Research Article | Vol. 27, No. 26 / 23 December 2019 / Optics Express 38312
Optics EXPRESS

Spectral-depth imaging with deep learning based reconstruction

MINGDE YAO,[1] ZHIWEI XIONG,[1,*] LIZHI WANG,[2] DONG LIU,[1] AND XUEJIN CHEN[1]

[1] University of Science and Technology of China, Hefei 230027, China
[2] Beijing Institute of Technology, Beijing 100081, China
[*] zwxiong@ustc.edu.cn

Abstract: We develop a compact imaging system to enable simultaneous acquisition of the spectral and depth information in real time. Our system consists of a spectral camera with low spatial resolution and an RGB camera with high spatial resolution, which captures two measurements from two different views of the same scene at the same time. Relying on an elaborate computational reconstruction algorithm with deep learning, our system can eventually obtain a spectral cube with a spatial resolution of 1920×1080 and a total of 16 spectral bands in the visible light section, as well as the corresponding depth map with the same spatial resolution. Quantitative and qualitative results on benchmark datasets and real-world scenes show that our reconstruction results are accurate and reliable. To the best of our knowledge, this is the first attempt to capture 5D information (3D space + 1D spectrum + 1D time) with a miniaturized apparatus and without active illumination.

1. Introduction

For decades, high dimensional imaging has attracted wide attention from both academia and industry communities, among which spectrum and depth constitute two essential dimensions. By capturing dozens of images in different electromagnetic bands, spectrum provides refined information about texture and reflectance of an object, while depth describes the geometric appearance of an object. Together they offer a nearly complete description of the target scene, which facilitates object rendering with any given illumination and from any given perspective. However, while stereo cameras are getting popular on smartphones, the acquisition of real-time, high-resolution spectral images generally requires highly customized hardware. It can be predicted that tremendous new applications would be opened up if fast and accurate spectral-depth imaging can be realized with portable imagers that can be hand-held or easily integrated into consumer electronics.

A few recent efforts have shown 3D imagers can be integrated with spectral imagers for capturing high dimensional characteristics of the target scene, e.g., 3D imaging spectroscopy [1], cross-modal stereo [2], and 5D hyperspectral imaging [3]. These pioneer works explored the joint imaging of spectrum and depth and made encouraging progresses in this field. However, directly combining one 3D imager and one spectral imager as in previous works is brute-force, which either involves high-complexity spectral imager (e.g., CASSI) [1,2] or active illumination based depth imager [1,3]. The resulting systems are thus not convenient to carry on or sensitive to ambient light, both prohibiting the outdoor usage. Moreover, existing spectral-depth imagers suffer from a limited resolution either in spatial or temporal dimension.

In this paper, we propose a novel compact system for spectral-depth imaging in real time and with high resolution. Our system consists of an off-the-shelf spectral camera with low spatial resolution (LR) and an RGB camera with high spatial resolution (HR), which captures two measurements from two different views of the same scene at the same time, as shown in Fig. 1. We then propose a novel two-stage scheme for jointly reconstructing a high-resolution spectral

#377857 https://doi.org/10.1364/OE.27.038312
Journal © 2019 Received 16 Sep 2019; revised 1 Nov 2019; accepted 2 Nov 2019; published 17 Dec 2019

Optics EXPRESS

cube along with a depth map from the above one-shot measurements. The first stage is conducted as follows. The obtained LR spectral measurement is super-resolved to the same spatial resolution as the RGB measurement, and the output is called SR-spectral cube. The SR-spectral cube is then synthesized into an RGB image to estimate a disparity map with the cooperation of the HR RGB measurement. Before disparity estimation, a color transfer operation is performed here to eliminate the color inconsistency between the synthesized RGB and the HR RGB measurement. Once the disparity is obtained, the second stage is to warp the RGB measurement to the view of the SR-spectral cube, i.e., they are aligned pixel to pixel. With the guidance of the aligned RGB information, the SR-spectral cube can be further enhanced on texture details. Since deep learning has demonstrated the strong capability of modeling the highly non-linear mapping in related problems and achieved promising performance as well as fast inference speed during test time, we implement the above super-resolution, disparity estimation, and texture enhancement procedures through deep neural networks tailored for these tasks. External spectral and stereo image datasets are used for training the networks [4,5]. Evaluation results on both simulated and real-world scenes prove the effectiveness of our proposed deep learning based methods.

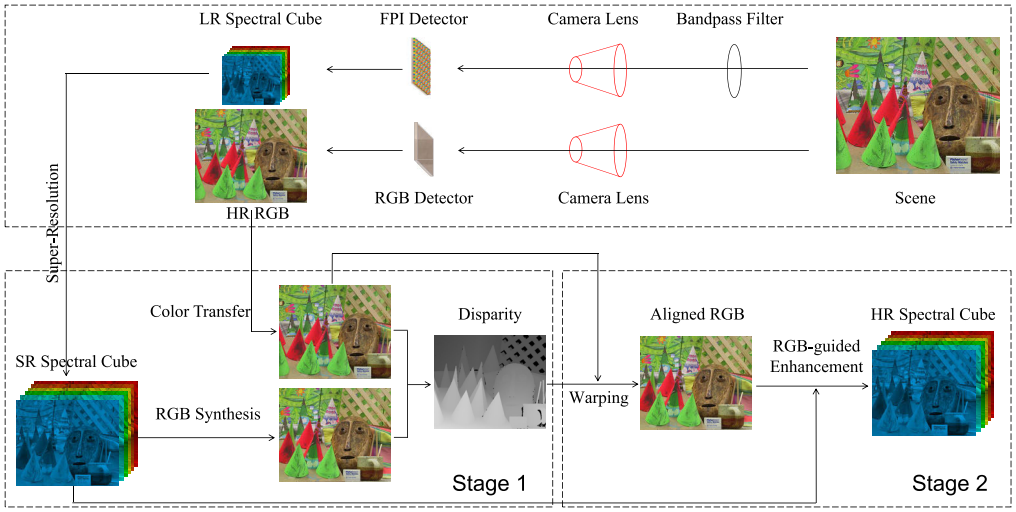


Fig. 1. Overview of the proposed system and reconstruction algorithm.

