MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Unrolled Projected Gradient Descent for Multi-Spectral Image Fusion

Lohit, S.; Liu, D.; Mansour, H.; Boufounos, P.T.

TR2019-010 March 29, 2019

Abstract

In this paper, we consider the problem of fusing low spatial resolution multi-spectral (MS) aerial images with their associated high spatial resolution panchromatic image. To solve this problem, various methods have been proposed, using either model-based or modelagnostic algorithms such as deep learning techniques. In this paper, we aim to utilize more interpretable architectures to solve the MS fusion problem by integrating existing ideas from image processing with deep learning. In particular, we develop a signal processinginspired learning solution, where we unroll the iterations of the projected gradient descent (PGD) algorithm, and each iteration contains a projection operation carried out by a deep convolutional neural network. We observe that our proposed method provides a new perspective on existing deep-learning solutions, and under certain circumstance it reduces to current black-box deep learning methods. Our extensive experimental results show significant improvements of the proposed approach over several baselines.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

UNROLLED PROJECTED GRADIENT DESCENT FOR MULTI-SPECTRAL IMAGE FUSION

Suhas Lohit[⋆] *Dehong Liu*[†] *Hassan Mansour*[†] *Petros T. Boufounos*[†]

[⋆] Arizona State University, Tempe, AZ, USA
[†] Mitsubishi Electric Research Laboratories, Cambridge, MA, USA

ABSTRACT

In this paper, we consider the problem of fusing low spatial resolution multi-spectral (MS) aerial images with their associated high spatial resolution panchromatic image. To solve this problem, various methods have been proposed, using either model-based or model-agnostic algorithms such as deep learning techniques. In this paper, we aim to utilize more interpretable architectures to solve the MS fusion problem by integrating existing ideas from image processing with deep learning. In particular, we develop a signal processing-inspired learning solution, where we unroll the iterations of the projected gradient descent (PGD) algorithm, and each iteration contains a projection operation carried out by a deep convolutional neural network. We observe that our proposed method provides a new perspective on existing deep-learning solutions, and under certain circumstance it reduces to current black-box deep learning methods. Our extensive experimental results show significant improvements of the proposed approach over several baselines.

Index Terms— multi-spectral image fusion, deep learning, projected gradient descent

1. INTRODUCTION

Multi-spectral (MS) imaging, widely used in remote sensing and its related areas, allows sensing of images across a wider range of wavelengths compared to conventional optical imagers. The bands of interest in MS imaging cover RGB, near infra-red (NIR) and short-wave IR (SWIR) in general. The advantage of MS imaging lies in several aspects such as (a) better discrimination of objects with different material properties which may otherwise be very similar in the RGB bands, and (b) more information gathering capability in the presence of harsh atmospheric conditions such as haze and fog, as infra-red waves can travel more easily through these media, compared to visible light.

Multi-spectral sensing presents an interesting challenge. It is necessary in many applications to have both high spatial and spectral resolutions. However, there is a fundamental trade-off between the bandwidth of the sensor and the spatial resolution it can have. High spatial resolution typically can be achieved by panchromatic (PAN) image covering the visible bands but without rich spectral information. This leads to the problem of MS image fusion. Given a set of low resolution MS images obtained at different wavelengths as well as a high resolution panchromatic image which does not have spectral information, we would like to fuse these two modes of information in order to produce a set of images which have both high spectral and high spatial resolutions.

This work was completed when Suhas Lohit was an intern at Mitsubishi Electric Research Laboratories.

MS image fusion is essentially an under-determined ill-posed problem. To solve this problem, various methods have been proposed, either model-based [1–4] or data-driven methods [5–8]. Model-based methods are generally simple to design and have theoretical guarantees but with relative poor performance compared to data-driven methods, especially deep learning based methods. On the other hand, purely data-driven methods operate as a black box and are hence less interpretable. Following recent studies on model-based deep learning [9–11], we formulate a combination of model-based and data-driven solution based on deep learning in order to solve the multi-spectral image fusion problem. We unroll the iterations of the projected gradient descent (PGD) algorithm, and replace the projection step of PGD with a convolutional neural network (CNN). Compared to other existing purely data-driven techniques, our work is based on well studied signal processing frameworks and guaranteed to converge to a meaningful point, and also provides superior performance compared to the various baselines considered. Our contributions are summarized as follows.

