Joint Reconstruction and Super Resolution of Hyper-Spectral CTIS Images

Mazen Mel[1]
mazen.mel@dei.unipd.it
Alexander Gatto[2]
alexander.gatto@sony.com
Pietro Zanuttigh[1]
zanuttigh@dei.unipd.it

[1] Department of Information Engineering
University of Padova
Padova, Italy

[2] Sony Europe B.V.
R&D Center - Stuttgart Laboratory 1
Stuttgart, Germany

Abstract

Computed Tomography Imaging Spectrometers (CTIS) capture dense spectrum of dynamic scenes as compressed 2D sensor measurements. Model-based Hyper-Spectral (HS) image reconstruction algorithms devised for such systems are typically very slow, sensitive to the selected data and noise models, and can only restore HS images with poor spatial resolution. On the other hand, deep learning-based approaches, once trained, are capable of performing the reconstruction in real-time and are more suitable for high frame-rate applications but generally suffer from limited generalization capabilities. In this paper for the first time, we jointly address the issues of reconstruction speed and spatial resolution of CTIS through a simple and interpretable deep learning architecture partially inspired by the Filtered Back-Projection (FBP) algorithm used in conventional CT scans. Our model is able to exploit aliased pixel information in CTIS images to recover spatially super-resolved HS cubes. Experimental results on simulated and real data demonstrate the effectiveness of our approach not only in reconstruction quality, but also in computation time and generalization ability.

1 Introduction

Different from scanning-based imaging spectrometers, snapshot systems offer greater flexibility and can capture full spectrum of still as well as dynamic scenes in a single coded 2D measurement achieved by multiplexing spatial and spectral information using a combination of lenses, dispersive elements, and coded aperture masks. Further processing is required to reconstruct a 3D HS cube which is usually a time consuming operation that impedes the real-time applicability of theses systems. Model-based reconstruction approaches exploit iterative schemes along with some prior knowledge constraints. However, even with hardware-enabled acceleration, the time required to reconstruct a single HS image does not meet the basic requirements for high frame-rate applications. Alternatively, deep learning-based methods learn complex non-linear mapping between pairs of 2D measurements and the corresponding 3D HS cubes in a supervised manner and once trained these networks can be used to infer HS data in real-time.

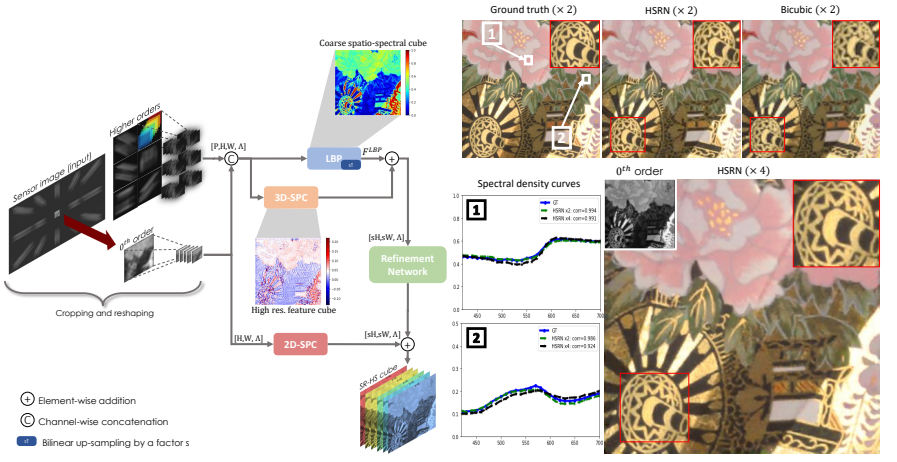

Figure 1: Left: Our proposed network (HSRN) reconstructs coarse spatio-spectral cubes generated by LBP. These are added to a high resolution residual generated by the 3D Sub-Pixel Convolution module (3D-SPC) in order to build a spatially super resolved HS cube. Right: Reconstructed HS images with $\times 2$ and $\times 4$ the resolution of the 0^{th} diffraction order are shown in sRGB space and compared with a bicubic up-sampled reference cube. Spectral density curves are shown along with the Pearson correlation coefficient between the predicted and ground truth curves.

At the same time, it is well known that the generalization capability of such models is limited by the training environment further hindering their usefulness in practice. We propose a lightweight network architecture (we named it HSRN) to efficiently reconstruct spatially super-resolved HS cubes from 2D measurements generated by a CTIS system [6, 30] owing to its high spectral resolution and the availability of multiple tomographic projections each carrying distinct and complimentary spatial and spectral information needed to reconstruct the latent HS cube. To this end, we propose to learn in an end-to-end fashion the Filtered Back-Projection (FBP) algorithm used in traditional CT scans inside the feature space thus enabling greater interpretability of our network. In addition, we propose a HS image Super-Resolution (SR) module exploiting side information present in higher order projections through 3D deconvolution layers. To the best of our knowledge, this is the first work to handle jointly hyper-spectral image reconstruction and super-resolution for CTIS systems. Our contributions can be summarized as the following:

- A simple yet efficient network architecture capable of reconstructing spatially super-resolved HS cubes in real-time (up to 30 fps for a cube of size $400 \times 400 \times 31$) from CTIS measurements.

- A novel end-to-end Learned Back-Projection (LBP) layer that enables superior reconstruction quality.

