HSGAN: Hyperspectral Reconstruction From RGB Images With Generative Adversarial Network

Yuzhi Zhao, *Graduate Student Member, IEEE*, Lai-Man Po, *Senior Member, IEEE*,
Tingyu Lin, Qiong Yan, Wei Liu, and Pengfei Xian

***Abstract*— Hyperspectral (HS) reconstruction from RGB images denotes the recovery of whole-scene HS information, which has attracted much attention recently. State-of-the-art approaches often adopt convolutional neural networks to learn the mapping for HS reconstruction from RGB images. However, they often do not achieve high HS reconstruction performance across different scenes consistently. In addition, their performance in recovering HS images from clean and real-world noisy RGB images is not consistent. To improve the HS reconstruction accuracy and robustness across different scenes and from different input images, we present an effective HSGAN framework with a two-stage adversarial training strategy. The generator is a four-level top-down architecture that extracts and combines features on multiple scales. To generalize well to real-world noisy images, we further propose a spatial–spectral attention block (SSAB) to learn both spatial-wise and channel-wise relations. We conduct the HS reconstruction experiments from both clean and real-world noisy RGB images on five well-known HS datasets. The results demonstrate that HSGAN achieves superior performance to existing methods. Please visit https://github.com/zhaoyuzhi/HSGAN to try our codes.**

***Index Terms*— Generative adversarial network (GAN), hyperspectral (HS) reconstruction, spatial–spectral attention.**

I. Introduction

HYPERSPECTRAL (HS) imaging technology analyzes a wide spectrum for each pixel in the image of a scene instead of only primary colors (red, green, and blue). Normally, HS images are sampled at more than 20 equally distributed wavelengths. The spectral range in HS images can extend beyond the human visible range (e.g., ultraviolet and infrared). Since HS images contain much richer information than RGB images, there are many specific applications, e.g., remote sensing [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], food processing [11], anomaly detection [12], and medical imaging [13], [14]. However, applying common HS imagers (e.g., HS spectrometers [15], [16], [17], [18], [19] and HS snapshot imaging [20], [21], [22], [23], [24], [25], [26], [27]) often encounters two challenges. First, HS spectrometers take a long operation time due to spatial-wise or spectral-wise scanning, which is not suitable for real-time applications and moving scenes. Second, HS snapshot imaging achieves video rates but reduces spatial and spectral resolutions. In addition, those devices are of high complexity and expensive in terms of consumer usage. Therefore, acquiring high-quality HS images at low cost in complex scenes becomes in high demand. Compared with HS imagers, RGB cameras are well developed and more robust to different scenes, i.e., RGB images are much easier to obtain than HS images. Based on this observation, many efforts have been made to recover spectra from RGB images recently.

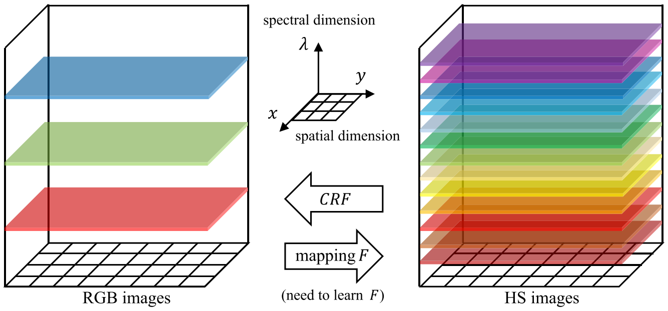

Fig. 1. Illustration of RGB and HS images. The RGB images have three channels: red (long-wavelength lights), green (medium-wavelength lights), and blue (short-wavelength lights), which are what human eyes can perceive. The HS images are in the form of a hypercube including n-dimensional image data. "CRF" denotes the camera response function [28], which relates image irradiance to image brightness.

HS reconstruction from RGB images can be formulated as learning a mapping function F, as shown in Fig. 1. It is challenging since the information of RGB images is much less than the HS spectrum. The algorithms need to learn to fill in the information of missing spectral bands from only three channels, which makes the task ill-posed. In addition, RGB images captured in real-world scenes are often noisy. The practical algorithms should have good denoising ability. Recent approaches to estimate F typically fall into one of the two learning-based categories.

1) Sparse-coding-based approaches [29], [30], [31].
2) Convolutional-neural-network-based approaches (a.k.a. CNN-based approaches) [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42].

Manuscript received 24 March 2021; revised 6 December 2021 and 29 November 2022; accepted 25 July 2023. *(Corresponding author: Yuzhi Zhao.)*

Yuzhi Zhao, Lai-Man Po, Tingyu Lin, and Pengfei Xian are with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong (e-mail: yzzhao2-c@my.cityu.edu.hk; eelmpo@cityu.edu.hk; tingyulin3-c@my.cityu.edu.hk; xian.pf@my.cityu.edu.hk).

Qiong Yan is with the SenseTime Research and Tetras.AI, Hong Kong (e-mail: sophie.yanqiong@gmail.com).

Wei Liu is with the ByteDance, Beijing 100086, China (e-mail: liujikun63@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3300099.

Digital Object Identifier 10.1109/TNNLS.2023.3300099

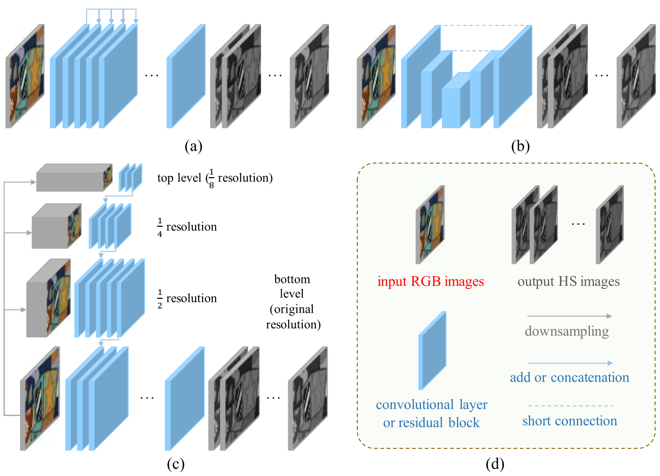

Fig. 2. Comparison of three network architectures used in HS reconstruction from RGB images. (a) Sequential convolutions with residual connections [32], [33], [34], [35], [36], [53]. (b) U-Net [37], [38], [50], [52]. (c) Hierarchical architecture. (d) Descriptions of layers, blocks and lines [39], [40], and the proposed HSGAN. We simplify the blocks in those papers to highlight architectures.

Among them, CNN-based approaches often achieve better performance, where the mapping F is formulated by a CNN. It is trained by maximizing a posteriori of HS images conditioned on RGB images. Though they could produce good HS images without any manual control or extra hardware support, there still exist the following issues.

1) Poor generalization ability across different RGB images and scenes (e.g., in different HS datasets).
2) Inconsistent performance on clean RGB images and real-world noisy RGB images.

To address the issues, we propose an HSGAN framework. To improve the generalization quality, we adopt the adversarial training [43] in addition to the traditional regression loss functions such as L1 and L2 loss. It improves the performance while does not introduce additional computational cost at the inference. We adopt two training stages to stabilize the adversarial training. For the HSGAN generator, we design a four-level top-down hierarchical architecture, as shown in Fig. 2 (c). It has three main advantages over previous architectures [e.g., Fig. 2(a) and (b)]: 1) the top level greatly enlarges the perceptive field, which models long-range relations for far pixels in the image; 2) the levels from top to bottom extract features from global to local regions; and 3) the connections between levels are lossless operations. To further improve the feature representation ability, we propose a spatial–spectral attention block (SSAB) at every level of the generator. It considers both the spatial importance of every pixel and the interchannel relations of feature maps. Since the feature representation ability is improved and long-range relations are modeled [44], HSGAN has a stronger denoising ability when receiving noisy inputs.

To demonstrate the good generalization ability and HS reconstruction accuracy achieved by HSGAN, we conduct experiments on five well-known HS datasets: NUS [45], Harvard [46], ICVL [30], CAVE [47], and ARAD [48]. Each of them includes paired clean RGB and HS images captured by different types of cameras (also with different CRFs). To evaluate the generalization ability, we apply different CRFs to the same dataset. To evaluate the HS reconstruction ability from real-world RGB images, we additionally add realistic noises and JPEG compression [48], [49] to the clean RGB images. Extensive experiments show that the proposed HSGAN achieves higher accuracy and consistent performance across different application cases than the state-of-the-art methods [32], [33], [34], [35], [38], [39], [40], [50], [51], [52], [53].

We summarize the main contributions of this article below.

1) We propose an HSGAN with a two-stage adversarial training scheme to improve HS reconstruction accuracy.
2) We propose a top-down four-level generator architecture to extract and combine features in different scales.
3) We propose a SSAB to further enhance feature representation.

II. Related Work

A. HS Image Acquisition

Early HS imaging devices [15], [17], [19], [54], [55] record the spatial and spectral information based on scanning. It can be divided into spatial scanning (reading images over time) and spectral scanning (acquiring images at different wavelengths). Due to hardware design, the acquisition of high-resolution HS images is extremely slow. To accelerate the HS acquisition, snapshot HS imaging devices [20], [21], [22], [23], [24], [25], [26], [27] capture HS images during a single integration time of a detector array. It is normally a two-stage imaging process including spectral data collection and 3-D HS cube reconstruction. However, they still rely on labor-intensive post-processing. More recently, coded aperture snapshot spectral imagers (CASSIs) [56], [57], [58], [59] take advantage of compressive sensing theory, which capture essential information with a reduced amount of measurements.

B. HS Reconstruction From RGB Images

Considering the high cost of HS imaging devices and the long operation time, HS reconstruction from RGB images has attracted more research and industrial interest recently. There are many learning-based approaches have been developed, e.g., correlation regression [60], radial basis function network [45], principal component analysis [61], [62], manifold-based mapping [63], Gaussian process [64], and sparse coding [30], [59], [65], [66]. Recently, CNN-based methods have been more common since they learn the mapping based on rich data prior from large datasets. They can be briefly categorized into three classes according to network architectures, as shown in Fig. 2: 1) sequential convolutions with residual connections [32], [33], [34], [35], [36], [53], [67]; 2) U-Net [37], [38], [50], [51], [52]; and 3) hierarchical architecture [39], [40].

For instance, Choi et al. [68] built an implicit HS prior based on CASSI reconstruction by the latent representation of a learned autoencoder. It was enhanced by spatial–spectral prior [69], [70] and self-attention [39], [71]. In recent NTIRE 2020 contest [48], Li et al. [53] proposed an adaptive weighted attention network (AWAN) with eight compact dual residual attention blocks (DRABs). DRAB consists of two convolution paths that represent long and short skip connections simultaneously. Zhao et al. [40] proposed a hieratical regression network (HRNet) with residual dense block [72] and global block [73]. It achieved better performance when recovering HS images from real-world noisy RGB images but not produced comparable results on clean RGB images. Though they applied self-ensemble or model-ensemble strategies, they are still hard to obtain consistent results on different types of input images.

