Learnable Reconstruction Methods from RGB Images to Hyperspectral Imaging: A Survey

Jingang Zhang, Runmu Su, Wenqi Ren, Qiang Fu, Yunfeng Nie

***Abstract*—Hyperspectral imaging enables versatile applications due to its competence in capturing abundant spatial and spectral information, which are crucial for identifying substances. However, the devices for acquiring hyperspectral images are expensive and complicated. Therefore, many alternative spectral imaging methods have been proposed by directly reconstructing the hyperspectral information from lower-cost, more available RGB images. We present a thorough investigation of these state-of-the-art spectral reconstruction methods from the widespread RGB images. A systematic study and comparison of more than 25 methods has revealed that most of the data-driven deep learning methods are superior to prior-based methods in terms of reconstruction accuracy and quality despite lower speeds. This comprehensive review can serve as a fruitful reference source for peer researchers, thus further inspiring future development directions in related domains.**

***Index Terms*—Spectral reconstruction, hyperspectral imaging, RGB images, deep learning.**

I. Introduction

HYPERSPECTRAL imaging refers to the dense sampling of spectral features with many narrow bands. Unlike traditional RGB images, each pixel of hyperspectral images (HSIs) contains a continuous spectral curve to identify the substance of the corresponding objects. Since spectral information can distinguish different materials, hyperspectral imaging has been used in many fields, such as remote sensing [1]–[12], agriculture [13], geology [14], astronomy [15], earth sciences [16], medical imaging [17]–[19], and so on. In recent years, HSIs have been further investigated in the emerging fields of computer vision by utilizing more advanced image processing tools, such as image segmentation [20], [21], recognition [22]–[25], tracking [26], pedestrian detection [27], [28], and anomaly detection [29]. By virtue of the highly widespread applications, hyperspectral imaging has attracted considerable attention and intensive research.

This work was supported by the Equipment Research Program of the Chinese Academy of Sciences (NO. YJKYYQ20180039 and NO. Y70X25A1HY), and the National Natural Science Foundation of China (NO. 61775219 , NO. 61771369 and NO. 61640422). Jingang Zhang and Runmu Su contributed equally to this work. (Corresponding author: Yunfeng Nie.)

J. Zhang is with School of Future Technology, The University of Chinese Academy of Sciences, Beijing, China, 100039 (email: zhangjg@ucas.ac.cn).

R. Su is with School of Future Technology, The University of Chinese Academy of Sciences, Beijing, China, 100039, also with Department of Computer Science and Technology, The Xidian University, Xi'an, China, 710071 (email: surunmu@stu.xidian.edu.cn).

W. Ren is with State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093 (email: rwq.renwenqi@gmail.com).

Q. Fu is with King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia. (email: qiang.fu@kaust.edu.sa).

Y. Nie is with Vrije Universiteit Brussel, 1050 Brussels, Belgium (email: Yunfeng.Nie@vub.be)

On one hand, the devices for acquiring HSI are typically bulkier and more expensive than common RGB cameras [30]–[32]. Figure 1 shows the schematics of a common RGB camera and a hyperspectral imager. Besides the more components in the optical path, the mainstream hyperspectral imaging technique relies on the scanning (e.g., pushbroom or whiskbroom scanners) of one spatial or spectral dimension to generate its three-dimensional (3D) datacube, i.e., 2D spatial data (x, y) and 1D spectral information (w). Different strategies have been developed to obtain snapshot HSI (e.g. Wagadarikar et al. [33], Cao et al. [34]), while either spatial or spectral resolution is lower than the mainstream methods. From this perspective, we realize that the progress of this technique by purely increasing opto-mechanical components is usually a disturbing compromise of spatial or spectral resolutions and system complexity.

On the other hand, common RGB cameras have been exponentially arising in almost ubiquitous domains, enabling great potentials towards a low-cost, wide-ranging spectral imaging strategy by directly reconstructing hyperspectral information from RGB images. Motivated by these great potentials, CVPR (conference on computer vision and pattern recognition) has launched two competitions on this topic, known as the spectral reconstruction challenges NTIRE-2018 [35] and NTIRE-2020 [36], resulting in many brilliant ideas that promote the reconstruction accuracy and quality.

In general, recovering hyperspectral information from RGB images is an inverse problem, which is usually ill-posed. Here, we categorize the spectral reconstruction (SR) approaches into two different kinds: prior-based and data-driven methods. The first type needs a known knowledge of the image features, while less data are need. Different algorithm models can use the identified priors to constrain the solution space, so as to obtain an approximate optimal solution. The representative methods in this type are sparse coding, SR A+, SR Gaussian process, and SR manifold mapping. The second type, known as the data-driven methods, becomes more prevalent since deep learning models can get more accurate solutions and perform better than prior-based methods when a large database is available. Various neural networks have been proposed to improve the reconstruction accuracy, from simple network models to more advanced neural networks using a variety of techniques (such as residual learning, dense connection, channel attention mechanism, etc.). Compared to prior-based methods, these methods are enhanced by more and more datasets to their high learning ability and good adaptability.

In this work, we take a comprehensive survey of the above-mentioned spectral reconstruction methods. In the following

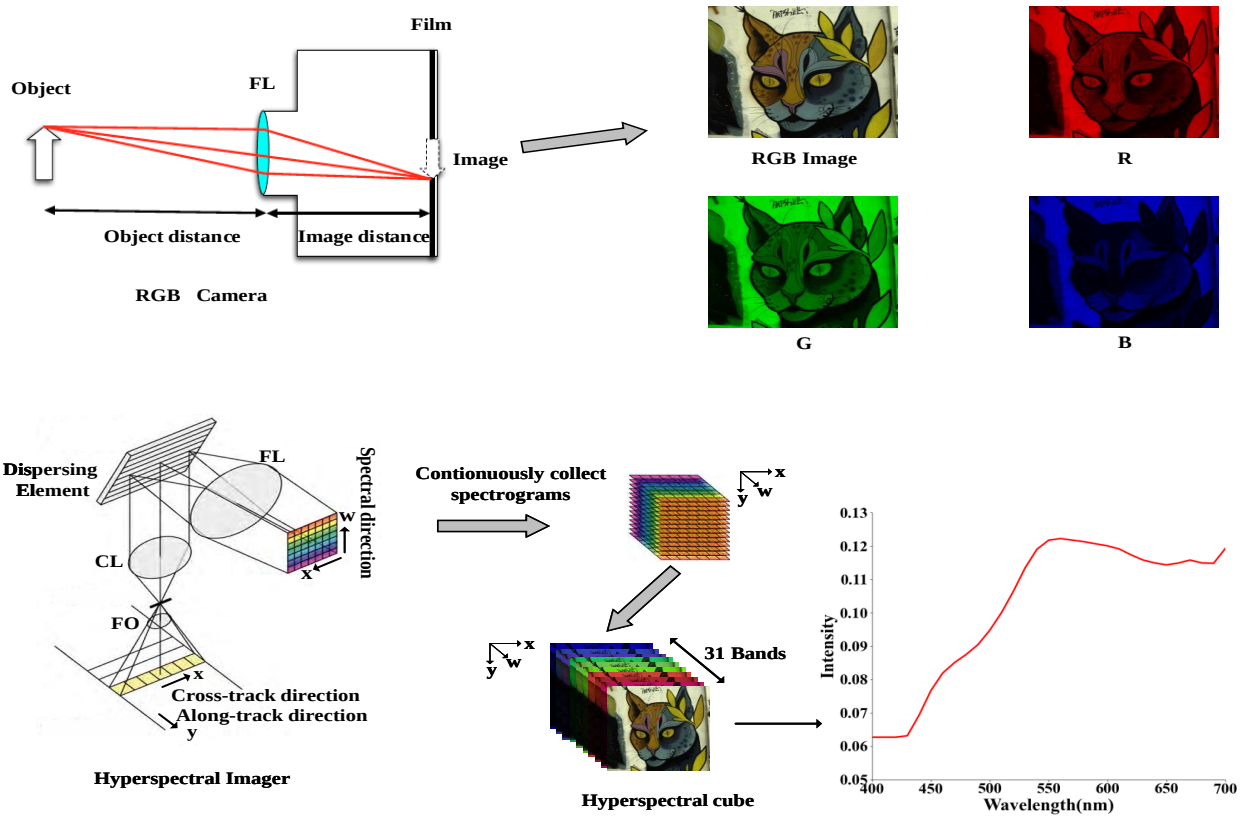


Fig. 1. Schematic diagrams of a RGB camera with a focusing lens group (top) and a typical hyperspectral imager with a collimating lens group, dispersive element(s) and a focusing lens group (bottom). An RGB image consists of three color channels, whereas in each pixel of a hyperspectral image, a spectral curve can be obtained to identify the specified substance.

TABLE I
FEATURES OF FOUR OPNE-SOURCE HYPERSPECTRAL DATASETS

Dataset	Amount	Resolution	Spectral range/(nm)	Scene
ICVL [37]	203	$1392 \times 1300 \times 31$	400-700	urban, suburban, rural, indoor and plant-life
CAVE [38]	32	$512 \times 512 \times 31$	400-700	studio images of various objects
BGU-HS [35]	286	$1392 \times 1300 \times 31$	400-700	urban, suburban, rural, indoor and plant-life
ARAD-HS [36]	510	$512 \times 482 \times 31$	400-700	various scenes and subjects

sections, we first describe the spectral reconstruction datasets and performance metrics mainly used in those methods. In Sec. 3, a detailed review of prior-based methods and data-driven methods for RGB image spectral reconstruction is given. In Sec. 4, we systematically compare a few representative algorithms using available open-source datasets. In the end, a summary is given, and outlooks are drawn.

II. FUNDAMENTALS

A. Image formation model

As we know, an RGB image has three color channels, while a typical HSI has dozens of spectral bands, as shown in Figure 1. The reconstruction of hyperspectral images is to recover the 'missing' spectral information from RGB images. In principle, an HSI is obtained by the interaction between the spectral reflectance and the illumination spectrum. The spectral reflectance is an essential attribute of the object, thus the HSIs obtained under different illumination conditions are different.

When under the same illumination An RGB image is obtained by integrating the product of the HSI and camera sensitivity function over the spectral range. In this case, since the illumination spectrum is not given, the relation between the RGB image and the HSI is expressed as

$$I_c(x,y) = \int_w H(x,y,w) S_c(w) dw, \tag{1}$$

where I represents the RGB image (c = r, g, b), H is HSI, (x, y) represents pixel space coordinates, S_c denotes camera

sensitivity function, and w refers to spectral coordinates. In the discrete form, Eq.1 can be rewritten as

$$I_c(x,y) = \sum_{w=1}^{\Omega} H(x,y,w)\, S_c(w), \tag{2}$$

where Ω refers to the number of spectral bands. Most SR algorithms are based on solving the inverse mapping from RGB images to HSIs.

When under different illuminations Given different illumination spectra, the relation between the RGB image and the HSI is expressed as

$$I_c(x,y) = \int_w R(x,y,w)\, L(w)\, S_c(w) dw, \tag{3}$$

where R represent the spectral reflectance, and L represent the illumination spectrum. The discrete form of the Eq.3 is

$$I_c(x,y) = \sum_{w=1}^{\Omega} R(x,y,w)\, L(w)\, S_c(w). \tag{4}$$

Only a few algorithms are based on this model, such as multiple non-negative sparse dictionaries. Under this condition, the reconstruction of HSI is divided into two steps, one is to solve the spectral reflectance, and the other is to obtain the illumination spectrum.

B. Dataset and performance evaluation metrics

A few HSI datasets are available as open-source, which are used as the datasets for training and verifying the following deep-learning networks. The performance metrics for evaluating and comparing different algorithms are introduced in this section.

