λ-net: Reconstruct Hyperspectral Images from a Snapshot Measurement

Xin Miao[1,†] Xin Yuan[2,*] Yunchen Pu[3] Vassilis Athitsos[1]
[1]University of Texas at Arlington, TX, USA [2]Nokia Bell Labs, NJ, USA [3]Facebook, CA, USA
xin.miao@mavs.uta.edu xyuan@bell-labs.com pyc40@fb.com athitsos@uta.edu

Abstract

*We propose the λ-net, which reconstructs hyperspectral images (*e.g.*, with 24 spectral channels) from a single shot measurement. This task is usually termed snapshot compressive-spectral imaging (SCI), which enjoys low cost, low bandwidth and high-speed sensing rate to capture the three-dimensional (3D) signal* i.e.*, (x, y, λ), using a 2D snapshot. Though proposed more than a decade ago, the poor quality and low-speed of reconstruction algorithms preclude wide applications of SCI. To address this challenge, in this paper, we develop a dual-stage generative model to reconstruct the desired 3D signal in SCI, dubbed λ-net. Results on both simulation and real datasets demonstrate the significant advantages of λ-net, which leads to $>$4dB improvement in PSNR on simulation data compared to the current state-of-the-art. Furthermore, λ-net can finish the reconstruction task within sub-seconds instead of hours taken by the most recently proposed DeSCI algorithm, thus speeding up the reconstruction $>$1000 times.*

1. Introduction

Snapshot compressive-spectral imaging (SCI) refers to compressive imaging systems where multiple hyperspectral frames are mapped into a single measurement [6, 12, 22, 39, 61, 62, 84]. The first SCI system, called coded aperture snapshot spectral imaging (CASSI), was developed in [22], which modulates signals at different wavelengths by a coded aperture (physical mask) and a disperser [61]. In this manner, a two-dimensional (2D) monochromatic camera can sample the hyperspectral scenes at video rate [62] and thus saves memory, bandwidth and cost significantly compared with that using a traditional spectrometer in addition to the high-speed sensing. While enjoying all these advantages, similar to other computational imaging systems [4], one important step in SCI is that *algorithms* are required to reconstruct the 3D hyperspectral data-cube from every snapshot measurement after the sensing process. Existing algorithms are either too slow or the performance is not high. Inspired by the recent advances of deep learning for inversion problems [13, 36, 43, 74, 82], in this paper, we propose λ-net for SCI reconstruction.

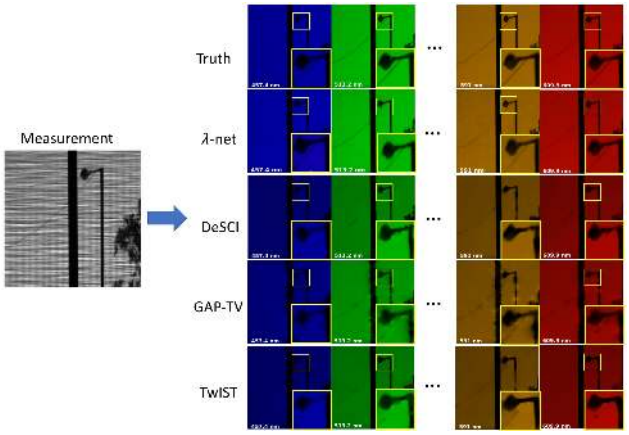

Figure 1. Hyperspectral images (right) reconstructed from a single shot measurement (left) using various algorithms: the proposed λ-net (PSNR: 30.0dB), DeSCI (PSNR: 22.4dB), GAP-TV (PSNR: 16.6dB) and TwIST (PSNR: 13.1dB), compared with the ground truth. Four out of 24 reconstructed frames at different wavelengths are shown. Notice that only λ-net can recover the continuous wire.

Fig. 1 depicts that different algorithms lead to various quality reconstructed images, where TwIST (Two-Step Iterative Shrinkage/Thresholding) [10] and GAP-TV (Generalized Alternating Projection based Total Variation) [76] are used in previous CASSI systems [39, 61, 84] as baselines and the most recently proposed algorithm, decompress SCI (DeSCI), introduced in [39], has achieved state-of-the-art results in both video and spectral SCI. Note that all these algorithms reconstruct 24 spectral images (channels) from a single measurement (Fig. 1 left) captured by the camera (Fig. 2 top), and 4 selected channels are plotted in the right of Fig. 1. It can be observed that TwIST leads to blurry results and GAP-TV provides unpleasant artifacts; DeSCI offers higher quality images than both of them but leads to over-smooth phenomenon. By contrast, our proposed λ-net has led to significant better results ($>$7dB in PSNR for the scene in Fig. 1) than all these previous methods, and only λ-net can reconstruct the continuous wire in Fig. 1. Further-

[†]Part of this work was performed when Xin Miao was a summer intern at Nokia Bell Labs in 2018. *Corresponding author. The code is available at https://github.com/xinxinmiao/lambda-net.

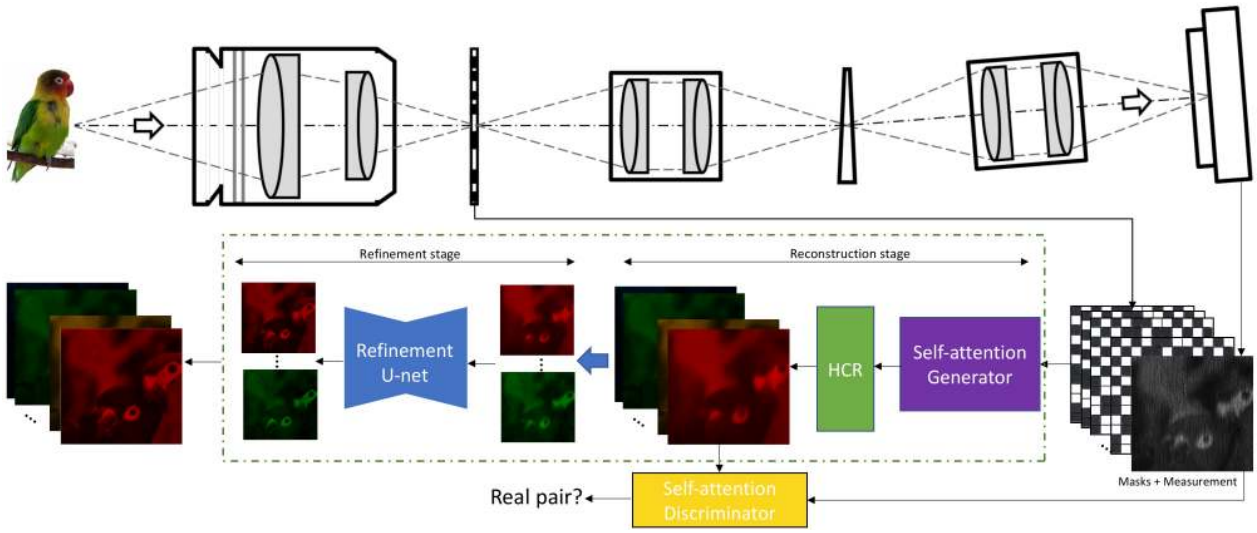

Figure 2. Top: Imaging process of SCI. The hyperspectral scene is imaged onto the coded aperture and after relayed by lenses, the prism spreads the light to different spatial locations for different wavelength and then captured by the camera (top right). Though a single mask is used, because of the disperser, signals at different wavelengths are modulated by shifted versions of the mask, thus differently. Bottom: These masks (shifted versions of the fixed mask in the upper part) along with the measurement (bottom right) are fed into the λ-net to reconstruct the hyperspectral data-cube (bottom left). Two stages exist in the proposed λ-net, where the first stage (reconstruction stage) consists of a self-attention GAN plus a hierarchical channel reconstruction (HCR) strategy to generate the 3D cube from masks and measurements and the second stage (refinement stage) refines the hyperspectral images in each channel.

more, speed is an important metric for different algorithms, especially for imaging. DeSCI, GAP-TV and TwIST are optimization based algorithms; while GAP-TV can finish the task, *e.g.*, reconstructing a $256 \times 256 \times 24$ pixels data-cube from a 256×256 measurement, in 30 seconds and TwIST usually needs 10 minutes, DeSCI requires about an hour on a desktop with a 12-core i7 CPU and 64G RAM. By contrast, our proposed λ-net can reconstruct the hyperspectral data-cube within 1 second on the same CPU, and within 33 milliseconds (ms) on a NVIDIA GTX 1080 Ti GPU. We understand these algorithms are running on different platforms and λ-net needs pre-training before performing the task. However, bearing these numbers in mind, we can anticipate that λ-net along with CASSI can provide *real-time* 3D hyperspectral imaging and reconstruction when the camera is working at 30 frames per second, and thus can be applied in our daily life.