C. Generative Adversarial Network

GAN was firstly proposed by Goodfellow et al. [43]. It has been utilized in many low-level vision areas like deblurring [74] and colorization [75] as an auxiliary loss to improve the reconstruction accuracy and sharpness. The original training scheme of generative adversarial network (GAN) often leads to loss fluctuation and mode collapse. To address these issues, researchers have developed new loss functions (e.g., LSGAN [76], WGAN [77], WGAN-GP [78]), effective training techniques (e.g., two time-scale update rule (TTUR) [79]), and normalization (e.g., spectral normalization [80]). For the HSGAN, we adopt the WGAN training scheme with spectral normalization on the discriminator.

III. Problem Formulation

Given an RGB image $x = [r, g, b]$, HS reconstruction algorithms learn a mapping function F to obtain an HS estimation $\hat{s} = [\hat{s}_1, \hat{s}_2, \hat{s}_3, \dots, \hat{s}_n]$, where n denotes the maximum index of spectral bands and every specific $\hat{s}_k$ records the information of a narrow band of the whole spectrum. For CNN-based algorithms [32], [33], [34], [35], [36], [37], [38], [39], [40], [50], [52], [53], the HS reconstruction is formulated as maximizing a posteriori of HS images conditioned on RGB images

$$\Theta^* = \arg\max_{\Theta} p(s|x, \Theta) \tag{1}$$

where x and s are input RGB images and ground-truth HS images, respectively. Θ^* is the theoretically optimal parameters of the CNN, which serves as the mapping function F.

We include five datasets in the training and evaluation: NUS [45], Harvard [46], ICVL [30], CAVE [47], and ARAD [48]. We name the HS reconstruction from original RGB images in the datasets as *"HS from clean images with non-fixed CRF"* since the CRFs [28] are normally unknown (expect ARAD).

To demonstrate that the proposed HSGAN can fit different CRFs, we apply a real known CRF extracted from ARAD [48] to CAVE [47] and ICVL [30] since HS images of these three datasets have the same bandwidths, the number of bands, and spectrum ranges. We name the HS reconstruction from these generated RGB images as *"HS from clean images with fixed CRF."* Since the CRF from ARAD [48] is linear, the data synthesis can be formulated as follows:

$$x = s \times \mathrm{CRF} \tag{2}$$

where CRF is a $n \times 3$ matrix [48] and s are ground-truth HS images. x is the generated three-channel RGB image. The colors of the generated x might be different from original RGB images in the datasets, e.g., color tones of images in Fig. 3(a) and (b) are different since the CRFs of CAVE [47] and ARAD [48] datasets might be different, though they share the same HS spectrum.

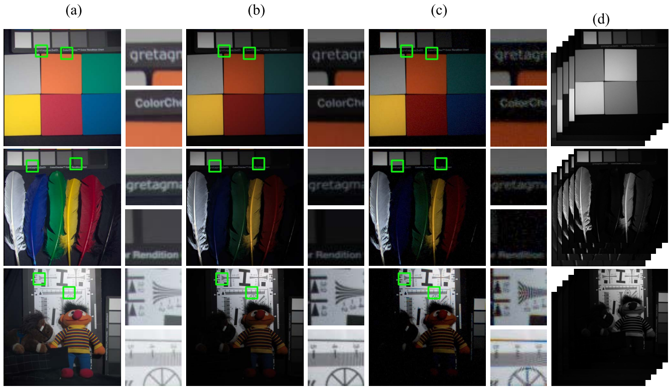

Fig. 3. Illustration of three settings of HS reconstruction. (a) HS from clean images with non-fixed CRF. (b) HS from clean images with fixed CRF. (c) HS from real-world images with fixed CRF. (d) Corresponding HS images. The samples are "beads," "feathers," and "chart_and_stuffed_toy" in the CAVE dataset. There are 31 channels of HS images and we only show five channels for visualization.

In addition, there often exist noises in real-world captured RGB images. In this situation, the algorithms need to learn HS reconstruction and denoising simultaneously. We name such setting as *"HS from real-world images with fixed CRF."* To model the real-world RGB images, there are three sequential steps based on the image signal processor pipeline [48].

1) We apply a real CRF [48] to HS data s to obtain an RGB image x [please refer to (2)].
2) Since real noises emerge on RAW data [49], we convert RGB image x to RAW space by mosaicking operation and then add noises. The noises include shot noise and read noise, which aligns with the setting of [48]. The shot noise is firstly added and then the read noise is added. Next, we convert the noisy RAW back to RGB space by demosaicking operation.
3) We save the RGB image to JPEG format.

The aforementioned steps can be represented as follows:

$$x = \mathrm{JPEG}(\mathrm{Demos}(\mathrm{Mos}(s \times CRF) + n)) \tag{3}$$

where Mos(∗) and Demos(∗) are mosaicking and demosaicking operations, respectively. The additive noise n includes shot noise and read noise, where the former follows Poisson distribution (signal-related) and the latter follows Gaussian distribution (signal-independent). The JPEG operation models JPEG compression artifacts for demosaicked images. In conclusion, (3) results in a noisy RGB image x from an HS images s. Some generated real-world noisy samples are illustrated in Fig. 3(c).

IV. Methodology

A. HSGAN Architecture

HSGAN consists of a generator F and a discriminator D, as shown in Figs. 4 and 5, respectively. The generator is a top-down hierarchical architecture that can be divided into four levels, as shown in Fig. 4(a). The inputs for levels 1–3

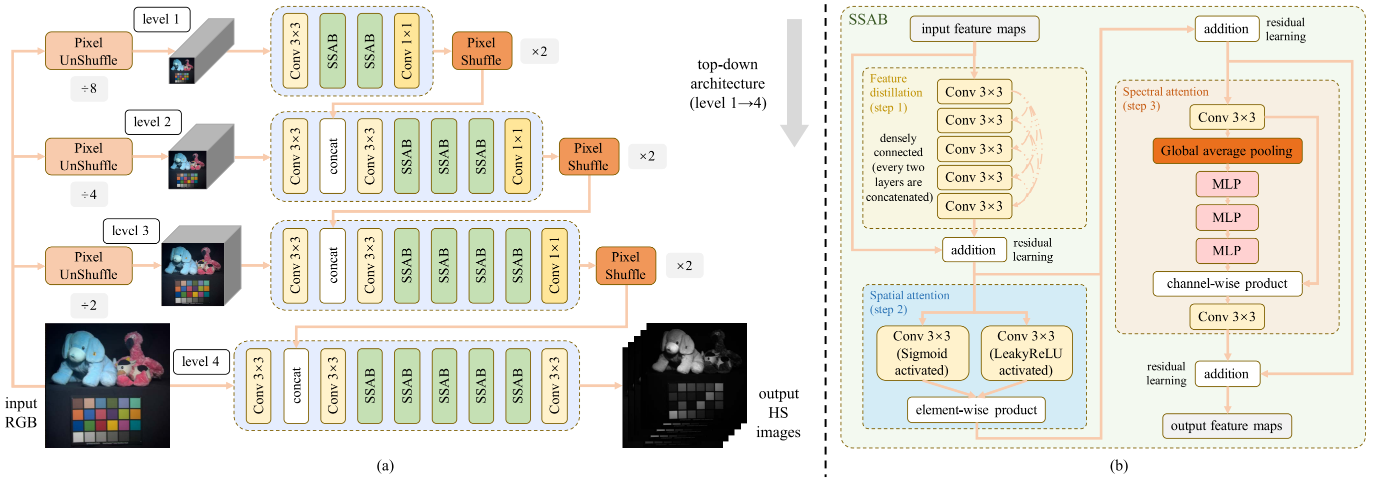

Fig. 4. Illustration of (a) HSGAN generator F, where the input is an RGB image and the output is multichannel HS images and (b) SSAB detailed structure.

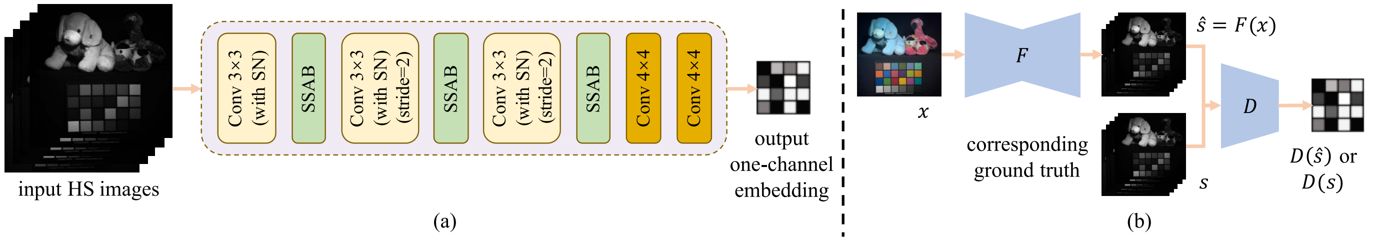

Fig. 5. Illustration of (a) HSGAN discriminator D, where HS images are input and the output is a one-channel embedding and (b) training pipeline of HSGAN.

are obtained by applying PixelUnShuffle [81] to the input RGB image. The features from the previous levels are passed to the next levels by PixelShuffle (reverse operation of PixelUnShuffle) and concatenation, which makes the next levels have distilled information with a larger perceptive field. The (SSABs, which will be discussed in Section IV-B) are used at every level to improve the feature representation. The 1 × 1 convolutional layers are further used to refine the channel information in the first three levels to refine the channel information before passing the features to the next levels. We conclude five advantages for the architecture.

1) Multilevel (multiscale) feature extraction and interlevel relation modeling.
2) Improved feature representation ability by SSABs.
3) Refined channel information at each shallower level by a 1 × 1 convolutional layer.
4) Enlarged perceptive field by PixelUnShuffle with limited computational consumption.
5) Lossless downsampling and upsampling by PixelUnShuffle and PixelShuffle.

The discriminator is PatchGAN-styled [82], i.e., the output is a one-channel embedding, as shown in Fig. 5(a). We use convolutional layers and SSABs alternately. The resolution of output embedding is 30 × 30 when the input has a size of 256 × 256. At training, the discriminator is used to compute adversarial loss and the relationship between the generator and discriminator is shown in Fig. 5 (b).

B. Spatial–Spectral Attention Block

HS reconstruction from RGB images aims to interpolate missing compact spectral information from sparse three visible-color channels, which is highly ill-posed. Therefore, good generalization capability and reconstruction accuracy highly rely on the feature representation ability of CNNs. To improve it, we have two assumptions: 1) different pixels have different significance, e.g., some pixels might have very small values or noisy and 2) different channels might provide different amount of information, e.g., red channel might have less relations to far spectral bands. Built upon them, we propose a SSAB by emphasizing both spatial correlations and interchannel relations, whose structure is shown in Fig. 4(b).