- The effectiveness of our model is validated through exhaustive experimenting on synthetic as well as real CTIS images: it outperforms state-of-the-art approaches.

2 Related Work

HSI systems. Early spectrometers were predominantly scanning devices such as pushbroom [31], whiskbroom [5], and tunable filter cameras [16] which are capable of capturing images with high spatial and spectral resolution but at the same time they are fairly large and cumbersome devices incorporating multiple moving parts and requiring long acquisition times. Owing to the quick advancements in compressive sensing and deep learning, snapshot spectrometers or even conventional RGB cameras [18, 34] became widely used to capture dense spectrum of dynamic scenes, see [3, 12] for comprehensive surveys. Coded Aperture Snapshot Spectral Imaging (CASSI) systems [19, 36] stand out as one of the most used devices for HSI. However, they offer poor image quality as the spatial resolution is degraded due to the use of coded aperture masks. In addition, its spectral resolution is limited by the sensor pixel pitch along with the non-linear dispersion introduced by the prism leading to a trade-off between spatial and spectral resolution. Alternatively, in a CTIS system [6, 13, 30] light is dispersed into multiple tomographic projections via a Diffractive Optical Element (DOE) forming multiple projections of the latent HS cube on the image sensor. Indeed, CTIS practical applicability is reduced by the poor spatial resolution of its 0^{th} diffraction order which determines the resolution of the reconstructed HS image [10, 11, 15]. Furthermore, no previous work has tackled this problem so far, at least from a computational point of view. In this paper, we exploit sub-pixel displacements present in higher diffraction orders to perform image SR and reconstruct HS cubes with up to $\times 4$ the resolution of the 0^{th} order hoping to pave the way for more research into CTIS technology.

HS image reconstruction. Several approaches proposed to solve the problem of reconstructing a 3D HS cube from CASSI measurements iteratively with image priors in a Maximum A Posteriori (MAP) estimation framework. IST [9] and TwIST [4] incorporated a TV-norm regularization term to encourage sparsity of the solution. Liu *et al*. proposed DeSCI [26], exploiting a weighted nuclear norm regularizer solving a rank minimization problem. Aside from hand-crafted priors, a new class of optimization algorithms based on variable splitting techniques such as the Alternating Direction Method of Multipliers (ADMM) and Half Quadratic Splitting (HQS) were proposed to decouple data fidelity and prior terms treating the latter as a plug-and-play denoiser module using off-the-shelf powerful denoisers such as DBM3D [8] or even a trained CNN as in [7, 27, 47]. Even though impressive performance has been achieved using model-based approaches, reconstruction time is exorbitantly high reaching up to 4.6 hours in [27]. In an attempt to combine the interpretability and flexibility of model-based approaches and the reconstruction speed of CNNs at inference time, unrolled network architectures have been introduced by [37, 38, 44, 46]. For CTIS the Expectation-Maximization (EM) algorithm has been predominantly used in reconstruction [35] as in most CT based systems. The EM is a Maximum-Likelihood (ML) solver that cannot handle priors and is very sensitive to the presumed noise and system models leading to sub-optimal performance and poor reconstruction quality, Other approaches, such as low rank based estimation and superiorization have been proposed in [22] and [15]. Recently a GPU accelerated EM variant has been introduced by White *et al*. [39] exploiting spatial shift invariance of the system matrix reaching a significant speedup in reconstruction time but still with very low spatial resolution. Lately, deep learning-based approaches started gaining attraction for CTIS systems: Huang *et al*. [17] proposed to learn end-to-end mapping through a multi-branch CNN. They have also introduced a follow-up hybrid approach [1]

combining a CNN sequentially with an EM solver. Zimmermann *et al*. [48] implemented an initial reshaping layer enabling 3D processing of high dimensional input data to account for spatio-spectral correlations within multiple higher diffraction orders which is then followed by a U-Net like architecture used to refine the estimated HS cube.

Image SR. Recent approaches for recovering a High-Resolution (HR) image from a down-sampled and corrupted Low-Resolution (LR) measurement are mostly based on deep learning models due to their high capacity to learn complex non-linear LR/HR mappings and effectively reconstruct visually appealing high frequency details [21, 25, 33]. Multi-frame image SR on the other hand exploit side information provided by different frames as aliasing that is caused by relative camera movement or object movement within the scene. The complementary information present in multiple frames is combined and mapped into a higher resolution pixel grid with high spatial fidelity: such methods can be divided into model-based approaches [24, 40] and deep learning-based ones [12, 20].

3 Network Architecture

We propose to reformulate, in a learning context, intuitive yet effective algorithms used in CT scans, i.e., the back-projection and in particular its enhanced version FBP: by exploiting the representation capacity and flexibility of CNNs we show that it is possible to achieve superior spectral reconstruction performance with a lighter network architecture. Furthermore, in this way we also preserve some degree of network interpretability which is usually hard to achieve in standard learning-based approaches.

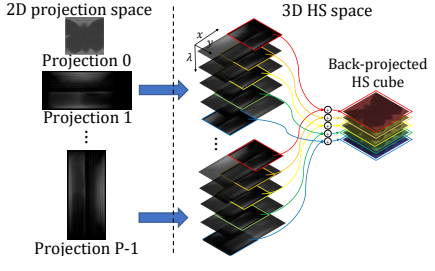

Figure 2: Back-projection of multiple 2D CTIS projections into a 3D HS cube.