- We unroll the iterations of PGD and use a CNN as the projection operator of PGD to solve the MS fusion problem. Our approach provides a signal processing-based perspective with superior performance and convergence guarantee.
- We learn not only the projection operator CNN with training data, but also the forward operator to overcome the challenge of the unknown forward operator in MS image fusion.
- Our method generalizes existing purely data-driven methods. When the forward operator is the identity operator and with suitable parameter settings, our method reduces to a purely deep-learning based method.

2. PRIOR ART

In this section, we briefly review relevant algorithms for multi-spectral fusion as well as other inverse problems.

Model-based iterative methods: In order to solve ill-posed problems, there is a rich literature on simple prior models of the desired signal. In the case of the multi-spectral fusion, priors include sparsity in the gradient domain – total-variation regularization, low-rank models [1, 2], over-complete dictionary learning with regularizer on the coefficients [3, 4]. These methods are generally simple to design and have theoretical guarantees. However, in terms of recovery performance as well as computational complexity during testing, these methods fare poorly compared to purely data-driven methods described next.

Purely data-driven approaches: In recent years, the resurgence of deep learning [5] has led to feed-forward non-iterative approaches for solving inverse problems in low-level vision including computational imaging, single-image super-resolution, deblurring [6] and

multi-spectral fusion [7, 8]. These methods are model-agnostic and simply learn a mapping from the measurements to the desired signal in a purely data-driven fashion. Compared to the model-based iterative methods, these methods generally yield superior results, and are also computationally faster owing to their non-iterative nature (a feed-forward operation at test time) as well as the ease of implementation on Graphics Processing Units (GPUs).

Model-based deep learning: Although purely data-driven approaches using deep learning perform very well compared to model-based shallow methods, they are less interpretable than the latter. In order to bridge the gap between understanding and performance, many methods recently combine iterative methods with the deep learning. This can be achieved in several ways. Sun *et al.* first proposed the ADMM-Net for MRI [9]. Here, the iterations of ADMM are unrolled and the projection operator as well as the shrinkage function are learned from data. Chang *et al.* proposed the OneNet [10] for inverse problems like super-resolution and restoration. It unrolls the ADMM algorithm such that projection operator is a deep learning method. More recently, Gupta *et al.* [11] propose a similar approach in the case of PGD and also provide theoretical guarantees for convergence. In this paper, we combine PGD with deep learning for the problem of multi-spectral image fusion. We unroll the iterations of PGD such that the projection operator is computed using a trained convolutional neural network (CNN) and all the parameters are learned end-to-end using a training dataset. This problem is different from other inverse problems in two aspects – (a) we are given the pan-chromatic image which acts as important side information, and (b) the forward operator $\mathbf{A}$ is usually unknown.

3. UNROLLING PGD USING A CNN AS THE PROJECTION OPERATOR

We first consider a general problem where we have measurements $\mathbf{y} \in \mathbb{R}^m$ of unknowns $\mathbf{x} \in \mathbb{R}^n$ via a forward operator $\mathbf{A} \in \mathbb{R}^{m \times n}$, with the goal of recovering $\mathbf{x}$, *i.e.*

$$\mathbf{y} = \mathbf{A}\mathbf{x}. \tag{1}$$

In real applications, we typically have $m < n$, leading to an underdetermined linear system with infinite solutions in general. In order to have a unique solution, we solve a constrained optimization problem as follows

$$\mathbf{x}^* = \text{argmin}\, \frac{1}{2}||\mathbf{y} - \mathbf{A}\mathbf{x}||_2^2 \text{ s.t. } \mathbf{x} \in \mathcal{C}, \tag{2}$$