It contains three sequential operations, i.e., feature distillation, spatial attention, and spectral attention. First, we use five densely connected convolutional layers [67], [72] for feature distillation. It provides refined features for the following spatial and spectral attention operations. Second, we perform spatial attention to reweight the importance of every pixel. It is implemented by the element-wise product of two branches from the same layer. The first branch is activated by Sigmoid, which serves as the weighting feature since its range is in [0,1]. Third, we perform spectral attention to learn interchannel correlations. The significance of channels is computed by a global average pooling layer with three MLP layers as in [73].

C. Training Strategy and Loss Functions

We define two training stages for HSGAN generator F, where the first stage learns good initial weights for it, and we perform adversarial training in the second stage. With this training strategy, adversarial training becomes more stable and effective.

TABLE I
SUMMARIZATION OF DIFFERENT DATASETS USED IN THE EXPERIMENT

Method	Bandwidth / nm	Spatial resolution	Spectrum range / nm	Number of bands	HS from clean images with non-fixed CRF	HS from clean images with fixed CRF [48]	HS from real-world images with fixed CRF [48]
NUS [45]	6.25	1392×951	400 - 700	49	✓	-	-
Harvard [46]	10	1392×1040	420 - 720	31	✓	-	-
ICVL [30]	10	1392×1300	400 - 700	31	✓	✓	✓
CAVE [47]	10	512×512	400 - 700	31	✓	✓	✓
ARAD [48]	10	512×482	400 - 700	31	✓	✓	✓

In the first stage, we adopt L1 loss, which measures the disparity between generated HS images and ground truth at each pixel. It is defined as follows:

$$L_1 = \mathbb{E}[||\hat{s} - s||_1] = \sum_{k=1}^{n} \mathbb{E}[||\hat{s}_k - s_k||_1] \tag{4}$$

where $\hat{s}$ and s are the generated and ground-truth HS images, respectively. The s_k is kth narrow band of the whole spectrum.

In the second stage, we perform GAN training, i.e., the HSGAN discriminator and generator are trained alternatively. We adopt WGAN loss [77] to promote the decrease of Wasserstein distance between data distributions of generated HS images and ground truth. Also, we attach spectral normalization [80] to each convolutional layer of the discriminator to ensure that HSGAN meets 1-Lipschitz continuity. The GAN loss in the discriminator D training step is defined as follows:

$$L_D = \mathbb{E}[D(\hat{s})] - \mathbb{E}[D(s)] \tag{5}$$

where $D(*)$ is the HSGAN discriminator, which outputs an embedding. The operator $\mathbb{E}(*)$ computes the average across all pixels of the embedding and outputs a scalar. The weights of the generator are not updated during the discriminator training step. Then, the GAN loss in the generator F training step is defined as follows:

$$L_G = -\mathbb{E}[D(\hat{s})] \tag{6}$$

where $\hat{s} = F(x)$. D is used to obtain the embedding for computing generator loss, whose weights are fixed. In addition, we use L1 loss for training, and the full loss function of the HSGAN generator is given by

$$\text{Loss} = \lambda_1 L_1 + \lambda_G L_G \tag{7}$$

where λ_1 and λ_G are trade-off parameters.

D. Optimization Details

For all datasets and settings, we extract 256 × 256 patches from input RGB images and ground-truth HS images, which are then normalized into range of [0,1]. The parameters of HSGAN are initialized with Xavier [83]. HSGAN is trained for 200 000 iterations, where 50 000 iterations are for the first stage and the remaining 150 000 iterations are for the second stage. The learning rate for the HSGAN generator is initially set to 1×10^{-4} and halved every 50 000 iterations. The learning rate for the HSGAN discriminator is the same as the generator. The trade-off parameters λ_1 and λ_G are empirically set to 10 and 1, respectively. We use Adam optimizer [84] with parameter $\beta_1 = 0.5$, $\beta_2 = 0.999$. The batch size is fixed to 1. We implement the proposed HSGAN with PyTorch 1.0.0 and train it on a single NVIDIA TITAN Xp GPU.

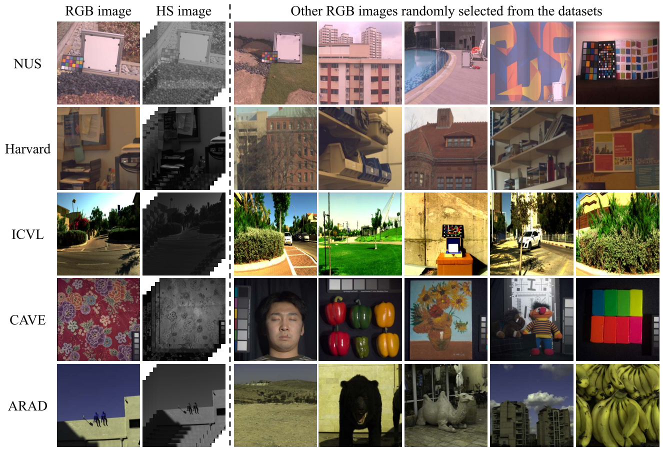

Fig. 6. Illustration of RGB images sampled from five datasets with HS images for the first RGB image. Only five channels of HS images are illustrated for visualization; the actual number of channels are larger than five.

V. EXPERIMENT

A. Experiment Settings

1) Datasets: We use five well-known datasets (NUS [45], Harvard [46], ICVL [30], CAVE [47], and ARAD [48]). They include different spectrum ranges (Harvard and other datasets), bandwidth (NUS and other datasets), and the number of bands (NUS and other datasets), as shown in Table I. We show some samples selected from the datasets in Fig. 6, which vary in scenes (e.g., park, lab, university, and field) and light conditions (e.g., natural light and indoor light). We randomly divide the number of training and validation images according to the 14:1 ratio. We introduce the details for them as follows.

1) *NUS [45]:* It contains 66 HS images with their illuminations (where four images are for validation). Different color temperatures were considered in capturing process. The raw data ranges from 400 to 1000 nm and has 396 bands overall. We extract a downsampled version, where the bands are between 400 and 700 nm. The RGB images were mapped by the Canon 1-D Mark III CRF. However, in our experiment, we assume the function is blind so that the reconstruction is non-fixed.
2) *Harvard [46]:* It includes overall 50 HS images ranging from 420 to 720 nm (where four images are for validation). They were taken by a commercial HS camera (Nuance FX, CRI Inc.) in daylight illumination. The camera is equipped with an apochromatic lens

(CoastalOpt UV-VIS-IR 60 mm Apo Macro, Jenoptik Optical Systems, Inc.) It includes both indoor and outdoor images with a variety of objects and materials. All HS images were rendered into RGB format by a specific mapping function.

3) *ICVL [30]:* It includes 201 HS images over 519 spectral bands (where 11 images are for validation). The images were taken by a line scanner (Specim PS Kappa DX4 HS camera). The authors downsampled the original data to 31 bands from 400 to 700 nm. It includes many different indoor and outdoor scenes. The corresponding RGB images were mapped by the CIE 1964 color matching functions [85].

4) *CAVE [47]:* It contains 32 HS images (where two images are for validation). Each HS image ranges from 400 to 700 nm with 10 nm increments. The photos were captured by a cooled CCD camera (Apogee Alta U260). They were taken in the laboratory including textiles, skin, drinks, paints, vegetables, etc.

5) *ARAD [48]:* It contains 510 HS images (where ten images are for validation) that are much larger than previous datasets. They were collected with a Specim IQ mobile HS camera, which is a push–broom imaging system. There are 204 spectral bands in the 400–1000 nm range, but they are downsampled to 400–700 nm to fit previous settings.

2) HS Reconstruction Settings: There are three settings: 1) HS reconstruction from clean images with non-fixed CRF; 2) HS reconstruction from clean images with fixed CRF from ARAD [48]; and 3) HS reconstruction from real-world images with fixed CRF from ARAD [48]) in the experiment (please refer to Section III for more details). Note that the RGB images in different settings share the same ground-truth HS images.

3) Baselines: We include 11 most recent approaches for comparisons, including encoder–decoder architecture [38], [50], [51], [52], sequential convolution architecture [32], [33], [34], [35], [53], and hieratical architecture [39], [40]. We follow original papers to define their hyperparameters. At training, the baselines share the same data with HSGAN. For evaluation, we generate a full HS spectrum patch-by-patch, where the patch size equals 256 × 256 for all baselines and HSGAN.

4) Evaluation Metrics: There are four metrics utilized for the evaluation of HS reconstruction algorithms. We utilize mean relative absolute error (MRAE) and root mean square error (RMSE) to measure the pixel-wise disparity between the generated and ground truth. We also adopt spectral angle mapper (SAM) [86] to measure spectral fidelity. We use back projection MRAE (BPAE) to compute differences between recovered RGB images (obtained by multiplying generated HS images and CRF) and input RGB images. Note that, BPAE is only available if CRF is known (i.e., only for evaluating HS from clean images with fixed CRF setting). All the aforementioned metrics are defined as follows:

$$\text{MRAE} = \frac{1}{n}\sum_{k=1}^{n}\frac{|\,\hat{s}_k - s_k\,|}{s_k} \tag{8}$$

TABLE II

HS RECONSTRUCTION FROM CLEAN RGB IMAGES WITH NON-FIXED CRF COMPARISONS ON ARAD VALIDATION SET. THE RED, BLUE, GREEN COLORS DENOTE THE BEST, THE SECOND BEST, AND THE THIRD BEST PERFORMANCE, RESPECTIVELY

Method	Architecture	ARAD			
		MRAE	RMSE	SAM	BPAE
U-Net [51]	Enc-Dec	0.0475	0.0142	0.0505	0.0425
LSS [50]	Enc-Dec	0.0458	0.0162	0.0560	0.0412
RSCNN [38]	Enc-Dec	0.0461	0.0157	0.0572	0.0413
MXR-U-Net [52]	Enc-Dec	0.0476	0.0162	0.0582	0.0428
HSCNN [32]	Seq Conv	0.0458	0.0155	0.0553	0.0406
HSCNN++ [35]	Seq Conv	0.0459	0.0157	0.0557	0.0417
2d-3d CNN [33]	Seq Conv	0.0470	0.0159	0.0575	0.0422
EffCNN [34]	Seq Conv	0.1755	0.0504	0.1480	0.1599
AWAN [53]	Seq Conv	0.0425	0.0139	0.0508	0.0382
LWRDANet [39]	Hierarchical	0.0754	0.0435	0.1240	0.0703
HRNet [40]	Hierarchical	0.0424	0.0135	0.0498	0.0378
HSGAN	Hierarchical	0.0416	0.0129	0.0501	0.0350

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(\hat{s}_k - s_k)^2} \tag{9}$$

$$\text{SAM} = \cos^{-1}\left(\frac{\hat{s}\cdot s}{||\hat{s}||\cdot||s||}\right) \tag{10}$$

$$\text{BPAE} = \frac{1}{n}\frac{|\;\hat{s}\times\text{CRF}-x\;|}{x} \tag{11}$$

where $\hat{s}_k$ and s_k are the kth channel of the generated HS images $\hat{s}$ and ground-truth HS images s, respectively.