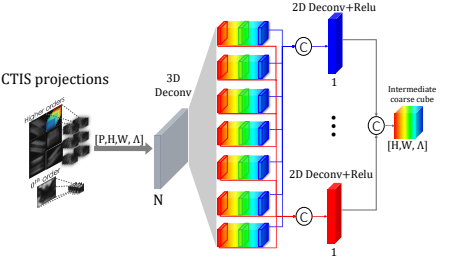

Figure 3: LBP architecture inspired by the filtered back projection algorithm.

3.1 Learned Back Projection (LBP)

In CTIS systems the underlying image formation model can be written as:

$$g = Hf + \varepsilon \tag{1}$$

where $g \in \mathbb{R}^{(MN)\times 1}$ and $f \in \mathbb{R}^{(HW\Lambda)\times 1}$ are respectively the vectorized coded 2D sensor image and the latent 3D HS cube to be reconstructed with Λ spectral channels. $H \in \mathbb{R}^{(MN)\times(HW\Lambda)}$ is the system matrix, and ε is an additive noise term. The back-projection operation (see Fig. 2) can be approximated by mapping a low dimensional 2D projection into the higher

dimensional 3D HS space by simply repeating the projection across the spectral dimension and summing all back-projected spectral slices from multiple projections and stacking them channel-wise producing a rendition of the latent 3D HS cube. Mathematically this operation correspond to the transpose of the system matrix H^T.

Given $\{g_p\}_{p=0}^{P-1}$ projections (including the 0^{th} order), we denote G_p the reshaped version of g_p in 3D space were we crop Λ slices of size $H \times W$ via a sliding window (see Fig. 3) and stack them channel-wise ending up with a 3D cube of shape $H \times W \times \Lambda$. Notice that each channel of this cube contains the latent spectral band to be reconstructed, in the case of the 0^{th} order we just repeat the image across the spectral dimension. The back-projected image f^{BP} is formed by summing up all $\{G_p\}_{p=0}^{P-1}$ for a given spectral band λ:

$$f^{BP}(x,y,\lambda) = \sum_p G_p(x,y,\lambda) \quad \text{for} \quad \lambda = 1,...,\Lambda \tag{2}$$

f^{BP} contains coarse spatio-spectral information of the latent cube but at the same time it is heavily blurred with halo-like effects and has a low SNR. To overcome this issue, the projections are first filtered out using a high pass Ramp filter [32] with kernel w, then the back-projection is performed, the combined operation is known as the filtered back-projection. Rewriting (2) to account for w results in:

$$f^{FBP}(x,y,\lambda) = \sum_p w * G_p(x,y,\lambda) \quad \text{for} \quad \lambda = 1,...,\Lambda \tag{3}$$

where $*$ represents the 2D convolution. Notice that w is a fixed filter that mainly enhances the contrast within each projection to avoid blurring, but at the same time it introduces high frequency noise and ringing artifacts due to the structure of such filter. Furthermore, the back-projection evenly maps 2D projected data back into HS space as it is a global operation and through the summation in Eq. (3) it does not take into account the different contributions of each projection, e.g., the fact that the amount of dispersion differs for each projection. We propose LBP with the aim to learn more complex non-linear relationships among CTIS projections but also within each projection. In particular, intra-projection correlations are learned by means of a 3D deconvolution layer [43]: we chose deconvolution instead of a normal convolution layer to restore high order image features and "reverse" the spatio-spectral multiplexing in the input. In more detail, as illustrated in Fig. 3, we form a 4D hyper-cube $G \in \mathbb{R}^{P \times H \times W \times \Lambda}$ with P channels corresponding to $\{g_p\}_{p=0}^{P-1}$ and apply a 3D deconvolution with N 3D filters $W^3 = \{w_i^3\}_{i=1}^{N}$ that produces a feature map $F \in \mathbb{R}^{N \times H \times W \times \Lambda}$. Inter-projection correlations are learned via 2D deconvolution layers: for each spectral band, all sub-channels from F carrying distinct spatial and spectral information of the same band are concatenated to form $\{F_i\}_{i=1}^{N} \in \mathbb{R}^{N \times H \times W}$ and fed into Λ 2D deconvolution layers with 1 output channel each. Lastly, the Λ output bands are concatenated channel-wise producing a coarse spatio-spectral version of the latent HS cube (see Fig. 1).

3.2 Hyper-spectral Image Super-resolution

Each local PSF generated by the DOE for a given projection in the CTIS sensor image differs slightly for each wavelength and for each higher order projection: in addition to sub-sampling by the sensor pixel grid, each diffracted order contains aliasing which provides distinct spatial information needed to reconstruct a spatially super-resolved HS image. The setting can be viewed as if each projection was a unique view of the smeared latent HS cube.