where $\mathcal{C}$ is the constraint set. In our case, $\mathcal{C}$ is the set of feasible images. Generally, the set $\mathcal{C}$ is chosen based on domain knowledge, *e.g.*, the set of images with small ℓ_1 norm of the wavelet coefficients. A popular approach to solving the above problem in image processing is by employing the PGD algorithm, which consists of two alternating steps:

$$\mathbf{w}^{k+1} = \mathbf{x}^k + \alpha \mathbf{A}^T(\mathbf{y} - \mathbf{A}\mathbf{x}^k), \tag{3}$$

$$\mathbf{x}^{k+1} = \Pi_{\mathcal{C}}(\mathbf{w}^{k+1}), \tag{4}$$

where $\Pi_{\mathcal{C}}$ is a projection operator onto the set $\mathcal{C}$. The first step in the above optimization process is gradient descent, which is guaranteed to reduce the value of the cost function given a suitable value of the learning rate α. However, the output of the gradient descent step is not guaranteed to be a feasible point. The second step is to map the intermediate output from gradient descent to the closest point in the set of feasible solutions through the projection operator $\Pi_{\mathcal{C}}$.

The MS image fusion problem can be formulated as an inverse problem. Let I_P, I_L, and I_H represent the vectorized versions the Pan, low resolution MS, and high resolution MS images, respectively. We denote $\mathbf{y} = (I_P; I_L)$ and $\mathbf{x} = (I_P; I_H)$. The forward operator $\mathbf{A}$ models the mapping from high resolution to the low resolution MS images.

However, there are several challenges to solve this MS image fusion problem. First, in the case of MS aerial images considered in this paper (as well as natural images in general), it is difficult to provide a precise mathematical definition of a feasible set and it is also unclear what a good approximate to the constraint set is. The goodness of approximate constraint set may also depend on the properties of $\mathbf{A}$. Second, although we know that the forward operator $\mathbf{A}$ can be represented as a combination of blurring and down sampling, the exact coefficients of the blur kernel are unknown.

In existing methods, some of the widely used hand-crafted priors include the sparsity priors in wavelet and gradient domains. In the case of dictionary learning, an over-complete sparsifying basis is also learned from the data and sparsity priors are used on the coefficients. However, these techniques fall short in terms of providing high quality solutions and have given way to purely data-driven non-linear methods using deep learning, as explained in Section 2. A main drawback of deep learning methods is that they are not easily interpretable and it is unclear what functions they perform in terms of signal processing.

Following similar works for compressive imaging, MRI, and super-resolution, etc., we propose the following framework. We ***unroll*** the iterations of PGD (Eqs. (3) and (4)), and use a trained CNN in each step to replace the projection operator. We choose *apriori* the number of iterations to unroll and train the entire pipeline end-to-end. The whole framework is illustrated in Fig. 1. Therefore, although the core architecture of the CNN is hand-crafted, at a higher level, the algorithm is based on well-studied methods.

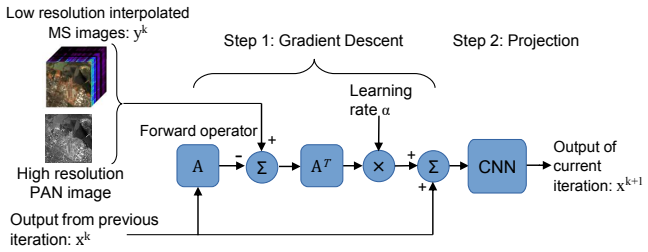

Fig. 1. A single stage of unrolled PGD framework we propose in this paper. The gradient descent step is carried out as usual and a CNN is used as the projection operator onto the set of high resolution multi-spectral images. Note that both forward operator $\mathbf{A}$ and the layers in the CNN are learned end-to-end jointly.

In order to overcome the challenge of unknown $\mathbf{A}$, we explore two options with either given or learned $\mathbf{A}$.