Though deep learning based algorithms have started being used in computational imaging systems [55, 36, 46, 54, 13, 74, 82], significant challenges and questions exist in SCI reconstruction using deep learning.

1) Limited training dataset is available. Though some datasets [1, 2, 5] are available, the spectral wavelengths are usually different for different imaging systems. In order to overcome this challenge, in addition to the generally used data argumentation techniques, we further use the *spectral interpolation* to unify the datasets to the same set of wavelengths.
2) The measurement of SCI is a single frame, while more than 20 spectral channels (24 is used in our experiments) are to be generated (reconstructed). Therefore, a deep (generative) model is expected to be used. However, this is challenging due to the large number of parameters in the network and the limited dataset mentioned above.
3) The third question this paper aims to address is that is it possible to adopt a small network to boost up the quality of SCI reconstruction?

Bearing these challenges and questions in mind, this paper makes the following contributions.

i) A generative model based on U-net [52] is developed to reconstruct the 3D spectral cube from the SCI measurement and masks. The self-attention generative adversarial network (GAN) [87] is integrated with the U-net to exploit the non-local correlation in the spectral images.

ii) A hierarchical channel reconstruction (HCR) strategy is proposed to *progressively* reconstruct spectral channels based on the features extracted by the neural network and previous reconstructed channels. This HCR strategy plus self-attention GAN constitutes the *reconstruction stage* of our λ-net (Fig. 2).

iii) A *refinement stage* composed of a small U-net and residual learning [24] is developed to boost up the quality of reconstructed images from the first stage. In this stage, each channel is performed independently.

iv) We have verified our proposed λ-net on extensive "real-mask-in-the-loop" simulation data and also the *real data* captured by the CASSI camera [62]. λ-net offers much better results than DeSCI (and other deep learning methods) and can finish the reconstruction in sub-seconds.

2. Snapshot Compressive-spectral Imaging

As demonstrated in the top part of Fig. 2, in CASSI [22, 61], the spectral scene is collected by the objective lens and spatially coded by a fixed mask. Then the coded scene is spectrally dispersed by the disperser. Following this, the spatial-spectral coded scene is detected by the charge-coupled device (CCD). A snapshot on the CCD thus encodes tens of spectral bands of the scene. The number of coded frames for a snapshot is determined by the dispersion property of the dispersive element and the pixel sizes of the mask and the CCD. Consider B-frames (spectral channels) are modulated and encoded in SCI and each frame has n (=$n_x \times n_y$) pixels. Without considering optical details, mathematically, the measurement in SCI can be modeled by [61]

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{g}\,, \tag{1}$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{n\times nB}$ is the sensing matrix, $\boldsymbol{x} \in \mathbb{R}^{nB}$ is the desired signal, and $\boldsymbol{g} \in \mathbb{R}^n$ denotes the noise.

Though Eq. (1) has the formulation similar to compressive sensing (CS) [17, 15], unlike traditional CS, the sensing matrix considered here is not a dense matrix, and it does not satisfy the restricted isometry property. In SCI, the matrix $\boldsymbol{\Phi}$ has a very specific structure and can be written as $\boldsymbol{\Phi} = [\boldsymbol{D}_1, \ldots, \boldsymbol{D}_B]$, where $\{\boldsymbol{D}_k\}_{k=1}^B$ are diagonal matrices defined by the following mask. Specifically, consider that B spectral frames $\{\boldsymbol{X}_k\}_{k=1}^B \in \mathbb{R}^{n_x\times n_y}$ are modulated by shifted versions of the fixed mask, $\{\boldsymbol{C}_k\}_{k=1}^B \in \mathbb{R}^{n_x\times n_y}$, correspondingly (Fig. 2, bottom right). The measurement $\boldsymbol{Y} \in \mathbb{R}^{n_x\times n_y}$ is given by

$$\boldsymbol{Y} = \textstyle\sum_{k=1}^B \boldsymbol{X}_k \odot \boldsymbol{C}_k + \boldsymbol{G}\,, \tag{2}$$

where $\odot$ denotes the element-wise product, and $\boldsymbol{D}_k = \mathrm{diag}(\mathrm{vec}(\boldsymbol{C}_k))$, for $k = 1, \ldots, B$. For all B pixels (in the B frames) at position (i, j), $i = 1, \ldots, n_x$; $j = 1, \ldots, n_y$, they are collapsed to form one pixel in the snapshot measurement as $y_{i,j} = \sum_{k=1}^B c_{i,j,k}x_{i,j,k} + g_{i,j}$. By defining $\boldsymbol{x} = \left[\boldsymbol{x}_1^\mathsf{T}, \ldots, \boldsymbol{x}_B^\mathsf{T}\right]^\mathsf{T}$, where $\boldsymbol{x}_k = \mathrm{vec}(\boldsymbol{X}_k)$, we have the vector formulation of Eq. (1). Thus, $\boldsymbol{x} \in \mathbb{R}^{n_x n_y B}$, $\boldsymbol{\Phi} \in \mathbb{R}^{n_x n_y \times (n_x n_y B)}$, and the compressive sampling rate in SCI is equal to $1/B$. It has been proved recently in [30, 31] that the reconstruction of SCI is bounded even when $B > 1$.

3. λ-net

The target of λ-net is to reconstruct the hyperspectral image cube from the single measurement captured by the SCI camera. Recently, GAN [23] and variational autoencoder (VAE) [33] become the most convincing generative models and are denominating the recent emerging researches in deep learning [14, 63]. It has also been suggested that using U-net as the generative model in GAN is capable of solving diverse problems [29, 47, 89]. In our task, in addition to the U-net plus GAN, the most recently proposed self-attention mechanism is adapted to exploit both the non-local similarity of spatial textures and the long-range spectral similarity. Furthermore, we propose an additional HCR strategy to gradually reconstruct all channels which guarantees the quality of result and the accuracy of spectral information.

λ-net reconstructs hyperspectral images with B (24 in our experiments) spectral channels that is high dimensional data. Even a deep U-net and HCR are used, it still does not guarantee to reconstruct high quality images due to the large number of parameters and the limited training data. In order to overcome this challenge, we propose to use another *refinement* U-net which is shallower than the first U-net in the reconstruction stage. This refinement stage improves the image quality of each spectral channel separately.

3.1. Reconstruction Stage

The reconstruction stage outputs the hyperspectral images and it aims to extract both spatial and spectral information from the measurement.