Relying on the efficient computational reconstruction algorithm with deep learning instead of customized hardware, our proposed system can eventually obtain spectral images with a temporal resolution up to 50*fps*, a spatial resolution of 1920×1080, and a total of 16 spectral bands covering the wavelength range of $470 - 630nm$, as well as the corresponding depth map of the target scene with the same spatial resolution. The above resolutions are either competitive or higher than existing systems, while neither high-complexity hardware nor active illumination is required. Considering its compactness and easy manipulation, our proposed system offers a practical solution for ubiquitous spectral-depth imaging in the wild, e.g., on smartphones or UAVs.

2. Related work

Due to the hardware restriction, it is difficult to directly capture spectral images with high spatial and temporal resolution [6,7]. Scanning spectrometers sacrifice the temporal resolution and cannot measure dynamic scenes. Snapshot spectral imagers generally multiplex the sensor pixels for a number of spectral bands and suffer from low spatial resolution. As representative

Optics EXPRESS

computational systems, CASSI [8] and PMIS [9], especially their updated dual-camera versions [10–13] enable fast and accurate spectral reconstruction. However, the hardware involved is of high complexity and not easy for manipulation. Another mainstream approach is called pansharpening, which can be regarded as spectral image super-resolution guided by higher resolution panchromatic/RGB images of the same scene [14]. Recently, relying on the power of deep learning and external hyperspectral image datasets, spectral image super-resolution has seen a notable improvement in performance [15].

By combining a 3D imager and a spectral imager together, geometry and spectrum of the target scene can be captured simultaneously. Kim et al. first developed such a system by integrating a laser scanner and the highly customized CASSI to capture high-resolution images for modeling the appearance of birds, named 3D imaging spectroscopy [1]. While huge efforts are dedicated to the system calibration and characterization, it takes hours for capturing a single scene, which prohibits usage of this system in dynamic conditions. To get higher temporal resolution, Wang et al. proposed a cross-modal stereo system [2] that keeps the CASSI for spectral imaging and replaces the laser scanner with an ordinary grayscale camera. An iterative scheme is then proposed to reconstruct the depth under the stereo configuration and improves the spectral reconstruction from CASSI. On the other hand, Heist et al. utilized a structured light based depth camera and two snapshot spectral imagers to realize joint spectral and depth imaging in real time [3]. Specifically. the two spectral imagers can accommodate different wavelength ranges with different band-pass filters. Despite of the encouraging results as achieved by the above systems, the high-complexity hardware (e.g., CASSI) for spectral imaging or active illumination (e.g., structured light) for depth imaging greatly restricts their application scope especially for outdoor usage. Relying on computational reconstruction algorithms with deep learning, the compact imaging system developed in this paper not only relieves these restrictions but also achieves the best performance when jointly considering the three dimensions, as summarized in Table 1.

Table 1. Comparison of representative spectral imagers.

Imager	Resolution			Depth
	spectrum	pixel	framerate	
3DIS [1]	~12*nm*	4.0*Mega*	0.5*hours*	✓
Stereo-CASSI [2]	~10*nm*	0.4*Mega*	15*fps*	✓
5Dhyperspectral [3]	~10*nm*	0.1*Mega*	17*fps*	✓
Dual-CASSI [10]	~10*nm*	0.4*Mega*	100*fps*	×
Dual-PMIS [12]	~5*nm*	0.8*Mega*	15*fps*	×
Ours	~10*nm*	2.1*Mega*	50*fps*	✓

3. Proposed algorithm

As shown in Fig. 1, our system consists of two branches, and the reflected light from the scene is captured by both branches at the same time. The spectral camera captures an LR spectral cube Y_{LR} and the RGB camera captures an HR RGB image Y_{RGB}. From these two measurements, we finally reconstruct the HR spectral cube Y_{HR} along with the HR depth map D of the target scene. Each step of our proposed reconstruction algorithm is described in detail below.

3.1. Spectral image super-resolution

We first adopt an end-to-end convolutional neural network, named SISR-Net, to enhance the acquired spectral measurement in the spatial dimension. As shown in Fig. 2, SISR-Net takes a spatially interpolated spectral cube as input and directly outputs a spectral cube with the same spatial resolution as the RGB measurement. Compared with conventional RGB images,

Optics EXPRESS

spectral images have more channels and larger dynamic range, which increase the learning costs of the network. Therefore, we employed a residual network by only learning the difference between the spatially interpolated spectral data and the ground-truth HR spectral data, where the mean-square-error is adopted as the loss function.

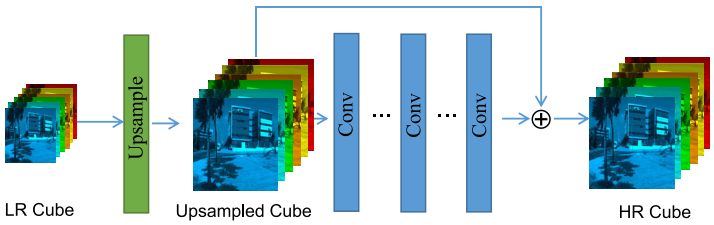


Fig. 2. The SISR-Net architecture.

The SISR-Net is trained on a large public dataset named ICVL [4] with spectral cubes at an original resolution of $1392 \times 1296 \times 519$. To accommodate the spectral resolution (i.e., 16 bands) of the spectral camera in our system, we first synthesize spectral images from original images with 519 bands by employing the spectral response function of our own spectral camera, as shown in Fig. 3 (left part). These spectral images are then spatially down-sampled to generate the LR-HR spectral image pairs for training the SISR-Net. Afterwards, we obtain an SR-spectral cube Y_{sr} from Y_{LR} with SISR-Net, which has the same spatial resolution as Y_{RGB}. We can then calculate the depth map from Y_{sr} and Y_{RGB} under a cross-modal stereo configuration.