1) Open-source datasets: Table I lists four HSI datasets commonly used in the SR community. As we can see, different datasets have various amounts of HSIs, resolutions, spectral ranges, and image scenes. More details of each dataset are introduced as follows.

- **ICVL** [37] is collected and published by Arad and Ben Shahar. This dataset contains 203 scenes acquired using a line scanner camera (Specim PS Kappa DX4 hyperspectrometer). The camera is mounted on a kinematic platform for 1D spatial scanning. Various indoor and outdoor scenes are captured, ranging from man-made objects to natural objects. The spatial resolution is 1392×1300 with 519 spectral bands over 400-1,000 nm wavelength range, but it has been downsampled to 31 spectral channels from 400 nm to 700 nm in 10 nm increments.
- **CAVE** [38] is a frequently used hyperspectral dataset. Unlike other datasets using a linear scanner, this dataset was captured using a tunable filter instead of a dispersive grating or prism to sequentially record the hyperspectral bands. It contains 32 different images with a spatial resolution of 512×512 pixels and 31 different spectral bands between 400 and 700 nm, captured by a cooled CCD camera (Apogee Alta U260). CAVE is a collection of various objects, including faces, fake and real fruits, candies, paintings, textiles and so forth.
- **BGU-HS** [35] is the largest and most detailed natural HSI database collected so far. For the spectral reconstruction challenge NTIRE-2018, the database has been expanded to include 286 images. This dataset was divided into 256 training images, 10 verification images, and 20 test images for a fair evaluation of all the participants in this challenge. Each HSI has a spatial resolution of 1392×1300, and is composed of 31 continuous spectral bands ranging from 400 nm to 700 nm with an interval of 10 nm.
- **ARAD-HS** [36] is a newer HSI dataset. In this challenge, a total of 510 pictures were divided into 450 training images, 30 validation images, and 30 test images. The spatial resolution of each image is 512×482, and the spectral band is 31. This dataset was collected by Specim IQ mobile hyperspectral camera, which is an independent, battery-powered, push-broom spectral imaging system. Its size is the same as that of a traditional SLR camera ($207 \times 91 \times 74$ mm), and it can operate independently without an external power supply. The use of such a compact mobile system facilitates the collection of extremely diverse datasets with h a large variety of scenes and subjects.

2) Performance evaluation metrics: In the field of SR, quantitative analysis is needed to evaluate the performance of various algorithms. Currently, there are many indicators without a generalized criterion ,such as Mean Relative Absolute Error (MRAE), Root Mean Square Error (RMSE), relative Root Mean Square Error (rRMSE), which are defined by the equations below.

$$\text{MRAE} = \frac{1}{N}\sum_{c=1}^{N}\left(\left|H_{GT}^{c} - H_{SR}^{c}\right| / H_{GT}^{c}\right), \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{c=1}^{N}\left(H_{GT}^{c} - H_{SR}^{c}\right)^2}, \tag{6}$$

$$\text{rRMSE} = \sqrt{\frac{1}{N}\sum_{c=1}^{N}\left(\left(H_{GT}^{c} - H_{SR}^{c}\right) / \overline{H_{GT}}\right)^2}, \tag{7}$$

where H_{GT}^{c} represents the c-th pixel value of the ground truth HSI, H_{SR}^{c} denotes the c-th pixel value of the reconstructed HSI, and N is the total number of pixels.

The above three metrics are the most commonly used for the performance evaluation of different SR methods, while some minority metrics such as Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM) and others are also used sometimes. The PSNR is calculated as

$$\text{PSNR} = 20 \cdot \log_{10}\left(\frac{H_{MAX}}{\text{RMSE}}\right). \tag{8}$$

where H_{MAX} represents the maximum pixel value of the HSI.

SSIM is a classic indicator of image quality evaluation, which is more suitable for human visual perception systems.

TABLE II
AN OVERVIEW OF PRIOR-BASED SPECTRAL RECONSTRUCTION METHODS.

Method	Category	Priors
Sparse Coding [37]	Dictionary Learning	sparsity
SR A+ [39]	Dictionary Learning	sparsity, local euclidean linearity
Multiple Non-negative Sparse Dictionaries [40]	Dictionary Learning	spatial structure similarity, spectral correlation
Local Linear Embedding Sparse Dictionary [41]	Dictionary Learning	color and texture, local linearity
Spatially Constrained Dictionary Learning [42]	Dictionary Learning	spatial context
SR Manifold Mapping [43]	Manifold Learning	low-dimensional manifold
SR Gaussian Process [44]	Gaussian Process	spectral physics, spatial structure similarity

SSIM uses a combination of brightness, contrast, and structure to evaluate image quality, described by

$$\mathrm{SSIM} = \frac{\left(2\mu_r\mu_g + 6.5\right)\left(2\sigma_{r,g} + 58.5\right)}{\left(\mu_r^2 + \mu_g^2 + 6.5\right)\left(\sigma_r^2 + \sigma_g^2 + 58.5\right)}, \tag{9}$$

where, μ_r is the mean of the reconstructed HSI, μ_g is the mean of the ground truth HSI, σ_r^2 is the variance of the reconstructed HSI, σ_g^2 is the variance of the ground truth HSI and $\sigma_{r,g}$ is the covariance of the reconstructed and the ground truth HSI .

III. Algorithm Survey

Here, we divide the SR algorithms into two categories: prior-based and deep-learning-based methods. Among those prior-based methods, three types have been identified as Dictionary Learning, Manifold Learning, and Gaussian Process by the difference in their priors. Next, we investigate various network models in the deep learning methods, namely Linear CNN, U-Net model, GAN nodel, Dense Network, Residual Network, Attention Network, and Multi-branch Network. The overall taxonomy and the full lists of the algorithms to be analyzed are shown in Figure 2.

A. Prior-based methods

The target is to reconstruct a HSI $\widetilde{\mathbf{H}} \in \mathbb{R}^{x\times y\times L}$ from an RGB image $\widetilde{\mathbf{Y}} \in \mathbb{R}^{x\times y\times l}$. If not specified otherwise, $L = 31$ and $l = 3$. For convenience, we use a 2D matrix form to represent the image, where $\widetilde{\mathbf{Y}}$, $\widetilde{\mathbf{H}}$ respectively are written as $\mathbf{Y} \in \mathbb{R}^{l\times N}$ and $\mathbf{H} \in \mathbb{R}^{L\times N}$, where $N = xy$ is the total number of pixels. In this case, each column of the matrix represents the spectral band, and each row corresponds to the entire image of the spectral band.

It is well known that each pixel of a HSI is a mixture of the spectral responses of different materials in the scene in a certain proportion. Therefore, we can use the spectral response of the pure material (base spectrum, also called endmember) and the corresponding proportion (called abundance) to represent the HSI. According to the linear mixed model, $\mathbf{H}$ is described as $\mathbf{H} = \mathbf{EA}$, where $\mathbf{E}$ refers to the base spectrum, and $\mathbf{A}$ denotes the proportion. The RGB image is obtained by applying the spectral response function to the HSI, i.e., $\mathbf{Y} = \mathbf{SH}$ can be obtained and can also be expressed $\mathbf{Y} = \mathbf{SEA}$, where $\mathbf{S}$ refers to the spectral response function (SRF). To obtain HSIs from RGB images, it is essentially to solve the following optimization problem

$$\min_{\mathbf{E},\mathbf{A}} \|\mathbf{H} - \mathbf{EA}\|^2 + \|\mathbf{Y} - \mathbf{SEA}\|^2. \tag{10}$$

Recovering the lost spectral information from RGB images is an ill-posed problem, that is, the solution is not unique. In order to narrow down the solution space, we need to use prior knowledge to constrain Eq.10, which can be rewritten as

$$\min_{\mathbf{E},\mathbf{A}} \|\mathbf{H} - \mathbf{EA}\|^2 + \|\mathbf{Y} - \mathbf{SEA}\|^2 + \lambda P(\mathbf{A}), \tag{11}$$

where $P(\mathbf{A})$ is the regularization term (prior term), and λ denotes the weighting factor.

Prior knowledge is the statistical analysis of the data to obtain the inherent attributes and characteristics of the image. Prior knowledge about HSI usually includes sparsity [37], spatial structure similarity [44], correlation between spectra [40], etc. We summarize the most recent SR methods based on various priors in Table II.

1) Dictionary Learning: Statistical analysis of HSIs shows that it is sparse in space and spectrum [45], and the spectral signal is expressed as a sparse combination of base spectra. The base spectra are stored in a dictionary, which leads to the SR methods based on dictionary learning. The prior knowledge of the spatial structure similarity and high correlation across spectra are used as proper regularizations under the learned dictionary, which leads to several SR methods based on improved dictionary learning.

Sparse Coding Based on the sparsity of HSIs, Arad et al. [37] propose a sparse coding method to reconstruct HSIs from RGB images. According to Eq.11 , the objective is to solve the dictionary and the sparse coefficients.

Sparse Coding firstly calculates the overcomplete spectral dictionary of the HSI. Once obtained, the spectral dictionary will be projected into the RGB space via the SRF to form an RGB dictionary. In the reconstruction step, given a test RGB image, the dictionary representation (sparse coefficient) of each pixel of the image is calculated. Once the dictionary representation is found, it is combined with the hyperspectral dictionary to obtain reconstructed hyperspectral pixels.

Compared with hardware acquisition of hyperspectral information, the SR method based on dictionary learning is

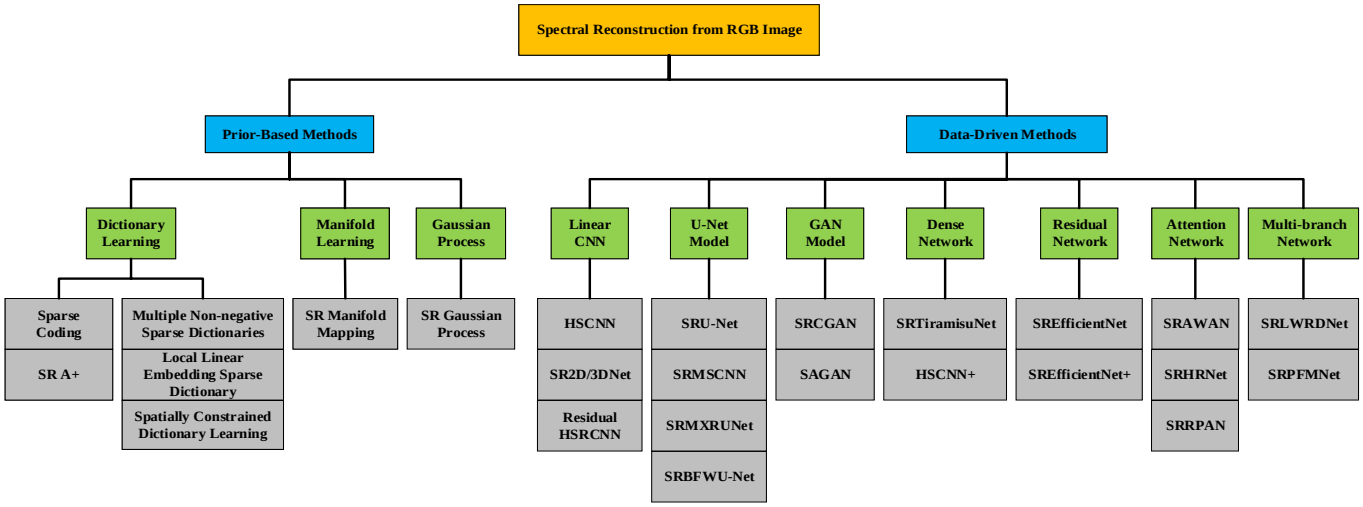


Fig. 2. The overall taxonomy of the spectral reconstruction methods and the full lists for each category.

low-cost and fast. As the hyperspectral dataset grows, the dictionary capacity increases, and the reconstruction timeis prolonged. The SR based on the sparse dictionary [37] completes the reconstruction task from a single pixel without considering the spatial correlation, so the quality of the reconstructed image is limited.