3.1.1 Conditional GAN

Unlike the unconditional (original) GAN, the discriminator in conditional GAN (cGANs) [45] can also observe the inputs from the generator. cGAN is appropriate for our SCI reconstruction as we aim to generate corresponding output hyperspectral images conditional on the input measurement and masks. Specifically, the inputs masks are fixed (in a prebuilt SCI system) while the input measurement depends on the captured scene. Thereby, the masks are not necessary to be observed by the discriminator. The objective function of our cGAN can be expressed as

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}}(G, D) = \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}}[\log D(\boldsymbol{y}, \boldsymbol{x})] \\ + \mathbb{E}_{\boldsymbol{y}}[\log(1 - D(\boldsymbol{y}, G(\boldsymbol{y}, \boldsymbol{\Phi})))],\end{aligned} \tag{3}$$

where G and D denotes the generator and discriminator, respectively.

3.1.2 Deeper U-net with Self-Attention

The U-net architecture detailed in Fig. 3 is used as the generator in our cGAN. As mentioned before, our output hyperspectral images and the input measurement share similar spatial structures, *e.g.*, location of edges. The encoder and decoder can help capture the shared low level information between input and output and remove the noise; but it may also lose the location information from the measurement. To tackle this challenge, we add the skip connection to help the location information pass through the network. Furthermore, since we are reconstructing high dimensional hyperspectral images, we employed a deeper U-net. In particular, we have 3 times convolution operations with stride 1 after

the downsampling or upsampling (which is 2 in [52]); we also have 5 times downsampling and upsampling in the encoder and decoder of U-net instead of 4. Experiment results in Sec. 5.2 (Table 2) show that our deeper U-net achieves better (1.92dB in PSNR) results than the original U-net.

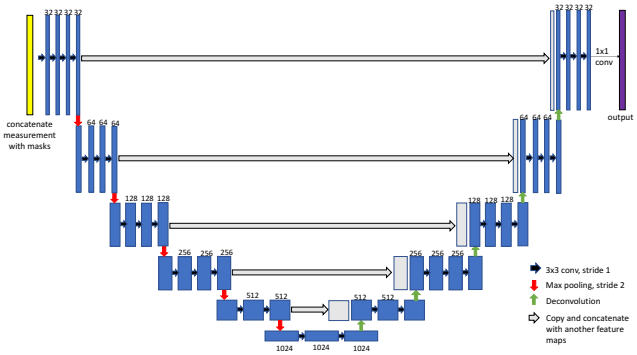

Figure 3. U-net architecture used in the reconstruction stage of our network.

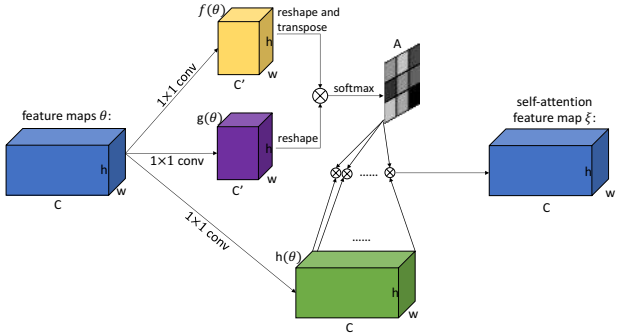

Figure 4. The self-attention module in our framework; softmax is applied to each row.

Attention module has been widely used in many computer vision tasks [86, 44, 38, 27]. Since the convolution operator in U-net has a local receptive field, only multiple convolutional layers can capture the long range dependencies. Via adding the self-attention, the network can learn the long range similarity in one layer easily. In our self-attention layer, all spectral channels share the same attention map [60], as we not only want to capture the long range dependencies in space but also to keep the spectral similarity in SCI reconstruction. This self-attention (Fig. 4) is not only used in the generator but also in the discriminator. We have performed the experiments by adding self-attentions to different layers of the network and found that imposing it on the middle-to-high layer feature maps will lead to better results, but with larger attention maps. Limited by the GPU memory, we show results by imposing the self-attention to the layer who has 256 feature maps before the deconvolution in the decoder of the U-net in Fig.3.

As depicted in Fig. 4, let $\boldsymbol{\theta} \in \mathbb{R}^{c\times h\times w}$ denote the feature map that we want to impose the self-attention. By using 1×1 convolutions on $\boldsymbol{\theta}$, we can get three feature spaces

$$f(\boldsymbol{\theta}) \in \mathbb{R}^{c'\times h\times w},\ \ g(\boldsymbol{\theta}) \in \mathbb{R}^{c'\times h\times w},\ h(\boldsymbol{\theta}) \in \mathbb{R}^{c\times h\times w},$$

where c' is an integer and we set $c' = \frac{c}{8}$ in our experiments. We now use $\{f(\boldsymbol{\theta}),\ g(\boldsymbol{\theta})\}$ to calculate the attention map. First, we reshape them to 2D matrices $\{f'(\boldsymbol{\theta}),\ g'(\boldsymbol{\theta})\} \in \mathbb{R}^{c'\times N}$, with $N = h \times w$; then each entry of the attention map $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is calculated by

$$a_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N}\exp(s_{ij})},\ \text{with}\ \ s_{ij} = f'(\theta_i)^{\mathsf{T}} g(\theta_j). \tag{4}$$

Here, $a_{j,i}$ represents that the extent of the model depends on the i^{th} location when generating the j^{th} region. This attention map $\boldsymbol{A}$ is then incorporated with the feature space $h(\boldsymbol{\theta})$. We first reshape $h(\boldsymbol{\theta})$ to $h'(\boldsymbol{\theta}) \in \mathbb{R}^{c\times N}$ and then impose $\boldsymbol{A}$ on it, which arrives

$$\boldsymbol{\xi}' = \boldsymbol{A}\, h'(\boldsymbol{\theta})^{\mathsf{T}} \in \mathbb{R}^{N\times c}. \tag{5}$$

Following this, we reshape each channel (column) in $\boldsymbol{\xi}'$ to get the output of the attention layer $\boldsymbol{\xi} \in \mathbb{R}^{c\times h\times w}$. Lastly, we multiply the output of the attention layer $\boldsymbol{\xi}$ by a *scale learnable* parameter γ and add it back to the input feature map $\boldsymbol{\theta}$. This leads to the final result

$$\boldsymbol{z} = \gamma\boldsymbol{\xi} + \boldsymbol{\theta}. \tag{6}$$

3.1.3 Hierarchical Channel Reconstruction

It is challenging to reconstruct all 24 channels images from a single measurement in one shot. Therefore, we propose a progressive reconstruction scheme, *i.e.*, Hierarchical Channel Reconstruction (HCR). HCR tries to recover a fraction of the spectral channels and then reconstruct the entire channels based on the information we have recovered.

In our experiment, 24 spectral channels need to be reconstructed. We first reconstruct $[\boldsymbol{x}_1, \boldsymbol{x}_5, \boldsymbol{x}_9, \boldsymbol{x}_{13}, \boldsymbol{x}_{17}, \boldsymbol{x}_{21}]$ spectral channels with an interval of 4. Then we reconstruct the $[\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_5, \ldots, \boldsymbol{x}_{23}]$ spectral channels with an interval of 2. Finally, all the 24 channels are reconstructed. The residual learning method is also employed. Details of the proposed HCR are showed in Fig. 5. In this manner, our λ-net reconstructs the hyperspectral images gradually, where we have decomposed the $1 \rightarrow 24$ problem to $1 \rightarrow 6 \rightarrow 12 \rightarrow 24$ cascaded problems. In other words, if we can reconstruct partial spectral channels with correct spectral information, a simple interpolation method should be qualified to reconstruct the entire channels. Table 2 shows HCR has improved the performance of λ -net.