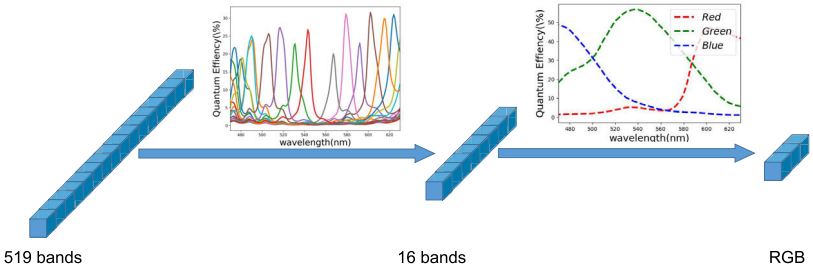


Fig. 3. Image synthesis. Left part: from 519 bands to 16 bands using the spectral response function of our spectral camera. Right part: from 16 bands to RGB using the spectral response function of our RGB camera.

It is worth mentioning that the ICVL dataset provides two types of spectral images, one is 519-band raw spectral image and the other is 31-band synthesized spectral image. The 31-band one is synthesized from the 519-band one by averaging adjacent bands instead of using any specific spectral response function, which is mostly used in related research works. Here we synthesize 16-band spectral images with the spectral response function of our own spectral imager for training the correspondingly networks. This elaborate synthesis process guarantees the generalization ability of the trained networks on real captured spectral data, which is justified in the experiments (see Section 5.1.2).

3.2. Depth estimation

Depth provides the geometric characteristics of the target scene, which can be derived from the disparity map according to $D = \frac{b \cdot f}{d}$, where b denotes the baseline of the binocular system, f denotes the focal length of the camera lens, and d is the disparity calculated from the stereo matching algorithm.

Optics EXPRESS

Before stereo matching, we should first deal with the difference of spectral resolution between Y_{sr} and Y_{RGB}. Here we conduct a color synthesis operation to get the stereo matching pair. Since the spectral response function of our own RGB camera is known, the SR-spectral cube Y_{sr} can be synthesized to an RGB image Y_{syn}, as shown in Fig. 3 (right part). However, due to the different built-in camera configurations of the spectral and RGB branches in practice, there is still a color discrepancy between Y_{syn} and Y_{RGB}, as can be seen from Fig. 4(a) and 4(b). To address this issue, we then conduct a color transfer procedure by employing the NDFlow algorithm [16] to transfer the color style of Y_{RGB} to that of Y_{syn}.

(a) (b) (c)

Fig. 4. Color transfer. (a) Image captured by the RGB branch (source color). (b) Synthetic RGB image from the spectral branch (target color). (c) NDFlow is applied to transfer source color to target color.

Specifically, NDFlow is a nonlinear intensity normalization scheme based on density matching, where the histograms are modeled as Gaussian mixtures. By minimizing the divergence between source and target mixture models, NDFlow ensures the intensity distribution of source agrees with that of target. The color transferred image, as shown in Fig. 4(c), is named Y_{tr}. It is worth mentioning that, while transferring the color style of Y_{syn} to that of Y_{RGB} is also feasible, we choose the current way to facilitate the following RGB-guided texture enhancement. The reason will be explained in Section 3.3.

So far, a synthetic RGB image Y_{syn} deriving from the spectral camera branch and its paired image Y_{tr} deriving from the RGB camera branch are obtained. Assuming that the two cameras have been calibrated in advance, the disparity will only exist in the horizontal direction of the image. We then adopt the MC-CNN algorithm [17] for stereo matching due to its robust performance among deep learning based methods, which involves two steps: matching cost calculation and post-processing. A convolution neural network is employed to get the matching cost between the two images, and the network is trained on a large public stereo image dataset [5]. The post-processing step includes semi-global matching, left-right consistency check, sub-pixel enhancement, median filtering, and bilateral filtering. Afterwards, we obtain the final disparity map and the depth map D can be derived as mentioned above.

3.3. RGB-guided texture enhancement

The SR-spectral cube Y_{sr} is obtained through learning from external spectral images. Although it is at the target spatial resolution, the HR texture details in Y_{sr} may still be missing, especially for scenes drastically different from the training data. The HR image captured by the RGB camera can provide supplementary spatial information of the target scene itself, relying on which we then conduct RGB-guided texture enhancement, the most important step in our reconstruction algorithm. Different from traditional pansharpening methods, this step is realized in a deep learning way with registered spectral and RGB image pairs. The registration includes two aspects: spatial alignment and spectral alignment.

After obtaining the disparity between Y_{syn} and Y_{tr}, we can use the disparity map to warp Y_{tr} to the perspective of the spectral camera, resulting in a warped RGB image Y_{wp}. However, due to

Optics EXPRESS

the inherent occlusion effect under the stereo configuration, there will be missing pixel values in certain areas around abrupt depth changes. Here we employ the image inpainting method in [18] to repair the holes in Y_{wp} caused by occlusion. After warping and inpainting, the RGB image Y_{wp} and the SR-spectral cube Y_{sr} can be considered well aligned in the spatial dimension.

Algorithm 1 Spectral-depth reconstruction

Input: LR spectral image Y_{LR} and HR RGB image Y_{RGB}

Stage 1:
super-resolve spectral image Y_{LR} on spatial dimension, get Y_{sr};
synthesize spectral image Y_{sr} to RGB, get Y_{syn};
color transfer from Y_{RGB} to Y_{tr};
stereo matching between Y_{syn} and Y_{tr}, get disparity map D.
Stage 2:
warp Y_{tr} to the view of spectral branch using D, get Y_{wp};
use Y_{wp} to guide Y_{sr} for texture enhancement, get Y_{HR}.

Output: HR spectral image Y_{HR} and disparity map D

To train a deep network for RGB-guided texture enhancement (EnhanceNet for short hereafter), we need a large number of spectral images and their corresponding RGB images. The dataset we use for spectral image super-resolution does contain such image pairs, however, the spectral response function for generating the RGB images in the dataset is unknown and it is unlikely to match that of the RGB camera in our system. To address this issue, we synthesize the required RGB images using the spectral response function of our own RGB camera, as shown in Fig. 3. Note that, similar to spectral image super-resolution, the RGB images are synthesized from the spectral images with 16 bands. Since the EnhanceNet is trained on synthetic RGB images, it is better to be consistent in the inference phase. Therefore, we transfer Y_{RGB} to the style of Y_{syn} (instead of the opposite) during depth estimation, and the spectral alignment is guaranteed in this way.