B. Evaluation of HS Reconstruction From RGB Images With Non-Fixed CRF

In this section, we evaluate the performance of HSGAN and 11 baselines [32], [33], [34], [35], [38], [39], [40], [50], [51], [52], [53] on original data from five datasets [30], [45], [46], [47], [48].

The quantitative results are concluded in Tables II and III. The HSGAN obtains the best performance on the Harvard, ICVL, CAVE, and ARAD datasets and the top-3 performance on the NUS dataset. It demonstrates that the proposed HSGAN reconstructs the HS spectrum with higher fidelity than other methods. Also, it generalizes well to different HS datasets, i.e., different HS imaging devices or CRFs. However, other methods cannot obtain consistently good results on all the datasets, e.g., RSCNN obtains the best MRAE on the Harvard dataset, but its performance is not good enough on other datasets. Though the state-of-the-art HRNet [40] achieves the best results on the NUS dataset, it does not perform well on both Harvard and ICVL datasets. The proposed HSGAN has better results in the most of datasets and metrics.

We illustrate HS reconstruction results on the ICVL dataset in Fig. 7. In the figure, the error map of a spectral band (i.e., error map e_k) for a specific method is computed as $e_k = |\hat{s}_k - s_k|$, where e_k, $\hat{s}_k$ and s_k are the kth error map, kth channel of the generated and ground-truth HS images, respectively. Compared with other methods, HSGAN yields better results since the error maps on different spectral channels approach 0 (e.g., if the color more approaches blue, the value approaches 0, as shown in the color bar of Fig. 7).

TABLE III

HS Reconstruction From Clean RGB Images With Non-Fixed CRF Comparisons on Four Validation Sets. The Red, Blue, Green Colors Denote the Best, the Second Best, and the Third Best Performance, Respectively

Method	NUS			Harvard			ICVL			CAVE		
	MRAE	RMSE	SAM	MRAE	RMSE	SAM	MRAE	RMSE	SAM	MRAE	RMSE	SAM
U-Net [51]	0.3360	0.0747	0.2593	0.0780	0.0102	0.0541	0.2902	0.0868	0.2168	0.1094	0.0248	0.0962
LSS [50]	0.3775	0.0896	0.4188	0.0676	0.0102	0.0555	0.1348	0.0324	0.1807	0.2404	0.0235	0.0936
RSCNN [38]	0.3982	0.1216	0.3317	0.0655	0.0096	0.0520	0.1903	0.0453	0.2379	0.0915	0.0268	0.1055
MXR-U-Net [52]	0.4162	0.1217	0.3285	0.0673	0.0095	0.0513	0.1675	0.0410	0.2163	0.1603	0.0282	0.1117
HSCNN [32]	0.4187	0.1316	0.3562	0.0721	0.0098	0.0524	0.1480	0.0360	0.1993	0.1787	0.0283	0.1781
HSCNN++ [35]	0.3916	0.1298	0.3249	0.0659	0.0095	0.0516	0.1450	0.0360	0.1943	0.1507	0.0283	0.1137
2d-3d CNN [33]	0.4604	0.1277	0.3735	0.0667	0.0092	0.0501	0.1657	0.0399	0.2166	0.1421	0.0284	0.1136
EffCNN [34]	0.4779	0.0913	0.4273	0.1970	0.0233	0.1177	0.2206	0.0479	0.2269	0.2138	0.0467	0.1569
AWAN [53]	0.4678	0.1489	0.4847	0.0679	0.0100	0.0564	0.3357	0.0335	0.1866	0.2714	0.0579	0.1832
LWRDANet [39]	0.4560	0.0974	0.4308	0.0707	0.0107	0.0589	0.1973	0.0457	0.2173	0.1486	0.0512	0.1891
HRNet [40]	0.2796	0.0739	0.2768	0.0667	0.0095	0.0509	0.1417	0.0356	0.1879	0.0838	0.0263	0.1047
HSGAN	0.2838	0.0845	0.2750	0.0658	0.0092	0.0508	0.1256	0.0319	0.1833	0.0736	0.0254	0.0958

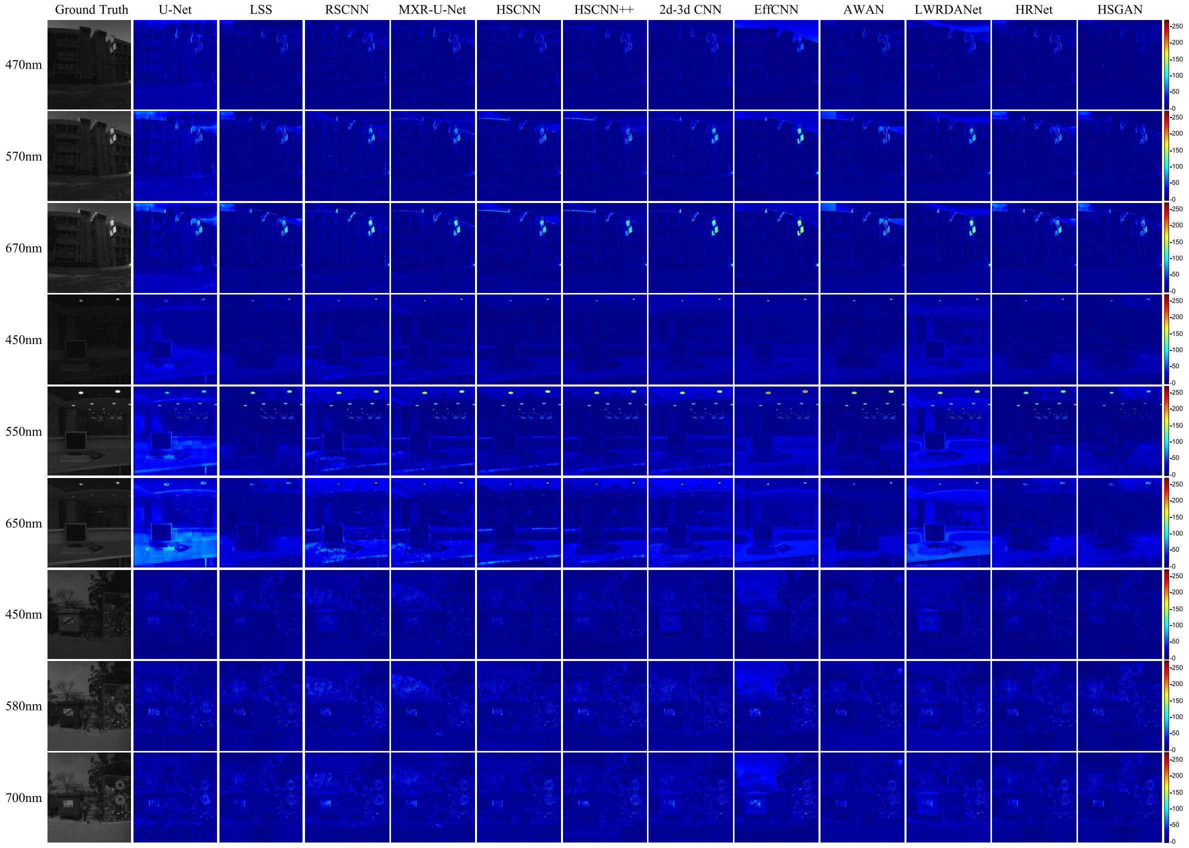

Fig. 7. Illustration of error maps of HS reconstruction from clean RGB images with non-fixed CRF on ICVL validation set. There are three scenes ("bguCAMP_0514-1718," "lst_0408-1004," and "omer_0331-1118") and we select different bandwidths for them. The left column includes one channel from the ground-truth spectrum. The other columns include the error maps of 11 baselines and HSGAN. Please zoomed-in view for a better view.

The top-ranked methods LSS, AWAN, and HRNet produce more artifacts on the shown samples than HSGAN. It demonstrates that HSGAN architecture performs better than others in terms of HS reconstruction quality. Similarly, we illustrate the performance of five top-ranked methods on the other four datasets in Fig. 8. HSGAN obtains better results than other methods on the shown samples. The state-of-the-art methods AWAN and HRNet produce more noisy results than HSGAN, e.g., HRNet on "balloons" sample of the CAVE dataset, and AWAN on "ARAD_HS_0455" sample of the ARAD dataset. It also proves that HSGAN is general to different datasets.

Also, we draw MRAE curves on randomly chosen points for the generated HS images from different methods in Fig. 9, where a line with a lower position denotes the corresponding method is more accurate. HSGAN (red lines) obtains lower MRAE values on most of the sampled points compared with other top-ranked methods (other color lines). It further proves that HSGAN has good generalization ability and achieves good performance on different HS datasets.

C. Evaluation of HS Reconstruction From RGB Images With Fixed CRF

In this section, we evaluate the performance of HSGAN and 11 baselines [32], [33], [34], [35], [38], [39], [40], [50], [51], [52], [53] on ICVL, CAVE, and ARAD datasets [30], [47],

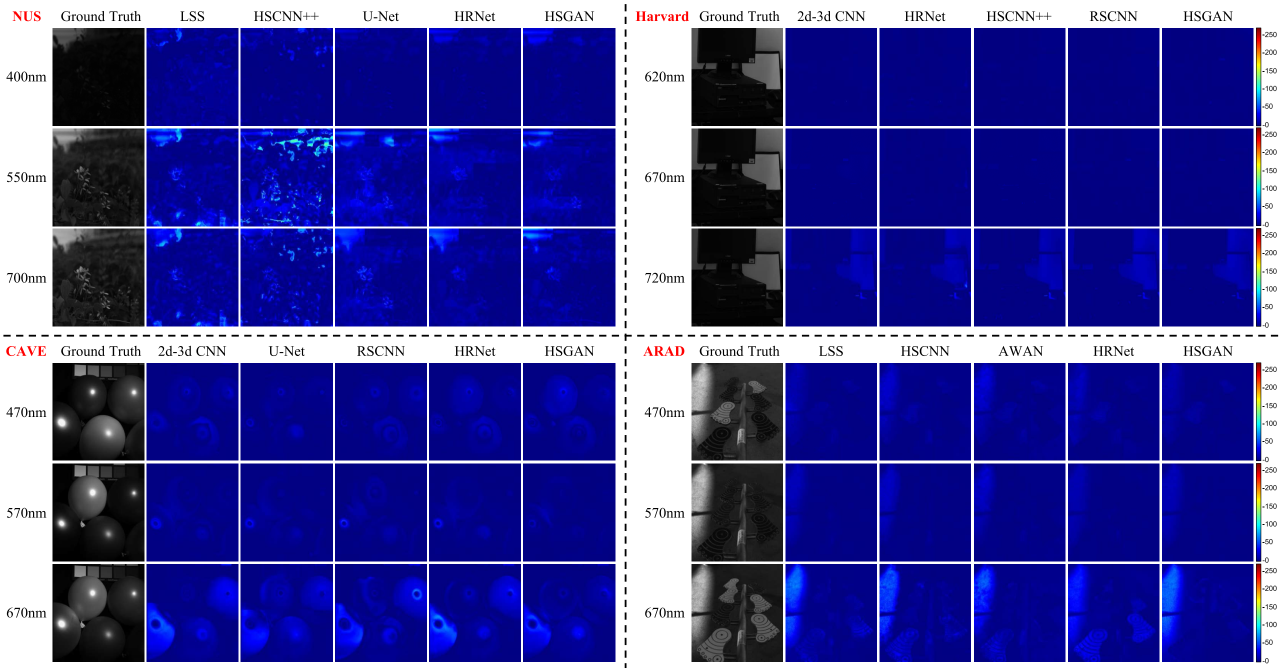

Fig. 8. Illustration of error maps of HS reconstruction from clean RGB images with non-fixed CRF on NUS, Harvard, CAVE, and ARAD datasets, respectively. Only the five top-ranked methods are shown for each dataset. There are four groups (“Scene06” from the NUS dataset, “imgh1” from the Harvard dataset, “balloons” from the CAVE dataset, and “ARAD_HS_0455” from the ARAD dataset) and we select different bandwidths and ranges for them. In each group, the left column includes one channel from the ground-truth spectrum. The other columns include the results of five baselines and HSGAN.