We define the intermediate outputs and the final output as $I_1(\boldsymbol{y}, \boldsymbol{\Phi})$, $I_2(\boldsymbol{y}, \boldsymbol{\Phi})$ and $G(\boldsymbol{y}, \boldsymbol{\Phi})$, respectively. The target of λ-net is to reconstruct the signal and thus it is reasonable to add the ℓ_2 loss into our objective function,

$$\begin{aligned}\mathcal{L}_{\ell_2}(G) =& \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}}[\|\boldsymbol{x}^1 - I_1(\boldsymbol{y}, \boldsymbol{\Phi})\|_2 + \|\boldsymbol{x}^2 - I_2(\boldsymbol{y}, \boldsymbol{\Phi})\|_2 \\ &+ \|\boldsymbol{x} - G(\boldsymbol{y}, \boldsymbol{\Phi})\|_2],\end{aligned} \tag{7}$$

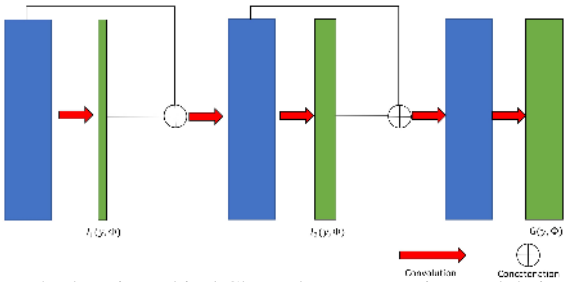

Figure 5. The Hierarchical Channel Reconstruction module in our experiment.

where $\boldsymbol{x}^1 \overset{\text{def}}{=} [\boldsymbol{x}_1, \boldsymbol{x}_5, \boldsymbol{x}_9, \boldsymbol{x}_{13}, \boldsymbol{x}_{17}, \boldsymbol{x}_{21}]$ and $\boldsymbol{x}^2 \overset{\text{def}}{=} [\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_5, \ldots, \boldsymbol{x}_{23}]$. Eq. (7) denotes that the generator not only aims to fool the discriminator but also enforces the output close to the ground truth. Our final objective is

$$(G^*, D^*) = \arg\min_G \max_D \mathcal{L}_{\text{GAN}}(G, D) + \alpha \mathcal{L}_{\ell_2}(G), \quad (8)$$

where α is a parameter to balance these two terms. Via integrating this HCR strategy with self-attention GAN, we have the output of the reconstruction stage

$$\boldsymbol{x}' = G^*(\boldsymbol{y}, \boldsymbol{\Phi}) = \left[(\boldsymbol{x}'_1)^\mathsf{T}, \ldots, (\boldsymbol{x}'_B)^\mathsf{T}\right]^\mathsf{T}, \quad (9)$$

which is the desired 3D hyperspectral image.

3.2. Refinement Stage

The reconstruction stage can capture the spectral information of the hyperspectral image cube but it doesn't have sufficient capability to offer high quality images, especially the spatial resolution. Otherwise, an even deeper network should be used but this will require larger training datasets. To overcome this challenge, we propose the refinement stage to enhance the reconstruction quality. The input for refinement stage is a single frame instead of all spectral channels in one shot. In this manner, the network treats each spectral channel as an independent image, and it can extract the information across all spectral channels. Given the fact that the input and output images share the same structure, we use another U-net as the basic architecture in the refinement stage, but this time we output a single frame with high quality. Since each frame is of a small size, a shallow U-net is sufficient for this task, *i.e.*, 4 times down-sampling or deconvolution in the encoder and decoder, respectively. Furthermore, we also add the *residual learning* to the input image, which has improved (1.27dB in PSNR in Table 2) the final results.

We pass every single frame in the hyperspectral image cube obtained by the reconstruction stage to the refinement stage. The ℓ_2 loss between the ground truth and the output of the refinement stage is used as the objective function

$$\mathcal{L}_{\ell_2}(\text{refine}) = \mathbb{E}_{\boldsymbol{x}_i, \boldsymbol{x}'_i}[\|\boldsymbol{x}_i - \boldsymbol{x}'_i\|_2], \quad \forall i = 1, \ldots, B. \quad (10)$$

We train the network in the reconstruction stage first and fix the parameters; then we sent the results to the refinement stage to train the second U-net. This separate training strategy is mainly due to the size difference of the data. As mentioned above, the reconstruction stage outputs the 3D hyperspectral image cube but the refinement stage processes each spectral frame independently. It is possible to train both networks jointly. However, since each batch in the reconstruction stage contains all channels of the same scene, while in the refinement stage, we hope each batch consisting of different scenes (probably at different spectral channels, too), we may need a huge memory to save these data and parameters. Limited by the GPU memory, we perform our experiments via separate training.

4. Related Work

Generally speaking, the SCI problem we are interested in this paper belongs to computational imaging (CI) [4]. Different from traditional imaging, where the user captures the desired signal directly, in CI, the captured measurement is usually not the signal itself, but includes the signal in a complicated way and reconstruction algorithms are required to recover the signal from the measurement. Inspired by CS, various compressive imaging systems [16] have been built to capture high-dimensional data, from videos [21, 26, 50, 40, 79, 56, 57, 58, 81, 80, 83, 85], hyperspectral images [12, 22, 61, 62, 84], dynamic range [75] to depth [41, 49, 77] and polarization [59], *etc*. Regarding the spectral compressive imaging, following CASSI, which used a coded aperture and a prism to implement the wavelength modulation, other modulations such as occlusion mask [11], spatial light modulator [84] and digital-micromirror-device [70] have also been used. Meanwhile, advances of CASSI have also been developed by using multiple-shots [34], dual-channel [64, 65, 66, 67] and high-order information [9]. A parallel research is the mask design for spectral compressive imaging [7, 8, 19, 20, 25], which usually requires multiple measurements.

Another important research direction is the algorithm design. In addition to the NeAREst proposed in [62], various optimization algorithms, such as TwIST [10], GPSR [18] and GAP-TV [76] have been utilized. Other algorithms, such as Gaussian mixture models and sparse coding [51, 73, 67] have also been developed. Most recently, DeSCI proposed in [39] to reconstruct videos or hyperspectral images in SCI has led to state-of-the-art results. The only drawback of DeSCI is the running time, which usually takes hours to reconstruct 24 channel spectral images. To address this and inspired by the recent advances of deep learning on image restoration [71, 88], researchers have started using deep learning in computational imaging [13, 28, 32, 36, 46, 42, 54, 72, 82, 78]. Most recently, deep learning models have been used to reconstruct hyper-

spectral images from RGB images [3, 35, 37, 48, 53]. However, this is different from our problem, which aims to reconstruct hyperspectral images from a coded measurement as designed in CASSI (Fig. 2). A recent paper related to our work is [68], which employed convolutional neural networks to jointly learn the mask design and reconstruction. However, similar to the pioneer work in [36], a *repeated pattern* is used, which is very challenging or even unrealistic in real cameras [82].

5. Experiments

We compare λ-net with several state-of-the-art methods including TwIST [10], GAP-TV [76], and DeSCI [39]. We have also tried the sparse coding algorithms in [66, 67]; they perform worse than DeSCI and take even longer time to run. Similar cases exist in other algorithms [73, 84] and thus ignored here due to space limit. Both peak-signal-to-noise-ratio (PSNR) and structural similarity (SSIM) [69] are used as metrics to evaluate the performance. As mentioned earlier, the most recently proposed DeSCI algorithm delivers state-of-the-art results [39]. The λ-net consistently produces high performance results and surpasses DeSCI in the "Real-Mask-in-the-Loop" (MIL) simulation data (Figs 7-8 and Table 1). Hereby the MIL-simulation denotes that we generate the measurement using *real masks* captured by the CASSI camera, rather than randomly generated ones. It is well known that the real captured data have noise inside and thus the problem is more challenging. On real data (we can only have a single real data with ground truth from the authors of CASSI), our λ-net has also achieved better results than DeSCI (Figs 9-10).

Though our λ-net is the first network developed for CASSI reconstruction for *real* data, we do compare with some other networks even they are developed for other tasks. With some modifications, we have compared λ-net with the networks developed in [35, 37, 53] for CASSI reconstruction.

Figure 6. 16 testing scenes used in the experiments.

