

Received March 2, 2021, accepted March 13, 2021, date of publication March 15, 2021, date of current version March 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3066472

U-Net-Based Multispectral Image Generation From an RGB Image

TAO ZENG¹, CHANGYU DIAO², AND DONGMING LU¹

¹College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

²Cultural Heritage Institute, Zhejiang University, Hangzhou 310027, China

Corresponding author: Changyu Diao (dcy@zju.edu.cn)

This work was supported in part by the Key Research and Development Projects in Zhejiang Province of China in 2018 under Grant 2018C03051, in part by the Research on the Core Technology of Digitalization of Cave Temples and Ancient Buildings, in part by the National Social Science Fund of China under Grant 14ZDB057, in part by the City and School Strategic Development Research Platform of Hangzhou and Zhejiang University of China under Grant SX201904, and in part by the Key Scientific Research Base for Digital Conservation of Cave Temples in Zhejiang University.

ABSTRACT Multispectral images have lower spatial resolution than RGB images. It is difficult to obtain multispectral images with both high spatial resolution and high spectral resolution because of expensive capture setup and sophisticated acquisition processes. In this paper, we propose a deep neural network structure based on U-Net to convert ordinary RGB images into multispectral images with high spectral resolution. Our variant U-Net neural network structure not only preserves detailed features of RGB images, but also promotes the fusion of different feature scales, enhancing the quality of multispectral image generation. Apart from the training stage, our proposed method does not require low-resolution multispectral images, as do some earlier learning-based methods; multispectral images can be obtained using only the corresponding RGB high-resolution images. We also employ the Inception block to achieve richer image features and the feature loss function to optimize the non-local features. Our proposed algorithm achieves state-of-the-art visual effects and quantitative measurements such as RMSE and rRMSE on several different public datasets.

INDEX TERMS Convolutional neural networks, multispectral image, U-Net.

I. INTRODUCTION

Because multispectral images have more spectral information than RGB images, spectral analysis is an important research method in the natural sciences. It plays an important role in geological research [1], cultural heritage protection [2], [3], and astronomical research [4]. Spectral technology can detect the physical structure and chemical composition of measured objects. It can also be used for qualitative and quantitative analysis of detected objects as well as for positioning analysis. At present, multispectral and hyperspectral techniques play an important role in remote sensing technology [5], agricultural production research [6], geology [7], astronomy [8], and earth science [9]. In computer vision, multispectral or hyperspectral images provide more spectral information references for image classification and detection [10]. Furthermore, research on dimensionality reduction for hyperspectral images has been conducted, which aids better analysis of the spectral information required in applications [11], [12].

The associate editor coordinating the review of this manuscript and approving it for publication was Ahmed A. Zaki Diab.

Light-field imaging devices have been developed for traditional modeling of spectral and polarimetric radiance [13].

The application of multispectral technology in the field of computer science is still in its infancy because multispectral images are much lower in spatial resolution than RGB images and they are expensive and complicated to produce. It is difficult to rapidly obtain low-cost images with high spatial resolution and high spectral resolution.

The research conducted on multispectral image reconstruction is broadly divided into two categories based on the input-image type:

(1) Interpolation or super-resolution with low-spectral-resolution multispectral image with the help of RGB image priors: Low-resolution multispectral images are the primary input, and RGB images are used for optimization [14], [15]. This type aims to increase the resolution of multispectral images to better utilize spectral information. Because it is difficult to obtain available multispectral images, the second type of research, explained next, is more common.

(2) Multispectral image generation using learning-based methods from only RGB images: This type aims to generate high-resolution multispectral images from available RGB

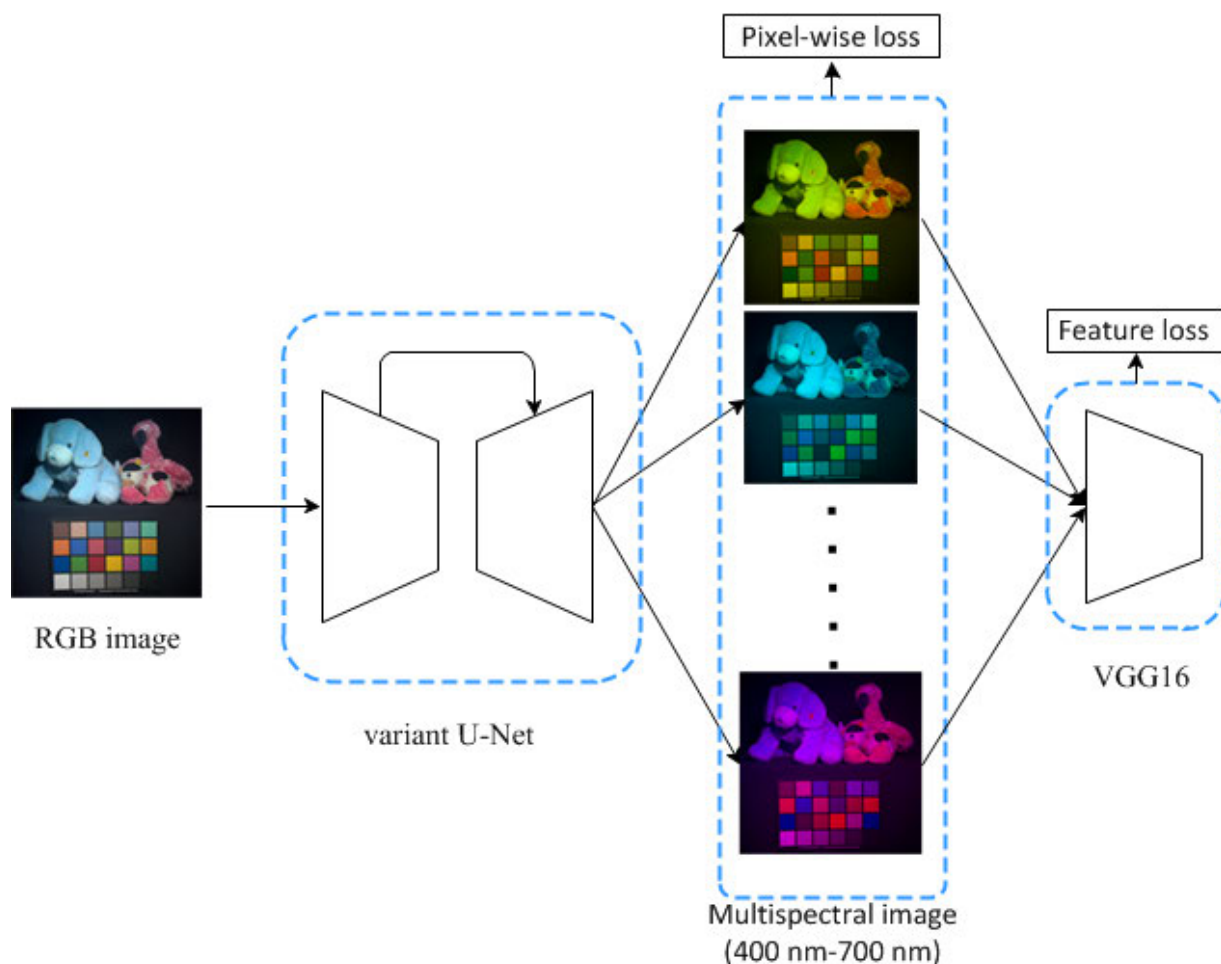


FIGURE 1. The pipeline of our CNN method. The network has two primary parts: the variant U-Net converts RGB images to multispectral images (not depicted explicitly in this figure, but explained completely in the network structure section); the loss function is comprised of pixel-wise loss and feature loss to measure local and non-local similarities between the output image and reference image.