As shown in Fig. 5, the EnhanceNet can be divided into two stages: feature extraction and feature fusion. The features of the spectral and RGB images are first extracted individually, and then fused together to obtain the final reconstruction. In both stages, we use the advanced channel attention residual blocks [19], and the mean-square-error is adopted as the loss function. This EnhanceNet, once trained, takes the SR-spectral cube Y_{sr} and the warped RGB image Y_{wp} as input, and generates the HR spectral cube Y_{HR} as output. The whole reconstruction process is now complete, which is summarized in Algorithm 1.

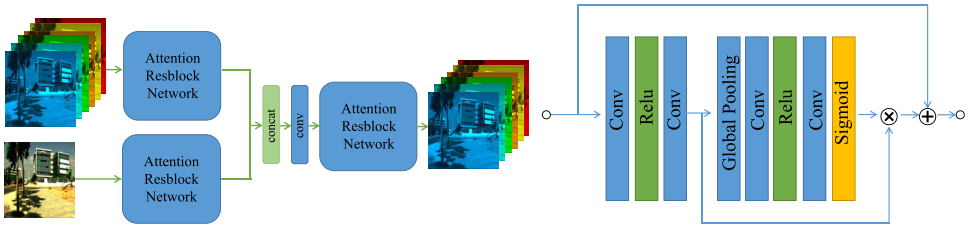


Fig. 5. (a) The architecture of EnhanceNet. (b) Channel attention residual block.

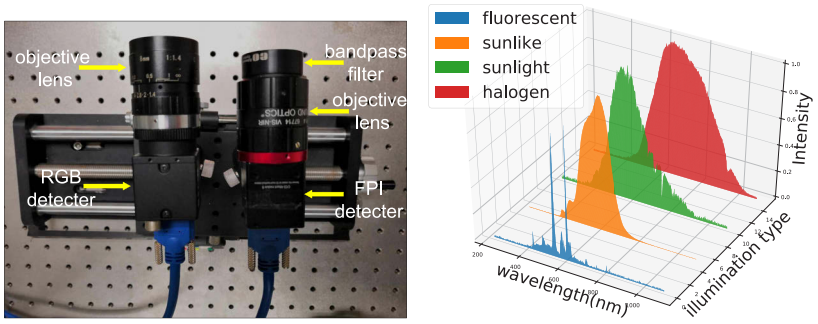
4. Hardware implementation

4.1. Hardware system setup

As shown in Fig. 6(a), our system merely consists of a spectral camera and an RGB camera. Both cameras are of small size and the whole system is compact. The spectral camera is

Optics EXPRESS

a commercial product with model XIMEAMQ022HG-IM-SM4X4-VIS [20]. Based on the Fabry-Pérot interference (FPI) principle [21], the spectral camera is snapshot [22,23] and operates at a speed up to 170*fps*. The sensor has a total spatial resolution of 2048×1024 and a spectral resolution of 16 bands. The spatial resolution is actually multiplexed by the 16 spectral bands in the wavelength range of $470 - 630nm$ (restricted by the band-pass filter) at an interval of $\sim 10nm$. That is to say, for each band, the effective spatial resolution is 512×256. The focal length of the objective lens on the spectral camera is fixed at $8mm$. The RGB camera is PointGrey FL3-U3-32S2C-CS with a spatial resolution of 1920×1080 and a temporal resolution up to 50*fps*, which is equipped with a $8mm$ fixed-focus lens. The two cameras are placed in parallel with a baseline distance of $5cm$, and the target scene is about $50 - 70cm$ in front of the system. The final reconstruction generates a spectral cube with a resolution of $1920 \times 1080 \times 16$ and a disparity map of the same spatial resolution. The temporal resolution of the whole system is up to 50*fps*, given sufficient ambient illumination.



(a) (b)

Fig. 6. (a) Hardware prototype. (b) Spectral curves of different light sources.

4.2. Illumination selection

We consider the following requirements for choosing an optimal light source for laboratory illumination: (a) sufficient intensity to illuminate different surface structures and textures in the target scene; (b) active and flat spectral response within the effective wavelength range of the spectral camera; and (c) easy manipulation and low cost. To this end, we test several alternative light sources as shown in Fig. 6(b) for comparison. Among them, sunlight is closest to the ideal light source, but it is not easy to control and much influenced by weather conditions. The spectral curve of the fluorescent lamp is not smooth and uniform enough, and the spectral curve of the halogen lamp in the visible cold light section is not as good as that of the white LED. The white LED employing the sunlike technology [24] can approach natural light as far as possible in the visible light section, while the intensity and direction of light can be easily controlled. Therefore, we choose the white LED as the light source for laboratory illumination in the following experiments.

4.3. Calibration and correction

In a stereo configuration, the spectral camera and the RGB camera needs to be calibrated in advance to facilitate the subsequent disparity calculation. Here we adopt a classical strategy similar to the one in [25] to calibrate the two cameras with a black-white checkerboard, and get intrinsic matrix, extrinsic matrix, and distortion coefficients of the two cameras. During the calibration process, we also obtain the scaling factor between the two cameras. This factor reflects the real resolution gap between the spectral and RGB measurements, and once obtained,

Optics EXPRESS

will be fixed for generating the training data of the corresponding networks. The correction includes reflection correction and spectral correction. The former can eliminate the influence of sensor transmission efficiency, and the latter can reduce the influence of the channel crosstalk, so as to obtain the real spectral attributes of the object surface [26]. We have conducted an additional experiment to justify the necessity of spectral correction for preventing crosstalk (see Section 5.1.3).

5. Experimental results

5.1. Spectral output evaluation

The evaluation of spectral reconstruction includes two aspects: spectral image super-resolution and RGB-guided texture enhancement. We quantitatively and qualitatively evaluate the reconstruction performance of spectral images from benchmark datasets as simulation and real-world scenes captured by the proposed system. The evaluation is conducted in terms of both spatial and spectral fidelity metrics.