3.1. Using the denoising formulation i.e., $\mathbf{A} = \mathbf{I}$

In practice, we typically formulate the MS fusion problem as an image denoising problem by setting the measurements $\mathbf{y}$ to be upsampled low resolution images using bicubic interpolation and setting $\mathbf{A} = \mathbf{I}$ for simplicity, where $\mathbf{I}$ is the identity filter. We note that under this new formulation with the learning rate $\alpha = 1$, we have $\mathbf{w}^{k+1} = \mathbf{y}$, meaning that the unrolled PGD reduces to a projection step using a trained CNN. Therefore, this formulation, depending on the network architecture, reduces to the framework of deep-learning-

based MS fusion presented in [7] and [8], but using a signal processing methodology.

3.2. Using the general formulation i.e., A is learned

As a natural extension to the above, we investigate the possibility of *jointly* learning the forward operator $\mathbf{A}$, the learning rate α and the CNN. After bicubic interpolation, the low resolution images are of the same size, albeit blurred versions of the high resolution images. This suggests modeling $\mathbf{A}$ as a blurring operator. We assume that the same blurring operator, which we represent as a square convolutional kernel, operates on the entire image as well as on different spectral channels. We structure the kernel to be of the form

$$\mathbf{K_A} = \mathbf{K_B} + \mathbf{K_I}, \text{ s.t. } \sum_{i=1}^{S}\sum_{j=1}^{S} \mathbf{K_A}(i,j) = 1$$
$$\mathbf{K_A}(i,j) \geq 0, \forall i,j \in \{1,\ldots,S\}, \quad (5)$$

where the coefficients of $\mathbf{K_B}$ are learned and $\mathbf{K_I}$ is the identity filter. This encourages the 2D filter coefficients to be centrally dominant and the constraints on $\mathbf{K_A}$ ensure that the corresponding operator $\mathbf{A}$ is a valid blurring operator. In our experiments, we have chosen the size of $\mathbf{K_A}$ to be 9×9. The convolutional kernel as well as the whole convolutional filters are trained with training datasets by minimizing the error between the network output and the desired output using the Adam optimizer [12].

4. EXPERIMENTS

For all our experiments, we use a training dataset of 138 high resolution aerial MS images with 16 channels and a panchromatic image of size 256×256 pixels. These images are synthesized using the AVIRIS hyper-spectral image database [15]: each MS image channel is generated by a weighted linear combination of a band of hyper-spectral images. For training, we also need access to the low resolution images. The low resolution MS images of size 128×128 are produced by first low-pass filtering (anti-aliasing) and then downsampling by a factor of 2 (we focus on $2\times$ super-resolution, however the method can be extended to other factors).

The test set consists of four low resolution MS images with the same 16 channels – covering four areas of Moffett, Cambria Fire, Cuprite, and Los Angeles, respectively – of size 512×512 after interpolation and each with a panchromatic image, also of size 512×512. During test time, the images are split into overlapping patches and the fed through the trained networks. As before, we form 256×256 images of the test set which serve as the input to the algorithm. Our goal is to fuse the lower resolution MS image with the 512×512 panchromatic image to produce a high resolution MS output of resolution 512×512.

As described in Section 3, the projection operator (Eq. (4)) is learned from training data, and we choose to implement it using a CNN. Based on the work of Wei *et al.* [8], the architecture of this network is simple with 4 layers of convolutions plus Rectified Linear Units (ReLUs) with a residual connection connecting the bicubic interpolated low resolution MS images ($\mathbf{y}$) to the output of the penultimate layer of the CNN. We set the the filter size in all layers to be 9×9 and we use 32 feature maps for layers 1 and 2. Layer 3 produces 17 feature maps in order to be compatible with the number of channels of the input, and the output produces the desired 16 multi-spectral channels.