SR A+ The SR A+ method proposed by Aeschbacher et al. [46] is the same as the sparse coding method in that, it uses the same method to create hyperspectral and RGB dictionary. The difference is that SR A+ establishes a mapping from RGB to hyperspectral in a local dictionary.

In SR A+, the dictionary atoms are called anchor points. The linear combination of the neighborhood (D_{RGB}) of the anchor I_i is used to represent the RGB image. The linear combination coefficient is obtained by optimizing the following least square error

$$\arg\min_{\delta} \|I - D_{\mathrm{RGB}}\delta\|_2^2 + \lambda \|\delta\|_2^2 . \tag{12}$$

Combining the coefficient with the neighborhood of the anchor of the hyperspectral dictionary, the projection matrix from RGB to the hyperspectral (the projection matrix corresponding to the anchor I_i) can be obtained. The HSI dictionary, RGB image dictionary, and projection matrix are completed in the training phase. In the test phase, only the nearest neighbor search (the dictionary atom whose pixel in the RGB image is close to the RGB dictionary atom) is involved, and the projection matrix corresponding to the dictionary atom is multiplied by the RGB pixel to obtain an HSI.

SR A+ does not establish the RGB to hyperspectral mapping relationship from the global dictionary. Instead, it builds RGB-to-hyperspectral projection from local anchor points, which runs faster.

Multiple Non-negative Sparse Dictionaries HSIs are formed by mixing the spectral reflectance of the objects in the scene with the illumination spectrum. Fu et al. [40] proposed a method that solves the SR from RGB images task by reconstructing the spectral reflectance and the illumination spectrum. They introduced multiple non-negative sparse dictionaries, which are divided into three parts, spectral reflectance estimation, illumination spectral estimation, and HSI reconstruction respectively.

The hyperspectral data is clustered firstly. In the phase of spectral reflectance estimation, the spectral reflectance dictionary and the corresponding RGB dictionary are created on each cluster. Then given a test white balancing RGB image, the nearest clusters to the image pixels are searched, and the dictionary representations of the corresponding image pixels are calculated for each cluster. In each cluster, the dictionary representations are combined with the spectral reflectance dictionaries to get the estimated spectral reflectance. Combining the dictionary representations with the spectral reflectance dictionaries to get the estimated spectral reflectance. The spectral reflectance can then be obtained by aggregating estimated spectral reflectance from all clusters. Once obtained, according to Eq.4, the RGB camera spectral sensitivity function and the RGB image are known to estimate the illumination spectrum. Finally, the spectral reflectance is combined with the illumination spectrum to obtain a reconstructed HSI.

In multiple non-negative sparse dictionaries, the introduction of hyperspectral prior knowledge greatly improves the performance of sparse representation. Multiple sparse dictionaries are used to provide a more compact base representation for each cluster, effectively describing the spectral information of various materials in the scene. The disadvantage, however, is that, their algorithm may not continue to work when the SRF is unknown.

Local Linear Embedding Sparse Dictionary The SR of the sparse dictionary only considers the sparseness of spectral information and does not use local linearity. The drawback is that, the reconstruction is not accurate, and the reconstructed image has metamerism. Li et al. [41] have made three improvements. First, the HSI is divided into several cubes, and the optimal sample is selected using a selection strategy based on maximum volume [47] in each cube image, and finally a sparse dictionary is built using these samples. Secondly, the process of learning the dictionary introduces

local linearity. Thirdly, in the reconstruction process,the dictionary representation calculation of the test RGB image pixels uses texture information as regularization.

In this method, the locally best samples are selected to reduce the redundancy of the samples in a global space. In the process of dictionary learning, the local linearity of the spectrum is introduced to make the dictionary compact and improve the expression ability of the dictionary. The texture information is introduced in the reconstruction to ensure the reconstructed HSI quality and reduce metamerism [48].

Spatially Constrained Dictionary Learning In the early stage, the SR methods based on dictionary learning are pixel-wise, so that the reconstruction results are not accurate. Geng et al. [42] introduce spatial context information into sparse representation, which leads to a spatially constrained dictionary learning SR method. The algorithm is divided into two steps. The first step is to use the K-SVD algorithm to calculate the hyperspectral sparse dictionary. In the second step, the neighbouring pixels are used to constrain the sparse representation of the RGB space. The authors use a parallel orthogonal matching pursuit (SOMP) algorithm [49] to estimate the sparse coefficients. After estimating the sparse coefficients, the coefficients are combined with the HSI dictionary, and finally the reconstructed HSI is obtained.

Compared with the pixel-wise reconstruction, the introduced spatial context information can preserve spatial structures of the image, and the physical object distributions in the image is guaranteed.

2) Manifold Learning: Manifold learning is to find a low-dimensional manifold [50]–[55] that can uniquely represent high-dimensional data. The low-dimensional manifold has the properties of Euclidean space. Hyperspectral data can be represented by a set of low-dimensional manifolds. With this prior knowledge, an RGB image SR model based on manifold learning can be established.

SR Manifold Mapping The authors use the manifold learning method [43] to simplify the three-to-many mapping problem into a three-to-three problem to obtain accurate reconstruction results.

Specifically, the isometric mapping algorithm [56] is used to reduce the dimensionality of the spectrum to a three-dimensional subspace, by training the radial basis network to predict the mapping from the RGB space. In order to reconstruct the original spectrum from its three-dimensional embedding, the dictionary learning method is used to learn the dictionary pair of the high-dimensional spectrum and the three-dimensional embedding, and their relationship can be used to recover the high-dimensional spectrum data from the embedding space.

SR Manifold Mapping simplifies the SR problem and can get accurate reconstruction results.

3) Gaussian Process: The spectral signal is a relatively smooth function of the wavelength, which can be modelled by the Gaussian process. Following this, the Gaussian Processes [57] are used in the RGB space to reconstruct hyperspectral information with the following example.

SR Gaussian Process The Gaussian Process models the spectral signal of each material, and combines them with the RGB image to obtain a reconstructed HSI. According to Eq.11, this method [44] solves a set of Gaussian Processes and corresponding coefficients by two steps.

In training, firstly, the HSI is divided into several image patches which are clustered to obtain C clusters via the K-means. Then, Gaussian Processes are established on each cluster, which transforms the mean parameters to match the spectral quantization of the RGB image via the SRF. In the testing, RGB image patches are extracted and assigned the RGB transformations of the HSI clusters. The transformed Gaussian process means that the matched clusters are used to represent the patch. Finally, the representation coefficients are combined with the original Gaussian processes to reconstruct the desired HSI.

Note that, the clusters of similar image patches are extracted from training images in this method. It introduces spatial similarity and spectral correlation into SR. The physical characteristics of the spectral signal are incorporated into the Gaussian processes through its kernel and the use of non-negative mean prior probability distribution, which makes the reconstruction accuracy higher. However, this algorithm is relatively more complicated to solve.

B. Data-driven methods

From the perspective of the most distinctive features in the network architectures, we divide data-driven deep learning based methods into eight groups, as seen in Table III.

1) Linear CNN: CNN has a strong learning ability with outstanding achievements in the field of SR. Linear CNN is a stack of convolutional layers, and the input sequentially flows from the initial layer to the later layers. This kind of network design only has a single path and does not include multiple branches. Several different CNN designs are described as follows.

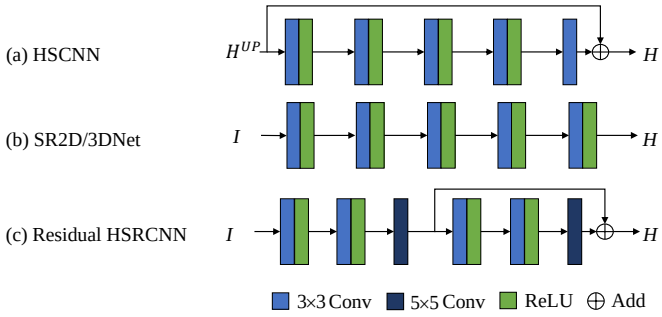


Fig. 3. These are the three methods in the Linear CNN category. (a) HSCNN [58] proposed HSCNN. H^{UP} represents the spectral image that is consistent with the H spectral dimension after spectral upsampling. (b) SR2D/3DNet [59]. (c) The improved SRCNN [76] is used as the benchmark for spectral reconstruction and combined with the residual recovery network to form Residual HSRCNN [60].

HSCNN HSCNN [58] is a unified deep learning framework for restoring hyperspectral information from spectrally undersampled images, such as RGB and compressed sensing

TABLE III
OVERVIEW OF SPECTRAL RECONSTRUCTION OF RGB IMAGES BASED ON DATA-DRIVEN DEEP LEARNING.

Category	Method	Depth [1]	Filters [2]	Loss Function	Basic Idea
Linear CNN	HSCNN [58]	5	64	$\mathcal{L}_{MSE}$	The proposed a unified deep learning framework for recovering spectral information from spectrally undersampled images.
	SR2D/3DNet [59]	5	-	$\mathcal{L}_2$	The model is a linear 2D-3D CNN architecture.
	Residual HSRCNN [60]	6	64	$\mathcal{L}_{MSE}$	This is a two-step linear network used to recover low-frequency and high-frequency spectral information.
U-Net Model	SRU-Net [61]	5	128	$L_{\Delta E}$	This network is based on the U-Net paradigm.
	SRMSCNN [62]	10	1024	$\mathcal{L}_{MSE}$	This model is based on the multi-scale CNN framework.
	SRMXRUNet [63]	56	4096	$\mathcal{L}_{mper}$	This network contains modified versions of encoders and decoders.
	SRBFWU-Net [64]	4	-	$\mathcal{L}_{MRAE}$, $\mathcal{L}_1$	A encoder-decoder structure has supervised and unsupervised learning.
GAN Model	SRCGAN [65]	8	512	$\mathcal{L}_{adv}$, $\mathcal{L}_1$	The first method to introduce conditional Generation Adversarial networks into the topic.
	SAGAN [66]	40	1024	$\mathcal{L}_{adv}$, $\mathcal{L}_1$	The working is based on the conditional generation network of scale feature attention.
Dense Network	SRTiramisuNet [67]	23	16	$\mathcal{L}_{eu}$	This network uses a Dense Network-based framework to learn inverse mapping.
	HSCNN+ [68]	160	64	$\mathcal{L}_{MRAE}$	This working has two types of networks based on dense structure with forward path expansion and residual networks.
Residual Network	SREfficientNet [69]	9	128	$\mathcal{L}_2$	This framework has global residual learning and local residual learning.
	SREfficientNet+ [70]	13	128	$\mathcal{L}_2$	The network can be applied to themes in the wild.
Attention Network	SRAWAN [71]	61	200	$\mathcal{L}_{MRAE}$, $\mathcal{L}_{\text{CSS}}$	The framework is Adaptive Weighted Attention Network with Camera Spectral Sensitivity Prior.
	SRHRNet [72]	57	256	$\mathcal{L}_1$	A 4-level hierarchical regression network that extracts features of different scales, which PixelShuffle is used as an inter-level interaction.
	SRRPAN [73]	135	64	$\mathcal{L}_{MRAE}$	This network is based on the pixel attention network with residual structure.
Multi-branch Network	SRLWRDNet [74]	40	32	$\mathcal{L}_2$, $\mathcal{L}_{SSIM}$	A multi-branch paradigm network contains two parallel subnets.
	SRPFMNet [75]	9	64	$\mathcal{L}_1$	A multi-branch network that adaptively determines the size of the receptive field and the mapping function for each pixel.

