institute of Technology Madras, Chennai

StaffsCVL

Title: RGB to Spectral Reconstruction via Learned Basis Functions and Weights
Members: *Biebele Joslyn Fubara*[1] *(fubarabjs@yahoo.co.uk)*, Dave Dyke[1], Mohamed Sedky[1]
Affiliations:
[1] Staffordshire University, UK

Pixxel AI

Title: MXR-U-Nets for Real Time Hyperspectral Reconstruction
Members: *Akash Palrecha*[1] *(akashpalrecha@gmail.com)*, Atmadeep Banerjee[1]
Affiliations:
[1] Pixxel (https://www.pixxel.space)

Image Lab

Title: Light Weight Residual Dense Attention Net for Spectral Reconstruction
Members: *Sabarinathan*[1] *(sabarinathantce@gmail.com)*, K Uma[2], D Synthiya Vinothini[2], B Sathya Bama[2] ,S M Md Mansoor Roomi[2]
Affiliations:
[1] Couger Inc
[2] Thiagarajar college of engineering

References

[1] A. Abdelhamed, M. Afifi, R. Timofte, M. Brown, et al. NTIRE 2020 challenge on real image denoising: Dataset, methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[2] J. Aeschbacher, J. Wu, and R. Timofte. In defense of shallow learned spectral reconstruction from rgb images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 471–479, 2017. 1, 4

[3] F. Agahian, S. A. Amirshahi, and S. H. Amirshahi. Reconstruction of reflectance spectra using weighted principal component analysis. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 33(5):360–371, 2008. 12

[4] A. P. Aitken, C. Ledig, L. Theis, J. Caballero, Z. Wang, and W. Shi. Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize. *ArXiv*, abs/1707.02937, 2017. 13

[5] C. O. Ancuti, C. Ancuti, F.-A. Vasluianu, R. Timofte, et al. NTIRE 2020 challenge on nonhomogeneous dehazing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[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. 1, 2, 4

[7] B. Arad, O. Ben-Shahar, R. Timofte, et al. NTIRE 2018 challenge on spectral reconstruction from rgb images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 1, 2, 3, 4, 10

[8] B. Arad and Y.-T. Lin. NTIRE 2020 challenge on spectral reconstruction from rgb images - challenge code. https://github.com/boazarad/NTIRE2020_spectral, 2020. 3

[9] B. Arad, R. Timofte, Y.-T. Lin, G. Finlayson, O. Ben-Shahar, et al. NTIRE 2020 challenge on spectral reconstruction from an rgb image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[10] A. Banerjee and A. Palrecha. Mxr-u-nets for real time hyperspectral reconstruction, 2020. 13

[11] E. Beletkaia and J. Pozo. More than meets the eye: Applications enabled by the non-stop development of hyperspectral imaging technology. *PhotonicsViews*, 17(1):24–26, 2020. 1

[12] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 3

[13] Y. B. Can and R. Timofte. An efficient cnn for spectral reconstruction from rgb images. *arXiv preprint arXiv:1804.04647*, 2018. 1

[14] C.-I. Chang. *Hyperspectral data exploitation: theory and applications*. John Wiley & Sons, 2007. 1

[15] R. L. et al. An intriguing failing of convolutional neural networks and the coordconv solution. In *Neural Inf. Process. System*, December 2018. 14

[16] H. Fu, L. Bian, X. Cao, and J. Zhang. Hyperspectral imaging from a raw mosaic image with end-to-end learning. *Optics Express*, 28(1):314–324, 2020. 1

[17] B. J. Fubara, M. Sedky, and D. Dyke. Rgb to spectral reconstruction via learned basis functions and weights. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 5, 12, 13

[18] D. Fuoli, Z. Huang, M. Danelljan, R. Timofte, et al. NTIRE 2020 challenge on video quality mapping: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[19] S. Galliani, C. Lanaras, D. Marmanis, E. Baltsavias, and K. Schindler. Learned spectral super-resolution. *arXiv preprint arXiv:1703.09470*, 2017. 1

[20] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *ArXiv*, abs/1508.06576, 2015. 13

[21] B. Geelen, N. Tack, and A. Lambrechts. A compact snapshot multispectral imager with a monolithically integrated per-pixel filter mosaic. In *Advanced Fabrication Technologies for Micro/Nano Optics and Photonics VII*, volume 8974, page 89740L. International Society for Optics and Photonics, 2014. 1

[22] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010. 9

[23] P. Gonzalez, K. Tack, B. Geelen, B. Masschelein, W. Charle, B. Vereecke, and A. Lambrechts. A novel cmos-compatible, monolithically integrated line-scan hyperspectral imager covering the vis-nir range. In *Next-Generation Spectroscopic Technologies IX*, volume 9855, page 98550N. International Society for Optics and Photonics, 2016. 1

[24] 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. 8, 13

[25] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li. Bag of tricks for image classification with convolutional neural networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 558–567, 2018. 13

[26] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 8

[27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 8

[28] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 13

[29] B. Kaya, Y. B. Can, and R. Timofte. Towards spectral estimation from a single rgb image in the wild. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3546–3555. IEEE, 2019. 1

[30] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 13

[31] F. Lahoud, R. Zhou, and S. Susstrunk. Multi-modal spectral image super-resolution. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018. 1