Differently from standard multi-frame image SR where sub-pixel shift can be estimated for correct image registration, spectral and spatial multiplexing makes it harder to work with

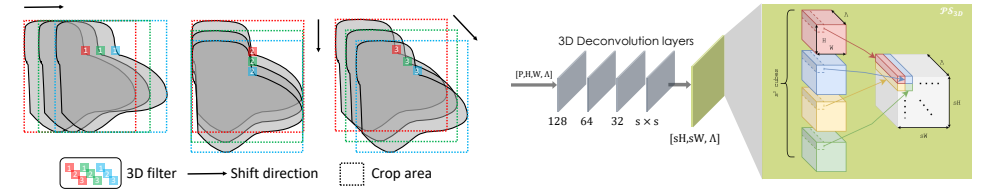

Figure 4: Simplified schematic of the 3D deconvolution operation with the filter's receptive field (left) and 3D-SPC module with the sub-pixel shuffling layer (right).

such information in CTIS. However, given the fact that sub-pixel displacements are mainly detected across image edges and that the smearing direction for each projection preserves image edges along said direction, e.g., vertical edges are preserved in vertical higher order projections and so on (see Fig. 4 (left)), we exploit such observations and treat the SR sub-problem in a residual learning context where our SR module restores high resolution image features like edges, i.e., high spatial frequencies, that are summed-up with the coarse spatio-spectral cube generated by LBP.

Still, the finite number of projections provides limited aliased information which further motivates the use of deep-learning based approaches in this case. In particular, inspired by spatio-temporal processing in video SR [23] we use 3D deconvolutions [43] to link spatial information scattered across multiple higher diffraction orders and learn more complex high level features. We propose an adaptation of ESPCN originally introduced by Shi *et al.* [53] for single image SR. Because convolutions are carried out in low resolution space, such approach is very efficient yet it achieves competitive image restoration results. More precisely, we propose a new 3D Pixel Shift ($\mathcal{PS}_{3D}$) operation exploiting side spatial information to perform SR via 3D periodic shuffling:

$$\mathcal{PS}_{3D}(T)\Big(x,y,\lambda\Big) = T\Big[s\cdot mod(y,s)+mod(x,s)+\lambda, \lfloor x/s\rfloor, \lfloor y/s\rfloor, \lambda\Big] \quad \text{for} \quad \lambda = 1,...,\Lambda \tag{4}$$

Where $T \in \mathbb{R}^{s\cdot s\times H\times W\times\Lambda}$ is a 4D feature map obtained from $G \in \mathbb{R}^{N\times H\times W\times\Lambda}$ (see Section 3.1) by applying multiple 3D deconvolution layers with 128, 64, 32, and $s\times s$ output channels. We refer to the whole block as 3D Sub-Pixel Convolution (3D-SPC). Eq. (4) implies that for a given spectral band λ, $s\times s$ SR pixels are obtained by periodically shuffling low-resolution pixels from $s\times s$ feature cubes (see Fig. 4 (right)). As illustrated in Fig. 4 (left) each 3D filter's receptive field sees a different 3D signal of the same HS image region each containing aliased pixel information and distinct spectral and spatial cues depending on the smearing direction. Mathematically such convolution can be expressed as:

$$Out = w_3 * \{G_p\}_{p=0}^{P-1} \tag{5}$$

where p is the projection index, w_3 is a 3D filter, $G_p \in \mathbb{R}^{H\times W\times\Lambda}$ is the reshaped tensor from a given projection $g_p = DH_pW_pf_{HR}$ where D is a down-sampling operator, H_p is a dispersion matrix, W_p is an affine warping matrix for sub-pixel displacement, f_{HR} is the latent cube to be reconstructed: notice that each kernel of w_3 is applied on $\{G_p\}_{p=0}^{P-1}$ with each channel carrying distinct yet complimentary spatial information.

A refinement network consisting of 7 convolution layers with 64 filters each and ReLU activations further refines the intermediate prediction from LBP and 3D-SPC stages. The

output is summed up with a super-resolved 0^{th} order using the 2D-SPC layer from [33]. Notice that even without such residual connection the network output will not be heavily affected. Rather, we observe that such connection introduces robustness to noise and leads to more stable training with faster convergence on noisy data in accordance with [45].

4 Experimental Evaluation

4.1 Data and Training Setup

We evaluate the performance of our approach (HSRN) on synthetic CTIS data generated from three publicly available datasets: TokyoTech-31 [29], CAVE [41], and ICVL [2]. We randomly choose $\sim 75\%$ of images as training data and the rest for testing. Results on a fourth dataset (Hyperspectral Video [28], used to assess real-time performance) are in the *suppl. mat*. We simulate CTIS images with 200 spectral bands spanning the range from $420nm$ to $720nm$ for TokyoTech-31 and $400nm$ to $700nm$ for CAVE and ICVL using Fourier optics (refer to the *suppl. mat* for further details on CTIS image simulation). In particular, a ground truth HS cube interpolated across the spectral dimension is convolved with a wavelength-dependent PSF to generate a CTIS sensor image with 14 higher diffraction orders (see Fig. 1). In case of noisy inputs we introduce shot noise simulating a quantum full well capacity of $1e^3$ photons. The training data is augmented using random rotation and flipping of the ground truth HS cubes before simulating the sensor image.

First of all we evaluate spectral reconstruction performance without the SR task. Then, we evaluate the network generalization capability via cross dataset validation. Later, we report results of HSRN for joint HS reconstruction and SR with $\times 2$ and $\times 4$ spatial resolution. Finally, we present some reconstruction results on a real image taken by our CTIS system. For all setups, except otherwise specified, we train the network using Adam optimizer with a learning rate of $1e^{-4}$ for 500 epochs with the following loss function:

$$\mathcal{L}(I_{SR}, I_{SR}^{GT}) = MSE(I_{SR}, I_{SR}^{GT}) + \gamma \cdot MAE(I_{SR}, I_{SR}^{GT}) + MSE(I_{LR}, I_{LR}^{GT}) \tag{6}$$

I.e., the first loss component is given by the MSE between the ground truth super-resolved HS cube I_{SR}^{GT} and the estimated one I_{SR}. The additional use of the MAE is motivated by the fact that it better preserves high spatial frequencies, we balance between the two components with a parameter γ set to 0.1. In order to force LBP to produce coarse spatio-spectral images we introduce an additional MSE loss term between the output of LBP I_{LR} and the s-fold down-sampled reference cube I_{LR}^{GT}.