[1] Depth refers to the depth of the network, which is the number of convolutional layers.
[2] Filters are the width of the network, which are the number of convolution kernels.

[77], [78] images. HSCNN inherits the spatial super-resolution algorithm VDSR [79].

The difference between HSCNN and VDSR is in the first and last layers. Because the HSIs are three-dimensional data, the first layer has 64 convolution kernels of size $3 \times 3 \times L$, and the last layer has L convolution kernels of size $3 \times 3 \times 64$. The other layers have the same configuration as the VDSR. See Figure 3(a). The spectrally up-sampled [80] RGB image is consistent with the spectral dimensions of the original HSI, and then the up-sampled image is used as the input of HSCNN to restore the missing high-frequency image information.

Mean square error $\mathcal{L}_{MSE}$ is used to train HSCNN to reduce the error between reconstructed and ground truth HSI. Although HSCNN has a simple network structure, it can achieve good reconstruction fidelity. However, HSCNN also has shortcomings. In the spectral upsampling operation of the method, it is necessary to know the SRF, otherwise it will limit the performance of the network. Besides, HSCNN fails

to improve performance by increasing the network depth.

SR2D/3DNet SR2D/3DNet [59] is a solution to the challenge of NTIRE-2018 RGB image spectral reconstruction. The authors use CNN based on 2D and 3D convolution kernels to solve the problem of SR. The difference is that 2DNet operates independently on channels and only considers the spatial domain, while 3D-Net considers the relationship between channels. The two network structure diagrams are shown in Figure 3(b).

In this SR2D-3DNet, the numbers of the 2D and 3D network layers are both 5 layers. During training, the RGB image and the HSI are divided into $6s \times 64$ image patches, and the 2D-3D network is trained with L2 loss $\mathcal{L}_2$ under the paired image patches.

Residual HSRCNN The authors draw on SRCNN [76] to propose the HSRCNN model [60]. HSRCNN has three convolutional layers and the kernel sizes are respectively 3×3, 3×3, 5×5 . Except for the last layer, the other convolutional layers are followed by a ReLU.

It is difficult for HSRCNN to recover high-frequency spectral information (residual). The authors overwrite the baseline CNN of HSRCNN to form Residual HSRCNN model to recover residual information. Finally, the output of HSRCNN and the output of Residual HSRCNN are added to get the final reconstructed HSI. The network is shown in Figure 3(c).

The RGB image is divided into 15×15 overlapping image patches as the network input, and the $\mathcal{L}_{MSE}$ is used to train the model.

2) U-Net model: The U-Net model is composed of an encoder and a decoder. The encoder encodes features of different scales of the image through continuous down-sampling operations. The up-sampling operation of the decoder restores the feature map to the original image size, and finally obtains the reconstruction result. Several SR methods based on the U-Net model are explained as follows.

SRU-Net *Sparse Coding, SR A+, and SR manifold mapping* methods are from a single pixel to establish RGB to hyperspectral mapping. These methods only consider the prior knowledge in the spectral domain and ignore the spatial context information. U-Net was originally used for semantic segmentation tasks, because it can focus on the local information of the image. The authors use it as the main framework to create SRU-Net and SRMU-Net [61]. The overall network architecture diagram is shown in Figure 4(a).

In contrast to the original U-Net, SRU-Net removes the pooling layer. The model focuses on local context information to enhance the reconstruction results. The network accepts 32×32 image patches as input to further enforce this. SRMU-Net was further proposed to handles "real world" images with noise and different levels of blur, by adding 5×5 convolutional layer at the very start for pre-processing. The model takes the color error metric as the objective function $L_{\triangle E}$.

SRMSCNN To solve the SR problem, the local and non-local similarity of RGB images can help improve the reconstruction accuracy. The authors propose a multi-scale CNN based on U-Net called SRMSCNN [62]. The network structure is shown in Figure 4(b).

The encoder and decoder of SRMSCNN are symmetrical and connected at the same layer by skip connection operation. In the encoder part, each down-sampling step consists of convolution block with max-pooling. The convolution block consists of two 3×3 convolutions. Each of them is followed by batch normalization, leaky ReLU and dropout. The max-pooling completes downsampling. In the decoder part, the up-sampling step consists of pixel shuffle [81] (eliminating checkerboard artifacts) with a convolution block. Finally, a 1×1 convolutional layer is used to reconstruct HSI. The continuous down-sampling of the encoder has two functions. The first is to increase the feature map, and the second is to expand the receptive field and encode local and non-local information. The decoder uses compact features to reconstruct HSIs.

The network takes 64×64 image patch as input and the $\mathcal{L}_{MSE}$ as the loss function.

SRMXRUNet Based on the basic U-Net, the authors use the XResnet family model [82] with Mish activation function [83] as an encoder. The decoder is based on the structure proposed by Howard and Gugger's [84], while the difference is made in sub-pixel up-sampling convolution, and a blur and a self-attention layer. Therefore, this model is called SRMXRUNet [63], as shown in Figure 4(c).

The modification on decoder reduces loss of pixels, so more information can be kept. However, sub-pixel convolution will produce checkerboard artifacts. So ICNR [85] with weighted normalization is used for weight initialization. The sub-pixel convolutional layer is followed by a blur layer [86]. The blur layer is composed of an average pooling layer, which is used to eliminate artifacts. The decoder adds a self-attention layer [87], mainly to help the network pay attention to the relevant parts of the image. These improvements enhance the learning ability of the network, thus improving the reconstruction accuracy. This article uses the improved perceptual loss $\mathcal{L}_{mper}$ [88] as the loss function.

SRBFWU-Net The spectrum can be obtained by a weighted combination of a set of basis functions (basis spectra). In the early research, only 10 basis functions are needed to accurately generate spectral features [89]–[92]. SRBFWU-Net [64] predicts 10 weights for each pixel as well as learns a set of 10 basis functions which are then combined to form the reconstructed HSI. The basis functions are learned as a 10×31 matrix variables during training. There are two learning methods, supervised and unsupervised.

In supervised learning, the 2×2 max-pooling layers are replaced with linear downsampling. 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 network structure is shown in Figure 4(d).

In unsupervised learning, adding two modules to the supervised learning network are image generation module and photometric reconstruction loss module, as see in Figure 4(d)

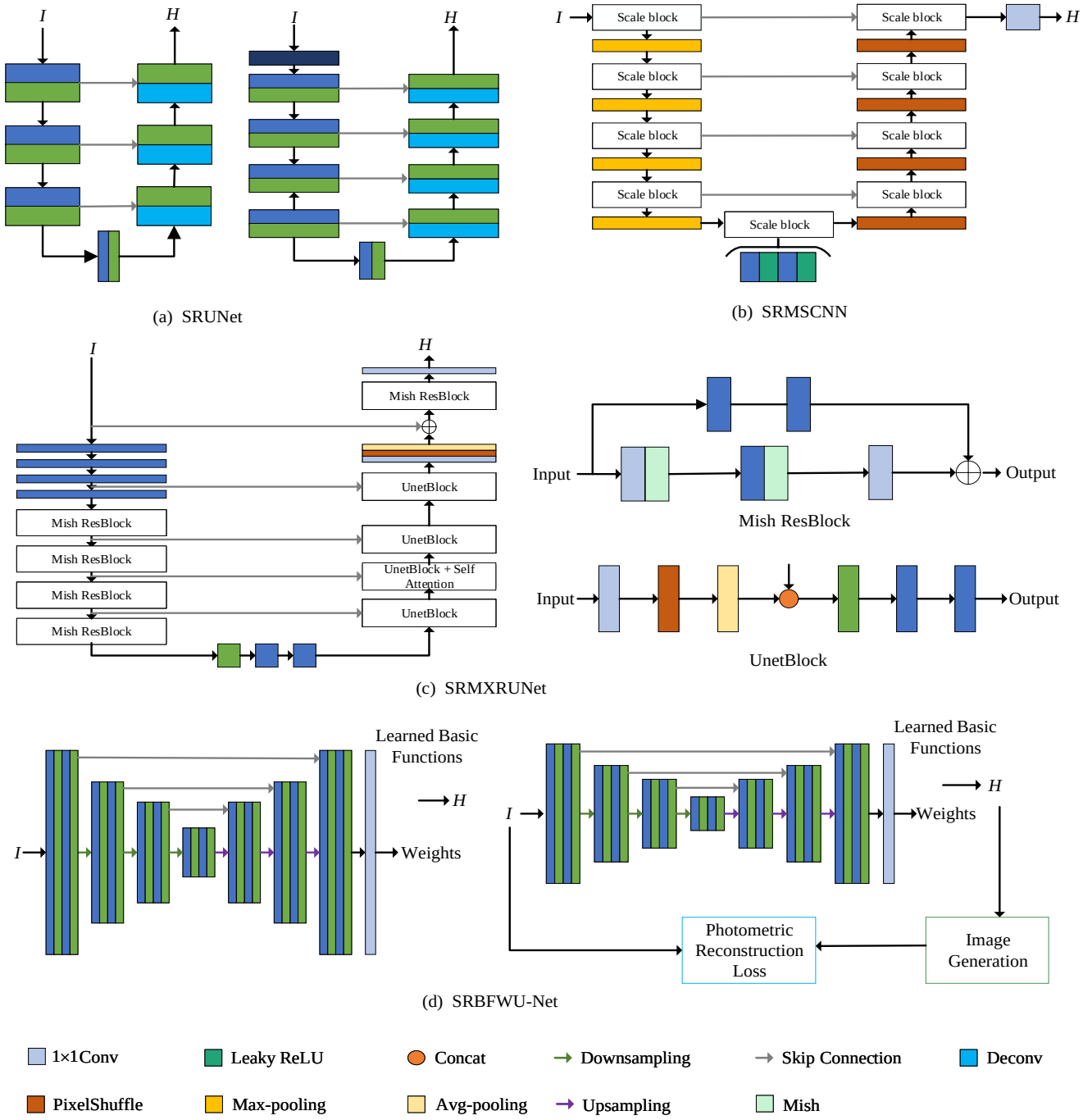


Fig. 4. Spectral reconstruction of RGB image based on U-Net model. (a) SRUNet [61]. (b) SRMSCNN [62]. (c) SRMXRUNet [63]. (d) SRBFWU-Net [64], on the left is supervised learning, and on the right is unsupervised learning.

. The image generation module restores the reconstructed HSI to an RGB image. The photometric reconstruction loss module compares the error between the restored RGB image and the original RGB image, and determines whether the spectrum reconstructed by the network is correct.

In these two methods, before the network training, the RGB image and the ground truth HIS are resized to 512×512, and the RGB image is cropped into several 64×64 image patches, which are used as the inputs of the model. The loss functions of the two methods are $\mathcal{L}_1$ and mean relative absolute error $\mathcal{L}_{\text{MRAE}}$.

3) GAN model : The generative adversarial network (GAN) model [93] is composed of a generator and a discriminator. The generator generates a reconstructed image, and the discriminator discriminates if the reconstructed image is real or fake. The final result is obtained when the game between the two reaches a balance. Several SR methods based on the GAN model are described as follows.