5.1. Training

All experiments are performed on a NVIDIA GTX 1080 Ti GPU. For a testing scene with size $256 \times 256 \times 24$, our framework can finish the reconstruction stage in 23ms (0.6s on CPU). In the refinement stage, every frame of the scene can be processed in parallel and finished within 10ms (0.4s on CPU). Without using the GPU, λ-net can finish both stages on an i7 CPU within 1 second.

5.1.1 Data Augmentation

The data to train and validate the model is downloaded from [5]. We manually chose 80 hyperspectral images as our training data to avoid the test scenes (Fig. 6) and training data having the same content. Besides randomly flipping the image, we also randomly rotate, scale, and translate the training images. The original dataset have a uniform resolution of $1392 \times 1300 \times 31$ in wavelength range from 400nm to 700nm with a 10nm interval, while the real data captured by the SCI camera has 24 channels from 400nm to 700nm, but with different intervals, *i.e.*, with wavelengths: {398.62, 404.40, 410.57, 417.16, 424.19, 431.69, 439.70, 448.25, 457.38, 467.13, 477.54, 488.66, 500.54, 513.24, 526.8., 541.29, 556.78, 573.33, 591.02, 609.93, 630.13, 651.74, 674.83, 699.51}nm. To mitigate this issue, we use the *spectral interpolation* to unify the datasets to the same wavelength set as in [61]. Specifically, we perform data interpolation for every spatial location. The hyperspectral images generated by our data augmentation are of size $1392 \times 1300 \times 24$.

5.1.2 Training Details

Table 1. PSNR in dB (left entry in each cell) and SSIM (right entry) of 16 different scenes reconstructed by different algorithms.

Algorithm	λ-net	GAP-TV	TwIST	DeSCI
Scene 1	**36.29**, **0.925**	29.48, 0.800	26.77, 0.772	31.51, 0.896
Scene 2	**30.07**, **0.929**	16.58, 0.805	13.14, 0.753	22.39, 0.806
Scene 3	**34.19**, **0.940**	21.48, 0.769	23.66, 0.738	24.92, 0.822
Scene 4	28.90, 0.899	26.49, 0.822	26.08, 0.861	**29.78**, **0.907**
Scene 5	**34.58**, **0.890**	26.63, 0.688	22.45, 0.695	29.02, 0.844
Scene 6	**28.09**, **0.858**	22.81, 0.614	20.11, 0.662	24.75, 0.797
Scene 7	**36.15**, **0.942**	24.95, 0.699	26.20, 0.753	29.68, 0.881
Scene 8	**32.64**, **0.909**	21.26, 0.695	18.38, 0.643	25.58, 0.823
Scene 9	**33.83**, **0.912**	29.94, 0.812	28.09, 0.807	32.86, 0.937
Scene 10	**28.63**, **0.877**	23.04, 0.706	20.84, 0.620	24.00, 0.748
Scene 11	**35.21**, **0.946**	24.07, 0.754	21.75, 0.785	28.19, 0.912
Scene 12	**34.77**, **0.823**	28.99, 0.758	26.75, 0.699	31.80, 0.863
Scene 13	**32.07**, **0.844**	27.57, 0.650	24.54, 0.718	30.91, 0.823
Scene 14	**33.73**, **0.869**	28.54, 0.764	26.27, 0.765	29.69, 0.852
Scene 15	**29.88**, **0.913**	25.80, 0.801	23.84, 0.765	27.45, 0.864
Scene 16	**30.54**, **0.855**	11.99, 0.293	20.50, 0.511	19.42, 0.305
average	**32.29**, **0.896**	24.35, 0.715	23.09, 0.722	27.62, 0.818

We randomly crop $256 \times 256 \times 24$ patches from the data obtained by the data augmentation. The batch size is set to 20. We alternately update the parameters in G and D in the reconstruction stage; α in Eq. (8) is set to 200. The input of the generator is the concatenation of measurement and masks (Fig. 2 bottom-right). We have performed the experiments to show that this performs better (2.09dB PSNR improvement) than only input the measurement to the network in Table 2. During testing, we input every single channel of the hyperspectral image cube obtained by the reconstruction stage to the refinement stage. Then we collect these $B = 24$ channel high quality images as the final output result.

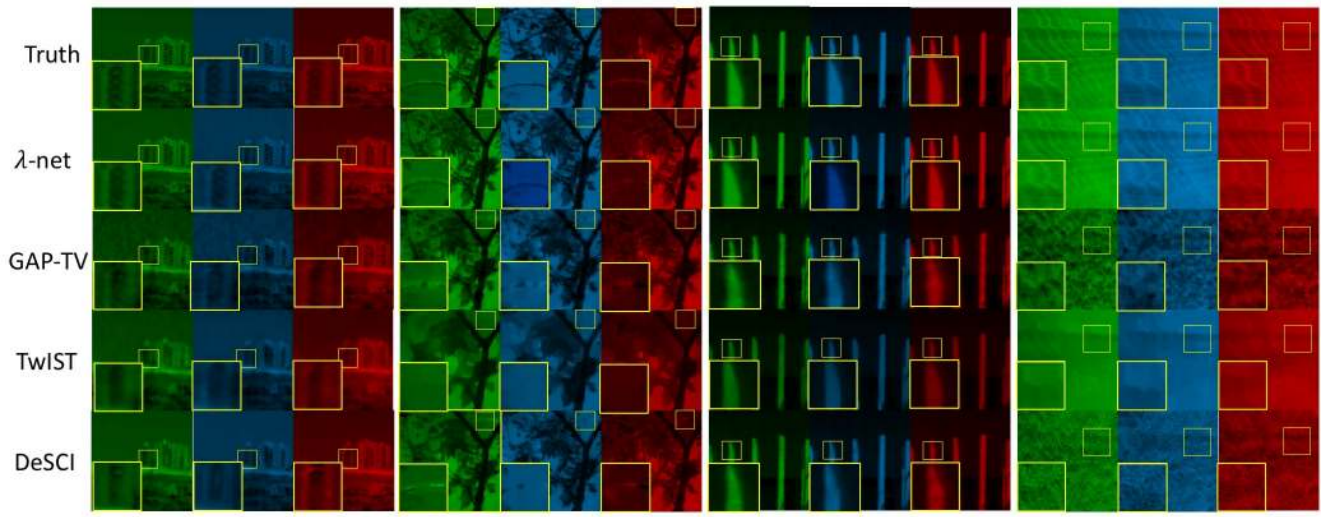

Figure 7. Example reconstructed images by 4 algorithms for four scenes (from left to right: Scene 7, 10, 11, 16 in Fig. 6). The three frames are at wavelengths 477.5nm, 526.8nm, and 630.1nm. Results of all channels can be found in the supplementary material (SM).

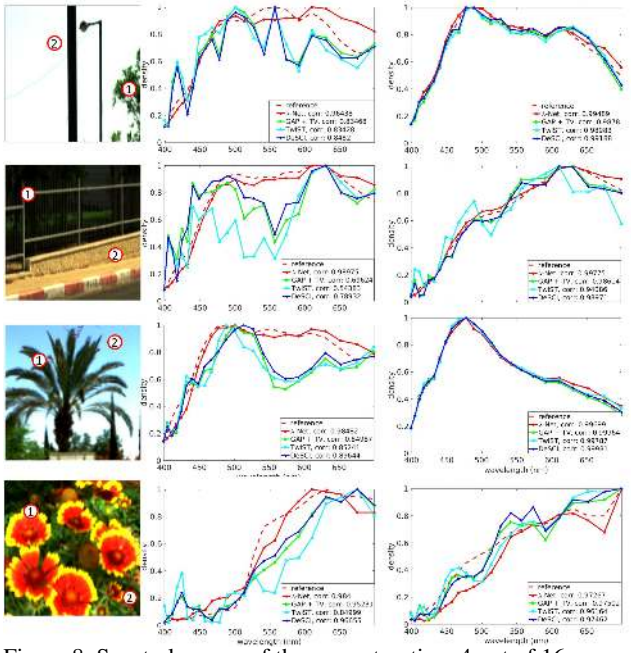

Figure 8. Spectral curves of the reconstruction, 4 out of 16 scenes are selected as examples with two regions in each scene.