5.1.1. Training settings

For spectral image super-resolution, we adopt the SISR-Net which enhances the spatial resolution of a spectral cube with a scaling factor of 4.45. This scaling factor is obtained during the calibration process, which reflects the real resolution gap between the spectral and RGB measurements after calibration. As mentioned above, we use the ICVL spectral image dataset [4] with high spatial (1392×1296) and spectral (519 bands) resolution to train the SISR-Net. We synthesize data from 519 bands to 16 bands using the spectral response function of our spectral camera. The synthesized dataset is divided into two groups for training and validation. There are 70 images in the training set and 14 images in the validation set.

For RGB-guided texture enhancement, we adopt the EnhanceNet to introduce the HR spatial information from the RGB image into the spectral image. The dataset for training the EnhanceNet is also generated from ICVL [4]. As mentioned above, we first synthesize spectral images from 519 bands to 16 bands and then synthesize corresponding RGB images from 16 bands to RGB. There are 70 images selected for training and 14 images for validation. Since the output of the SISR-Net will be the input of the EnhanceNet in practice, the training images for the two networks should be strictly non-overlapping. The validation images are kept the same as SISR-Net.

5.1.2. Simulation results

To verify the reconstruction algorithm in principle, we first conduct a simulation on the ICVL dataset, where the spectral images in the validation set are spatially down-sampled to serve as the LR test images. Each test image is super-resolved by the SISR-Net followed by the EnhanceNet. The input of the EnhanceNet consists of the output of the SISR-Net (i.e., the SR-spectral cube) and its corresponding HR RGB image, and the output of EnhanceNet is an HR spectral image with enriched texture information.

Note that the warped RGB image may suffer from occlusion in practice, and here we simulate the occlusion effect by adding random masks with 30% occluded (missing) pixels when synthesizing RGB images from 16-band spectral cubes. These pixels are then recovered by the inpainting method as mentioned above. This 30% occlusion ratio is an approximation obtained from a real stereo dataset [27]. Besides this model trained from data with simulated occlusion, We also synthesize RGB images without any occluded pixels for training an ideal model, which serves as the upperbound.

Table 2 gives the quantitative results while Fig. 7 shows the qualitative results at different reconstruction stages. We adopt the PSNR and RMSE metrics for evaluating the spatial and

Optics EXPRESS

spectral accuracy respectively. The PSNR between cube x and cube x' is defined as

$$PSNR(x,x') = \frac{1}{\lambda}\sum_{\lambda}(20log\frac{MaxValue}{\sqrt{\frac{1}{HW}\sum_H\sum_W(x-x')^2}}) \tag{1}$$

and RMSE is defined as

$$RMSE(x,x') = \frac{1}{HW}\sum_H\sum_W\sqrt{\frac{1}{\lambda}\sum_{\lambda}(x-x')^2} \tag{2}$$

where H, W and λ denote the spatial and spectral resolutions of the cube, respectively. As can be seen, the output of SISR-Net improves 1.97dB in PSNR and 19.36% in RMSE averagely over the bicubic interpolation result, which demonstrates the effectiveness of spectral image super-resolution through deep learning. Moreover, the output of EnhanceNet significantly improves the SISR-net result by 3.06dB in PSNR and 24.42% in RMSE averagely, which validates the effectiveness of RGB-guided texture enhancement through deep learning. The qualitative results in Fig. 7 provide consistent evidences. As shown in Fig. 7(a), compared to interpolation, SISR-Net notably improves the spatial details in the spectral image, yet not as significantly as EnhanceNet. Figure 7(b) shows the spectral reconstruction error at different wavelengths, which demonstrates the superiority of EnhanceNet again.

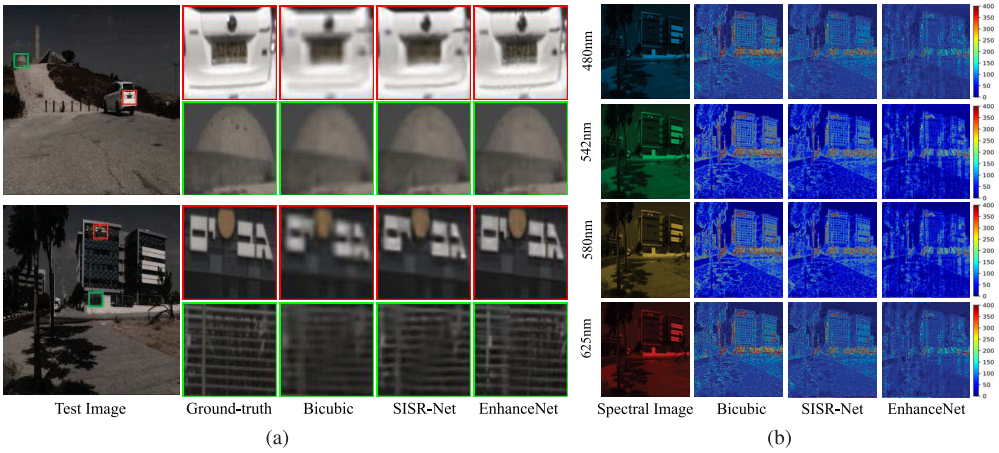


Fig. 7. (a) Reconstruction results on the ICVL dataset. Spectral images are synthesized into RGB for visualization. (b) Spectral reconstruction error on the ICVL dataset.

We also conduct another experiment to justify the causality of the spectral data from [4] and our own spectral imager. Following [4], we synthesize a new set of 16-band images by averaging the adjacent bands from the 519-band raw images, relying on which we retrain the SISR-Net and get a new model denoted as AVE. The model trained from synthesized data using the spectral response function of our own spectral imager is denoted as SRF. On the same test set in Table 2, the SRF model outperforms the AVE model by 0.33dB in PSNR, which proves the necessity of our data synthesis process.

5.1.3. Real-world results

To further verify the reconstruction algorithm on our proposed system, we conduct extensive experiments on representative real-world scenes. Figure 8 shows the reconstruction results of three captured scenes containing drastically different object surfaces. As can be seen, compared

Optics EXPRESS

Table 2. Quantitative results of spectral reconstruction through different methods on the ICVL dataset.