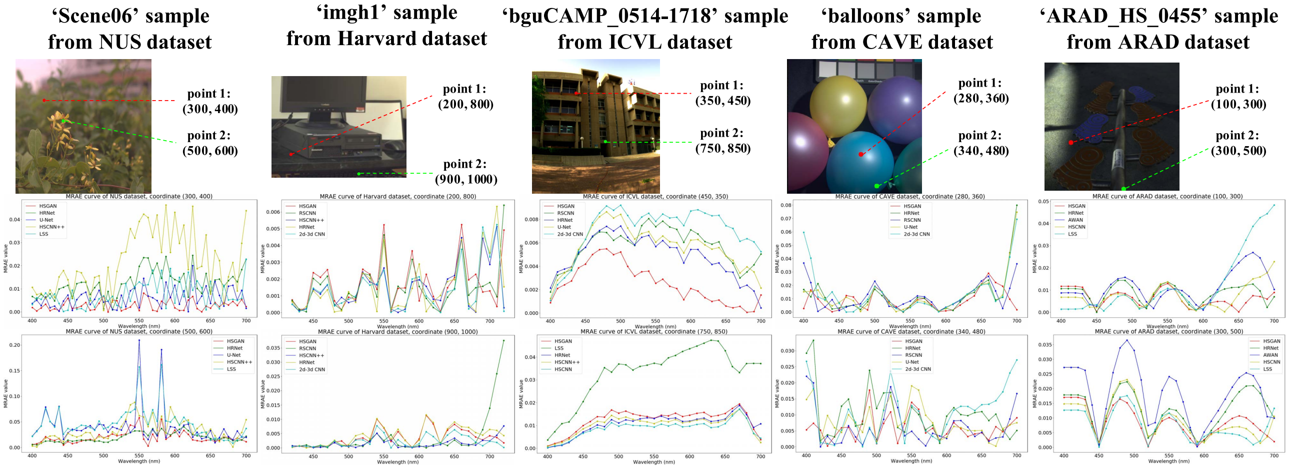

Fig. 9. Illustration of MRAE curves on a single point of 11 baselines and HSGAN. In the first row, the RGB images are inputs of the methods, where two sample points are highlighted. The following two rows include the MRAE values at each wavelength. Only the top-performed five methods of each dataset are illustrated, which is the same as Fig. 8.

[48] from clean or real-world RGB images with fixed CRF. The synthetic clean or real-world RGB images are generated using the CRF in the ARAD dataset [48] and (2) and (3).

The quantitative analysis is included in Tables IV and V. HSGAN obtains the best HS reconstruction accuracy on ICVL, CAVE, and ARAD datasets from both clean and real-world RGB images with fixed CRF. It demonstrates that HSGAN architecture generalizes well to real-world noises. HSGAN generator extracts features at four scales from compact to fine. It effectively expands the perceptive field that is beneficial to image denoising. Compared with existing methods, HSGAN adopts SSAB, which models both spatial and spectral attention. By concatenating SSABs, HSGAN obtains superior feature representation ability, which is helpful for HS reconstruction and image denoising. In addition, a tone mapping 1 × 1 convolution is used to refine the features at levels 1–3 before feeding into the next level. The channel information is better reweighted, which could be valuable for spectral interpolation. Therefore, HSGAN achieves better results than state-of-the-art methods.

We illustrate error maps of HS reconstruction on the ICVL validation set with clean and real-world settings in Fig. 10. The error maps of HS reconstruction results of three bands (450, 550, and 650 nm) and from clean or real-world RGB images are illustrated. It is clear that there are lower errors in HSGAN’s results, which shows that

TABLE IV

HS RECONSTRUCTION FROM RGB IMAGES WITH FIXED CRF FROM ARAD DATASET COMPARISONS ON ICVL AND CAVE VALIDATION SETS. THE "CLEAN" DENOTES THE RGB IMAGES ARE CLEAN [OBTAINED BY (2)] AND "REAL-WORLD" REPRESENTS THEY ARE NOISY [OBTAINED BY (3)]. THE RED, BLUE, GREEN COLORS DENOTE THE BEST, THE SECOND BEST, AND THE THIRD BEST PERFORMANCE, RESPECTIVELY

Method	ICVL (clean)				CAVE (clean)				ICVL (real-world)			CAVE (real-world)		
	MRAE	RMSE	SAM	BPAE	MRAE	RMSE	SAM	BPAE	MRAE	RMSE	SAM	MRAE	RMSE	SAM
U-Net [51]	0.0196	0.0056	0.0375	0.0174	0.1226	0.0234	0.0908	0.0905	0.0225	0.0061	0.0398	0.2245	0.0244	0.0936
LSS [50]	0.0156	0.0042	0.0287	0.0129	0.1284	0.0276	0.1098	0.1025	0.0198	0.0053	0.0350	0.1323	0.0238	0.0916
RSCNN [38]	0.0158	0.0044	0.0301	0.0134	0.1159	0.0236	0.0922	0.0892	0.0196	0.0056	0.0367	0.1720	0.0266	0.1022
MXR-U-Net [52]	0.0160	0.0043	0.0296	0.0135	0.0812	0.0241	0.0939	0.1601	0.0214	0.0056	0.0373	0.1472	0.0275	0.1066
HSCNN [32]	0.0179	0.0048	0.0311	0.0157	0.1189	0.0235	0.0909	0.0941	0.0354	0.0082	0.0509	0.1428	0.0284	0.1707
HSCNN++ [35]	0.0160	0.0044	0.0299	0.0137	0.1158	0.0230	0.0900	0.0927	0.0196	0.0052	0.0342	0.1339	0.0264	0.1034
2d-3d CNN [33]	0.0181	0.0051	0.0339	0.0158	0.1205	0.0256	0.1023	0.0959	0.0225	0.0061	0.0393	0.1401	0.0276	0.1079
EffCNN [34]	0.0156	0.0042	0.0376	0.0129	0.1284	0.0276	0.1956	0.1025	0.0198	0.0053	0.0701	0.1323	0.0238	0.1969
AWAN [53]	0.0198	0.0044	0.0299	0.0181	0.1067	0.0242	0.0974	0.1030	0.0207	0.0049	0.0330	0.2354	0.0245	0.1566
LWRDANet [39]	0.0205	0.0071	0.0449	0.0179	0.1391	0.0489	0.1768	0.1135	0.0262	0.0080	0.0500	0.1532	0.0492	0.1790
HRNet [40]	0.0162	0.0047	0.0322	0.0139	0.0721	0.0221	0.0857	0.0901	0.0273	0.0078	0.0489	0.0254	0.0244	0.0946
HSGAN	0.0147	0.0041	0.0279	0.0124	0.0757	0.0219	0.0914	0.0899	0.0185	0.0048	0.0326	0.0232	0.0235	0.0933

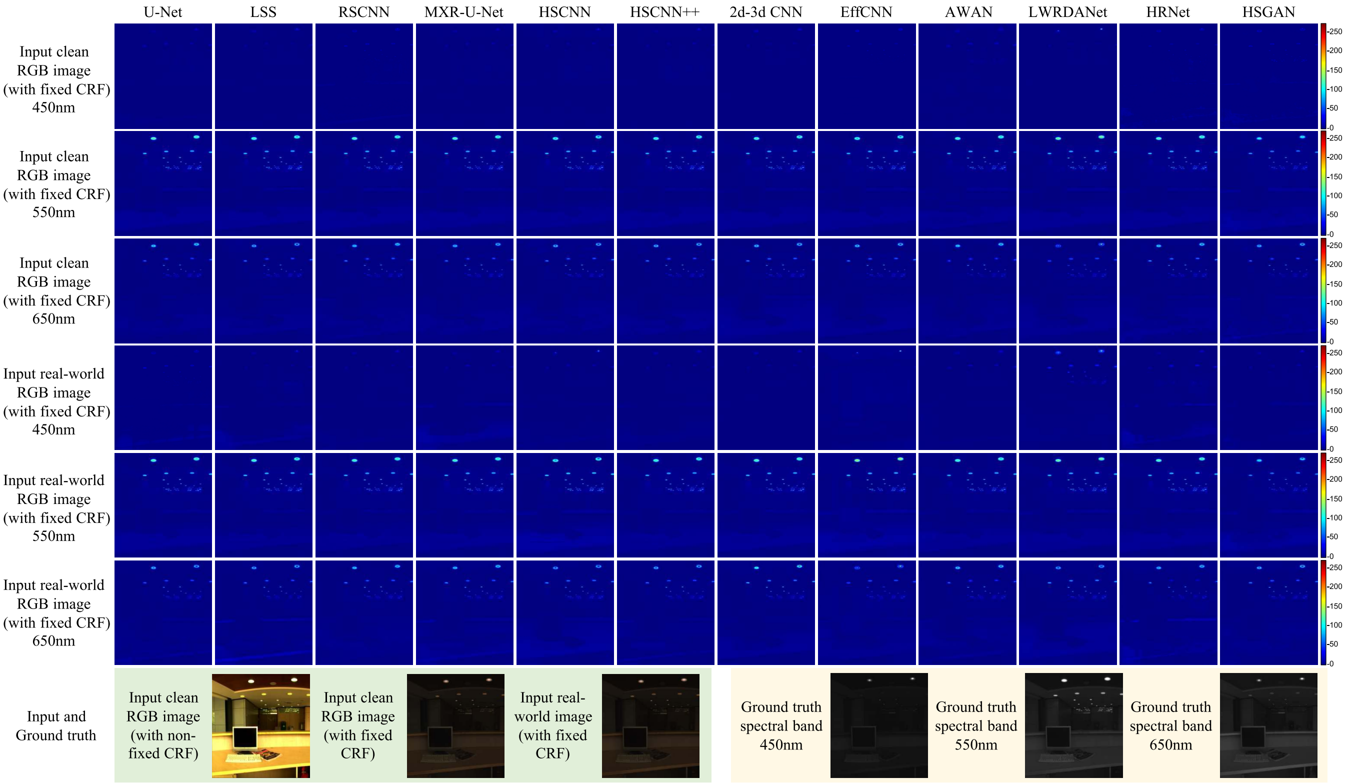

Fig. 10. Illustration of error maps of HS reconstruction from clean and real-world RGB images with fixed CRF on "lst_0408-1004" sample with three bands from the ICVL validation set. The columns include the error maps of 11 baselines and HSGAN from left to right. The 1–3 rows and the 4–6 rows include the error maps of HS reconstruction from clean and real-world RGB images with fixed CRF with 450, 550, and 650 nm bands, respectively. The last row represents the input RGB images of three settings and ground-truth spectral bands. Please zoomed-in view for a better view.