5.2. "Real-Mask-in-the-Loop" Simulation Results

As mentioned above, in the MIL-simulation, we generate the measurements using the real captured mask and the hyperspectral images consist of 24 spectral frames with each of size 256×256 pixels. We have 16 testing scenes (Fig. 6) from the dataset [5]. The generated measurements and masks are used to reconstruct the hyperspectral images by different algorithms. Table 1 lists the average PSNR and SSIM of these 16 scenes by using all four algorithms. It can be seen that in average, our λ-net surpasses the best previous method DeSCI 4.67dB. The only exception is Scene 4, which is a simple scene with a large area being the same white screen. This fits the rank minimization model in DeSCI and thus DeSCI offers 0.88dB higher PSNR. λ-net performs better than DeSCI on all other scenes. Exemplar reconstructed frames of various algorithms compared with the truth are shown in Fig. 7. Obviously, λ-net can provide both large-scale structures and fine details of the scene. GAP-TV usually leads to blob artifacts and TwIST provides blocky artifacts. DeSCI offers better results than GAP-TV and TwIST; however, as observed in [39], it usually leads to over-smooth reconstruction. One important metric to evaluate the SCI algorithm is how good the spectral information they can reconstruct as different objects have different spectral information, *e.g*. sky, tree, wall, *etc*. We plot the spectral curves of a small region and calculate the correlation between the reconstruction and ground truth in Fig. 8. Compared with other methods, λ-net provides higher correlation values for different objects. This clearly demonstrates that λ-net can extract more spectral information than other methods.

To quantitatively investigate different blocks of our proposed λ-net, we performed experiments with partial components in λ-net, *e.g*., without GAN, without self-attention, with results summarized in Table 2. It can be seen that all components play important roles in our λ-net; *e.g*., without GAN, the results degraded 2.81dB in PSNR; without self-attention, the results degraded 3.52dB in PSNR, and without the refinement stage, the results degraded 1.62dB in PSNR. As mentioned before, masks contain useful information, and thus using masks along with the measurement improved the results by 2.09dB in PSNR. Furthermore, residual learning in the refinement U-net has led to 1.27dB improvement in PSNR and HCR imporves the result for 0.48dB.

Table 2. Comparison using different components of the model. For each column, $\surd$ means used and $\times$ means not used.

Reconstruction stage	U-net [52]	×	×	×	×	×	×	√	×
	Deep U-net1	√	√	√	√	√	√	×	√
	GAN	√	√	×	√	√	√	√	√
	Self-attention	√	√	√	×	√	√	√	√
	HCR	√	×	√	√	√	√	√	√
Refinement stage	U-net2	√	√	√	√	×	√	√	√
	residual learning	√	√	√	√	×	√	√	×
inputs	measurement+masks	√	√	√	√	√	×	√	√
	measurement	×	×	×	×	×	√	×	×
result PNSR result SSIM		32.29 0.896	31.81 0.882	29.48 0.860	28.77 0.854	30.67 0.873	30.20 0.866	30.37 0.870	31.02 0.878

As mentioned before, we have also compared our λ-net with other networks, with our modifications for CASSI reconstruction. The results are summarized in Table 3, where we can observe that λ-net provides significant better results than other networks.

Table 3. Compare with other deep networks

Network	Simu PSNR	Simu SSIM	Real PSNR
λ-net	32.29	0.896	25.59
[35]	27.42	0.750	21.42
[37]	26.78	0.735	21.09
[53]	29.07	0.836	23.77

5.3. Real Data Results

The bird measurement data is captured by the CASSI system [61]. consisting of 24 spectral frames with each of size 1021×703 pixels. Due to the limitation of GPU memory, we used 416×416 pixels to perform our experiments[1]. In Fig. 9, we visualize the reconstruction results of 6 channels using 4 algorithms. We can see that λ-net can provide marginally better (0.4dB) results than DeSCI and about 1dB higher PSNR than GAP-TV and TwIST. Notably, only λ-net can reconstruct the last frame at wavelength 699.5nm. Exemplar spectral curves are shown in Fig. 10. Owing to the mismatch between the training dataset and this real data, the spectra are not perfect; even this, λ-net can still offer higher or comparable correlation values with other three algorithms.

As mentioned before, we only have one real data, *i.e.*, the bird data, with ground truth captured by CASSI. To further verify the universality of our λ-net, we have modified the network to the video CS system [40]. The results are comparable with DeSCI (shown in the SM).

6. Conclusions

This paper aims to address the challenging problem in spectral compressive imaging: the slow reconstruction. Inspired by the recent advances of deep learning, especially the emerging generative models, we have built a two-stage reconstruction network to recover the hyperspectral images from a snapshot measurement.

By integrating U-net into the self-attention GAN framework, we have incorporated the nonlocal similarity in the spectral images into the reconstruction network, thus have improved the performance of our model. The hierarchical channel reconstruction has been proposed to decompose the hard problem into several easier tasks. The experiment results proved that HCR can further improve the performance. To further enhance the quality of reconstructed images, we have adapted another small U-net with residual learning to refine the results of the first stage. By processing each spectral frame independently, the parameters in this second U-net have decreased dramatically and thus it is easy to train. The quality of reconstructed images has improved significantly due to this refinement stage.

Our proposed λ-net has been verified by the real data captured by the compressive spectral camera. It not only achieves better results than the current state-of-the-art, but also finishes the reconstruction in a short time. It is expected to use the CASSI camera with our λ-net to build an end-to-end video-rate 3D hyperspectral imaging system, while enjoying the benefits of low cost and low bandwidth.

[1]It is possible to train multiple λ-nets for different regions of the large area, since the mask values for different places are different. However, the training takes too long and multiple GPUs are required, which is beyond our capability. We believe this 416×416 region can demonstrate the performance of our proposed λ-net.

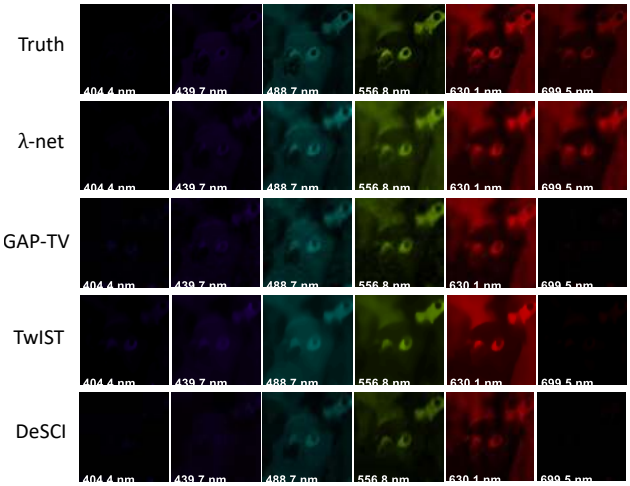

Figure 9. Real data results: reconstructed bird data from measurement captured by the real camera. Six ({404.4, 439.7, 488.7, 556.8, 630.1 and 699.5}nm) out of 24 spectral channels are shown to compare with the ground truth. It can be seen that only λ-net can recover the last channel (far right). PSNR: λ-net 25.59dB, GAP-TV 24.58dB, TwIST 24.33dB, and DeSCI 25.13dB.