Test Image	Spatial Metric (PSNR)				Spectral Metric (RMSE)			
	Bicubic	SISR-Net	EnhanceNet	Upperbound	Bicubic	SISR-Net	EnhanceNet	Upperbound
BGU_1113	34.47	37.41	39.94	56.44	34.86	25.26	21.81	5.30
BGU_1217	40.40	43.01	45.24	60.09	23.28	17.22	14.15	3.77
Lehavim_1627	37.51	39.51	42.59	58.55	25.22	21.19	16.41	4.37
Lehavim_1716	40.11	41.57	44.98	59.88	24.75	21.26	14.67	3.73
BGU_1439	37.58	39.80	43.41	59.02	27.01	21.48	15.58	3.99
Eve_1551	37.70	39.74	42.72	58.43	27.79	21.75	17.72	4.50
Gavyam_0944	33.77	35.84	38.95	57.81	43.31	33.53	24.77	4.63
Lehavim_1600	35.49	37.66	40.64	59.19	36.95	29.56	20.21	4.10
Nachal_1117	35.57	37.39	40.78	58.63	33.00	26.34	19.74	4.32
Negev_1003	33.30	33.98	37.50	56.67	48.24	44.37	29.85	5.15
Omer_1104	35.18	37.30	39.86	55.09	35.82	29.04	24.93	5.95
Omer_1135	35.52	37.51	40.56	58.16	31.77	25.32	19.69	4.36
Rsh_1356	33.46	35.24	38.41	55.77	50.80	40.74	30.01	5.69
Sat_1157_1135	37.54	39.18	42.48	57.45	29.94	25.16	18.64	4.75
Average	36.26	38.23	41.29	57.94	33.77	27.23	20.58	4.62

to the LR input, SISR-Net improves the spatial details to a certain extent, but also introduces artifacts sometimes. The underlying reason is that SISR-Net only learns from external data, which may be irrelevant with the target scene. By exploiting the RGB image of the same scene under a cross-modal stereo configuration, EnhanceNet reconstructs the spatial details much more faithfully.

To evaluate the spectral reconstruction quantitatively, we use a standard 24-colorchecker as the target scene (as shown in Fig. 4). The ground-truth spectral signatures at 24 locations corresponding to the center of each color are measured by a Stellar-Net BLK-CXR-SR-50 probe spectrometer with 1.3nm spectral resolution. The reconstructed spectral images are evaluated against the ground-truth signatures. As shown in Fig. 9, the proposed system along with the reconstruction algorithm achieves less than 3.5% RMSE averagely in terms of different colors on the checkerboard, which is competitive to existing snapshot spectral cameras [10]. This real-world result also confirms the fidelity of resolution enhancement through our proposed method.

Furthermore, to demonstrate the effectiveness of spectral correction as mentioned in Section 4.3, we conduct another experiment on the checkerboard scene by deliberately omitting the spectral correction. We then get an RMSE of spectral signature over 8.5% while the original result with spectral correction has an RMSE of less than 3.5%. Therefore, the necessity of spectral correction on minimizing the crosstalk error can be verified.

5.2. Depth output evaluation

Besides the spectral image with high spatial resolution, our system also generates the depth map of the target scene with the same spatial resolution. As mentioned above, we adopt the pretrained model of MC-CNN [17] on the KITTI dataset [5] to calculate the disparity from a pair of synthesized/color-transferred RGB images.

We conduct two experiments to evaluate the performance of depth estimation. The first experiment quantitatively evaluates the depth accuracy, which is widely adopted for depth camera evaluation in literature [28–30]. We place a set of planes in front of the system at different

Optics EXPRESS

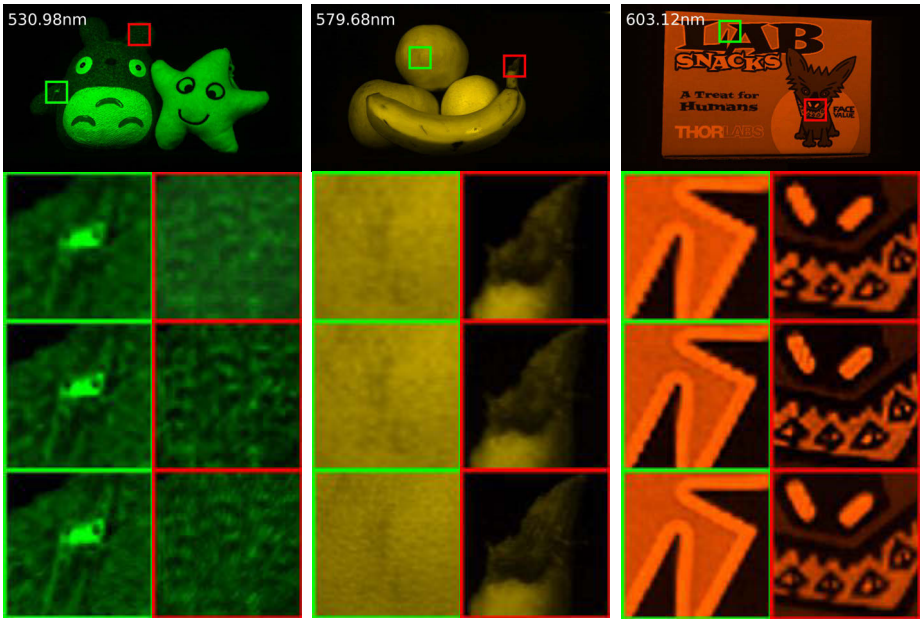


Fig. 8. Reconstructed spectral images of real-world scenes. From top to bottom: real-world scene at selected bands, zoom-in results of bicubic, SISR-Net, and EnhanceNet. (Please refer to the electronic version on a bright display for better visualization.)