images [16]–[18], and [19]. Our proposed work belongs to the second type of research; RGB images are the only input images required for multispectral image reconstruction. Researchers extracted spectral information from RGB images and recovered the spectra over a wide range. Because of the lack of spectral information from RGB images, learning-based methods are required to train some mapping functions between RGB and multispectral images from existing datasets.

In recent years, deep learning [20] has shown promising results in many visual tasks such as image classification [21], object detection [22], and image segmentation [23]. The convolutional neural network (CNN) can describe the detailed features of images effectively and learn through millions of characteristic parameters. It can also remove the shortcomings of human experience and better restore the geometric information of the image. Furthermore, recently, the use of deep convolutional neural networks to reconstruct multispectral images from RGB images has emerged prominently [14], [22], [24], and [25].

The neural network used in our study is based on the U-Net semantic segmentation network [26]. The first advantage of

using this network is the ability to utilize the features of different dimensions that the U-Net structure can extract, thus depicting low-dimensional features while retaining the original high-dimensional features and improving the quality of multispectral images. The second advantage is that we can avoid expensive multispectral acquisition equipment while still achieving expansion from RGB three-channel images to decade- or hundred-channel spectral images (8 channels in this paper).

Additionally, the original U-Net converts information such as image segmentation and depth estimation to equal or lower dimensions well. Because the generation of multispectral images is the conversion of low-dimensional data to high-dimensional data, it is inefficient to use the existing U-Net structure directly for describing the multi-channel spectral information of multispectral images. To solve this problem, we improve U-Net so that some convolutional layers can better learn the transformation between RGB images and multispectral images. The pipeline used in our method is depicted in Figure 1. The image of the stuffed toys is from the CAVE [27] multispectral dataset. The input is an RGB image and multispectral images (usually 400 nm – 700 nm in

the visible light range) are the output. In the training phase, we use a pixel-wise loss function to compute the distance between the result images and reference images, enhanced by image super-resolution. We also introduce the feature loss function [28] computed by VGG16 to obtain a more detailed performance of the resulting images. The variant U-Net we obtain by modifying the original U-Net is more suitable for the generation of high-channel images.

The next section of this paper explains the related work of generating multispectral image technology from RGB images. The third section introduces our variant U-Net deep neural network structure and optimization method. The fourth section presents our experimental results on different datasets and a comparison with two existing learning-based methods. The fifth section is a summary of this paper.

II. RELATED WORK

We divide the multispectral image reconstruction problem into two parts according to the input-image type. In the first type, a combination of high-resolution RGB images and low-resolution multispectral images is used (i.e., interpolation problem), and in the second type, only high-resolution RGB images are used (i.e., generation problem).

A. USE OF BOTH RGB AND LOW-RESOLUTION MULTISPECTRAL IMAGES

In [29], the researchers proposed the use of the sparse matrix decomposition method to extract features of high-resolution RGB images and low-resolution multispectral images. They used PCA extraction basis functions to reconstruct high-resolution multispectral images, but the calculations required by such methods are lengthy; an image usually takes hours to calculate. Reference [30] simplifies the calculations in matrix decomposition to obtain multispectral video data. References [21] and [31] used linear interpolation to combine high-resolution RGB images with multispectral images to improve the resolution of multispectral images, but they did not fully utilize the corresponding spectral information contained in RGB images. This can cause distortion of the spectrum through rough estimation and inaccurate spectral restoration.

In [14], [15], and [32], low-resolution multispectral images are processed with super-resolution (upsampled). Then a low-resolution to high-resolution corresponding dictionary is constructed by using sparse coding to optimize the upsampling process to solve the problem of spectral distortion. The resulting dictionary reflects the results of a typical mapping between a small number of RGB image blocks and multispectral image blocks. This process is equivalent to extracting a set of basis functions in the mapping relationship. High-resolution multispectral images are recovered by applying a set of basis functions for a linear combination that may experience some estimation and approximation errors. Reference [33] used an autoencoder to generate high-resolution multispectral images, but also encountered the disadvantage

of employing extra low-resolution multispectral images and a detail map of each spectral band.

The aforementioned methods use both low-resolution multispectral images and high-resolution RGB images as input sources to improve the spatial resolution of low-resolution multispectral images. However, current research is shifting its focus toward multispectral image generation from only RGB images (except the training phase in the learning methods) because initially acquiring low-resolution multispectral images is difficult. Our method avoids the inconvenience of using low-resolution multispectral images by using the spectral information contained in the RGB image itself, extracting multi-scale features through deep learning, and recovering information on multiple spectral channels. Thus, we obtain multispectral images with the same high-resolution as RGB images.

B. USE OF ONLY RGB IMAGES

An RGB image is the combination of irradiance information collected from the three broad spectral bands of red, green, and blue. RGB images span a greater bandwidth than multispectral images, and they also store considerable spectral information. In this study, we propose to recover multispectral information using only these three broad bands of information in RGB images. Whether the test data uses only RGB images or both RGB and low-resolution multispectral images when restoring multispectral information, establishing a mapping relationship of RGB/multispectral image pairs is a key issue. More researchers are attempting to avoid the use of multispectral devices and study methods that only use RGB images to recover spectral information; dictionary learning and neural network-based methods are the two main methods.

Mapping from RGB images to multiple spectral images requires pre-training or prior knowledge of spectral statistics. Some of the work extracts the typical color in the RGB/multispectral image pair, or the principal component, from the sample to construct a dictionary. Such a dictionary can have hundreds of atomic data (atoms) and, through linear combinations, can fit the test data, such as [17]–[19], [22], [34], and [35].