4.2 Experimental Results

HS reconstruction: Quantitative results on HS cubes of size $100 \times 100 \times 31$ for our approach are shown in Tab. 1 along with qualitative results in Fig. 5. We compared them with Zimmerman *et al*. [48] and Ahlebæk *et al*. [1]. Since both competing approaches did not tackle the problem of spatial super-resolution, for this experiment we set the scale factor s to 1, thus reducing the sub-pixel shift layer to an identity mapping. See the *suppl. mat.* for details on the implementation of competitors. HSRN is able to outperform competing approaches on all three datasets with fewer trainable parameters and faster reconstruction time. More in detail, Zimmermann *et al*. [48] is capable of outperforming [1] with a much lighter model size, but is in turn outperformed by our approach that is also lighter. Fig. 5 shows

three reconstructed HS cubes from TokyoTech-31, CAVE, and ICVL converted to sRGB space. HSRN is able to produce better spatial and spectral distributions with less artifacts such as color leakage and blurring.

Method	#Params (M)	Time (s) (CNN/EM)	TokyoTech-31 RMSE↓	TokyoTech-31 PSNR↑	TokyoTech-31 SSIM↑	CAVE RMSE↓	CAVE PSNR↑	CAVE SSIM↑	ICVL RMSE↓	ICVL PSNR↑	ICVL SSIM↑
Ahlebæk *et al.* [1] †	26.6	0.05 / $\geq$ 10	0.035	28.849	0.872	0.039	28.708	0.823	0.021	33.896	0.881
Zimmermann *et al.* [43]	1.5	0.017 / -	0.028	33.033	0.917	0.024	34.448	0.941	0.005	47.497	0.991
HSRN (ours)	**0.9**	0.010 / -	**0.025**	**33.809**	**0.941**	**0.018**	**37.282**	**0.964**	**0.004**	**48.470**	**0.995**

Table 1: Quantitative comparison on multiple HS datasets with the two competing CTIS approaches. (†) Network architecture modified to account for the new input/output dimensions.

HS & SR reconstruction: HSRN is able to reconstruct HS cubes with a $\times 2$ and $\times 4$ resolution w.r.t. the 0^{th} diffraction order. Quantitative performance results are shown in Tab. 2(a) along with reconstructed samples in Fig. 6. We compare HSRN performance with two sequential approaches where we use the HSRN reconstruction stage followed by either: (i) a bicubic up-sampling with refinement through multiple convolution layers trained separately or (ii) by the original ESPCN [33] network. In the easier case of $\times 2$ SR, both sequential approaches are able to achieve satisfactory performance but still significantly lower than the one achieved by HSRN and in the $\times 4$ SR case they fall short of achieving acceptable results while HSRN preserves high PSNR scores (up by roughly 7 dB on ICVL compared to [33]). Notice that the reconstruction speed on an A6000 GPU of a $400 \times 400 \times 31$ HS cube is about 0.033s.

Cross dataset validation: We further test the generalization ability of HSRN on TokyoTech-31 and CAVE by training the network on one dataset and testing on the other (we did the test in both directions). We compared results with [43] since it is the best competitor. We train both architectures to generate HS cubes with 29 spectral bands ($420 \rightarrow 700nm$) with the same spatial resolution of the 0^{th} order, furthermore we show the results of HSRN at $\times 2$ spatial-resolution. All evaluation metrics reported in Tab. 2(b) prove the generalization capability of HSRN where it achieves better performance with respect to [43] at the base resolution and maintains good performance at $\times 2$ SR.

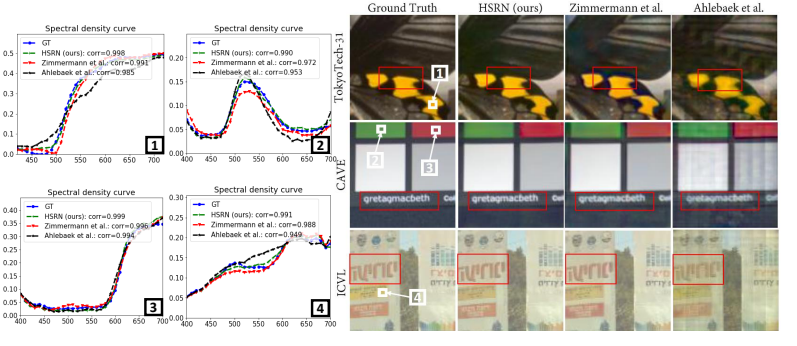

Figure 5: Reconstruction results of HSRN as well as the two competitors on three different benchmarks with HS cubes of size $100 \times 100 \times 31$.