HSGAN produces more accurate HS bands under these settings.

We also illustrate error maps of reconstructed RGB images for the "HS reconstruction from clean images with fixed CRF" setting in Fig. 11. The images are obtained by post-processing generated HS images with the known CRF and (2). In the figure, the error map of a generated RGB image (i.e., e_{rgb}) for a specific method is the cumulative errors of R, G, and B channels, which is computed as follows:

$$e_{\text{rgb}} = |\hat{r}-r| + |\hat{g}-g| + |\hat{b}-b| \tag{12}$$

where e_{rgb} is the error map. The $\hat{r}$, $\hat{g}$, $\hat{b}$, r, g, and b are the R, G, B channel of the generated and ground-truth RGB input images, respectively. It is clear that HSGAN produces relatively cleaner reconstruction results than other methods.

In conclusion, the proposed HSGAN obtains better HS reconstruction results on five datasets (NUS, Harvard, ICVL, CAVE, and ARAD) with three different settings (HS reconstruction from clean images with non-fixed CRF, from clean images with fixed CRF, and from real-world images with fixed CRF). The experiments demonstrate HSGAN has better generalization ability across different RGB images and scenes (i.e., consistent results on different datasets and clean or noisy images) than existing methods. They also demonstrate HSGAN obtains superior HS reconstruction accuracy than existing methods.

TABLE V

HS Reconstruction From Real-World RGB Images With Fixed CRF From ARAD Dataset Comparisons. Since CRF Is From ARAD Dataset, "ARAD (Clean)" Is the Same as "ARAD" of Table II. The Red, Blue, Green Colors Denote the Best, the Second Best, and the Third Best Performance, Respectively

Method	ARAD (real-world)		
	MRAE	RMSE	SAM
U-Net [51]	0.0742	0.0186	0.0696
LSS [50]	0.0731	0.0187	0.0702
RSCNN [38]	0.0750	0.0191	0.0718
MXR-U-Net [52]	0.0749	0.0191	0.0722
HSCNN [32]	0.0734	0.0192	0.0720
HSCNN++ [35]	0.0712	0.0185	0.0689
2d-3d CNN [33]	0.0741	0.0190	0.0713
EffCNN [34]	0.1723	0.0506	0.1548
AWAN [53]	0.0699	0.0184	0.0681
LWRDANet [39]	0.1036	0.0451	0.1357
HRNet [40]	0.0682	0.0179	0.0679
HSGAN	0.0676	0.0180	0.0691

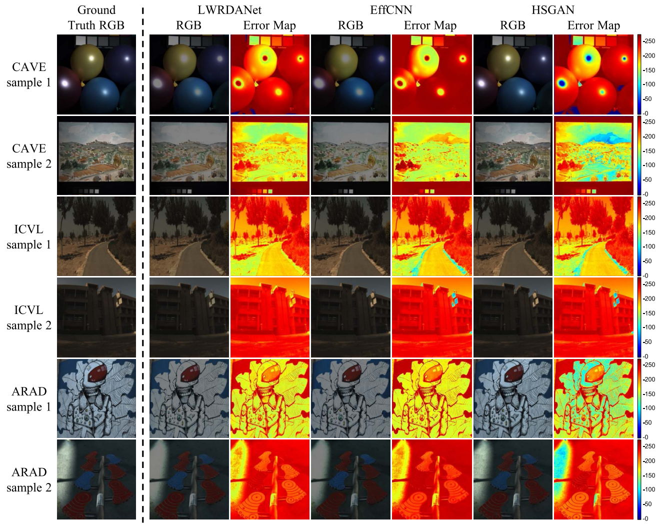

Fig. 11. Illustration of the generated RGB images ("RGB" in the figure) and error maps ("Error Map" in the figure) of HS reconstruction from clean RGB images with fixed CRF on selected samples on CAVE, ICVL, and ARAD validation sets. Please zoomed-in view for a better view.

D. Ablation Study

We perform the ablation study on adversarial training scheme and network architectures (including the number of blocks and the components of SSAB). There are overall 13 settings.

1) *w/o Adversarial Loss:* No adversarial training is used, i.e., only L1 loss is used for training HSGAN.
2) *w/o Feature Distill:* Feature distillation (step 1) in the SSAB is replaced by convolutional layers with the same numbers for fairness.
3) *w/o Spatial attn:* Spatial attention (step 2) in the SSAB is replaced by convolutional layers with the same numbers.
4) *w/o Spectral attn:* Spectral attention (step 3) in the SSAB is replaced by convolutional layers with the same numbers.
5) *w/o SSAB:* The full SSAB is replaced by convolutional layers with the same numbers.
6) *w/o 1 × 1 Conv:* No use of 1 × 1 convolutional layer at the end of each level of HSGAN generator;
7) *Kernel Size = 5 × 5:* The kernel sizes for all convolutional layers and blocks are changed to 5 × 5.
8) *w/o PixelUnShuffle:* PixelUnShuffle and PixelShuffle are replaced by bicubic sampling.

9)–13) *SSAB [n_1 n_2 n_3 n_4]:* The number of SSABs in levels 1–4 are changed to n_1, n_2, n_3, and n_4, respectively. We let [n_1 n_2 n_3 n_4] equal to [1234], [1345], [2245], [2344], and [2335]. Note that, the number of full HSGAN are [2345].

The experiments are conducted on the ICVL validation set with three settings. The results are concluded in Table VI. Compared with the full HSGAN, all ablation study settings lead to poorer values of metrics. It demonstrates that adversarial loss and all network components are beneficial to better HS reconstruction quality. In addition, we find that full HSGAN obtains the best performance across different input settings (i.e., clean images with non-fixed CRF, clean images with fixed CRF, and real-world images with fixed CRF) consistently. Therefore, adversarial loss and all network components contribute to consistent and high-quality HS reconstruction results.

For specific settings, setting 1) demonstrates the proposed adversarial training is beneficial to HS reconstruction quality across different input data. In addition, we find that HSGAN is hard to converge when kernel size is changed to 5 × 5 [setting 7)], e.g., the MRAE becomes "nan." The tone mapping 1 × 1 convolutions at the end of each level [setting 6)] are significant since the channel information is vital for HS reconstruction. Omitting 1 × 1 convolutions leads to much higher MRAE, i.e., from 0.1256 to 0.1869 for HS reconstruction from clean images with non-fixed CRF setting. PixelUnShuffle sampling [setting 8)] is also helpful since it is a lossless sampling method compared with bicubic sampling. The SSAB is also significant since settings 2)–5) show that SSAB has a much stronger feature extraction ability than normal convolutions. For settings 9)–13), we observe that the number of SSABs are closely related to the HS reconstruction performance. The proposed HSGAN adopts proper number of SSABs and achieves better results than the other settings.

In conclusion, the proposed adversarial training scheme and network architectures (including the number of blocks and the components of SSAB) are all significant for HSGAN.

E. Discussion on SSAB

To further demonstrate the proposed HSGAN with the SSAB is superior to the state-of-the-art method HRNet [40], we conduct the following additional experiments.

1) *HSGAN SSAB → GB:* We replace SSABs by the basic blocks in HRNet [40], where global attention is computed once at each level and spatial attention is not considered. The architecture of HSGAN is kept.
2) *HSGAN SSAB [2235]:* We keep the architecture of HSGAN but the number of SSABs in levels 1–4 are changed to 2, 2, 3, and 5, respectively. This setting uses the same number of blocks in HRNet [40].

TABLE VI
ABLATION STUDY RESULTS ON ICVL VALIDATION SET. THE TOP-PERFORMED METHOD IS LABELED IN RED

Ablation study setting	ICVL (clean, with non-fixed CRF)			ICVL (clean, with fixed CRF)				ICVL (real-world, with fixed CRF)		
	MRAE	RMSE	SAM	MRAE	RMSE	SAM	BPAE	MRAE	RMSE	SAM
1) w/o adversarial loss	0.1383	0.0329	0.1821	0.0153	0.0042	0.0288	0.0129	0.0186	0.0049	0.0326
2) w/o feature distill	0.1661	0.0401	0.2133	0.0152	0.0043	0.0292	0.0127	0.0188	0.0051	0.0341
3) w/o spatial attn	0.1398	0.0340	0.1855	0.0157	0.0042	0.0292	0.0135	0.0189	0.0049	0.0327
4) w/o spectral attn	0.1438	0.0348	0.1875	0.0163	0.0043	0.0287	0.0143	0.0188	0.0049	0.0329
5) w/o SSAB	0.1453	0.0335	0.1859	0.0182	0.0047	0.0307	0.0163	0.0189	0.0050	0.0330
6) w/o 1×1 Conv	0.1869	0.0458	0.2223	0.0157	0.0043	0.0299	0.0132	0.0188	0.0049	0.0330
7) kernel size = 5×5	cannot converge			cannot converge				cannot converge		
8) w/o PixelUnShuffle	0.1622	0.0394	0.2034	0.0155	0.0044	0.0297	0.0131	0.0193	0.0051	0.0344
9) SSAB [1234]	0.1537	0.0399	0.1922	0.0155	0.0043	0.0295	0.0129	0.0193	0.0051	0.0340
10) SSAB [1345]	0.1403	0.0344	0.1820	0.0149	0.0042	0.0289	0.0125	0.0193	0.0051	0.0334
11) SSAB [2245]	0.1426	0.0346	0.1836	0.0151	0.0041	0.0286	0.0127	0.0192	0.0051	0.0343
12) SSAB [2335]	0.1383	0.0335	0.1834	0.0155	0.0044	0.0296	0.0129	0.0195	0.0051	0.0339
13) SSAB [2344]	0.1515	0.0363	0.1907	0.0153	0.0044	0.0299	0.0127	0.0188	0.0050	0.0334
HSGAN (full)	0.1256	0.0319	0.1833	0.0147	0.0041	0.0279	0.0124	0.0185	0.0048	0.0326