Reference [17] reimplemented the work of [22], building dictionaries via K-SVD [36] and orthogonal matching pursuit (OMP) [37], and proposed a shallow learned method to reconstruct hyperspectral images (also called A+ method). The results indicate that the performance was comparable with the deep learning-based method presented in [38]. In [18], the radial basis function network was used to establish a mapping between camera-specific RGB values and scene reflectance spectra from a single RGB image captured by a camera with a known spectral response. Reference [39] proposes a nonlinear mapping between RGB and hyperspectral image pairs using class-based back propagation neural networks, and it delivers state-of-the-art performance.

Dictionaries are essential tools for obtaining mapping relationships, and the construction of dictionaries is an empirical problem. The quantity of atomic data varies from human

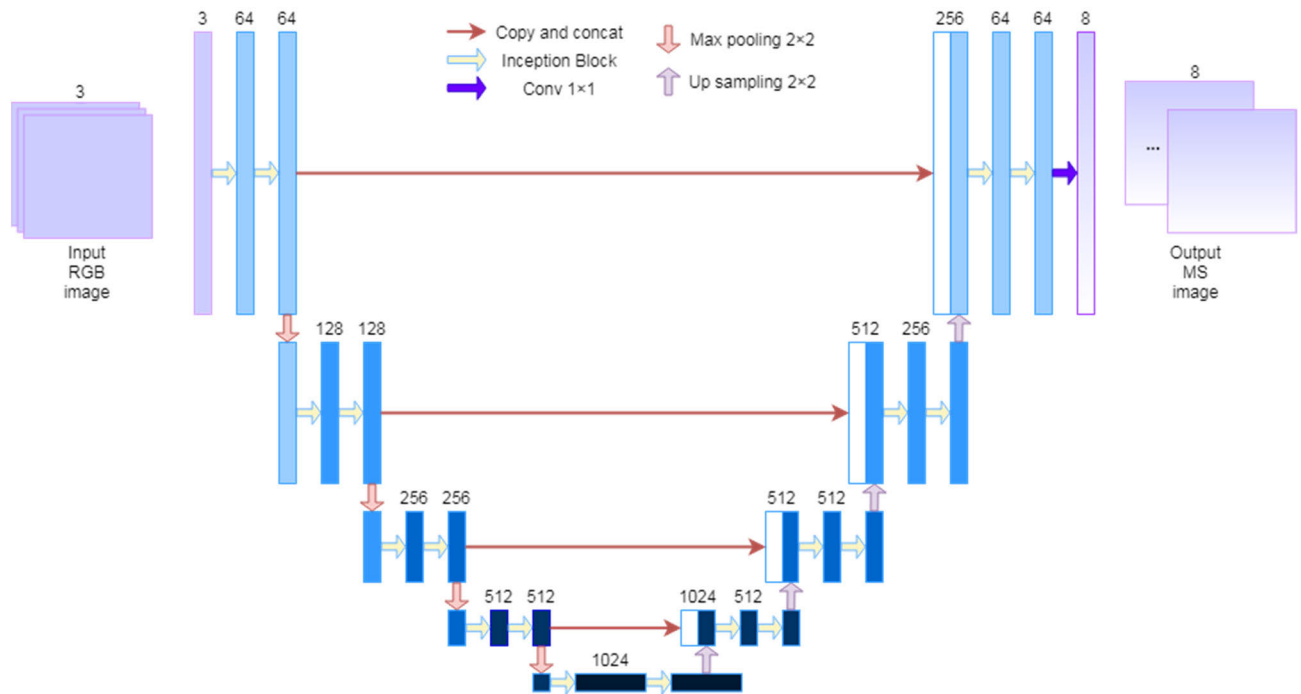


FIGURE 2. Our variant U-net network structure. The network consists of three parts: downsampling, upsampling and feature concatenation. We turn conventional convolution into Inception block (detailed in the next figure) to better produce the generation of high-dimensional images. Each rectangle represents a feature map, the number above it represents the number of feature channels.

experience. For multispectral data with large spectral bands and spectral resolutions, an empirical dictionary may not be complete and applicable.

Recently proposed methods based on deep learning of multispectral image reconstruction, including those presented in [14], [22], [24], [25], [38], [40] and [41], were studied. Reference [38] proposed the use of the Densenet structure [39] to map the end-to-end relationship between RGB images and corresponding multispectral images. The Densenet structure makes the connections between layers more compact, reducing the occurrence of the vanishing gradient problem and over-fitting, but increasing the number of learning parameters and calculations. Reference [41] proposes a moderately deep redundant-network-structure (Residual Block) CNN network to learn the correspondence between RGB and multispectral images [42]. The aim was to fuse features from different dimensions to obtain some accurate restoration with no pooling layer. Thus, high-level and low-level features alias together to a certain degree by ignoring some important features that implicitly impact the training results. Our convolutional layer will only concatenate same feature scales, reducing the computational complexity.

The deep neural network-based training method we propose in this paper obscures the mapping relationship of RGB/multispectral images in the neural network parameters, and thus avoids the limitation of artificial settings (such as the PCA function). This makes the deep U-Net neural network superior to the dictionary learning method. It also offers a more complete description of the mapping relationship

between RGB and multispectral images. Our simpler, more accurate method only requires high-resolution RGB images. Using the U-Net structure, after downsampling or upsampling, same-scale feature maps are connected instead of added, which preserves the independence and richness of each feature map and also avoids the vanishing gradient problem during the gradient descent process. The convergence result is achieved through the variant U-Net architecture.

In the next section, we specifically describe the variant U-net architecture and then discuss some experimental results on visual and quantitative comparisons between the work of [38] and [41] and the classic methods of [17] and [39].

III. VARIANT U-NET DEEP NEURAL NETWORK

We propose a deep neural network as well as loss function to generate more fine-grained multispectral images. We first describe the composition of the proposed variant U-Net neural network structure, and then introduce the loss function used in the training phase, focusing on improvements made.

A. NETWORK STRUCTURE

1) DOWNSAMPLING AND UPSAMPLING

The proposed network structure shown in Figure 2 is based on the original U-Net [26]. In addition to the vanilla convolution in the original U-Net, the encoder structure is superimposed on two Inception blocks [43] (with padding, using an activation function pReLU in each convolution layer) and max-pooling layer alternately, and the decoder is the combination

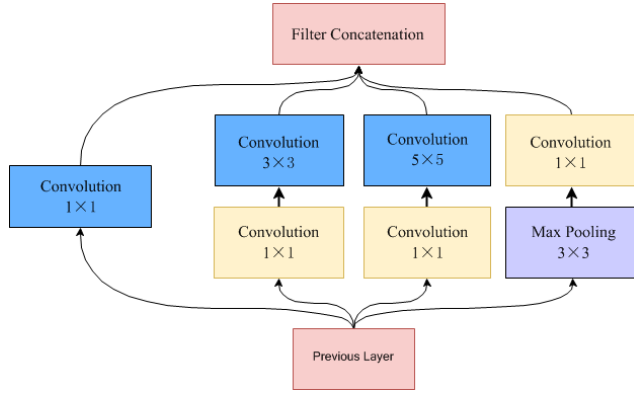


FIGURE 3. Inception block illustration. The output layer concatenates different convolutions with different receptive fields (convolution kernels) to merge features with multiple scales.

of an upsampling layer and Inception blocks. The 1×1 convolutional layer completes the output channel of the entire network (8 channels or 8 bands in this paper).