The low resolution 128×128 MS training images are first upsampled using bicubic interpolation to match the pixel resolution of the PAN image, *i.e.*, 256×256. Using the 138 pairs of low-res and high-res training images, we first create a dataset of about 60832 patches (of size $32 \times 32 \times 17$) of the interpolated low-res MS and PAN images which form the input to the fusion algorithm, and $32 \times 32 \times 16$ high-res multi-spectral images which form the desired output. About 1800 of these patches are used as the validation set in order to select the hyperparameters. For the case with $\mathbf{A} = \mathbf{I}$, the PGD algorithm reduces to simply applying the projection operator (the CNN, in our case) **once** on the low-res input $\mathbf{y}$, and thus, the algorithm essentially reduces to the one described by Wei *et al.* [8]. For this case, we train 3 networks of depths of $4, 12$ and 20 layers, repectively. When $\mathbf{A}$ is learned, we also need to choose the number of iterations, n_{iter} of PGD to unroll. We conduct experiments with $n_{iter} = 1, 3, 5$. The networks are trained using the Adam optimizer [12] for 2×10^5 iterations with a batch size of 32. We use the mean squared error over the batch between the desired high resolution patches and the output of the algorithm as the loss function. All the networks are trained using Tensorflow [16] on a GPU.

For comparison, we provide experimental results of the proposed approach with three baseline algorithms: (1) Bicubic interpolation, where the output of the algorithm is the channel-wise upsampling of the lower resolution images using bicubic interpolation, (2) Shrinkage field (SF) networks by Schmidt and Roth [13] which is a trainable architecture, but is applied to each channel independently, and (3) deep Coupled Analysis and Synthesis Dictionary (CASD), a recent work by Wen *et al.* [14] which uses channel-wise outputs from the SF network and a CASD framework in order to exploit the inter-channel relationship in order to improve fusion results.

The results of multi-spectral fusion on the test images using various algorithms are shown in Fig. 2 of Cambria Fire and Fig. 3 of Los Angeles. It is clear that our results are visually much sharper and preserve better spectral information compared to other results using existing baseline methods. To quantitatively analyze our results, we compute both the regressed Peak Signal-to-Noise Ratio (PSNR) [11] and the structural similarity index (SSIM) [17] to measure the performance of the algorithms, as shown in Table 1. The measures are computed channel-wise and averaged over the 16 MS channels. From the table, we clearly observe that the results using unrolled PGD are superior to all the baselines considered by 3-6dB when $\mathbf{A} = \mathbf{I}$. Further improvements of 0.6dB on average are achieved by learning the forward operator $\mathbf{A}$.

As regarding to computational time, we observed empirically that the validation error converges for the chosen number of iterations and it takes a few hours to train on an Nvidia Titan-X GPU. For the MS fusion test process, it takes less than a second on the same GPU with batch processing of all the patches at once.

5. CONCLUSION

In this paper, we developed an unrolled projected gradient descent (PGD) method for multi-spectral (MS) image fusion, with projection operator replaced by a trained convolutional neural network (CNN) to provide superior performance with convergence guarantee. Our method also generalizes the purely data-driven method by learning the unknown forward operator simultaneously with the CNN. When the forward operator is set to be the identity operator, our approach reduces to a purely data-driven deep learning method. Our experiments show that the learning-the-projection operation outperforms several baselines considered, and improves the results further with a learned forward operator.

Image Name	Bicubic	Shrinkage Fields [13]	DeepCASD [14]	Unrolled PGD					
				$\mathbf{A} = \mathbf{I}$ (reduces to [8])			$\mathbf{A}$ is learned		
				Number of Layers			Number of Iterations		
				4	12	20	1	3	5
Moffett	32.24 0.4788	34.21 0.6981	34.53 0.7185	37.44 0.9710	38.29 0.9768	37.46 0.9729	37.59 0.9706	38.52 0.9778	38.17 0.9776
Cambria Fire	35.32 0.5887	37.51 0.7941	37.62 0.7987	37.83 0.9734	38.91 0.9734	38.71 0.9696	37.99 0.9775	38.91 0.9765	39.33 0.9771
Cuprite	32.44 0.5060	34.33 0.7437	34.52 0.7616	36.88 0.9750	37.56 0.9842	36.82 0.9823	37.95 0.9794	38.56 0.9837	39.02 0.9840
Los Angeles	27.96 0.4888	30.39 0.7628	30.50 0.7761	36.27 0.9702	37.38 0.9755	37.28 0.9760	36.42 0.9712	37.79 0.9777	37.77 0.9790
Mean	31.99 0.5156	34.11 0.7497	34.29 0.7637	37.11 0.9760	38.03 0.9775	37.57 0.9752	37.49 0.9746	38.45 0.9789	38.57 0.9794