Data	Scale	HSRN (ours)			Bicubic+CNN			Shi et al. [ ]		
		RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑
TokyoTech-31	×2	**0.026**	**34.738**	**0.945**	0.033	33.495	0.914	0.029	33.030	0.928
CAVE	×2	**0.018**	**37.244**	**0.956**	0.024	35.313	0.942	0.022	35.538	0.944
ICVL	×2	**0.011**	**42.065**	**0.972**	0.025	39.371	0.958	0.018	40.623	0.965
TokyoT. + CAVE	×4	**0.033**	**32.731**	**0.907**	0.078	24.556	0.844	0.057	27.375	0.888
ICVL	×4	**0.012**	**39.661**	**0.955**	0.061	29.065	0.889	0.036	32.732	0.902

(a) Spatial SR and HS reconstruction results.

Method	Time	Checkerboard			Butterfly		
	(CPU-s)	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑
SS-CASSI [ ]	16911	0.08	18.536	0.611	0.025	29.322	0.799
GAP-TV [ ]	17	0.055	21.975	0.700	0.027	27.446	0.884
DeSCI [ ]	4465	0.055	21.975	0.700	0.019	29.191	0.909
HSRN (ours)	0.1	**0.022**	**29.186**	**0.898**	**0.010**	**35.944**	**0.956**

(c) Comparison with CASSI-based reconstruction approaches.

	Scale	TokyoTech-31 → CAVE			CAVE → TokyoTech-31		
		RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑
Zimm.*et al.* [ ]	×1	0.025 (↑13.6%)	33.539 (↓4.5%)	0.917 (↓3.2%)	0.058 (↑114.8%)	29.931 (↓10.7%)	0.895 (↓2.9%)
HSRN (ours)	×1	0.022 (↑4.5%)	35.164 (↓1.6%)	0.948 (↓2.1%)	0.034 (↑47.8%)	31.052 (↓10.2%)	0.918 (↓2.6%)
HSRN (ours)	×2	0.022 (↑37.5%)	34.912 (↓8.4%)	0.930 (↓3.4%)	0.033 (↑50%)	31.687 (↓13%)	0.922 (↓3%)

(b) Cross dataset validation of HSRN and [ ] (increase/decrease percentages in blue).

LBP	3D-SPC	Residual	TokyoTech-31 (w/ shot noise) PSNR↑
✗	✓	✓	30.214
✓	✗	✓	30.521
✓	✓	✗	31.501
✓	✓	✓	**31.832**

(d) Ablation experiments.

Table 2: Quantitative results in different settings.

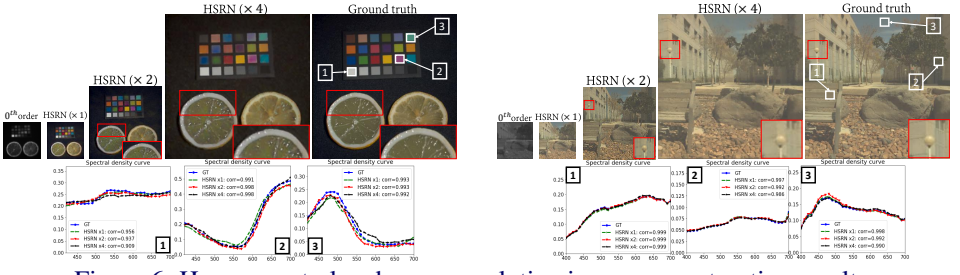

Figure 6: Hyper-spectral and super-resolution image reconstruction results.

Comparison with CASSI-based reconstruction methods: We showcase the suitability of CTIS systems for HSI beyond the spatial resolution limitations by comparing reconstruction performance with HS reconstruction methods designed for CASSI system of HS cubes of size $400 \times 400 \times 29$. Tab. 2(c) compares HSRN, that performs joint spectral reconstruction and $\times 4$ SR, with model-based approaches for CASSI, that optimize directly on a super-resolved measurement corrupted by an aperture mask. Even with a low resolution input HSRN is able to reconstruct HS cubes with higher spatial and spectral accuracy achieving a gain of 6 to 8 dB of PSNR. Visual results in Fig. 7 confirm the numerical ones showing how HSNR restores very fine image details on the *checkerboard* and *butterfly* images.

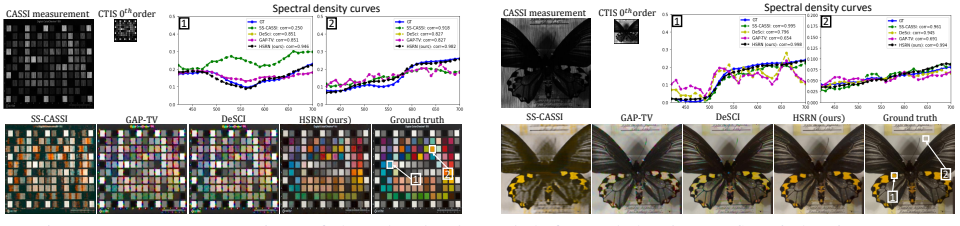

Figure 7: Reconstruction of the *checkerboard* (left) and the *butterfly* (right) images.