The Inception block is the concatenation of convolutional layers with different receptive fields, as depicted in Figure 3. The introduction of the Inception block increases the width of the network and improves the generalization ability of the original U-Net. Compared to the original U-Net, more non-linearity is introduced through the 1×1 convolutional layer of the Inception block. The entire network can characterize the image features more powerfully, and the complexity of the network is also controlled within a reasonable range.

The input image is a 3-channel RGB image of 512×512 pixels, and the output image is 8-channel image of the same resolution. The boundary of the image is lost during the convolution process in the original U-Net. However, in this study, we add a padding operation to each convolutional layer so that the size of the output image is unchanged. The result of each upsampling layer is linked to the feature map of the pooled layer of the corresponding scale for the next step (shown by the slender arrow in Figure 2). The 1×1 convolution is used in the last layer to simplify the calculation and reduce the dimensions of the parameters.

2) FEATURE CONCATENATION

After four maxpooling operations, the features of different scales from high dimension to low dimension can be preserved. Such feature maps are connected with the upscaled feature maps of the same scale to enhance the generalization ability of the model by merging local and non-local features. After feature fusion of multiple scales, the number of spectral bands is extended and end-to-end training is achieved.

B. LOSS FUNCTION

Our loss function consists of two parts. We combine the L1 distance function and the feature loss function to define the loss function of the entire network. We encountered difficulty in tuning the Euclidean regularization, which led to over-smoothness. The L1 loss function, as a first-order distance

function, can supervise the pixel response in the space and spectral domains more strongly than the L2 loss function. The feature loss function constrains the integrity of the image in a non-local manner.

1) PIXEL-WISE LOSS

The L1 function calculates the average Manhattan distance between the result feature map and the label image. L1 is a common loss function in training tasks and has a beneficial effect on the overall convergence of the model. The L1 function can be written as follows:

$$l_{L1} = \frac{1}{WHC} \sum_{i,j,k} \|F'(i,j,k) - F(i,j,k)\|, \quad (1)$$

where W , H , and C represent the width, height, and number of channels of the image, respectively; i , j , and k represent the index of the pixels in the image, $F'(i,j,k)$ represents the pixel value in the k -th feature map, and $F(i,j,k)$ represents the corresponding pixel value in the label image. We use the L1 function instead of the L2 function, which calculates the Euclidean distance, to reduce the influence of outliers after root-mean square error computation.

2) FEATURE LOSS

We also introduce feature loss functions to constrain high-level image feature representations [28]. We measure the feature loss function to encourage similar activations of the output image and ground truth image. The feature loss is computed through a VGG16 pre-trained network on the ImageNet dataset [44], [45] after the output image is reconstructed. We used the pre-trained parameters as the initial settings and retrained them in different frequency bands. We describe the feature loss function as follows:

$$l_f = \frac{1}{WHC} \sum \|G(F') - G(F)\|, \quad (2)$$

where $G(F')$ represents the activation of the output image F' through the VGG16 pre-trained network, and the other notations remain the same in as in (1).

3) LOSS FUNCTION

Therefore, the total loss function can be expressed as a linear combination of the L1 and feature loss functions as follows:

$$L = l_{L1} + \alpha l_f, \quad (3)$$

where α is the weight parameter. Such a loss function not only enables the training to converge, but also preserves the details in the image.

IV. EXPERIMENTAL RESULTS

In our experiment, we first introduce the training dataset and evaluation methods used. Then, we display a visual and quantitative comparison of the result image by several methods.

A. IMPLEMENTATION DETAILS

The proposed network is implemented in Tensorflow and an Adam optimizer [46] with Nesterov moment [47], [48]. We trained the network for 2000 epochs with a learning rate of 0.002 for the first 500 epochs and reduced by 0.5 every 500 epochs. The training and testing datasets were split as 80% and 20%. The testing images never enter the training process. PReLU activation functions are applied with a slope of 0.2 in the negative axis. Furthermore, the proposed structure utilizes batch normalization and a 50% dropout to tackle the overfitting problems. To create more training data, the input images were augmented via flips and rotations. We trained and tested the network on the Ubuntu 16.04 platform with Intel Xeon E5-2630 40-core 2.20 GHz CPU and Quadro M6000 GPU with 24 GB RAM.

We compared the CNN methods in the above platform. We implemented and trained the network in [38] with the same parameters as those in their paper for 400 epochs, an initial learning rate of 0.002, and a 50% dropout. Further, we trained the network in [41] with the same Adam optimizer as that in their study, an initial learning rate of 0.0005, and a momentum of 0.93.

B. DATASET AND EVALUATION METHOD

We used two public multispectral datasets, CAVE [27] and ICVL [22], for training, and compared the results qualitatively and quantitatively with two recent studies that used CNN networks for training [38], [41] and other classic methods.

CAVE dataset contained 32 scenes each composed of a 400 nm to 700 nm multispectral image measured in 10 nm increments and its corresponding RGB image. The total number of bands was 31, and images were captured by Apogee Alte U260 cooled CCD camera with a 16-bit PNG format output. The resolutions of the RGB image and its corresponding spectral image are 512×512 . Eight bands were selected on average from the multispectral image for training (i.e., the network output channel number is 8).

The ICVL dataset contains 201 images and corresponding spectral bands from 400 nm to 1000 nm at roughly 1.25 nm increments with 1392×1300 resolution. The images were captured by a Specim PS Kappa DX4 camera. ICVL also provides a downsampled version of 31 bands from 400 nm to 700 nm in 10 nm increments.