TABLE VII
EXPERIMENT RESULTS ON THE ANALYSIS OF THE SSAB. THE TOP-PERFORMED METHOD IS LABELED IN RED

Experiment setting	ICVL (clean, with non-fixed CRF)			ICVL (clean, with fixed CRF)				ICVL (real-world, with fixed CRF)		
	MRAE	RMSE	SAM	MRAE	RMSE	SAM	BPAE	MRAE	RMSE	SAM
HRNet	0.1417	0.0356	0.1879	0.0162	0.0047	0.0322	0.0139	0.0273	0.0078	0.0489
1) HSGAN SSAB → GB	0.1397	0.0352	0.1868	0.0162	0.0046	0.0309	0.0137	0.0202	0.0057	0.0374
2) HSGAN SSAB [2235]	0.1393	0.0351	0.1860	0.0153	0.0043	0.0294	0.0127	0.0185	0.0049	0.0329
3) HSGAN SSAB → GB [2235]	0.1581	0.0400	0.2053	0.0165	0.0046	0.0313	0.0140	0.0206	0.0058	0.0380
HSGAN (full)	0.1256	0.0319	0.1833	0.0147	0.0041	0.0279	0.0124	0.0185	0.0048	0.0326

TABLE VIII
EXPERIMENT RESULTS ON THE ANALYSIS OF THE SSAB ON DOUBLE NOISE PARAMETERS SETTING. THE TOP-PERFORMED METHOD IS LABELED IN RED

Experiment setting	ICVL (real-world, with fixed CRF, double noise parameters)		
	MRAE	RMSE	SAM
HRNet	0.0278	0.0080	0.0496
1) HSGAN SSAB → GB	0.0216	0.0058	0.0377
2) HSGAN SSAB [2235]	0.0215	0.0058	0.0376
3) HSGAN SSAB → GB [2235]	0.0225	0.0061	0.0398
HSGAN (full)	0.0203	0.0057	0.0372

TABLE IX
EXPERIMENT RESULTS ON THE ANALYSIS OF THE SSAB ON TRIPLE NOISE PARAMETERS SETTING. THE TOP-PERFORMED METHOD IS LABELED IN RED

Experiment setting	ICVL (real-world, with fixed CRF, triple noise parameters)		
	MRAE	RMSE	SAM
HRNet	0.0326	0.0083	0.0516
1) HSGAN SSAB → GB	0.0237	0.0063	0.0401
2) HSGAN SSAB [2235]	0.0238	0.0062	0.0401
3) HSGAN SSAB → GB [2235]	0.0242	0.0063	0.0404
HSGAN (full)	0.0237	0.0062	0.0400

3) *HSGAN SSAB → GB [2235]:* We replace SSABs by the basic blocks in HRNet [40]. Then, we reduce the number of blocks in levels 1–4 to 2, 2, 3, and 5, respectively.

The experiment results are concluded in Table VII. First, if replacing SSABs in HSGAN with the basic blocks in HRNet [i.e., setting 1)], the HS reconstruction accuracy becomes worse (e.g., obvious increases of MRAE, RMSE, and SAM values). It demonstrates the proposed SSAB outperforms the blocks in the HRNet [40] given different inputs (e.g., noisy/clean RGB images). Second, we use the same number of blocks as in HRNet [40] (i.e., setting 2). The results are still inferior to the full HSGAN. It shows that the proper number of blocks of HSGAN is significant, which matches the conclusion in Section V-D. We observe that the results of setting 2) are slightly better than setting 1), e.g., for "HS from clean images with non-fixed CRF," the MRAE value of setting 2) is 0.1393 while it is 0.1397 for setting 1). It represents that the block designs might be more significant for the HSGAN to obtain higher HS reconstruction accuracy than the number of blocks. Third, we replace the SSAB and use the same number of blocks as in HRNet [40], i.e., setting 3). This setting is still inferior to the full HSGAN, which demonstrates the proposed HSGAN architecture contributes to better performance. Setting 3) is also inferior to settings 1) and 2) since there are two degradations applied to HSGAN.

In addition, we conduct two additional experiments by increasing the noise parameters by double or triple used in [48] (i.e., double/triple noise parameters compared with the setting "HS from real-world images with fixed CRF.") The noise parameters include the gain parameter of the shot noise (Poisson noise) and the sigma value of the zero-mean read noise (Gaussian noise). The experiment results are concluded in Tables VIII and IX, respectively. On the one hand, settings 1)–3) obtain lower HS reconstruction accuracy than HSGAN. It demonstrates the proposed SSAB and the number of SSABs are significant for HSGAN. On the other hand, the performance of HRNet is inferior to settings 1)–3) and HSGAN if the input is noisy (please the last three columns in Tables VII–IX). It shows that the proposed SSAB has significant value when receiving real-world noisy data. It greatly improves the denoising ability of HSGAN.

In conclusion, the proposed HSGAN with the SSAB is significant for achieving a high HS reconstruction quality.

VI. Conclusion

In this article, we present an HSGAN framework for automatically reconstructing the HS spectrum from RGB images. It is a GAN-based architecture and we propose a two-stage adversarial training strategy. The generator is a top-down four-level hierarchical architecture that extracts and combines features at different scales. We propose SSABs at each level to improve the feature representation. It includes three sequential operations, i.e., feature distillation, spatial attention, and spectral attention. The discriminator is a patch-based architecture and we also use SSABs in it. Finally, we conduct experiments on five well-known HS datasets (NUS, Harvard, ICVL, CAVE, and ARAD) with three different settings (HS reconstruction from clean images with non-fixed CRF, from clean images with fixed CRF, and real-world images with fixed CRF). The performance of HSGAN is superior to existing methods on all the datasets, which demonstrates its good generalization ability across different RGB images and scenes. Also, HSGAN recovers consistently higher quality HS spectrum from both clean and real-world images.

Acknowledgment

The authors would like to thank Qinbin Li and Xuehui Wang for their kind discussions and hardware support. They would also like to thank the anonymous reviewers and editors for many helpful comments for this work.

References

[1] F. Melgani and L. Bruzzone, “Classification of hyperspectral remote sensing images with support vector machines,” *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[2] M. J. Mendenhall and E. Merenyi, “Relevance-based feature extraction for hyperspectral images,” *IEEE Trans. Neural Netw.*, vol. 19, no. 4, pp. 658–672, Apr. 2008.

[3] P. Ghamisi, M. Dalla Mura, and J. A. Benediktsson, “A survey on spectral–spatial classification techniques based on attribute profiles,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2335–2353, May 2015.

[4] Y. Chen, X. Zhao, and X. Jia, “Spectral–spatial classification of hyperspectral data based on deep belief network,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015.

[5] Y. Gao, X. Wang, Y. Cheng, and Z. J. Wang, “Dimensionality reduction for hyperspectral data based on class-aware tensor neighborhood graph and patch alignment,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 8, pp. 1582–1593, Aug. 2015.

[6] P. Zhong and R. Wang, “Jointly learning the hybrid CRF and MLR model for simultaneous denoising and classification of hyperspectral imagery,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 7, pp. 1319–1334, Jul. 2014.

[7] Q. Wang, J. Lin, and Y. Yuan, “Salient band selection for hyperspectral image classification via manifold ranking,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1279–1289, Jun. 2016.

[8] N. Akhtar and A. Mian, “Nonparametric coupled Bayesian dictionary and classifier learning for hyperspectral classification,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4038–4050, Sep. 2018.

[9] R. Dian, S. Li, A. Guo, and L. Fang, “Deep hyperspectral image sharpening,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5345–5355, Nov. 2018.

[10] S. Jia, Z. Lin, B. Deng, J. Zhu, and Q. Li, “Cascade superpixel regularized Gabor feature fusion for hyperspectral image classification,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1638–1652, May 2020.

[11] J. Qin, K. Chao, M. S. Kim, R. Lu, and T. F. Burks, “Hyperspectral and multispectral imaging for evaluating food safety and quality,” *J. Food Eng.*, vol. 118, no. 2, pp. 157–171, Sep. 2013.

[12] J. A. Jablonski, T. J. Bihl, and K. W. Bauer, “Principal component reconstruction error for hyperspectral anomaly detection,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 8, pp. 1725–1729, Aug. 2015.

[13] Y. Zhang, Y. Xi, Q. Yang, W. Cong, J. Zhou, and G. Wang, “Spectral CT reconstruction with image sparsity and spectral mean,” *IEEE Trans. Comput. Imag.*, vol. 2, no. 4, pp. 510–523, Dec. 2016.

[14] Y. Zhang, X. Mou, G. Wang, and H. Yu, “Tensor-based dictionary learning for spectral CT reconstruction,” *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 142–154, Jan. 2017.

[15] H. R. Morris, C. C. Hoyt, and P. J. Treado, “Imaging spectrometers for fluorescence and Raman microscopy: Acousto-optic and liquid crystal tunable filters,” *Appl. Spectrosc.*, vol. 48, no. 7, pp. 857–866, Jul. 1994.

[16] E. Herrala, J. T. Okkonen, T. S. Hyvarinen, M. Aikio, and J. Lammasniemi, “Imaging spectrometer for process industry applications,” *Proc. SPIE*, vol. 2248, pp. 33–40, Nov. 1994.

[17] N. Gat, “Imaging spectroscopy using tunable filters: A review,” in *Proc. SPIE, Wavelet Appl. VII*, vol. 4056. Bellingham, WA, USA: SPIE, 2000, pp. 50–64.

[18] J. James, *Spectrograph Design Fundamentals*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[19] A. Mohan, R. Raskar, and J. Tumblin, “Agile spectrum imaging: Programmable wavelength modulation for cameras and projectors,” *Comput. Graph. Forum*, vol. 27, no. 2, pp. 709–717, Apr. 2008.

[20] T. Okamoto and I. Yamaguchi, “Simultaneous acquisition of spectral image information,” *Opt. Lett.*, vol. 16, no. 16, pp. 1277–1279, 1991.

[21] M. Descour and E. Dereniak, “Computed-tomography imaging spectrometer: Experimental calibration and reconstruction results,” *Appl. Opt.*, vol. 34, no. 22, pp. 4817–4826, 1995.

[22] B. Ford, M. R. Descour, and R. M. Lynch, “Large-image-format computed tomography imaging spectrometer for fluorescence microscopy,” *Opt. Exp.*, vol. 9, no. 9, pp. 444–453, 2001.

[23] W. R. Johnson, D. W. Wilson, and G. Bearman, “Spatial–spectral modulating snapshot hyperspectral imager,” *Appl. Opt.*, vol. 45, no. 9, pp. 1898–1908, 2006.

[24] N. Hagen and E. L. Dereniak, “Analysis of computed tomographic imaging spectrometers I spatial and spectral resolution,” *Appl. Opt.*, vol. 47, no. 28, p. F85, 2008.