SRCGAN The authors use the conditional GAN [95] to capture spatial context information, so as to obtain accurate reconstructed HSIs. SRCGAN [65] is divided into a generator and a discriminator. The network diagram as shown in Figure

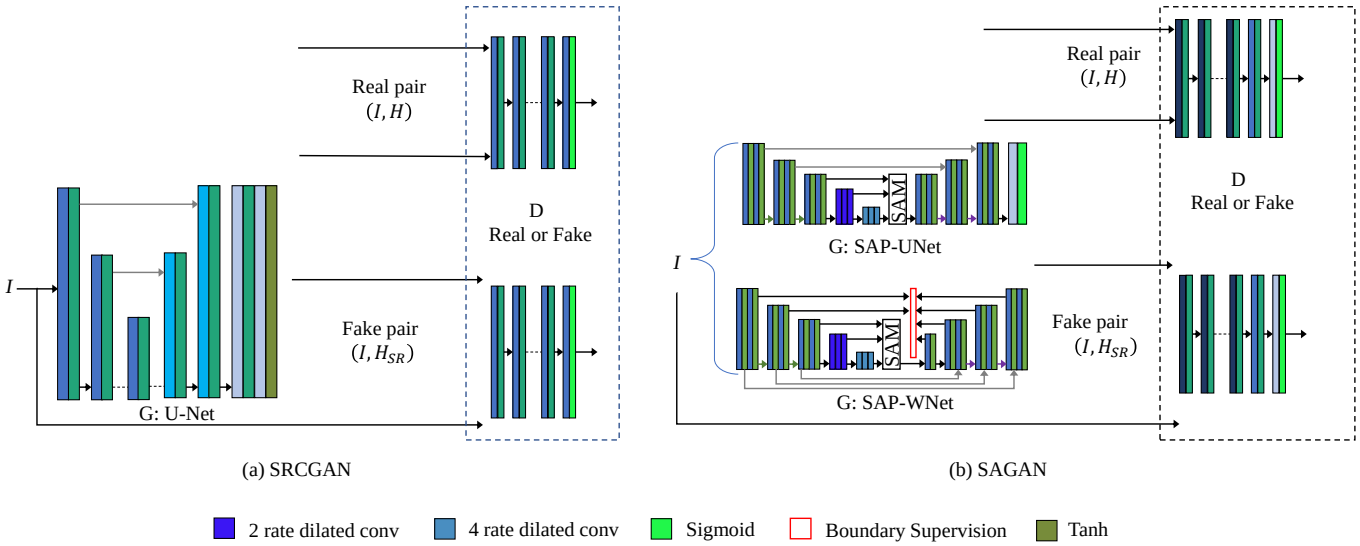


Fig. 5. (a) Conditional Generative Adversarial Network for spectral reconstruction from RGB image [65]. (b) The Generative Adversarial Network is used as the benchmark frame for RGB image spectral reconstruction. SAP-UNet is scale attention pyramid UNet. SAP-WNet is a scale attention pyramid UNet with boundary supervision branch. The discriminator of these two GANs is PatchGAN [94].

5(a).

The generator is based on U-Net and the batch normalization layer [96] is removed. In the encoding stage, eight successive 3×3 convolutions are cascaded, and each convolution is followed by a leaky ReLU. The decoding stage consists of eight deconvolution blocks, each of which includes the deconvolution, followed by dropupt and leaky ReLU. Finally, two 1×1 convolutions are added, and followed by leaky ReLU and tanh activation, respectively. PatchGAN [94] is used as the discriminator, which consists of five consecutive 3×3 convolutional layers. Except for the last convolutional layer followed by sigmoid, the others are followed by leaky ReLU.

Two pairs of images $[I, H_{SR}]$ and $[I, H]$ are input to discriminator to discriminate if they are real or fake. SRCGAN combines GAN loss $\mathcal{L}_{GAN}$ and $\mathcal{L}_1$ as loss function, preserving the global structure of the image and does not produce artifacts and blur [97], [98].

SAGAN The authors propose two improvements based on the GAN model to improve SR performance. SAPUNet [66] uses a U-Net with dilated convolution as the generator, and its encoding stage consists of five large residual blocks, as shown in Figure 5(b). The SAPUNet builds the feature pyramid from feature maps of the last three scale blocks, and uses the scale attention module for scale feature selection [99], [100]. SAPWNet establishes boundary supervision branch on the basis of SAPUNet, which is shown in Figure 5(b). Boundary supervision uses the Canny algorithm to extract the edge features of the image as depth supervision.

The improved PatchGAN is used as the discriminator of SAGAN, which is composed seven consecutive 3×3, one 3×3, and one 3×3 convolutional layers. The last layer is followed by a sigmoid activation. Each convolutional layer is activated by leaky ReLU. In this SAGAN, the $\mathcal{L}_{adv}$ adopts the objective function of WGAN [101], which is combined with the $\mathcal{L}_1$ to form the total loss function, and the model takes 256×256 image patches as input.

4) Dense Network: The core idea of dense network [102] is to densely connect all front and back layers to achieve higher flexibility and richer feature representation, which can reduce the vanishing of gradients and ensure the stability of the network. We discuss several SR methods based on the dense network below.

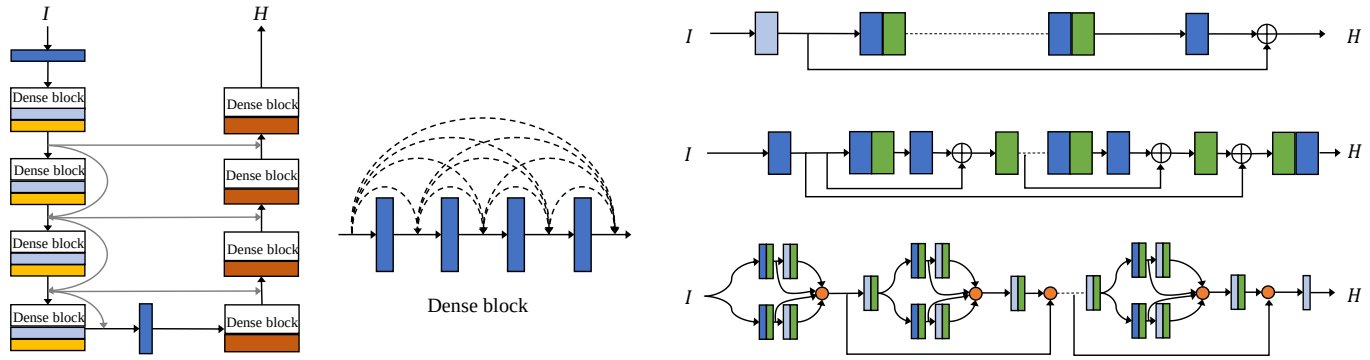
SRTiramisuNet SRTiramisuNet [67] uses a variant of the Tiramisu network [103], which belongs to the category of dense network, as shown in Figure 6(a). More importantly, its architecture is based on a multi-scale paradigm, which allows the network to learn the overall image structure while keeping the image resolution constant.

This network includes down-sampling and up-sampling. Each down-sampling step consists of a 1×1 convolutional layer and a max-pooling, while each scale is composed of dense blocks, with 4 convolutional layers and each layer has 16 convolution kernels of size. Sub-pixel convolution [81]completes up-sampling to ensure pixel fidelity. Skip connections, inside and across the dense blocks perform concatenation instead of summation of layers to speed up learning.

This SRTiramisuNet takes Euclidean loss $\mathcal{L}_{eu}$ as the objective function, and its input is an RGB image patch of size 64×64.

HSCNN+ HSCNN+ [68] has three improvements to HSCNN, namely HSCNN-U, HSCNN-R, and HSCNN-D. The three network structure diagrams can be seen in Figure 6(b).

HSCNN-U uses 1×1 convolutional layer to achieve spectral upsampling. Its method reduces the dependence on the SRF. HSCNN-U only changes the up-sampling operation, the overall network structure inherits HSCNN, and the performance is slightly improved. HSCNN-R replaces the plain convolutional



(a) SRTiramisuNet (b) HSCNN+

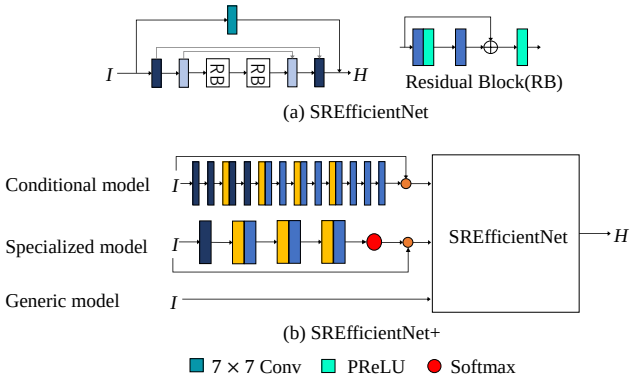
Fig. 6. The above figure shows two methods of spectral reconstruction from RGB image based on Dense Network. (a) SRTiramisuNet [103]. (b) HSCNN+ [68] proposed three network structures for spectral reconstruction from RGB image, which are HSCNN-U, HSCNN-R, and HSCNN-D from top to bottom.

layer of HSCNN with residual blocks while remaining global residual learning to further improve the accuracy of SR.

As the network gets deeper and wider, HSCNN-R is still affected by the vanishing of gradients. HSCNN-D replaces residual blocks by dense blocks with the path-widening fusion scheme, which can substantially alleviate the vanishing of gradients issue. The difference between HSCNN-D and HSCNN-R is that feature fusion is not an addition, but a concatention, which can better learn the SR inverse mapping. With much deeper networks, HSCNN-D can provide higher reconstruction fidelity.

The mean square error [79], [104], [105] causes the luminance deviation of the spectral bands and affects the reconstruction accuracy. The model uses $\mathcal{L}_{MRAE}$ as the loss function, and takes 50×50 RGB image patch as input.

5) Residual Network: Compared with linear CNN, with the deepening of the network, the residual network can avoid the vanishing of the gradient. The residual network makes the reconstructed image not only more detailed, but also remains the global structure. Several SR algorithms based on residual networks are described below.



(a) SREfficientNet

(b) SREfficientNet+

Fig. 7. These are two network models based on the residual network.(a) SREfficientNet [69]. (b) SREfficientNet+ [70].

SREfficientNet SREfficientNet [69] employs ResNet Blocks to learn the mapping of RGB images to HSIs. The model contains a backbone network and 7×7 convolutional layer. The architecture is shown in Figure 7(a).

The 7×7 convolutional layer learns the basic mapping from RGB to hyperspectral. The backbone network is firstly shallow feature extraction and shrinking, which reduces over-fitting and forces the network to learn more compact and relevant features. The second is complex feature extraction, which consists of stacking two residual blocks [106] to obtain more complex features. The third is the expansion and reconstruction, which expands the features and reconstructs the HSI. Skip connections to make full use of low-level features. Finally, the output of the 7×7 convolutional layer and the backbone network are added to obtain the result. In order to increase the network nonlinearity, the ReLU activation function is replaced with PReLU [107].

The input of the model is 36×36 image patches, and augmented using flipping and rotation operations, while the network is trained with $\mathcal{L}_2$.

SREfficientNet+ SRefficientNet+ [70] solves the problem of SR from RGB images acquired under unknown conditions (e.g.,unknown camera sensitivity function), in the wild. The authors propose three models for this, which are *generic model*, *conditional model*, and *specialized model*, respectively, as shown in Figure 7(b).

The general model is to train the SR network with RGB image inputs of different sensitivities without adding information about the sensitivity function. In the conditional model, firstly an approximate sensitivity function is obtained by the estimator (sensitivity function estimation network), and then use the estimated sensitivity function together with the RGB image as the input of the SR network. Specialized model trains the SR model for each function in a set of limited sensitivity functions. In the wild, the model is selected, which is achieved through a classifier (classification network).

The estimator is composed of 12 convolutional layers and 4 max-pooling layers, and ReLU is used as the activation function, and the sensitivity function is finally generated. The objective function of the estimator is composed of reconstruction loss ($\mathcal{L}_{\text{REC}} = \frac{1}{N}\left\|I - H\widetilde{S}\right\|_F^2$), supervision loss ($\mathcal{L}_s = \left\|S - \widetilde{S}\right\|_F^2$), and smooth regularization loss ($\mathcal{L}_{\text{REG}} = \|TS\|_F^2$). Where, $\widetilde{S}$ represent an estimated sensitivity function, and

$\|\cdot\|_F^2$ is the squared Frobenius norm, and T is the second-order derivative operator. The classifier uses the cross-entropy loss ($\mathcal{L}_{ce}$) to train the network. The network configuration of the SR network in the SREfficientNet+ method is consistent with SREfficientNet.