To quantify the spectral images, we chose three commonly used quality metrics to measure errors: root mean square error (RMSE), relative root mean square error (rRMSE), and peak signal-to-noise ratio (PSNR). The smaller the RMSE and rRMSE and the greater the PSNR, the better the image quality. The RMSE unit is the difference between two pixels. The rRMSE is a normalized measure describing the pixel similarity of two images. The PSNR unit is dB, describing the difference between the image and noise. Although there are different definitions of RMSE and rRMSE, we use the following formulas, where n is the total number of pixels in the

image, with element-wise subtraction. MAX is the maximum pixel value of an image, that is, 255 (8 bits) in our experiment. Other notations are the same as those in the previous section. We also use the spectral angle mapper (SAM) indicator to measure the performance of each method from the original papers.

$$RMSE = \sqrt{\frac{1}{n} \sum (F' - F)^2} \quad (4)$$

$$rRMSE = \sqrt{\frac{1}{n} \sum \frac{(F' - F)^2}{F^2}} \quad (5)$$

$$PSNR = 10 \log_{10} \left(\frac{MAX^2}{RMSE^2} \right) \quad (6)$$

C. RESULTS

1) CAVE DATASET

The number of images in the training set was 27, and the number in the test set was 5. We flipped and rotated the training images four times to quadruple the number of images. Figure 4 shows the difference in visual effects between our proposed method and the comparison methods. The first column is the ground truth (GT) image, the next three columns are the result images of different methods, and the last three columns are the respective error maps between the result and GT image of the various methods. We take into account three scenes (stuffed toys, thread spools, and sponges) in three rows and compare the output spectral images (440 nm) and the error maps of our method and the two other learning-based methods. We see better visual effects in the error maps as smooth pixels increase and outliers decrease. We display the red region at the bottom-right corner of every image to highlight the amplification effect. Using the error map, we can see that our method results in pixel values around zero, which means our output is closer to the ground truth image. Can's method is smooth but, due to the lack of efficiency in its use of conventional convolution blocks, it results in higher error map values. Our experimental results indicate that at some short wavelengths (near 400 nm), some image edge positions (specifically, dark areas) are significantly erroneous. Although there may be unsuitable places, our strategy is to arrange the network structure and optimization methods (including hyperparameters such as learning rate and batch size) such that the overall loss of the network is minimized.

In Table 1, we use some image quality metrics to evaluate the average difference between the 8-bit result image and the ground truth test image. Our method exhibits better behavior in terms of three quantitative metrics, except SAM. Galliani's and A+ method obtain more artifacts, resulting in a high RMSE and rRMSE. Can's method obtains a good visual effect but a larger variance than ours in pixel values. Han's method [39] (CBPNN) and our method achieved similar results in terms of RMSE and rRMSE, but RMSE is reduced 0.21 and rRMSE is reduced 0.016 by our method. However, SAM is reduced 1.8 in Han's method, and we took the second place. We achieved the best result in PSNR than the other methods.

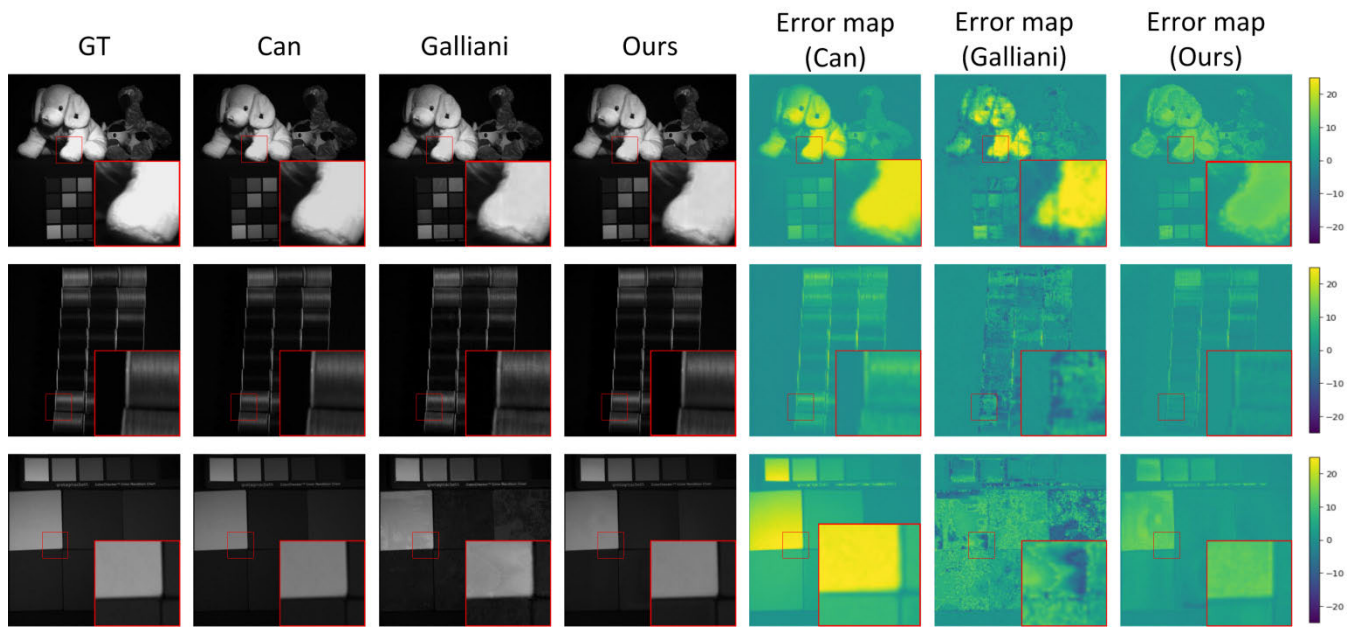


FIGURE 4. Visual comparison of result images and error maps of Can [41], Galliani [38] and our method in 440 nm w.r.t 8-bit image. Three different scenes are displayed in three rows. Ground truth is in the first column, then the spectral result image of Can's, Galliani's and our method, then the error maps of three methods, respectively. The red region is amplified in the bottom right corner.