[25] B. Geelen, N. Tack, and A. Lambrechts, “A compact snapshot multispectral imager with a monolithically integrated per-pixel filter mosaic,” *Proc. SPIE*, vol. 8974, Mar. 2014, Art. no. 89740L.

[26] P. Gonzalez et al., “A novel CMOS-compatible, monolithically integrated line-scan hyperspectral imager covering the VIS-NIR range,” *Proc. SPIE*, vol. 9855, May 2016, Art. no. 98550N.

[27] X. Cao et al., “Computational snapshot multispectral cameras: Toward dynamic capture of the spectral world,” *IEEE Signal Process. Mag.*, vol. 33, no. 5, pp. 95–108, Sep. 2016.

[28] M. D. Grossberg and S. K. Nayar, “What is the space of camera response functions?” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2003, p. 602.

[29] A. Robles-Kelly, “Single image spectral reconstruction for multimedia applications,” in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 251–260.

[30] B. Arad and O. Ben-Shahar, “Sparse recovery of hyperspectral signal from natural RGB images,” in *Proc. ECCV*, 2016, pp. 19–34.

[31] J. Wu, J. Aeschbacher, and R. Timofte, “In defense of shallow learned spectral reconstruction from RGB images,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 471–479.

[32] Z. Xiong, Z. Shi, H. Li, L. Wang, D. Liu, and F. Wu, “HSCNN: CNN-based hyperspectral image recovery from spectrally undersampled projections,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 518–525.

[33] S. Koundinya et al., “2D-3D CNN based architectures for spectral reconstruction from RGB images,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 844–851.

[34] Y. Baran Can and R. Timofte, “An efficient CNN for spectral reconstruction from RGB images,” 2018, *arXiv:1804.04647*.

[35] Z. Shi, C. Chen, Z. Xiong, D. Liu, and F. Wu, "HSCNN+: Advanced CNN-based hyperspectral recovery from RGB images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 939–947.

[36] H. Li, Z. Xiong, Z. Shi, L. Wang, D. Liu, and F. Wu, "HSVCNN: CNN-based hyperspectral reconstruction from RGB videos," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3323–3327.

[37] A. Alvarez-Gila, J. Van De Weijer, and E. Garrote, "Adversarial networks for spatial context-aware spectral image reconstruction from RGB," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 480–490.

[38] T. Stiebel, S. Koppers, P. Seltsam, and D. Merhof, "Reconstructing spectral images from RGB-images using a convolutional neural network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 948–953.

[39] D. S. Nathan, K. Uma, D. S. Vinothini, B. S. Bama, and S. M. M. Roomi, "Light weight residual dense attention net for spectral reconstruction from RGB images," 2020, *arXiv:2004.06930*.

[40] Y. Zhao, L.-M. Po, Q. Yan, W. Liu, and T. Lin, "Hierarchical regression network for spectral reconstruction from RGB images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 422–423.

[41] Y.-T. Lin and G. D. Finlayson, "Physically plausible spectral reconstruction from RGB images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2257–2266.

[42] J. Li, C. Wu, R. Song, Y. Li, and W. Xie, "Residual augmented attentional U-shaped network for spectral reconstruction from RGB images," *Remote Sens.*, vol. 13, no. 1, p. 115, Dec. 2020.

[43] I. Goodfellow et al., "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[44] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 60–65.

[45] R. M. Nguyen, D. K. Prasad, and M. S. Brown, "Training-based spectral reconstruction from a single RGB image," in *Proc. ECCV*, 2014, pp. 186–201.

[46] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. CVPR*, Jun. 2011, pp. 193–200.

[47] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010.

[48] B. Arad et al., "NTIRE 2020 challenge on spectral reconstruction from an RGB image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1806–1822.

[49] T. Brooks, B. Mildenhall, T. Xue, J. Chen, D. Sharlet, and J. T. Barron, "Unprocessing images for learned raw denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11028–11037.

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

[51] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Heidelberg, Germany: Springer, 2015, pp. 234–241.

[52] A. Banerjee and A. Palrecha, "MXR-U-Nets for real time hyperspectral reconstruction," 2020, *arXiv:2004.07003*.

[53] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 462–463.

[54] M. Yamaguchi et al., "High-fidelity video and still-image communication based on spectral information: Natural vision system and its applications," in *Proc. SPIE*, vol. 6062, pp. 129–140, Jan. 2006.

[55] M. Rosen and W. Jiang, "Lippman 2000: A spectral image database under construction," in *Proc. Int. Symp. Multispectral Imaging Color Reproduction Digital Arch.*, 1999, pp. 117–122.

[56] M. E. Gehm, R. John, D. J. Brady, R. M. Willett, and T. J. Schulz, "Single-shot compressive spectral imaging with a dual-disperser architecture," *Opt. Exp.*, vol. 15, no. 21, pp. 14013–14027, Oct. 2007.

[57] A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Appl. Opt.*, vol. 47, no. 10, p. B44, 2008.

[58] X. Lin, Y. Liu, J. Wu, and Q. Dai, "Spatial–spectral encoded compressive hyperspectral imaging," *ACM Trans. Graph.*, vol. 33, no. 6, pp. 1–11, Nov. 2014.

[59] L. Wang, Z. Xiong, G. Shi, F. Wu, and W. Zeng, "Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 10, pp. 2104–2111, Oct. 2017.

[60] N. Eslahi, S. H. Amirshahi, and F. Agahian, "Recovery of spectral data using weighted canonical correlation regression," *Opt. Rev.*, vol. 16, no. 3, pp. 296–303, May 2009.

[61] 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," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 23, no. 8, pp. 2020–2026, 2006.

[62] X. Zhang and H. Xu, "Reconstructing spectral reflectance by dividing spectral space and extending the principal components in principal component analysis," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 25, no. 2, pp. 371–378, 2008.

[63] Y. Jia et al., "From RGB to spectrum for natural scenes via manifold-based mapping," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4715–4723.

[64] N. Akhtar and A. Mian, "Hyperspectral recovery from RGB images using Gaussian processes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 100–113, Jan. 2020.

[65] Y. Li, C. Wang, and J. Zhao, "Locally linear embedded sparse coding for spectral reconstruction from RGB images," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 363–367, Mar. 2018.

[66] Y. Fu, Y. Zheng, L. Zhang, and H. Huang, "Spectral reflectance recovery from a single RGB image," *IEEE Trans. Comput. Imag.*, vol. 4, no. 3, pp. 382–394, Sep. 2018.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[68] I. Choi, D. S. Jeon, G. Nam, D. Gutierrez, and M. H. Kim, "High-quality hyperspectral reconstruction using a spectral prior," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–13, Dec. 2017.

[69] L. Wang, C. Sun, Y. Fu, M. H. Kim, and H. Huang, "Hyperspectral image reconstruction using a deep spatial–spectral prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8024–8033.

[70] T. Zhang, Y. Fu, L. Wang, and H. Huang, "Hyperspectral image reconstruction using deep external and internal learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8558–8567.

[71] X. Miao, X. Yuan, Y. Pu, and V. Athitsos, "Lambda-Net: Reconstruct hyperspectral images from a snapshot measurement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4058–4068.

[72] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[73] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[74] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8877–8886.

[75] Y. Zhao, L.-M. Po, K.-W. Cheung, W.-Y. Yu, and Y. A. U. Rehman, "SCGAN: Saliency map-guided colorization with generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3062–3077, Aug. 2021.

[76] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2813–2821.

[77] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. ICML*, 2017, pp. 214–223.

[78] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. NIPS*, 2017, pp. 5767–5777.

[79] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. NIPS*, 2017, pp. 6626–6637.

[80] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. ICLR*, 2018, pp. 1–26.

[81] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[82] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[83] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, 2010, pp. 249–256.

[84] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2014, pp. 1–15.

[85] K. Ohsawa, F. Koenig, M. Yamaguchi, and N. Ohyama, "Multiprimary display optimized for CIE1931 and CIE1964 color matching functions," in *Proc. 9th Congr. Int. Colour Assoc.*, Jun. 2002, pp. 939–942.

[86] F. A. Kruse et al., "The spectral image processing system (SIPS)-interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, nos. 2–3, pp. 145–163, May 1993.

Yuzhi Zhao (Graduate Student Member, IEEE) received the B.Eng. degree in electronic and information engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2018, and the Ph.D. degree in electronic engineering from the City University of Hong Kong (CityU), Hong Kong, in 2023.

His research interests include low-level vision, computational photography, generative models, and representation learning.

Dr. Zhao serves as a peer-reviewer in international conferences and journals, such as CVPR, ICCV, ECCV, WACV, ACCV, ICASSP, ICIP, and several IEEE Transactions.

Lai-Man Po (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from the City University of Hong Kong, Hong Kong, in 1988 and 1991, respectively.

He has been with the Department of Electronic Engineering, City University of Hong Kong, since 1991, where he is currently an Associate Professor with the Department of Electrical Engineering. He has authored more than 150 technical journal and conference papers. His research interests include image and video coding with an emphasis deep learning-based computer vision algorithms.

Dr. Po is a member of the Technical Committee on Multimedia Systems and Applications and the IEEE Circuits and Systems Society. He was the Chairperson of the IEEE Signal Processing Hong Kong Chapter in 2012 and 2013. He was an Associate Editor of HKIE Transactions in 2011 to 2013. He also served on the Organizing Committee, of the IEEE International Conference on Acoustics, Speech and Signal Processing in 2003, and the IEEE International Conference on Image Processing in 2010.

Tingyu Lin received the B.Eng. degree in software engineering from the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, in 2019, and the M.S. degree (Distinction) in multimedia information technology from the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, in 2020.

His research interests include computer vision, computational photography, and deep learning.

Qiong Yan received the bachelor's degree in computer science and technology from the University of Science and Technology of China, Hefei, China, in 2009, and the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2013.

She is currently a Research Director with SenseTime, Hong Kong, leading a group on computational imaging related research and production. Her research focuses on low-level vision tasks, such as image/video restoration and enhancement, and image editing and generation.

Wei Liu received the B.S. and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 2016 and 2020, respectively.

He was a visiting student with the Ohio State University, Columbus, OH, USA, for two years. He used to be an intern with SenseTime, Hong Kong, and currently works as an Algorithm Engineer with ByteDance, Beijing, China. His research interests include image generation, domain adaptation, semantic segmentation, and low-level computer vision.

Dr. Liu serves as a Peer Reviewer for IEEE Transactions on Image Processing, *ISPRS Journal of Photogrammetry and Remote Sensing*, and IEEE Transactions on Geoscience and Remote Sensing.

Pengfei Xian received the B.Eng. degree in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 2017. He is currently pursuing the Ph.D. degree in electrical engineering with the City University of Hong Kong, Hong Kong.

His research interests include instance and sematic segmentation on images and videos, together with reinforcement learning applications.