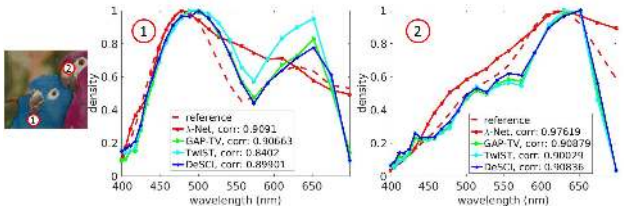

Figure 10. Real data results: reconstructed spectra of the bird data from measurement captured by the real SCI camera.

Acknowledgments

This work was partially supported by National Science Foundation grant IIS 1565328.

References

[1] Cave multispectral image database. http://www1.cs.columbia.edu/CAVE/databases/multispectral/. Accessed: 2018-11-05.

[2] Hyperspectral and color imaging. https://sites.google.com/site/hyperspectralcolorimaging/dataset/general-scenes. Accessed: 2018-11-05.

[3] Naveed Akhtar and Ajmal S Mian. Hyperspectral recovery from rgb images using gaussian processes. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[4] Yoann Altmann, Stephen McLaughlin, Miles J. Padgett, Vivek K Goyal, Alfred O. Hero, and Daniele Faccio. Quantum-inspired computational imaging. *Science*, 361(6403), 2018.

[5] Boaz Arad and Ohad Ben-Shahar. Sparse recovery of hyperspectral signal from natural rgb images. In *ECCV*. Springer, 2016.

[6] 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.

[7] Henry Arguello and Gonzalo R Arce. Rank minimization code aperture design for spectrally selective compressive imaging. *IEEE transactions on image processing*, 22(3):941–954, 2012.

[8] Henry Arguello and Gonzalo R Arce. Colored coded aperture design by concentration of measure in compressive spectral imaging. *IEEE Transactions on Image Processing*, 23(4):1896–1908, 2014.

[9] Henry Arguello, Hoover Rueda, Yuehao Wu, Dennis W. Prather, and Gonzalo R. Arce. Higher-order computational model for coded aperture spectral imaging. *Appl. Opt.*, 52(10):D12–D21, Apr 2013.

[10] 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.

[11] Xun Cao, Hao Du, Xin Tong, Qionghai Dai, and Stephen Lin. A prism-mask system for multispectral video acquisition. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2423–2435, 2011.

[12] Xun Cao, Tao Yue, Xing Lin, Stephen Lin, Xin Yuan, Qionghai Dai, Lawrence Carin, and David J Brady. Computational snapshot multispectral cameras: Toward dynamic capture of the spectral world. *IEEE Signal Processing Magazine*, 33(5):95–108, 2016.

[13] Jen-Hao Rick Chang, Chun-Liang Li, Barnabas Poczos, BVK Vijaya Kumar, and Aswin C Sankaranarayanan. One network to solve them all solving linear inverse problems using deep projection models. In *ICCV*, pages 5889–5898, Oct 2017.

[14] Xueqing Deng, Yi Zhu, and Shawn Newsam. What is it like down there?: generating dense ground-level views and image features from overhead imagery using conditional generative adversarial networks. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 43–52. ACM, 2018.

[15] David L Donoho et al. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[16] Marco F Duarte, Mark A Davenport, Dharmpal Takhar, Jason N Laska, Ting Sun, Kevin F Kelly, and Richard G Baraniuk. Single-pixel imaging via compressive sampling. *IEEE signal processing magazine*, 25(2):83–91, 2008.

[17] Candes Emmanuel, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, February 2006.

[18] Mário AT Figueiredo, Robert D Nowak, and Stephen J Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of selected topics in signal processing*, 1(4):586–597, 2007.

[19] Laura Galvis, Henry Arguello, and Gonzalo R. Arce. Coded aperture design in mismatched compressive spectral imaging. *Appl. Opt.*, 54(33):9875–9882, Nov 2015.

[20] Laura Galvis, Daniel Lau, Xu Ma, Henry Arguello, and Gonzalo R. Arce. Coded aperture design in compressive spectral imaging based on side information. *Appl. Opt.*, 56(22):6332–6340, Aug 2017.

[21] Liang Gao, Jinyang Liang, Chiye Li, and Lihong V Wang. Single-shot compressed ultrafast photography at one hundred billion frames per second. *Nature*, 516(7529):74, 2014.

[22] ME Gehm, R John, DJ Brady, RM Willett, and TJ Schulz. Single-shot compressive spectral imaging with a dual-disperser architecture. *Optics express*, 15(21):14013–14027, 2007.

[23] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[25] Carlos Hinojosa, Jorge Bacca, and Henry Arguello. Coded aperture design for compressive spectral subspace clustering. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1589–1600, 2018.

[26] Yasunobu Hitomi, Jinwei Gu, Mohit Gupta, Tomoo Mitsunaga, and Shree K Nayar. Video from a single coded exposure photograph using a learned over-complete dictionary. In *2011 International Conference on Computer Vision*, pages 287–294. IEEE, 2011.

[27] Yuanjun Huang, Xianbin Cao, Xiantong Zhen, and Jungong Han. Attentive temporal pyramid network for dynamic scene classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8497–8504, 2019.

[28] Michael Iliadis, Leonidas Spinoulas, and Aggelos K. Katsaggelos. Deep fully-connected networks for video compressive sensing. *Digital Signal Processing*, 72:9–18, 2018.

[29] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[30] Shirin Jalali and Xin Yuan. Compressive imaging via one-shot measurements. In *ISIT*, 2018.

[31] Shirin Jalali and Xin Yuan. Snapshot compressed sensing: performance bounds and algorithms. *IEEE Transactions on Information Theory*, 2019.

[32] Kyong Hwan Jin, Michael T McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *TIP*, 26(9):4509–4522, Sept 2017.

[33] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[34] 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, December 2010.

[35] Sriharsha Koundinya, Himanshu Sharma, Manoj Sharma, Avinash Upadhyay, Raunak Manekar, Rudrabha Mukhopadhyay, Abhijit Karmakar, and Santanu Chaudhury. 2d-3d cnn based architectures for spectral reconstruction from rgb images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[36] Kuldeep Kulkarni, Suhas Lohit, Pavan Turaga, Ronan Kerviche, and Amit Ashok. Reconnet: Non-iterative reconstruction of images from compressively sensed random measurements. In *CVPR*, 2016.

[37] Huiqun Li, Zhiwei Xiong, Zhan Shi, Lizhi Wang, Dong Liu, and Feng Wu. Hsvcnn: Cnn-based hyperspectral reconstruction from rgb videos. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3323–3327, Oct 2018.

[38] Yan Li, Zehao Xiao, Xiantong Zhen, and Xianbin Cao. Attentional information fusion networks for cross-scene power line detection. *IEEE Geoscience and Remote Sensing Letters*, 2019.

[39] Yang Liu, Xin Yuan, Jinli Suo, David Brady, and Qionghai Dai. Rank minimization for snapshot compressive imaging. *PAMI*, in press, 2018.

[40] Patrick Llull, Xuejun Liao, Xin Yuan, Jianbo Yang, David Kittle, Lawrence Carin, Guillermo Sapiro, and David J Brady. Coded aperture compressive temporal imaging. *Optics Express*, 21(9):10526–10545, 2013.

[41] Patrick Llull, Xin Yuan, Lawrence Carin, and David J Brady. Image translation for single-shot focal tomography. *Optica*, 2(9):822–825, 2015.

[42] Jiawei Ma, Xiaoyang Liu, Zheng Shou, and Xin Yuan. Deep tensor admm-net for snapshot compressive imaging. In *IEEE/CVF Conference on Computer Vision (ICCV)*, 2019.

[43] Morteza Mardani, Qingyun Sun, David Donoho, Vardan Papyan, Hatef Monajemi, Shreyas Vasanawala, and John. Pauly. Neural proximal gradient descent for compressive imaging. In *NIPS*, pages 9573–9583, 2018.