Table 1. The table shows the experimental results of multi-spectral image fusion in terms of PSNR in dB (the top number in each cell) and SSIM (the bottom number in each cell) on the test set. Clearly, the results using unrolled PGD are superior to all the baselines considered. Also observe that the results improve further when $\mathbf{A}$ is learned. Note that when $\mathbf{A} = \mathbf{I}$, PGD reduces to a single projection operator as in [8]. Then number of layers refers to the number of layers in the projection CNN. In the case where $\mathbf{A}$ is learned, the "number of iterations" refers to the number of steps of PGD we unroll. The CNN in each projection operation contains 4 layers of convolutions plus ReLU.

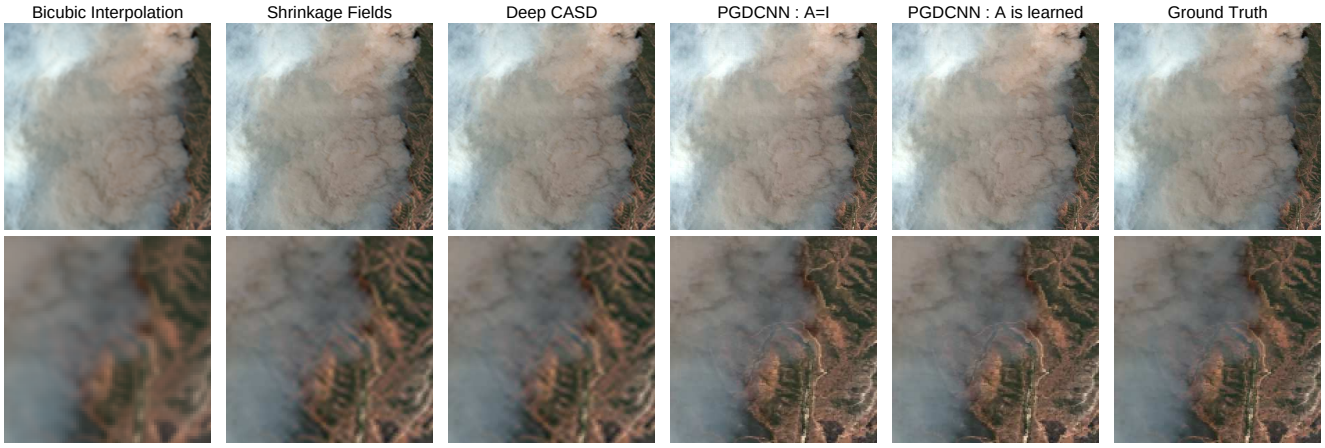

Fig. 2. Visual comparison of results for the "Cambria Fire" image (top row) and the zoomed in portions (bottom row). It is clear that the unrolled PGD (PGDCNN) provides much sharper spatial resolution and preserves better spectral information compared to the baselines.