6) Attention network: The previously discussed network treats the spatial location and channel equally. In some cases, we have to selectively focus on a few features in a given layer. The attention-based model allows this flexibility and takes into account that not all features are important for SR. Following are the examples of the CNN algorithms using the attention mechanism.

SRAWAN The authors propose an adaptive weighted attention network(SRAWAN), which explores the camera spectral sensitivity (CSS) prior and the interdependence among intermediate features , to reconstruct more accurate HSIs. SRAWAN is composed of Shallow feature extraction, deep feature extraction, and reconstruction module ,as shown in Figure 8.

Shallow feature extraction includes a 3×3 convolutional layer. Deep feature extraction stacks several dual residual attention modules (DRAB) [108]. Each DRAB contains a basic residual block, paired convolution layers with a large (5×5) and a small size (3×3) kernels, adaptive weighting Channel Attention Module (AWCA), long and short skip connections to form dual residual learning. The AWCA is inherited from SE [109], while the difference is that adaptive weighted feature statistics (convolutional layer) replaces global average pooling statistics, to strengthen feature learning.

The reconstruction module is composed of a patch-level second-order non-local module (PSNL) [110], which captures long-range spatial context information through second-order non-local operations, to obtain a more powerful feature representation. Then the deep features are passed through PSNL to generate the result.

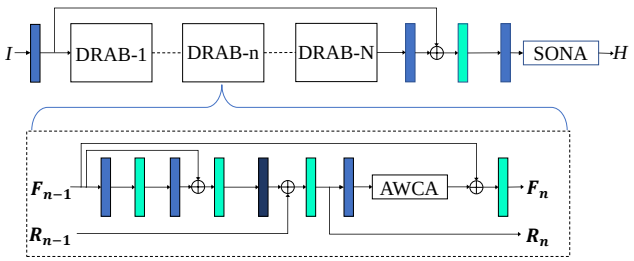


Fig. 8. Adaptive Weighted Attention Network with camera spectral sensitivity prior for spectral reconstruction from RGB Images. F_{n-1} and F_n represent the direct input and output of the n-th DRAB. R_{n-1} and R_n represent the residual input and output of the n-th DRAB.

The model removes the batch normalization layer, and replaces ReLU by PReLU. The loss function of this network is composed of the CSS loss $\mathcal{L}_{\text{CSS}}$ and $\mathcal{L}_{\text{MRAE}}$. CSS loss refers to the error between the restored RGB image obtained by applying the CSS function to the reconstructed HIS and the original RGB image. Note that this CSS function has the same effect as the SRF.

SRHRNet The authors proposed a 4-level hierarchical regression network (SRHRNet) [72] for SR from RGB image. The network structure is shown in Figure 9.

SRHRNet is a multi-scale structure that uses PixelUnShuffle and PixelShuffle to perform downsampling and upsampling respectively, while remaining pixel information. Each level consists of inter-layer integration, artifact removal and global feature extraction. The inter-layer integration means that the output features of the subordinate level are PixelShuffled, then concatenated to current level, finally processed by a 3×3 convolutional layer to unify the feature maps number. The artifact removal is completed by the residual dense block [106], [111], which includes five densely connected convolutional layers and the local residual. The global feature extraction is composed of the residual global block [106], [109] with skip connection of input, which is used to extract attention for every long-range pixels through MLP (Multilayer perceptron).

The integration of multiple modules at the top of the network can effectively remove artifacts and obtain high-quality reconstructed HSIs. SRHRNet uses $\mathcal{L}_1$ as the loss function and proposes an 8-setting ensemble strategy to further enhance generalization.

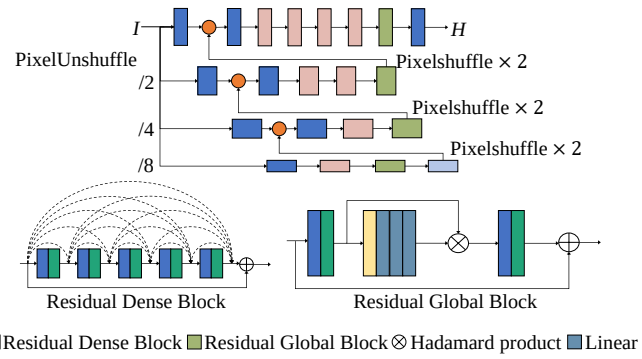


Fig. 9. 4-level hierarchical regression network for RGB image spectral reconstruction (SRHRNet) [72].

SRRPAN The authors propose the residual pixel attention network (SRRPAN) [73]for SR from RGB image, which can adaptively rescale each pixel-wise weights of all input feature maps, as shown in Figure 10.

SRRPAN contains a residual attention group (RAG) and a residual attention module (RPAB), which firstly goes through a 3×3 convolution layer, while the backbone consists of 4 stacked RAG blocks followed by a 1×1 convolutional layer. Each RAG is composed of 8 RPAN blocks and 3×3 convolutional layer, all with residual learning.

RPAB contains pixel attention (PA) block with skip connection of input to obtain pixel-level attention features. The PA block is similar to the channel attention (CA) block in RCAN [112], but the difference is that the global pooling is removed. As the network depth increases, each RAG extracts features of different scales [113] . In order to make full use of these feature maps, they are fused through the *concat* layer to improve the quality of the reconstructed HSI.

During training, the network takes 64×64 RGB image patches as inputs and $\mathcal{L}_{\text{MRAE}}$ as the loss function.

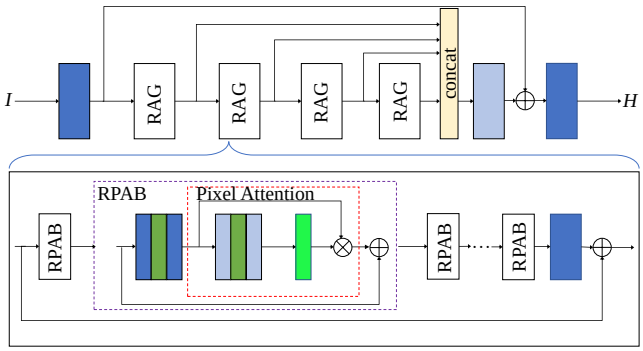


Fig. 10. Residual Pixel Attention Network for Spectral Reconstruction from RGB Images. [73]

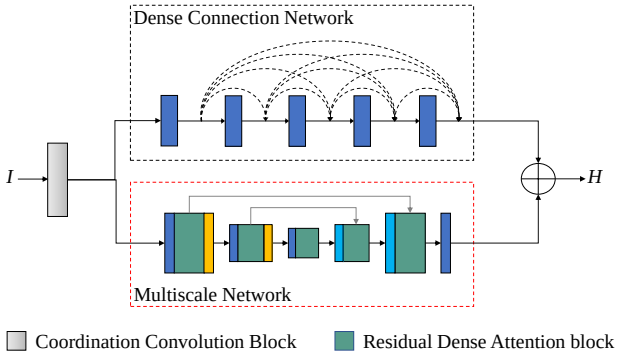


Fig. 11. Light Weight Residual Dense Attention Net for Spectral Reconstruction from RGB Images.

7) Multi-branch network: In contrast to single-stream CNN, the goal of multi-branch network is to obtain a diverse set of features on multiple context scales. This information is then fused to obtain better reconstructed HSI. This design can also achieve multi-path signal flow, leading to better information exchange during training. We explain the SR methods based on the multi-branch network below.

SRLWRDANet Nathan et al. [74] provide a lightweight residual dense attention network based on a multi-branch network to solve this problem, called SRLWRDANet, which has about 233059 parameters. SRLWRDANet consists of two parallel subnets, which are a densely connected network and a multi-scale network, respectively.

The network firstly goes through a coordinated convolution block [114] to extract shallow features and boundary information, and the parameters of which are shared by the latter two subnets. Densely connected network enable the global network to have strong feature representation ability and effectively alleviate the vanishing of gradient. The multi-scale network is composed of convolutional layer and a residual dense attention blocks (RDAB) [115] to extract scale-level features. RDAB combines the residual dense network with the attention mechanism to capture local hierarchical features. The subnet is a multi-scale connection of RDAB in the U-Net fashion , where the down-sampling is implemented by the max-pooling meanwhile the deconvolution completes the up-sampling. Finally, the outputs of the two subnets are fused to obtain the final result. The architecture of model is shown in Figure 11.

SRLWRDANet takes the sum of $\mathcal{L}_2$ and structural similarity loss ($\mathcal{L}_{\text{SSIM}}$) as the objective function to achieve the purpose of preserving structural features.

SRPFMNet The previous CNN-based SR methods are to map RGB to HSI in a size-specific receptive field centered on a certain pixel. Because of their different category and spatial position, pixels in hyperspectral usually require different sized receptive fields and distinct mapping functions. The authors propose a pixel-aware deep function-mixture network (SRPFMNet) [75] based on multi-branch network to solve this problem. The network architecture is shown in Figure 12.

The SRPFMNet is composed of a convolution layer followed by Relu and multiple function mixing (FM) modules, and fuses the intermediate features generated by the previous FM blocks with skip connection, while adopting global residual structure.

Each FM module includes mix function and several basis function subnets, and these networks are formed by stacking multiple convolutional layers followed by activations. The mix function subnet generates pixel-wise mix weights. These basis function subnets have different-sized convolution kernels to generate receptive fields of different size and learn distinct mapping schemes. The outputs of all basis function subnets are linearly mixed based on the generated pixel-wise weights. Finally, the RGB spectral up-sampled image is combined with the output of the backbone network to obtain a reconstructed HSI. The model uses $\mathcal{L}_1$ for training. The input patch size is 64×64.

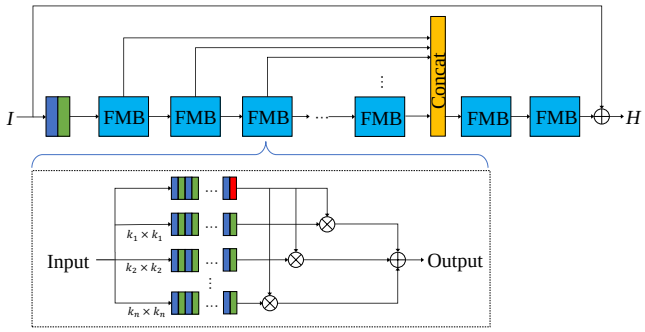


Fig. 12. Pixel-aware Deep Function-mixture Network for RGB Image Spectral Super-Resolution [75]

IV. Algorithm Analysis and Comparison

We compare the performance of different SR algorithms on the BGU-HS and ARAD-HS datasets in Table IV.

A. Comparison of Prior-based Methods

The sparse coding method considers the sparsity of HSIs, but ignores the characteristics of spatial structure similarity and correlation between spectra. Other methods in dictionary learning are to incorporate prior knowledge such as spectral feature correlation, spatial context information, and local

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT ALGORITHMS ON DIFFERENT DATASETS.