[44] Shaobo Min, Xuejin Chen, Zheng-Jun Zha, Feng Wu, and Yongdong Zhang. A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels. 2019.

[45] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014. cite arxiv:1411.1784.

[46] Ali Mousavi and Richard G Baraniuk. Learning to invert: Signal recovery via deep convolutional networks. In *ICASSP*, March 2017.

[47] Kamyar Nazeri, Eric Ng, and Mehran Ebrahimi. Image colorization using generative adversarial networks. In *International Conference on Articulated Motion and Deformable Objects*, pages 85–94. Springer, 2018.

[48] Shijie Nie, Lin Gu, Yinqiang Zheng, Antony Lam, Nobutaka Ono, and Imari Sato. Deeply learned filter response functions for hyperspectral reconstruction. In *CVPR*, June 2018.

[49] Mu Qiao, Yangyang Sun, Xuan Liu, Xin Yuan, and Paul Wilford. Snapshot optical coherence tomography. In *Digital Holography and Three-Dimensional Imaging 2019*, page W4B.3. Optical Society of America, 2019.

[50] Dikpal Reddy, Ashok Veeraraghavan, and Rama Chellappa. P2C2: Programmable pixel compressive camera for high speed imaging. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 329–336.

[51] Francesco Renna, Liming Wang, Xin Yuan, Jianbo Yang, Galen Reeves, Robert Calderbank, Lawrence Carin, and Miguel RD Rodrigues. Classification and reconstruction of high-dimensional signals from low-dimensional features in the presence of side information. *IEEE Transactions on Information Theory*, 62(11):6459–6492, Nov 2016.

[52] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

[53] Zhan Shi, Chang Chen, Zhiwei Xiong, Dong Liu, and Feng Wu. Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[54] Ayan Sinha, Justin Lee, Shuai Li, and George Barbastathis. Lensless computational imaging through deep learning. *Optica*, 4(9):1117–1125, Sep 2017.

[55] Liyan Sun, Zhiwen Fan, Yue Huang, Xinghao Ding, and John Paisley. Compressed sensing mri using a recursive dilated network. In *AAAI*, 2018.

[56] Yangyang Sun, Xin Yuan, and Shuo Pang. High-speed compressive range imaging based on active illumination. *Optics Express*, 24(20):22836–22846, Oct 2016.

[57] Yangyang Sun, Xin Yuan, and Shuo Pang. Compressive high-speed stereo imaging. *Opt Express*, 25(15):18182–18190, 2017.

[58] Tsung-Han Tsai, Patrick Llull, Xin Yuan, Lawrence Carin, and David J Brady. Spectral-temporal compressive imaging. *Optics Letters*, 40(17):4054–4057, Sep 2015.

[59] Tsung-Han Tsai, Xin Yuan, and David J Brady. Spatial light modulator based color polarization imaging. *Optics Express*, 23(9):11912–11926, May 2015.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[61] 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.

[62] Ashwin A Wagadarikar, Nikos P Pitsianis, Xiaobai Sun, and David J Brady. Video rate spectral imaging using a coded aperture snapshot spectral imager. *Optics Express*, 17(8):6368–6388, 2009.

[63] Chaojie Wang, Bo Chen, Sucheng Xiao, and Mingyuan Zhou. Convolutional Poisson Gamma Belief Network. In *ICML*, 2019.

[64] Lizhi Wang, Zhiwei Xiong, Dahua Gao, Guangming Shi, and Feng Wu. Dual-camera design for coded aperture snapshot spectral imaging. *Appl. Opt.*, 54(4):848–858, Feb 2015.

[65] Lizhi Wang, Zhiwei Xiong, Dahua Gao, Guangming Shi, Wenjun Zeng, and Feng Wu. High-speed hyperspectral video acquisition with a dual-camera architecture. In *CVPR*, pages 4942–4950, June 2015.

[66] Lizhi Wang, Zhiwei Xiong, Hua Huang, Guangming Shi, Feng Wu, and Wenjun Zeng. High-speed hyperspectral video acquisition by combining nyquist and compressive sampling. *PAMI*, pages 1–1, 2018.

[67] Lizhi Wang, Zhiwei Xiong, Guangming Shi, Feng Wu, and Wenjun Zeng. Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging. *PAMI*, 39(10):2104–2111, Oct 2017.

[68] Lizhi Wang, Tao Zhang, Ying Fu, and Hua Huang. Hyperreconnet: Joint coded aperture optimization and image reconstruction for compressive hyperspectral imaging. *IEEE Transactions on Image Processing*, 28(5):2257–2270, May 2019.

[69] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: From error visibility to structural similarity. *TIP*, 13(4):600–612, 2004.

[70] Yuehao Wu, Iftekhar O. Mirza, Gonzalo R. Arce, and Dennis W. Prather. Development of a digital-micromirror-device-based multishot snapshot spectral imaging system. *Opt. Lett.*, 36(14):2692–2694, Jul 2011.

[71] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 341–349. Curran Associates, Inc., 2012.

[72] Kai Xu and Fengbo Ren. CSVideoNet: A real-time end-to-end learning framework for high-frame-rate video compressive sensing. *arXiv: 1612.05203*, Dec 2016.

[73] Jianbo Yang, Xuejun Liao, Xin Yuan, Patrick Llull, David J Brady, Guillermo Sapiro, and Lawrence Carin. Compressive sensing by learning a Gaussian mixture model from measurements. *TIP*, 24(1):106–119, January 2015.

[74] Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Deep admm-net for compressive sensing mri. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *NIPS*, pages 10–18, 2016.

[75] Xin Yuan. Compressive dynamic range imaging via Bayesian shrinkage dictionary learning. *Optical Engineering*, 55(12):123110, 2016.

[76] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In *ICIP*, Sept 2016.

[77] Xin Yuan, Xuejun Liao, Patrick Llull, David Brady, and Lawrence Carin. Efficient patch-based approach for compressive depth imaging. *Applied Optics*, 55(27):7556–7564, Sep 2016.

[78] Xin Yuan, Patrick Llull, David J Brady, and Lawrence Carin. Tree-structure bayesian compressive sensing for video. *arXiv:1410.3080*, 2014.

[79] Xin Yuan, Patrick Llull, Xuejun Liao, Jianbo Yang, David J. Brady, Guillermo Sapiro, and Lawrence Carin. Low-cost compressive sensing for color video and depth. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3318–3325, 2014.

[80] Xin Yuan and Shuo Pang. Compressive video microscope via structured illumination. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1589–1593, Sept 2016.

[81] Xin Yuan and Shuo Pang. Structured illumination temporal compressive microscopy. *Biomedical Optics Express*, 7:746–758, 2016.

[82] Xin Yuan and Yunchen Pu. Parallel lensless compressive imaging via deep convolutional neural networks. *Optics Express*, 26(2):1962–1977, Jan 2018.

[83] Xin Yuan, Yangyang Sun, and Shuo Pang. Compressive temporal stereo-vision imaging. In *Computational Optical Sensing and Imaging (COSI)*, 2016.

[84] Xin Yuan, Tsung-Han Tsai, Ruoyu Zhu, Patrick Llull, David Brady, and Lawrence Carin. Compressive hyperspectral imaging with side information. *IEEE Journal of Selected Topics in Signal Processing*, 9(6):964–976, September 2015.

[85] Xin Yuan, Jianbo Yang, Patrick Llull, Xuejun Liao, Guillermo Sapiro, David J Brady, and Lawrence Carin. Adaptive temporal compressive sensing for video. *IEEE International Conference on Image Processing*, pages 1–4, 2013.

[86] Lei Yue, Xin Miao, Pengbo Wang, Baochang Zhang, Xiantong Zhen, and Xianbin Cao. Attentional alignment networks. In *BMVC*, volume 2, page 7, 2018.

[87] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[88] 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, July 2017.

[89] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.