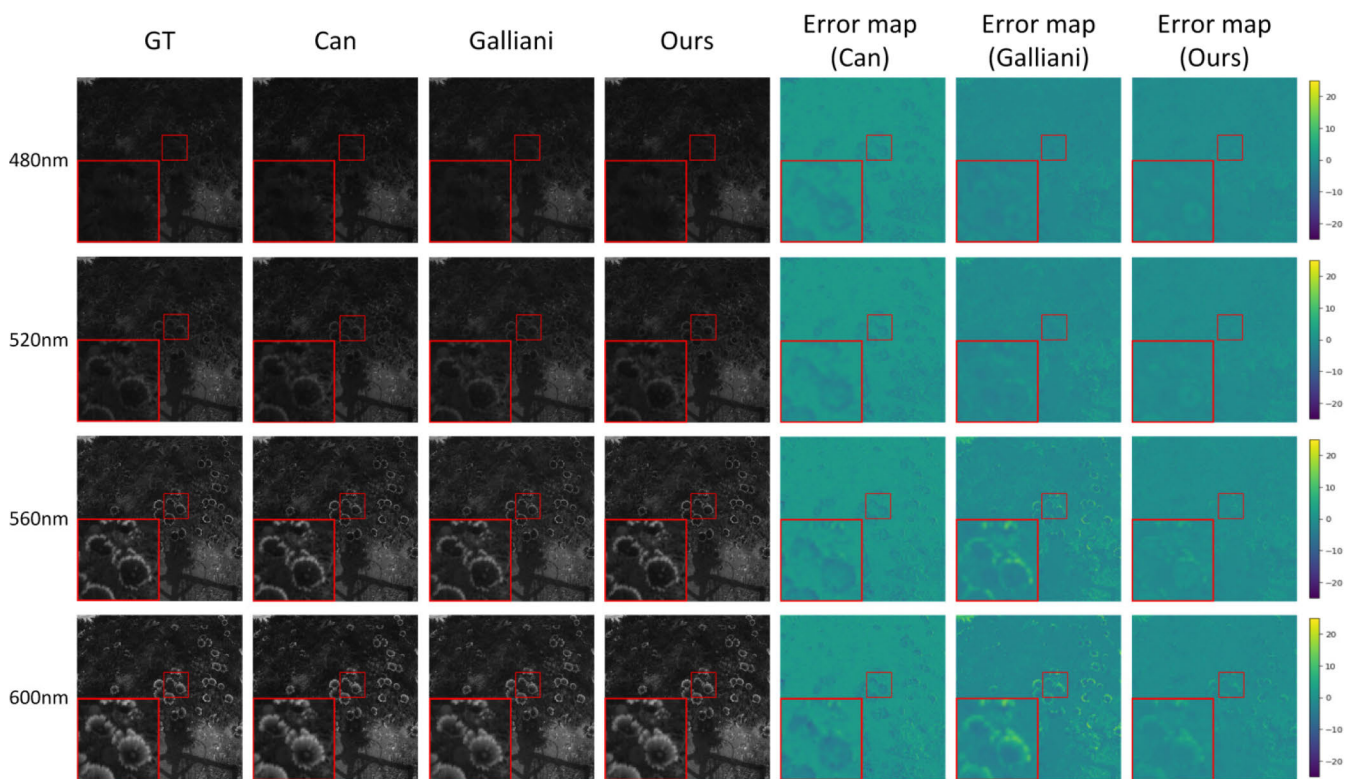


FIGURE 5. Visual comparison of result images and error maps of Can [41], Galliani [38] and our method in four non-consecutive spectral bands w.r.t 8-bit image. We display four different spectral bands (480 nm, 520 nm, 560 nm, 600 nm) in four rows. Ground truth is in the first column, then the spectral result images of Can's, Galliani's and our method, then the error maps of three methods, respectively. The red region is amplified in the bottom left corner.

2) ICVL DATASET

We resize the training and testing images into 1024×1024 resolution for more convenient convolution. We take out five images for the testing procedure and ignore data

augmentation preprocessing. Notably, flipping and rotating the training images like we did to the CAVE dataset did not improve the study results, but significantly increased the training time. Our proposed network's output channel is 8,

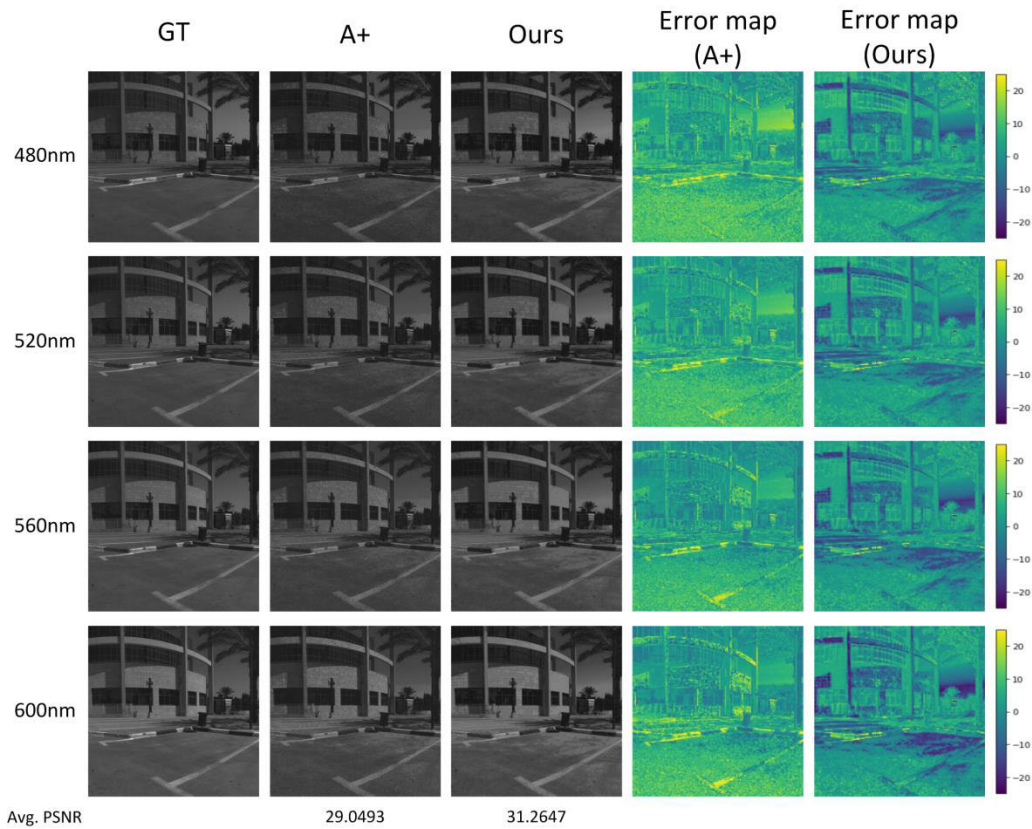


FIGURE 6. Visual comparison of resulting images and error maps of A+ [17] and our method in four non-consecutive spectral bands with respect to 8-bit image. We display four different spectral bands (480, 520, 560, and 600 nm) in four rows. Ground truth is presented in the first column; the spectral result images of A+ and our method are presented in the second and third columns; the error maps of three methods are presented in fourth and fifth columns. The last row presents the average PSNRs of the spectral images.