Results on real data: We built a CTIS prototype that features a DOE with 12 higher diffraction orders and a 1MP monochrome sensor (see the *suppl. mat* for further details on the real setup and on the data acquisition procedure). HSRN has also been trained and tested on real sensor measurements taken from it. The network generates HS cubes with spatial resolution of 278×278 pixels, that is $\times 2$ the resolution of the 0^{th} diffraction order and 25 spectral bands spanning the range from $455nm$ to $695nm$. We show a sample of a

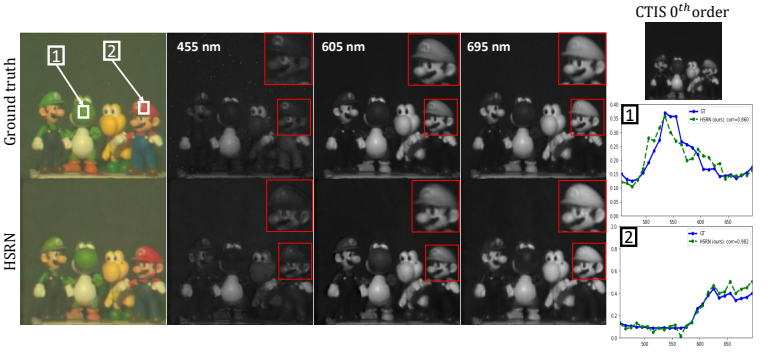

Figure 8: Sample of a reconstructed image from CTIS real data.

reconstructed cube in Fig. 8 in sRGB space together with 3 individual spectral bands and selected spectra of two image regions.

Ablation study: We evaluate the contribution of each module of HSRN by testing the network performance without said module. The network was trained for 250 epochs in all ablation experiments and tested on TokyoTech-31 dataset with shot noise. Tab. 2(d) shows how each component gives a relevant and non-overlapping contribution to the results.

5 Conclusion and Discussion

We proposed a joint approach for HS reconstruction and SR from CTIS data tackling for the first time the major shortcomings of such systems and providing an efficient model capable of performing reconstructions in real-time. By exploiting side information from higher diffraction orders HSRN was able to produce HS cubes with fine spatial details with up to $\times 4$ the spatial resolution of the 0^{th} diffraction order. That being said, the main limitations of this approach are two-fold: *Spectral-wise*, small angles of parallel projection, i.e., the amount the HS cube is smeared in a given projection, may hinder the reconstruction quality as spectral bands severally overlap each other at the sensor and the network struggles to accurately resolve them. *Spatial-wise*, enough higher order projections are needed to reach acceptable reconstruction accuracy specially for large SR factors, e.g.,$\times 4$, as more complementary information would be available which in turns require larger sensor area. Further research will focus on improving the real CTIS system and on evaluating HSRN on real data from the improved setup.

Acknowledgment: This collaborative work was funded by Sony Group Corporation's Research and Development Center (RDC).

References

[1] Mads J Ahlebæk, Mads S Peters, Wei-Chih Huang, Mads T Frandsen, René L Eriksen, and Bjarke Jørgensen. The hybrid approach–convolutional neural networks and expectation maximization algorithm–for tomographic reconstruction of hyperspectral images. *arXiv preprint arXiv:2205.15772*, 2022.

[2] Boaz Arad and Ohad Ben-Shahar. Sparse recovery of hyperspectral signal from natural rgb images. In *European Conference on Computer Vision*, pages 19–34. Springer, 2016.

[3] Gonzalo R Arce, David J Brady, Lawrence Carin, Henry Arguello, and David S Kittle. Compressive coded aperture spectral imaging: An introduction. *IEEE Signal Processing Magazine*, 31(1):105–115, 2013.

[4] José M Bioucas-Dias and Mário AT Figueiredo. A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image processing*, 16(12):2992–3004, 2007.

[5] Nicola Brusco, S Capeleto, M Fedel, Anna Paviotti, Luca Poletto, Guido Maria Cortelazzo, and G Tondello. A system for 3d modeling frescoed historical buildings with multispectral texture information. *Machine Vision and Applications*, 17(6):373–393, 2006.

[6] Theodor V Bulygin and Gennady N Vishnyakov. Spectrotomography: a new method of obtaining spectrograms of two-dimensional objects. In *Analytical Methods for Optical Tomography*, volume 1843, pages 315–322. SPIE, 1992.

[7] Inchang Choi, MH Kim, D Gutierrez, DS Jeon, and G Nam. High-quality hyperspectral reconstruction using a spectral prior. Technical report, 2017.

[8] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007.

[9] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

[10] Clément Douarre, Carlos F Crispim-Junior, Anthony Gelibert, Laure Tougne, and David Rousseau. On the value of ctis imagery for neural-network-based classification: a simulation perspective. *Applied optics*, 59(28):8697–8710, 2020.

[11] Clément Douarre, Carlos F Crispim-Junior, Anthony Gelibert, Gérald Germain, Laure Tougne, and David Rousseau. Ctis-net: a neural network architecture for compressed learning based on computed tomography imaging spectrometers. *IEEE Transactions on Computational Imaging*, 7:572–583, 2021.

[12] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5759–5768, 2022.

[13] Ralf Habel, Michael Kudenov, and Michael Wimmer. Practical spectral photography. In *Computer graphics forum*, volume 31, pages 449–458. Wiley Online Library, 2012.

[14] Nathan A Hagen and Michael W Kudenov. Review of snapshot spectral imaging technologies. *Optical Engineering*, 52(9):090901, 2013.

[15] Weizhe Han, Qianlong Wang, and Weiwei Cai. Computed tomography imaging spectrometry based on superiorization and guided image filtering. *Optics Letters*, 46(9): 2208–2211, 2021.