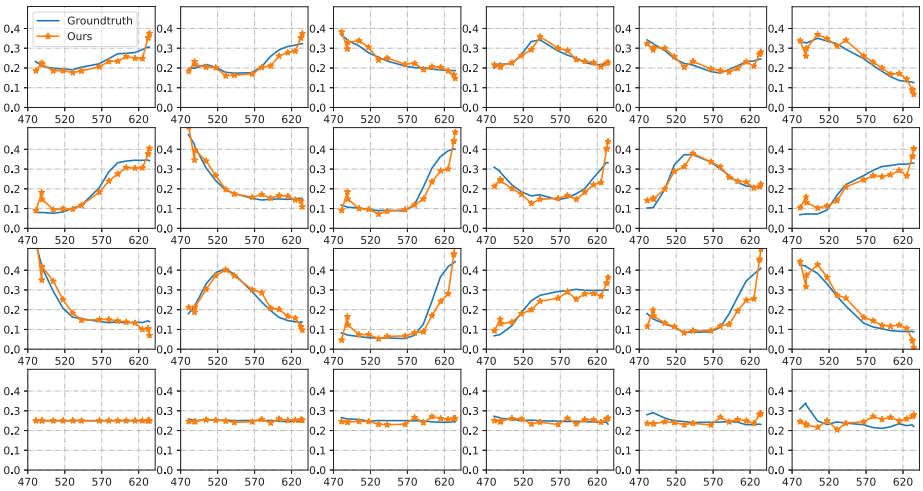
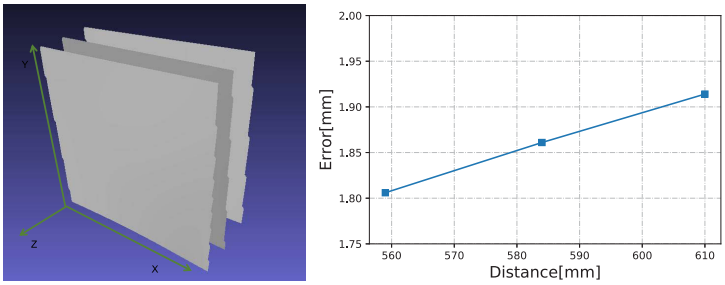


Fig. 9. Recovered spectral signatures of a standard 24-colorchecker.

distances (texture on the plane surface is shown in Fig. 8, last column). At each distance, a plane is fitted to the 3D points and the RMSE between the reconstructed points and the fitted plane is calculated. As can be seen from Fig. 10, the depth reconstruction error increases as the distance increases, yet the error is less than $2mm$ at a distance of $0.6m$, which is competitive to mainstream real-time depth cameras [31]. It is worth mentioning that, we add two steps (color transfer and RGB synthesis) before estimating disparity to eliminate the inconsistency between the cross-modal image pair, which could improve the reliability of stereo matching.

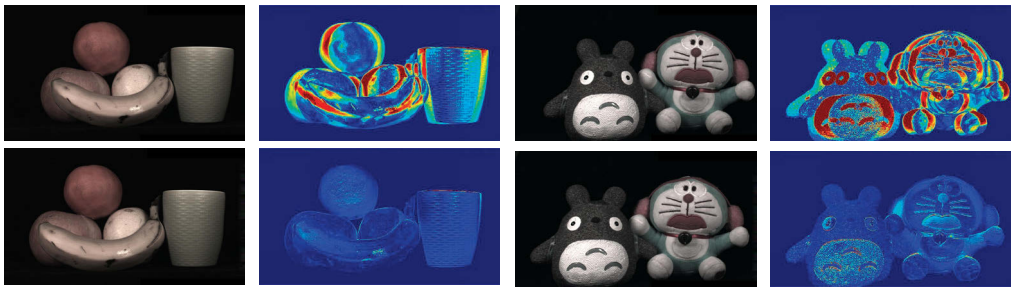
Optics EXPRESS



(a) (b)

Fig. 10. Evaluation of depth estimation. (a) Point clouds of a set of planes at different distances. (b) Accuracy of estimated depth maps in terms of RMSE.

In the second experiment, we adopt "warping error" as an auxiliary metric to evaluate the fidelity of depth estimation. The jet color map is used to display the error in Fig. 11. We can see that there is a distinct disparity between the spectral image and the RGB image before warping, which results in severe ghosting effect in the error maps. On the other hand, after warping with the disparity calculated through MC-CNN, the error between the two images is largely eliminated and there are no more ghosting in the error maps. Note that the remaining visible pixels in the error maps are due to color inconsistency between the spectral and RGB cameras even after color transfer (see Section 3.2). This experiment again demonstrates the superior accuracy of our depth estimation results. More point cloud results of real-world scenes are shown in Fig. 12.



(a) (b) (c) (d)

Fig. 11. Warping error for disparity evaluation. (a) RGB measurements before (top) and after (bottom) warping. (b) Error maps between two stereo views before (top) and after (bottom) warping. (c) & (d) for another scene.

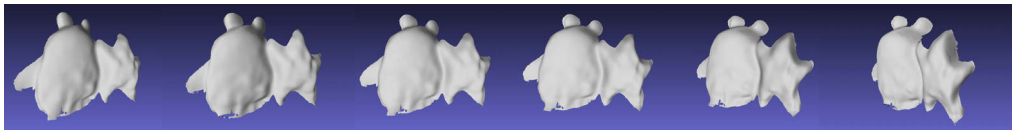


Fig. 12. Point cloud results of a real-world scene (texture on the surface shown in Fig. 8, first column) from different views.

Optics EXPRESS

6. Conclusion and discussion

In this paper, we present a compact system for joint spectral and depth imaging in real time and with high resolution. Instead of developing highly customized hardware that is generally of high complexity, our system only ensembles two off-the-shelf cameras and relies on deep learning based computational reconstruction to achieve state-of-the-art imaging performance. The proposed system can eventually obtain spectral images with a temporal resolution up to 50*fps*, a spatial resolution of 1920×1080, and a total of 16 spectral bands covering the wavelength range of $470 - 630nm$, as well as the corresponding depth map of the target scene with the same spatial resolution. This work allows 5D information (3D space + 1D spectrum + 1D time) of the target scene to be captured with a miniaturized apparatus and without active illumination. Considering its compactness and easy manipulation, the proposed system offers a practical solution for ubiquitous spectral-depth imaging in the wild, e.g., on smartphones or UAVs.

Still, there are several aspects to further improve our current system, which are considered as our future work. First, we now simulate the training data with digital downsampling which may not reflect the real degradation of LR images in practice. The networks trained in this way might have certain deviation from the groundtruth during the resolution enhancement process. This is a common problem for existing learning-based super-resolution methods, and has been studied in recent works [32]. It requires more efforts to address this issue in the scenario of our system, and we will follow this direction to improve the system accuracy. Second, we would like to test the generalization capability of the proposed reconstruction algorithm under different camera configurations, e.g., real image data with different spatial and spectral resolutions.