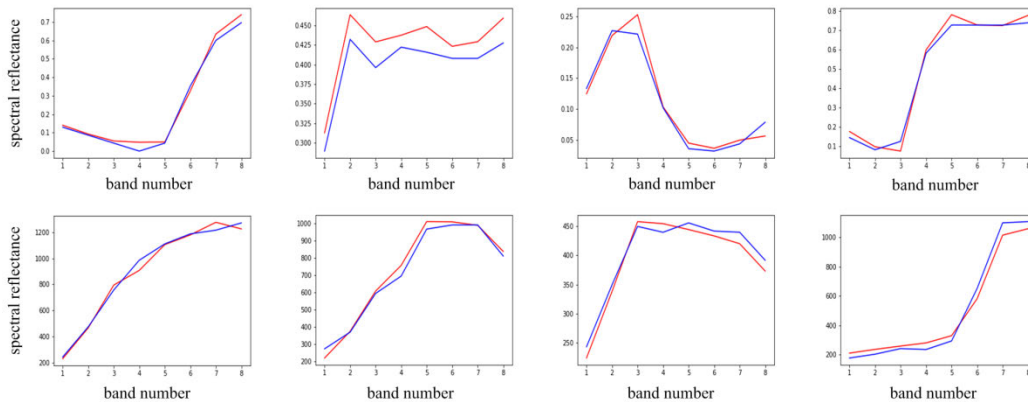


FIGURE 7. Spectral reflectance of the ground truth (red) and the reconstructed results with our method (blue) on some typical pixels of the CAVE (the first row) and ICVL (the second row) datasets.

as mentioned above, but we may modify it to the number of bands if needed as well. Figure 5 demonstrates the difference in visual effects between the comparison methods and ours. We take into account four different spectral bands (480 nm, 520 nm, 560 nm, and 600 nm) in four rows in the same scene to see how effective the network is in dealing with different spectra. We also amplified the red region to see the details in the spectral images and error maps. From the error map, we can see a better effect on the edge of the flowers

in Can's and our method, while in Galliani's method there is slightly lower clarity in the same location. Figure 6 presents a visual comparison of A+ [17] and our method; from the reconstructed images, it is clear that the error ranges are almost equivalent, but our average PSNR is slightly higher because of smoother textures.

In Table 2, we compute the image quality metrics to compare the average difference between the result image and the ground truth image in 8-bit. Our method exhibits better

TABLE 1. Error measurement in CAVE dataset on 8-bit image.

	RMSE	rRMSE	SAM	PSNR
Can	4.6158	0.1551	-	35.2630
Galliani	6.0155	0.1979	12.10	33.1612
A+	6.70	0.3034	-	-
CBPNN	3.8982	0.1396	7.3467	-
Ours	3.6828	0.1239	9.1536	37.2261

TABLE 2. Error measurement in ICVL dataset on 8-bit image.

	RMSE	rRMSE	SAM	PSNR
Can	2.0014	0.0545	-	42.1293
Galliani	2.0766	0.0558	2.04	41.8784
A+	1.8233	0.0599	-	-
CBPNN	1.5433	0.0410	1.1686	-
Ours	1.4579	0.0399	1.5526	44.8638

results in both visual and quantitative comparisons, except in terms of SAM where only inferior to the CBPNN method. We improved the RMSE metric and the rRMSE metric by 5% and the PSNR metric by 6%. Learning-based methods can extract features from different spectral bands well if they can handle multiple dimensions and scales of spectral images with the proper model and dataset. Furthermore, results indicate that our method can handle spectral image generation with different spectral bands.

Figure 7 presents the reconstructed spectral reflectance of our method and ground truth in both CAVE (first row) and ICVL (second row) datasets in some typical pixels. Our spectra fit the ground truth well in the eight band scale. Some of the curves, such as CAVE data in the first row, fit less well with the ground truth possibly owing to the insufficient generalization of training data.

D. FURTHER DISCUSSION

Since deep learning was introduced to generate multispectral and hyperspectral images, we can obtain large resolution spectral images more conveniently. However, the effectiveness of deep learning is highly related to the volume of the training dataset. Our experiment shows that we can see better overall visual effects on ICVL results than CAVE results because the ICVL dataset contains more training images and complex scenes for the extraction of image features. We currently focus on the reconstruction of multispectral images, but we can also expand the bandwidth of hyperspectral scale with some fine tuning and optimization schemes. Additional studies should focus on the combination of optical rules and deep learning abilities.

V. CONCLUSION AND FUTURE WORK

This paper proposes a variant U-Net deep learning neural network structure to convert an ordinary RGB image into multispectral images with high spectral resolution. There are two main contributions of our method: (1) with the introduction of Inception block, more low-level features are

retained along with high-level features, making the spectral information recovery of the image more accurate; and (2) we combine pixel-wise loss and feature loss functions to better measure local and non-local features. We improve the original U-Net and eliminate the inconvenience and expense of using multispectral imaging equipment. Such a network structure can also be extended to recover multispectral images of an arbitrary number of output bands, such as 31 bands (400 nm to 700 nm, bandwidth of 10 nm), and we only need to adjust the number of network output channels to fine-tune it. We have demonstrated that our variant U-Net deep learning method has achieved good results in both visual effects and quantitative analysis, and provides an effective method for the generation of multispectral images.

ACKNOWLEDGMENT

The authors would like to thank L. Zhao who provided valuable suggestions with this work.

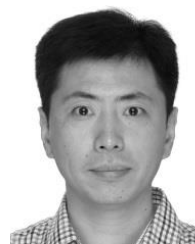
REFERENCES

- [1] R. Brigante, C. Cencetti, P. De Rosa, A. Fredduzzi, F. Radicioni, and A. Stoppini, "Use of aerial multispectral images for spatial analysis of flooded riverbed-alluvial plain systems: The case study of the paglia river (central Italy)," *Geomatics, Natural Hazards Risk*, vol. 8, no. 2, pp. 1126–1143, Apr. 2017.
- [2] M. Hain, J. Bartl, and V. Jacko, "Multispectral analysis of cultural heritage artefacts," *Meas. Sci. Rev.*, vol. 3, no. 3, pp. 9–12, Jan. 2003.
- [3] C. Simon Chane, A. Mansouri, F. S. Marzani, and F. Boochs, "Integration of 3D and multispectral data for cultural heritage applications: Survey and perspectives," *Image Vis. Comput.*, vol. 31, no. 1, pp. 91–102, Jan. 2013.
- [4] D. Nuzillard and A. Bijaoui, "Blind source separation and analysis of multispectral astronomical images," *Astron. Astrophys. Suppl. Ser.*, vol. 147, no. 1, pp. 129–138, Nov. 2000.
- [5] T. M. Lillesand, R. W. Kiefer, and J. W. Chipman, "Remote sensing and image interpretation (fifth edition)," *Geogr. J.*, vol. 146, no. 3, pp. 586–592, Jan. 2004.
- [6] D. Haboudane, "Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture," *Remote Sens. Environ.*, vol. 90, no. 3, pp. 337–352, Apr. 2004.
- [7] E. A. Cloutis, "Review article hyperspectral geological remote sensing: Evaluation of analytical techniques," *Int. J. Remote Sens.*, vol. 17, no. 12, pp. 2215–2242, Aug. 1996.
- [8] K. Hege, D. O'Connell, W. R. Johnson, S. Basti, and E. L. Dereniak, "Hyperspectral imaging for astronomy and space surveillance," *Imag. Spectrometry*, vol. 9, no. 5159, pp. 380–391, Jan. 2004, doi: [10.1117/1.2.506426](https://doi.org/10.1117/1.2.506426).
- [9] J. Mustard and J. M. Sunshine, "Spectral analysis for earth science: Investigations using remote sensing data," *Remote Sens. Earth Sci.*, vol. 3, pp. 251–306, Jan. 1999.
- [10] M. D. Mura, A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone, "Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 542–546, May 2011.
- [11] H. Yang, Q. Du, H. Su, and Y. Sheng, "An efficient method for supervised hyperspectral band selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 138–142, Jan. 2011.
- [12] 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.
- [13] L. Su, Y. Liu, and Y. Yuan, "Spectrum reconstruction of the light-field multimodal imager," *IEEE Access*, vol. 7, pp. 9688–9696, Jan. 2019, doi: [10.1109/ACCESS.2018.2890458](https://doi.org/10.1109/ACCESS.2018.2890458).
- [14] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang, "Hyperspectral image super-resolution with optimized RGB guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11653–11662.