Category	Method	BGU-HS		ARAD-HS	
		RMSE	MARE	RMSE	MARE
Dictionary Learning	Sparse Coding [37]	49.217	0.0816	0.0331	0.0787
Linear CNN	HSCNN [58]	17.27	0.0190	-	-
	SR-2DNet [59]	21.394	0.018	-	-
	SR-3Dnet [59]	20.010	0.020	-	-
U-Net Model	SRU-Net [61]	15.88	0.0156	0.0152	0.0395
	SRMUNet [61]	23.88	0.0312	0.0187	0.0698
	SRMSCNN [62]	19.28	0.0231	0.0235	0.0724
	SRMXRUNet [63]	-	-	0.0165	0.0444
	SRBFWU-Net [64]	-	-	0.0198	0.0440
Dense Network	SRTiramisuNet [67]	20.98	0.0272	0.0251	0.0850
	HSCNN-R [68]	13.911	0.0145	0.0143	0.0372
	HSCNN-D [68]	13.128	0.0135	-	-
Attention Network	SRAWAN [71]	**10.24**	**0.0114**	0.0129	**0.0301**
	SRHRNet [72]	13.5165	0.01369	0.0139	0.0323
Multi-branch Network	SRPFMNet [75]	-	-	**0.01267**	0.0308

[1] The first and second best results are shown in bold and underlined respectively

linear relationship into the dictionary learning to improve the representation ability and reconstruction performance of the dictionary. Manifold learning and Gaussian process build models based on the statistical information of HSIs.

Prior-based methods rely on manually-made priors, making the model's representation ability inferior to deep learning. In the case of unknown camera sensitivity function, the reconstruction performance and accuracy will be reduced.

B. Comparison between data-driven methods

The natural images have rich spectral structure information in the spectral domain, which can guide CNN to predict more accurate HSIs.

The powerful feature representation ability of neural net work is not possessed by traditional algorithms. So far, there are many SR methods based on CNN. Linear CNN is the simplest SR model. The methods based on GAN model make the reconstructed HSI close to reality, *e.g. SRCGAN [65], SAGAN [66]* . The U-Net-based model, which jointly encodes the local and non-local information of the image, to obtains an accurate HSI, *e.g. SRU-Net [61], SRMSCNN [62] , SRBFWU-Net [64], SRMXRUNet [63]*. The spectral reconstruction model of residual network and dense network can alleviate the vanishing of gradients during training and ensure that more accurate results can be obtained. The attention networks and multi-branch networks greatly enhance SR performance by increasing network complexity, *e.g. SRHRNet [72]* .

When the RGB image is noisy and compressed, the physical consistency of the network is less convincing. Physical consistency refers to the error between the RGB image regenerated from the HSI according to Eq. 1 and the original RGB image. The use of different brightness adjustments for the original RGB image proves that the network is unstable to brightness changes. The results are shown in the Table V.

TABLE V
STATE-OF-ART MODELS TESTED FOR GENERALIZATION IN THE ARAD-HS DATASET.

Method	SRAWAN [71]	SRHRNet [72]	SRPFMNet [75]
Original RGB	0.0301	0.0323	0.0308
0.5×Brightness	0.0327	0.0405	0.0356
2×Brightness	0.0397	0.0442	0.0339
Physica-Consistency	0.0329	0.0335	0.0356

However, in the algorithm of RGB image spectral reconstruction based on deep learning, it is difficult to declare that an algorithm is a clear winner. Because many factors are involved, such as network complexity, network depth, training data, training patch size, number of feature maps, etc. Only by maintaining the consistency of all parameters can a fair comparison be made.

C. Trends and Challenges

As people pay more and more attention to spectral reconstruction, many state-of-art deep learning methods have emerged. Despite the great success, there are still many unsolved problems. We point out these problems and introduce some promising trends in future development.

Network design Spatial context information plays an important role in network performance, so local and global information are combined in the design process. Low-frequency information and high-frequency information determine the quality of the reconstructed image, and this information is considered in the design. In different scenarios, people often pay attention to different features of things, and combine attention mechanisms to enhance their attention to key features and promote the production of real details. A good network design can ensure the best performance while reducing space and time complexity. How to do this is still a question.

Objective function After the network model is established, it is necessary to design a suitable objective function to obtain the optimal solution in the huge solution space. Existing network models use pixel-level objective functions, such as L2, L1. These may be useful for spectral reconstruction, but do not greatly improve the image perception quality. In the future, content loss, perceptual loss, and texture loss can be added.

Datasets Because the hyperspectral image acquisition equipment is expensive and complicated, there are only a few hyperspectral datasets available at this stage. Deep learning training often requires a large dataset. Existing methods are likely to overfit. In the training network, operations such

as cropping, flipping, zooming, rotating, and color dithering can be used to increase the training dataset. Future research directions can combine image statistics with deep learning.

Real Word The real-world RGB images are obtained from unknown, uncalibrated cameras. At this time, the network performed poorly in this situation. An important future research direction is that a well-designed network can reconstruct accurate spectral images in real-word RGB images.

Generalization For the same scene, the same object is exposed to different exposures, and it is difficult for the existing models to achieve a good reconstruction effect. For physical consistency, all methods have not found a spectrum consistent with the original RGB. Existing methods are difficult to deal with these problems.

In the future, spectral reconstruction can be used in the video field. The spectral reconstruction algorithm is applied to the medical field to improve the diagnosis rate of doctors.

V. Conclusion

We present a systematic review of spectral reconstruction from RGB image, including prior-based methods and data-driven deep learning approaches. The mathematical relationship between RGB images and hyperspectral images is firstly given as the fundamentals for the following reconstruction methods. Then, these two types of methods were compared.

On the BGU-HS and ARAD-HS datasets, we use MRAE and RMSE as the evaluation criteria to compare sparse coding, HSCNN, SRAWAN, SRHRNet, etc. The results show that the SR algorithm based on deep learning occupies a great advantage. In the future, a higher-performance neural network will be designed to improve SR performance.

We only summarize the reconstruction of hyperspectral images in the spectral domain, and the spatial super-resolution of hyperspectral images can also be performed, *e.g.* Dian et al. [116], Fu et al. [117]. The SR method has a profound impact in the field of remote sensing, geological exploration and medical treatment. We can perform SR on the endoscopic image to accurately locate the lesion.

Acknowledgment

The authors would like to thank the authors of Sparse Coding, HSCNN, SR2D/3DNet, SRUNet, SRMSCNN, and etc. for providing open-source code.

References

[1] T. M. Lillesand, R. W. Kiefer, and J. W. Chipman, "Remote sensing and image interpretation (fifth edition)," *Geographical Journal*, vol. 146, no. 3, 2004. 1

[2] N. Akhtar and A. Mian, "Non-parametric coupled bayesian dictionary and classifier learning for hyperspectral classification," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2017. 1

[3] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 1, no. 2, pp. 6–36, 2013. 1

[4] A. S. Charles, B. A. Olshausen, and C. J. Rozell, "Learning sparse codes for hyperspectral imagery," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 963–978, Sep. 2011. 1

[5] G. J. Edelman, E. Gaston, T. G. V. Leeuwen, P. J. Cullen, and M. C. G. Aalders, "Hyperspectral imaging for non-contact analysis of forensic traces," *Forensic science international*, vol. 223, no. 1-3, 2012. 1

[6] Y. Wang and R. Niu, "Hyperspectral urban remote sensing image smoothing and enhancement using forward-and-backward diffusion," in *Urban Remote Sensing Event*, 2009. 1

[7] L. Ojha, M. B. Wilhelm, S. L. Murchie, A. S. Mcewen, J. J. Wray, J. Hanley, M. Massé, and M. Chojnacki, "Spectral evidence for hydrated salts in recurring slope lineae on mars," *Nature Geoscience*, 2015. 1

[8] E. Belluco, M. Camuffo, S. Ferrari, L. Modenese, S. Silvestri, A. Marani, and M. Marani, "Mapping salt-marsh vegetation by multispectral and hyperspectral remote sensing," *Remote Sensing of Environment*, vol. 105, no. 1, pp. 54–67, 2006. 1

[9] M. Borengasser, W. S. Hungate, and R. Watkins, *Hyperspectral remote sensing: Principles and applications*, 2007. 1

[10] A. Castrodad, Z. Xing, J. Greer, E. Bosch, and G. Sapiro, "Discriminative sparse representations in hyperspectral imagery," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, 2010. 1

[11] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, 2004. 1

[12] E. Underwood, S. Ustin, and D. Dipietro, "Mapping nonnative plants using hyperspectral imagery," *Remote Sensing of Environment*, vol. 86, no. 2, pp. 150–161, 2003. 1

[13] D. Haboudane, J. R. Miller, E. Pattey, P. J. Zarco-Tejada, and I. B. Strachan, "Hyperspectral vegetation indices and novel algorithms for predicting green lai of crop canopies: Modeling and validation in the context of precision agriculture," *Remote Sensing of Environment*, vol. 90, no. 3, pp. 337–352, 2004. 1

[14] E. A. Cloutis, "Hyperspectral geological remote sensing : evaluation of analytical techniques," *Intl. J. Remote Sens*, vol. 14, 1996. 1

[15] K. Hege, D. O'Connell, W. Johnson, S. Basty, and E. Dereniak, "Hyperspectral imaging for astronomy and space surviellance," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 5159, 01 2004. 1

[16] J. Mustard and J. Sunshine, "Spectral analysis for earth science: Investigations using remote sensing data," *Remote Sensing for the Earth Sciences: Manual of Remote Sensing*, vol. 3, pp. 251–307, 01 1999. 1

[17] G. Lu and B. Fei, "Medical hyperspectral imaging: a review," *Journal of Biomedical Optics*, vol. 19, no. 1, p. 10901, 2014. 1

[18] Y. Zhou, H. Chang, K. Barner, P. Spellman, and B. Parvin, "Classification of histology sections via multispectral convolutional sparse coding," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 3081–3088. 1

[19] P. Andersson, S. Montan, and S. Svanberg, "Multispectral system for medical fluorescence imaging," *IEEE Journal of Quantum Electronics*, vol. 23, no. 10, pp. 1798–1805, 1987. 1

[20] Y., Tarabalka, , , J., Chanussot, , , A. J., and Benediktsson, "Segmentation and classification of hyperspectral images using watershed transformation," *Pattern Recognition*, 2010. 1

[21] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 45–54, 2014. 1

[22] M. Uzair, A. Mahmood, and A. Mian, "Hyperspectral face recognition using 3d-dct and partial least squares," in *BMVC 2013*, 2013. 1

[23] M. Uzair and A. Mahmood, "Hyperspectral face recognition with spatiospectral information fusion and pls regression," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 24, no. 3, pp. 1127–37, 2015. 1

[24] D. Zhang, W. Zuo, and F. Yue, "A comparative study of palmprint recognition algorithms," *Acm Computing Surveys*, vol. 44, no. 1, pp. 1–37, 2012. 1

[25] R. Ramanath, W. E. Snyder, and Hairong Qi, "Eigenviews for object recognition in multispectral imaging systems," in *32nd Applied Imagery Pattern Recognition Workshop, 2003. Proceedings.*, Oct 2003, pp. 33–38. 1

[26] H. V. Nguyen, A. Banerjee, P. Burlina, J. Broadwater, and R. Chellappa, *Tracking and Identification via Object Reflectance Using a Hyperspectral Video Camera*. Springer Berlin Heidelberg, 2011. 1

[27] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 1

[28] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," 2016. 1

[29] D. W. J. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *Signal Processing Magazine IEEE*, vol. 19, no. 1, pp. 58–69, 2002. 1

[30] X. Cao, H. Du, X. Tong, Q. Dai, and S. Lin, "A prism-mask system for multispectral video acquisition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 33, no. 12, pp. 2423–2435, 2011. 1

[31] M. Descour and E. Dereniak, "Computed-tomography imaging spectrometer: experimental calibration and reconstruction results," *Appl Opt*, vol. 34, no. 22, pp. 4817–4826, 1995. 1

[32] L. Gao, R. T. Kester, N. Hagen, and T. S. Tkaczyk, "Snapshot image mapping spectrometer (ims) with high sampling density for hyperspectral microscopy," *Optics Express*, vol. 18, no. 14, pp. 14 330–14 344, 2010. 1