Funding

Key Technologies Research and Development Program (2018YFC0307905); National Natural Science Foundation of China (61671419).

Disclosures

The authors declare no conflicts of interest.

References

1. M. H. Kim, T. A. Harvey, D. S. Kittle, H. Rushmeier, J. Dorsey, R. O. Prum, and D. J. Brady, "3d imaging spectroscopy for measuring hyperspectral patterns on solid objects," ACM Trans. Graph. **31**(4), 1–11 (2012).
2. L. Wang, Z. Xiong, G. Shi, W. Zeng, and F. Wu, "Simultaneous depth and spectral imaging with a cross-modal stereo system," IEEE Trans. Circuits. Syst. Video Technol. **28**(3), 812–817 (2018).
3. S. Heist, C. Zhang, K. Reichwald, P. Kühmstedt, G. Notni, and A. Tünnermann, "5d hyperspectral imaging: fast and accurate measurement of surface shape and spectral characteristics using structured light," Opt. Express **26**(18), 23366–23379 (2018).
4. B. Arad and O. Ben-Shahar, "Sparse recovery of hyperspectral signal from natural rgb images," in "*Proceedings of European Conference on Computer Vision*," 19–34 (2016).
5. A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in "*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*," 3354–3361 (2012).
6. L. Wang, Z. Xiong, H. Huang, G. Shi, F. Wu, and W. Zeng, "High-speed hyperspectral video acquisition by combining nyquist and compressive sampling," IEEE Trans. Pattern Anal. Machine Intell. **41**(4), 857–870 (2019).
7. 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. Machine Intell. **39**(10), 2104–2111 (2017).
8. A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," Appl. Opt. **47**(10), B44–B51 (2008).
9. X. Cao, H. Du, X. Tong, Q. Dai, and S. Lin, "A prism-mask system for multispectral video acquisition," IEEE Trans. Pattern Anal. Machine Intell. **33**(12), 2423–2435 (2011).
10. L. Wang, Z. Xiong, D. Gao, G. Shi, and F. Wu, "Dual-camera design for coded aperture snapshot spectral imaging," Appl. Opt. **54**(4), 848–858 (2015).
11. X. Yuan, T.-H. Tsai, R. Zhu, P. Llull, D. Brady, and L. Carin, "Compressive hyperspectral imaging with side information," IEEE J. Sel. Top. Sign. Process. **9**(6), 964–976 (2015).

Optics EXPRESS

12. C. Ma, X. Cao, X. Tong, Q. Dai, and S. Lin, "Acquisition of high spatial and spectral resolution video with a hybrid camera system," Int. J. Comput. Vis. **110**(2), 141–155 (2014).
13. C. Ma, X. Cao, R. Wu, and Q. Dai, "Content-adaptive high-resolution hyperspectral video acquisition with a hybrid camera system," Opt. Lett. **39**(4), 937–940 (2014).
14. I. Amro, J. Mateos, M. Vega, R. Molina, and A. K. Katsaggelos, "A survey of classical methods and new trends in pansharpening of multispectral images," EURASIP J. Adv. Signal Process **2011**(1), 79 (2011).
15. Z. Shi, C. Chen, Z. Xiong, D. Liu, Z.-J. Zha, and F. Wu, "Deep residual attention network for spectral image super-resolution," in "*Proceedings of the European Conference on Computer Vision Workshop*," 214–229 (2018).
16. D. C. Castro and B. Glocker, "Nonparametric density flows for mri intensity normalisation," in "*International Conference on Medical Image Computing and Computer-Assisted Intervention*," 206–214 (2018).
17. J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in "*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*," 1592–1599 (2015).
18. M. Bertalmio, A. L. Bertozzi, and G. Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in "*Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*," **1**, I–I (2001).
19. Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in "*Proceedings of the European Conference on Computer Vision*," 286–301 (2018).
20. https://www.ximea.com/en/products/hyperspectral-cameras-based-on-usb3-xispec/mq022hg-im-sm4x4-vis
21. M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light* (2013).
22. N. Tack, A. Lambrechts, P. Soussan, and L. Haspeslagh, "A compact, high-speed and low-cost hyperspectral imager," Proc. SPIE **8266**, 82660Q (2012).
23. B. Geelen, N. Tack, and A. Lambrechts, "A compact snapshot multispectral imager with a monolithically integrated per-pixel filter mosaic," Proc. SPIE **8974**, 89740L (2014).
24. M. Worku, Y. Tian, C. Zhou, S. Lee, Q. Meisner, Y. Zhou, and B. Ma, "Sunlike white-light-emitting diodes based on zero-dimensional organic metal halide hybrids," ACS Appl. Mater. Interfaces **10**(36), 30051–30057 (2018).
25. Z. Zhang, "A flexible new technique for camera calibration," IEEE Trans. Pattern Anal. Mach. Intell. **22**(11), 1330–1334 (2000).
26. J. Burger and P. Geladi, "Hyperspectral nir image regression part i: calibration and correction," J. Chemom. **19**(5-7), 355–363 (2005).
27. D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, W. Xi, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in "*German Conference on Pattern Recognition*," 31–42 (2014).
28. Y. Zhang, Z. Xiong, and F. Wu, "Unambiguous 3d measurement from speckle-embedded fringe," Appl. Opt. **52**(32), 7797–7805 (2013).
29. Y. Zhang, Z. Xiong, Z. Yang, and F. Wu, "Real-time scalable depth sensing with hybrid structured light illumination," IEEE Trans. Image Process. **23**(1), 97–109 (2014).
30. Z. Xiong, Y. Zhang, W. Feng, and W. Zeng, "Computational depth sensing : Toward high-performance commodity depth cameras," IEEE Signal Process. Mag. **34**(3), 55–68 (2017).
31. O. Wasenmüller and D. Stricker, "Comparison of kinect v1 and v2 depth images in terms of accuracy and precision," in "*Asian Conference on Computer Vision*," 34–45 (2016).
32. C. Chen, Z. Xiong, X. Tian, Z.-J. Zha, and F. Wu, "Camera lens super-resolution," in "*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*," 1652–1660 (2019).