[16] Jon Yngve Hardeberg, Francis JM Schmitt, and Hans Brettel. Multispectral color image capture using a liquid crystal tunable filter. *Optical engineering*, 41(10):2532–2548, 2002.

[17] Wei-Chih Huang, Mads Svanborg Peters, Mads Juul Ahlebaek, Mads Toudal Frandsen, René Lynge Eriksen, and Bjarke Jørgensen. The application of convolutional neural networks for tomographic reconstruction of hyperspectral images. *Displays*, 74: 102218, 2022.

[18] Yan Jia, Yinqiang Zheng, Lin Gu, Art Subpa-Asa, Antony Lam, Yoichi Sato, and Imari Sato. From rgb to spectrum for natural scenes via manifold-based mapping. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4705–4713, 2017.

[19] David Kittle, Kerkil Choi, Ashwin Wagadarikar, and David J Brady. Multiframe image estimation for coded aperture snapshot spectral imagers. *Applied optics*, 49(36):6824–6833, 2010.

[20] Bruno Lecouat, Jean Ponce, and Julien Mairal. Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2370–2379, 2021.

[21] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017.

[22] Qifeng Li, Yang Wang, Xiangyun Ma, Wenfang Du, Huijie Wang, Xinwei Zheng, and Da Chen. A low-rank estimation method for ctis image reconstruction. *Measurement Science and Technology*, 29(9):095401, 2018.

[23] Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. Fast spatio-temporal residual network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10522–10531, 2019.

[24] Xuelong Li, Yanting Hu, Xinbo Gao, Dacheng Tao, and Beijia Ning. A multi-frame image super-resolution method. *Signal Processing*, 90(2):405–414, 2010.

[25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 136–144, 2017.

[26] Yang Liu, Xin Yuan, Jinli Suo, David J Brady, and Qionghai Dai. Rank minimization for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):2990–3006, 2018.

[27] Ziyi Meng, Zhenming Yu, Kun Xu, and Xin Yuan. Self-supervised neural networks for spectral snapshot compressive imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2622–2631, 2021.

[28] Ajmal Mian and Richard Hartley. Hyperspectral video restoration using optical flow and sparse coding. *Optics express*, 20(10):10658–10673, 2012.

[29] Yusukex Monno, Sunao Kikuchi, Masayuki Tanaka, and Masatoshi Okutomi. A practical one-shot multispectral imaging system using a single image sensor. *IEEE Transactions on Image Processing*, 24(10):3048–3059, 2015.

[30] Takayuki Okamoto and Ichirou Yamaguchi. Simultaneous acquisition of spectral image information. *Optics letters*, 16(16):1277–1279, 1991.

[31] Wallace M Porter and Harry T Enmark. A system overview of the airborne visible/infrared imaging spectrometer (aviris). In *Imaging Spectroscopy II*, volume 834, pages 22–31. SPIE, 1987.

[32] GN Ramachandran and AV Lakshminarayanan. Three-dimensional reconstruction from radiographs and electron micrographs: application of convolutions instead of fourier transforms. *Proceedings of the National Academy of Sciences*, 68(9):2236–2240, 1971.

[33] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016.

[34] Adriano Simonetto, Pietro Zanuttigh, Vincent Parret, Piergiorgio Sartor, and Alexander Gatto. Semi-supervised deep learning techniques for spectrum reconstruction. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7767–7774. IEEE, 2021.

[35] Curtis Earl Volin. *Portable snapshot infrared imaging spectrometer*. PhD thesis, The University of Arizona, 2000.

[36] Ashwin Wagadarikar, Renu John, Rebecca Willett, and David Brady. Single disperser design for coded aperture snapshot spectral imaging. *Applied optics*, 47(10):B44–B51, 2008.

[37] Lizhi Wang, Chen Sun, Ying Fu, Min H Kim, and Hua Huang. Hyperspectral image reconstruction using a deep spatial-spectral prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8032–8041, 2019.

[38] Lizhi Wang, Chen Sun, Maoqing Zhang, Ying Fu, and Hua Huang. Dnu: Deep non-local unrolling for computational spectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1661–1671, 2020.

[39] Larz White, W Bryan Bell, and Ryan Haygood. Accelerating computed tomographic imaging spectrometer reconstruction using a parallel algorithm exploiting spatial shift-invariance. *Optical Engineering*, 59(5):055110, 2020.

[40] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38(4):1–18, 2019.

[41] Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K Nayar. Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. *IEEE Transactions on Image Processing*, 19(9):2241–2253, 2010.

[42] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2539–2543. IEEE, 2016.

[43] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition*, pages 2528–2535. IEEE, 2010.

[44] Jian Zhang and Bernard Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1828–1837, 2018.

[45] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

[46] Shipeng Zhang, Lizhi Wang, Lei Zhang, and Hua Huang. Learning tensor low-rank prior for hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12006–12015, 2021.

[47] Siming Zheng, Yang Liu, Ziyi Meng, Mu Qiao, Zhishen Tong, Xiaoyu Yang, Shensheng Han, and Xin Yuan. Deep plug-and-play priors for spectral snapshot compressive imaging. *Photonics Research*, 9(2):B18–B29, 2021.

[48] Markus Zimmermann, Simon Amann, Mazen Mel, Tobias Haist, and Alexander Gatto. Deep learning-based hyperspectral image reconstruction from emulated and real computed tomography imaging spectrometer data. *Optical Engineering*, 61(5):053103, 2022.