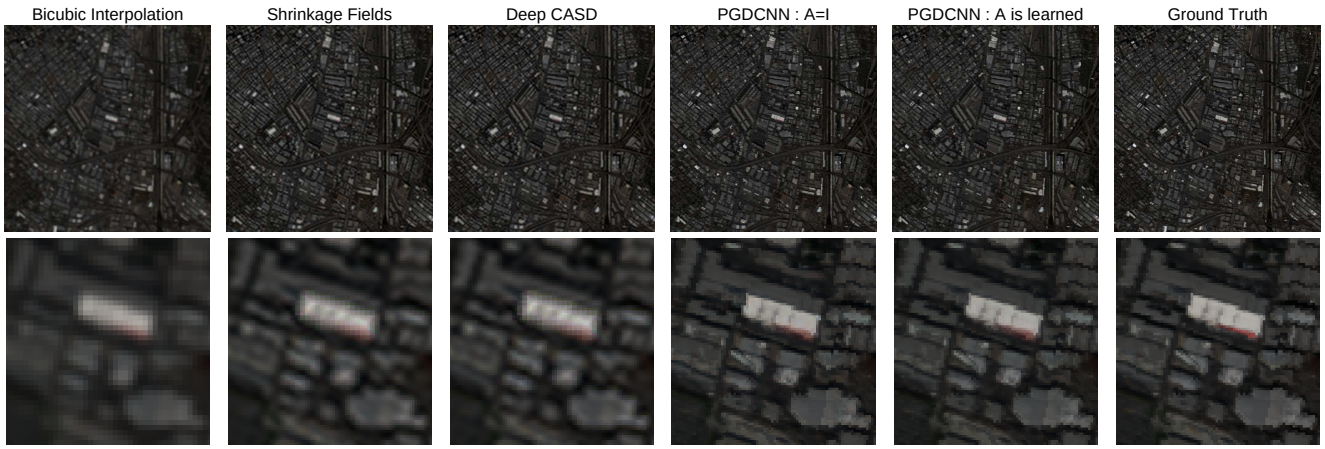

Fig. 3. Visual comparison of results for the "Los Angeles" image (top row) and the zoomed in portions (bottom row). It is clear that the unrolled PGD (PGDCNN) provides much sharper spatial resolution and preserves better spectral information compared to the baselines.

6. REFERENCES

[1] Hongyan Zhang, Wei He, Liangpei Zhang, Huanfeng Shen, and Qiangqiang Yuan, "Hyperspectral image restoration using low-rank matrix recovery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 8, pp. 4729–4743, 2014.

[2] Zhiyuan Zha, Xinggan Zhang, Qiong Wang, Lan Tang, and Xin Liu, "Analysis of the group sparsity based on the rank minimization methods," *arXiv preprint arXiv:1709.03979*, 2017.

[3] Michal Aharon, Michael Elad, Alfred Bruckstein, et al., "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311, 2006.

[4] Dehong Liu and Petros T Boufounos, "Dictionary learning based pan-sharpening," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 2397–2400.

[5] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436, 2015.

[6] Alice Lucas, Michael Iliadis, Rafael Molina, and Aggelos K Katsaggelos, "Using deep neural networks for inverse problems in imaging: beyond analytical methods," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 20–36, 2018.

[7] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa, "Pansharpening by convolutional neural networks," *Remote Sensing*, vol. 8, no. 7, pp. 594, 2016.

[8] Yancong Wei, Qiangqiang Yuan, Huanfeng Shen, and Liangpei Zhang, "Boosting the accuracy of multispectral image pansharpening by learning a deep residual network," *IEEE Geoscience Remote Sensing Letters*, 2017.

[9] Jian Sun, Huibin Li, and Zongben Xu, "Deep admm-net for compressive sensing MRI," in *Advances in Neural Information Processing Systems*, 2016, pp. 10–18.

[10] 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*, 2017, pp. 5889–5898.

[11] Harshit Gupta, Kyong Hwan Jin, Ha Q Nguyen, Michael T McCann, and Michael Unser, "CNN-based projected gradient descent for consistent CT image reconstruction," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1440–1453, 2018.

[12] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[13] Uwe Schmidt and Stefan Roth, "Shrinkage fields for effective image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2774–2781.

[14] Bihan Wen, Ulugbek Kamilov, Dehong Liu, Hassan Mansour, and Petros T Boufounos, "DeepCASD: An end-to-end approach for multi-spectural image super-resolution," *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*, 2018.

[15] "NASA AVIRIS data repository," `https://aviris.jpl.nasa.gov`, Accessed: 2018-08-11.

[16] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., "Tensorflow: a system for large-scale machine learning," .

[17] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.