[32] J. Li, C. Wu, R. Song, Y. Li, and F. Liu. Adaptive weighted attention network with camera spectral sensitivity prior for spectral reconstruction from rgb images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 5, 6, 7

[33] Y.-T. Lin and G. D. Finlayson. Exposure invariance in spectral reconstruction from rgb images. In *Color and Imaging Conference*, volume 2019, pages 284–289. Society for Imaging Science and Technology, 2019. 2, 4

[34] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 13

[35] A. Lugmayr, M. Danelljan, R. Timofte, et al. NTIRE 2020 challenge on real-world image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[36] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the International Conference on Machine Learning*, volume 30, page 3, 2013. 9

[37] X. Miao, X. Yuan, Y. Pu, and V. Athitsos. λ-net: Reconstruct hyperspectral images from a snapshot measurement. In *IEEE/CVF Conference on Computer Vision (ICCV)*, volume 1, 2019. 1

[38] P. Micikevicius, S. Narang, J. Alben, G. F. Diamos, E. Elsen, D. García, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu. Mixed precision training. *ArXiv*, abs/1710.03740, 2017. 13

[39] D. Misra. Mish: A self regularized non-monotonic neural activation function. *ArXiv*, abs/1908.08681, 2019. 13

[40] S. Nah, S. Son, R. Timofte, K. M. Lee, et al. NTIRE 2020 challenge on image and video deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[41] D. Nathan, K.Uma, D. S. Vinothini, B. S. Bama, and S. M. M. M. Roomi. Light weight residual dense attention net for spectral reconstruction from rgb images. In *arXiv preprint arXiv:2004.06930*, 2020. 13

[42] 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. 4

[43] M. Parmar, S. Lansel, 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. IEEE, 2008. 12

[44] H. Peng, X. Chen, and J. Zhao. Residual pixel attention network for spectral reconstruction from rgb images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 10

[45] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang. Gated fusion network for single image dehazing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 10

[46] 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. 1

[47] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 12, 13

[48] 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. 8, 13

[49] Z. Shi, C. Chen, Z. Xiong, D. Liu, and F. Wu. Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 1, 10, 12

[50] M. Shoeiby, A. Robles-Kelly, R. Timofte, R. Zhou, F. Lahoud, S. Susstrunk, Z. Xiong, Z. Shi, C. Chen, D. Liu, Z.-J. Zha, F. Wu, K. Wei, T. Zhang, L. Wang, Y. Fu, K. Nagasubramanian, A. K. Singh, A. Singh, S. Sarkar, and B. Ganapathysubramanian. Pirm2018 challenge on spectral image super-resolution: Methods and results. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018. 1

[51] M. Shoeiby, A. Robles-Kelly, R. Wei, and R. Timofte. Pirm2018 challenge on spectral image super-resolution: Dataset and study. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018. 1

[52] L. N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. *ArXiv*, abs/1803.09820, 2018. 13

[53] Y. Sugawara, S. Shiota, and H. Kiya. Super-resolution using convolutional neural networks without any checkerboard artifacts. *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 66–70, 2018. 13

[54] F. Tanriverdi, D. Schuldt, and J. Thiem. Dual snapshot hyperspectral imaging system for 41-band spectral analysis and stereo reconstruction. In *International Symposium on Visual Computing*, pages 3–13. Springer, 2019. 1

[55] 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. 1

[56] R. Timofte, R. Rothe, and L. Van Gool. Seven ways to improve example-based single image super resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1865–1873, 2016. 3, 7

[57] Woo, Sanghyun, and et al. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, September 2018. 13

[58] L. Wright and E. Shalnov. lessw2020/mish: Mxresnet release. https://doi.org/10.5281/zenodo.3727966, Mar. 2020. 13

[59] S. Xiang and H. Li. On the effects of batch and weight normalization in generative adversarial networks. *arXiv preprint arXiv:1704.03971*, 2017. 13

[60] S. Yuan, R. Timofte, A. Leonardis, G. Slabaugh, et al. NTIRE 2020 challenge on image demoireing: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[61] H. Zhang, Y. N. Dauphin, and T. Ma. Residual learning without normalization via better initialization. In *International Conference on Learning Representations*, 2019. 11, 12

[62] H. Zhang, I. J. Goodfellow, D. N. Metaxas, and A. Odena. Self-attention generative adversarial networks. *ArXiv*, abs/1805.08318, 2019. 13

[63] K. Zhang, S. Gu, R. Timofte, et al. NTIRE 2020 challenge on perceptual extreme super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[64] L. Zhang, Z. Lang, P. Wang, W. Wei, S. Liao, L. Shao, and Y. Zhang. Pixel-aware deep function-mixture network for spectral super-resolution. *arXiv preprint arXiv:1903.10501*, 2019. 8

[65] T. Zhang, Y. Fu, L. Wang, and H. Huang. Hyperspectral image reconstruction using deep external and internal learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8559–8568, 2019. 1

[66] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 10

[67] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image super-resolution. In *NProc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, December 2018. 14

[68] Y. Zhao, L.-M. Po, Q. Yan, W. Liu, and T. Lin. Hierarchical regression network for spectral reconstruction from rgb images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 5, 8