- [15] N. Akhtar, F. Shafait, and A. Mian, "Sparse spatio-spectral representation for hyperspectral image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2014, pp. 63–78.
- [16] X.-H. Han, B. Shi, and Y. Zheng, "Residual HSRCNN: Residual hyper-spectral reconstruction CNN from an RGB image," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2664–2669.
- [17] 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.
- [18] R. M. Nguyen, D. K. Prasad, and M. S. Brown, "Training-based spectral reconstruction from a single rgb image," in *Eur. Conf. Comput. Vis.*, pp. 186–201. Springer, 2014.
- [19] A. Robles-Kelly, "Single image spectral reconstruction for multimedia applications," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 251–260.
- [20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015, doi: [10.1038/nature14539](https://doi.org/10.1038/nature14539).
- [21] F. H. Imai and R. S. Berns, "High-resolution multispectral image archives—A hybrid approach," in *Proc. 6th Color Imag. Conf.*, Jan. 2001, pp. 224–227.
- [22] B. Arad and O. B. Shahar, "Sparse recovery of hyperspectral signal from natural RGB images," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 19–34.
- [23] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
- [24] Y. Yan, L. Zhang, W. Wei, and Y. Zhang, "Accurate spectral super-resolution from single RGB image using multi-scale CNN," in *Proc. Chin. Conf. Pattern Recog. Comput. Vis.*, Nov. 2018, pp. 206–217.
- [25] A. Rangnekar, N. Mokashi, E. Ientilucci, C. Kanan, and M. Hoffman, "Aerial spectral super-resolution using conditional adversarial networks," 2017, *arXiv:1712.08690*. [Online]. Available: <http://arxiv.org/abs/1712.08690>
- [26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent*, Nov. 2015, pp. 234–241.
- [27] 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.
- [28] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 694–711.
- [29] R. Kawakami, Y. Matsushita, J. Wright, M. Ben-Ezra, Y.-W. Tai, and K. Ikeuchi, "High-resolution hyperspectral imaging via matrix factorization," in *Proc. CVPR*, Jun. 2011, pp. 2329–2336.
- [30] X. Cao, X. Tong, Q. Dai, and S. Lin, "High resolution multispectral video capture with a hybrid camera system," in *Proc. CVPR*, Jun. 2011, pp. 297–304.
- [31] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of MS + Pan data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Oct. 2007.
- [32] H. Kwon and Y.-W. Tai, "RGB-guided hyperspectral image upsampling," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 307–315.
- [33] A. Azarang, H. E. Manoochehri, and N. Khehtarnavaz, "Convolutional autoencoder-based multispectral image fusion," *IEEE Access*, vol. 7, pp. 35673–35683, 2019, doi: [10.1109/ACCESS.2019.2905511](https://doi.org/10.1109/ACCESS.2019.2905511).
- [34] V. Heikkinen, R. Lenz, T. Jetsu, J. Parkkinen, M. Hauta-Kasari, and T. Jääskeläinen, "Evaluation and unification of some methods for estimating reflectance spectra from RGB images," *J. Opt. Soc. Amer. A, Opt. Image Sci., Vis.*, vol. 25, no. 10, pp. 2444–2458, Oct. 2008.
- [35] F. Ayala, J. F. Echavarri, 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, Aug. 2006.
- [36] M. Aharon, M. Elad, and A. Bruckstein, "rmK-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
- [37] Y. C. Pati, R. Rezaifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signals, Syst. Comput.*, Nov. 1993, pp. 40–44, doi: [10.1109/ACSSC.1993.342465](https://doi.org/10.1109/ACSSC.1993.342465).
- [38] S. Galliani, C. Lanaras, D. Marmanis, E. Baltsavias, and K. Schindler, "Learned spectral super-resolution," 2017, *arXiv:1703.09470*. [Online]. Available: <http://arxiv.org/abs/1703.09470>
- [39] X. Han, J. Yu, J.-H. Xue, and W. Sun, "Spectral super-resolution for RGB images using class-based BP neural networks," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Dec. 2018, pp. 721–727.
- [40] 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.
- [41] Y. B. Can and R. Timofte, "An efficient CNN for spectral reconstruction from RGB images," 2018, *arXiv:1804.04647*. [Online]. Available: <http://arxiv.org/abs/1804.04647>
- [42] 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.
- [43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
- [44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Apr. 2015.
- [46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [47] T. Dozat, "Incorporating nesterov momentum into adam," Stanford Univ., Stanford, CA, USA, Tech. Rep., Feb. 2016. [Online]. Available: http://cs229.stanford.edu/proj2015/054_report.pdf
- [48] I. Sutskever, J. Martens, G. E. Dahl, and G. E. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. ICML*, vol. 28, Apr. 2013, pp. 1139–1147.

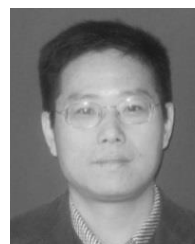


TAO ZENG received the B.S. degree from the University of Electronic and Technology of China, Chengdu, China, in 2010. He is currently pursuing the Ph.D. degree with Zhejiang University, Hangzhou, China. His current research interests include image deblurring and spectral imaging.



CHANGYU DIAO received the B.S., M.S., and Ph.D. degrees in computer science and technology from Zhejiang University, Hangzhou, China, in 2000, 2003, and 2008, respectively.

He is currently an Associate Professor with the Institute of Cultural Heritage, Zhejiang University. His current research interests include 3D modeling, image processing, and virtual reality.



DONGMING LU received the B.S., M.S., and Ph.D. degrees in computer science and technology from Zhejiang University, Hangzhou, China, in 1989, 1991, and 1994, respectively.

He is currently a Professor with the College of Computer Science and Technology, Zhejiang University. His current research interests include virtual reality, image processing, and sensor networks.

• • •