[33] A. A. Wagadarikar, N. P. Pitsianis, X. Sun, and D. J. Brady, "Video rate spectral imaging using a coded aperture snapshot spectral imager," *Optics Express*, vol. 17, no. 8, pp. 6368–6388, 2009. 1

[34] C. Xun, Y. Tao, L. Xing, S. Lin, Y. Xin, Q. Dai, L. Carin, and D. J. Brady, "Computational snapshot multispectral cameras: Toward dynamic capture of the spectral world," *IEEE Signal Processing Magazine*, vol. 33, no. 5, pp. 95–108, 2016. 1

[35] B. Arad, O. Ben-Shahar, R. Timofte, L. V. Gool, and M. H. Yang, "Ntire 2018 challenge on spectral reconstruction from rgb images," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018. 1, 2, 3

[36] B. Arad, R. Timofte, O. Ben-Shahar, Y. Lin, G. Finlayson, S. Givati, J. Li, C. Wu, R. Song, Y. Li, F. Liu, Z. Lang, W. Wei, L. Zhang, J. Nie, Y. Zhao, L. Po, Q. Yan, W. Liu, T. Lin, Y. Kim, C. Shin, K. Rho, S. Kim, Z. ZHU, J. HOU, H. Sun, J. Ren, Z. Fang, Y. Yan, H. Peng, X. Chen, J. Zhao, T. Stiebel, S. Koppers, D. Merhof, H. Gupta, K. Mitra, B. J. Fubara, M. Sedky, D. Dyke, A. Banerjee, A. Palrecha, S. sabarinathan, K. Uma, D. S. Vinothini, B. Sathya Bama, and S. M. Md Mansoor Roomi, "Ntire 2020 challenge on spectral reconstruction from an rgb image," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2020, pp. 1806–1822. 1, 2, 3

[37] B. Arad and O. Ben-Shahar, "Sparse recovery of hyperspectral signal from natural rgb images," in *European Conference on Computer Vision*, 2016. 2, 3, 4, 5, 14

[38] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 19, no. 9, p. 2241, 2010. 2, 3

[39] J. Wu, J. Aeschbacher, and R. Timofte, "In defense of shallow learned spectral reconstruction from rgb images," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017, pp. 471–479. 4

[40] F. Ying, Y. Zheng, Z. Lin, and H. Hua, "Spectral reflectance recovery from a single rgb image," *IEEE Transactions on Computational Imaging*, vol. PP, pp. 1–1, 2018. 4, 5

[41] Y. Li, C. Wang, and J. Zhao, "Locally linear embedded sparse coding for spectral reconstruction from rgb images," *IEEE Signal Processing Letters*, vol. PP, no. 99, pp. 1–1, 2017. 4, 5

[42] Y. Geng, S. Mei, J. Tian, Y. Zhang, and Q. Du, "Spatial constrained hyperspectral reconstruction from rgb inputs using dictionary representation," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019. 4, 6

[43] Y. Jia, Y. Zheng, L. Gu, A. Subpa-Asa, and I. Sato, "From rgb to spectrum for natural scenes via manifold-based mapping," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 4, 6

[44] N. Akhtar and A. S. Mian, "Hyperspectral recovery from rgb images using gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018. 4, 6

[45] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2011. 4

[46] R. Timofte, V. Desmet, and L. Vangool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian Conference on Computer Vision*, 2014.

[47] A. Çivril and M. Magdon-Ismail, "On selecting a maximum volume sub-matrix of a matrix and related problems," *Theoretical Computer Science*, vol. 410, no. 47-49, pp. 4801–4811, 2009. 5

[48] PALMER and E. Stephen, "Vision science: Photons to phenomenology," *Quarterly Review of Biology*, vol. 77, no. 4, pp. 233–234, 1999. 6

[49] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Simultaneous sparse approximation via greedy pursuit," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 5, March 2005, pp. v/721–v/724 Vol. 5. 6

[50] Cohen and Jozef, "Dependency of the spectral reflectance curves of the munsell color chips," *Psychonomic Science*, vol. 1, no. 1-12, pp. 369–370, 1964. 6

[51] Maloney and T. Laurence, "Evaluation of linear models of surface spectral reflectance with small numbers of parameters," *josaa/3/10/josaa-3-10-1673.pdf*, vol. 3, no. 10, pp. 1673–0, 1986. 6

[52] J. P. S. Parkkinen, "Characteristics spectra of munsell colors," *J Optical Society of America A*, vol. 6, 1989. 6

[53] T. Jaaskelainen, J. Parkkinen, and S. Toyooka, "Vector-subspace model for color representation," *J.optical Soc.america A*, vol. 7, no. 4, pp. 725–730, 1990. 6

[54] D. H. Marimont and B. A. Wandell, "Linear models of surface and illuminant spectra," *Journal of the Optical Society of America A Optics & Image Science*, vol. 9, no. 11, p. 1905, 1992. 6

[55] 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," *Journal of the Optical Society of America A*, vol. 23, no. 8, pp. 2020–2026, 2006. 6

[56] Tenenbaum, B. Joshua, D. Silva, Vin, Langford, and C. John, "A global geometric framework for nonlinear dimensionality reduction." *Science*, 2000. 6

[57] C. Rasmussen and C. Williams, "Gaussian process for machine learning," 01 2006. 6

[58] Z. Xiong, Z. Shi, H. Li, L. Wang, D. Liu, and F. Wu, "Hscnn: Cnn-based hyperspectral image recovery from spectrally undersampled projections," in *IEEE International Conference on Computer Vision Workshop*, 2017, pp. 518–525. 6, 7, 14

[59] S. Koundinya, H. Sharma, M. Sharma, A. Upadhyay, and S. Chaudhury, "2d-3d cnn based architectures for spectral reconstruction from rgb images," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018. 6, 7, 8, 14

[60] X. H. Han, B. Shi, and Y. Zheng, "Residual hsrcnn: Residual hyperspectral reconstruction cnn from an rgb image," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018. 6, 7, 8

[61] T. Stiebel, S. Koppers, P. Seltsam, and D. Merhof, "Reconstructing spectral images from rgb-images using a convolutional neural network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2018, pp. 1061–10 615. 7, 8, 9, 14

[62] Y. Yan, L. Zhang, W. Wei, and Y. Zhang, "Accurate spectral super-resolution from single rgb image using multi-scale cnn," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 2018. 7, 8, 9, 14

[63] A. Banerjee and A. Palrecha, "Mxr-u-nets for real time hyperspectral reconstruction," 04 2020. 7, 8, 9, 14

[64] B. J. Fubara, M. Sedky, and D. Dyke, "Rgb to spectral reconstruction via learned basis functions and weights," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2020, pp. 1984–1993. 7, 8, 9, 14

[65] A. Alvarez-Gila, J. Van De Weijer, and E. Garrote, "Adversarial networks for spatial context-aware spectral image reconstruction from RGB," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 480–490. 7, 9, 10, 14

[66] P. Liu and H. Zhao, "Adversarial networks for scale feature-attention spectral image reconstruction from a single rgb," *Sensors (Basel, Switzerland)*, vol. 20, no. 8, 2020. 7, 10, 14

[67] S. Galliani, C. Lanaras, D. Marmanis, E. Baltsavias, and K. Schindler, "Learned spectral super-resolution," 2017. 7, 10, 14

[68] Z. Shi, C. Chen, Z. Xiong, D. Liu, and F. Wu, "Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018. 7, 10, 11, 14

[69] Y. Can and R. Timofte, "An efficient cnn for spectral reconstruction from rgb images," 04 2018. 7, 11

[70] 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)*, 2020. 7, 11

[71] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2020, pp. 1894–1903. 7, 14

[72] Y. Zhao, L. M. Po, Q. Yan, W. Liu, and T. Lin, "Hierarchical regression network for spectral reconstruction from rgb images," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020. 7, 12, 14

[73] H. Peng, X. Chen, and J. Zhao, "Residual pixel attention network for spectral reconstruction from rgb images," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020. 7, 12, 13

[74] D. S. 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," 2020. 7, 13

[75] L. Zhang, Z. Lang, P. Wang, W. Wei, and Y. Zhang, "Pixel-aware deep function-mixture network for spectral super-resolution," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12 821–12 828, 2020. 7, 13, 14

[76] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 38, no. 2, pp. 295–307, 2016. 6, 8

[77] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006. 7

[78] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006. 7

[79] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," 2016. 7, 11

[80] B. Smits, "An rgb-to-spectrum conversion for reflectances." 7

[81] W. Shi, J. Caballero, F. Huszár, J. Totz, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," 2016. 8, 10

[82] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 558–567. 8

[83] D. Misra, "Mish: A self regularized non-monotonic neural activation function," 2019. 8

[84] J. Howard and S. Gugger, "fastai: A layered api for deep learning," *Information (Switzerland)*, 2020. 8

[85] A. 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," 07 2017. 8

[86] Y. Sugawara, S. Shiota, and H. Kiya, "Super-resolution using convolutional neural networks without any checkerboard artifacts," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 66–70. 8

[87] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 05 2018. 8

[88] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," 2016. 8

[89] J. K. Eem, H. D. Shin, and S. O. Park, "Reconstruction of surface spectral reflectances using characteristic vectors of munsell colors," in *Color & Imaging Conference*, 1994. 8

[90] D. Connah, S. Westland, and M. G. A. Thomson, "Recovering spectral information using digital camera systems," *Coloration Technology*, vol. 117, no. 6, pp. 309–312, 2001. 8

[91] R. Gershon, "From r,g,b to surface reflectance : Computing color constancy descriptors in images," *Proc. 10th Int. Joint Conf. on Artificial Intelligence, 1987*, 1987. 8

[92] M. Sedky, M. Moniri, and C. C. Chibelushi, "Spectral-360: A physics-based technique for change detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014. 8

[93] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *ArXiv*, 06 2014. 9

[94] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," 2016. 10

[95] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *Computer Science*, pp. 2672–2680, 2014. 9

[96] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015. 10

[97] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," 2016. 10

[98] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 2242–2251. 10

[99] L. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 3640–3649. 10

[100] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018. 10

[101] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," 03 2017. 10

[102] L. Zhao, J. Wang, X. Li, Z. Tu, and W. Zeng, "On the connection of deep fusion to ensembling," 11 2016. 10

[103] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 1175–1183. 10, 11

[104] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 38, no. 2, pp. 295–307, 2016. 11

[105] C. Chen, X. Tian, F. Wu, and Z. Xiong, "Udnet: Up-down network for compact and efficient feature representation in image super-resolution," in *2017 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2018. 11

[106] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778. 11, 12

[107] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," 2015. 11

[108] X. Liu, M. Suganuma, Z. Sun, and T. Okatani, "Dual residual networks leveraging the potential of paired operations for image restoration," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 7000–7009. 12

[109] J. Hu, L. Shen, G. Sun, and S. Albanie, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, 2017. 12

[110] Bryan, Xia, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," 2019. 12

[111] 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*, 2017, pp. 4700–4708. 12

[112] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 294–310. 12

[113] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution: 15th european conference, munich, germany, september 8-14, 2018, proceedings, part viii," in *European Conference on Computer Vision*, 2018. 12

[114] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the coordconv solution," *CoRR*, vol. abs/1807.03247, 2018. [Online]. Available: http://arxiv.org/abs/1807.03247 13

[115] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2020. 13

[116] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 15

[117] Y. Fu, Y. Zheng, H. Hua, S. Imari, and S. Yoichi, "Hyperspectral image super-resolution with a mosaic rgb image," *IEEE Transactions on Image Processing*, vol. PP, pp. 1–1, 2018. 15