ISTA-Net: Interpretable Optimization-Inspired Deep Network for Image Compressive Sensing

Jian Zhang, Bernard Ghanem

King Abdullah University of Science and Technology (KAUST), Saudi Arabia

jian.zhang@kaust.edu.sa, bernard.ghanem@kaust.edu.sa

Abstract

With the aim of developing a fast yet accurate algorithm for compressive sensing (CS) reconstruction of natural images, we combine in this paper the merits of two existing categories of CS methods: the structure insights of traditional optimization-based methods and the speed of recent network-based ones. Specifically, we propose a novel structured deep network, dubbed ISTA-Net, which is inspired by the Iterative Shrinkage-Thresholding Algorithm (ISTA) for optimizing a general ℓ_1 norm CS reconstruction model. To cast ISTA into deep network form, we develop an effective strategy to solve the proximal mapping associated with the sparsity-inducing regularizer using nonlinear transforms. All the parameters in ISTA-Net (e.g. nonlinear transforms, shrinkage thresholds, step sizes, etc.) are learned end-to-end, rather than being hand-crafted. Moreover, considering that the residuals of natural images are more compressible, an enhanced version of ISTA-Net in the residual domain, dubbed ISTA-Net$^+$, is derived to further improve CS reconstruction. Extensive CS experiments demonstrate that the proposed ISTA-Nets outperform existing state-of-the-art optimization-based and network-based CS methods by large margins, while maintaining fast computational speed. Our source codes are available: http://jianzhang.tech/projects/ISTA-Net.

1. Introduction

From much fewer acquired measurements than determined by Nyquist sampling theory, Compressive Sensing (CS) theory demonstrates that a signal can be reconstructed with high probability when it exhibits sparsity in some transform domain [6, 11]. This novel acquisition strategy is much more hardware-friendly and it enables image or video capturing with a sub-Nyquist sampling rate [34, 24]. In addition, by exploiting the redundancy inherent to a signal, CS conducts sampling and compression at the same time, which greatly alleviates the need for high transmission bandwidth and large storage space, enabling low-cost
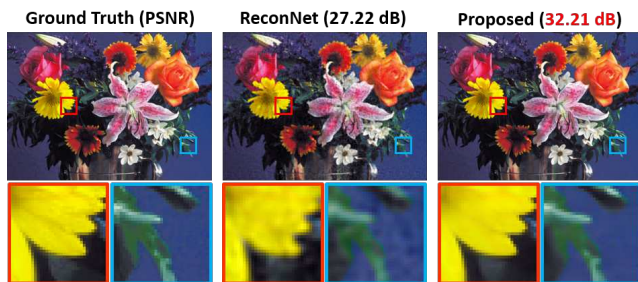

Figure 1. CS reconstruction results produced by our proposed ISTA-Net method and a recent network-based CS reconstruction method (ReconNet [21]), when the CS ratio is 25%. ISTA-Net clearly produces a higher fidelity reconstruction.

on-sensor data compression. CS has been applied in many practical applications, including but not limited to single-pixel imaging [11, 33], accelerating magnetic resonance imaging (MRI) [26], wireless tele-monitoring [50] and cognitive radio communication [36].

Mathematically, the purpose of CS reconstruction is to infer the original signal $\mathbf{x} \in \mathbb{R}^N$ from its randomized CS measurements $\mathbf{y} = \mathbf{\Phi}\mathbf{x} \in \mathbb{R}^M$. Here, $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ is a linear random projection (matrix). Because $M \ll N$, this inverse problem is typically ill-posed, whereby the CS ratio is defined as $\frac{M}{N}$. In this paper, we mainly focus on CS reconstruction of natural images. However, it is worth noting that our proposed framework can be easily extended to videos and other types of data.

In the past decade, a great deal of image CS reconstruction methods have been developed [31, 10, 15, 21]. Most traditional methods exploit some structured sparsity as an image prior and then solve a sparsity-regularized optimization problem in an iterative fashion [18, 22, 46, 45, 28]. Based on some well-studied image formation models and intrinsic image properties, these methods enjoy the advantages of strong convergence and theoretical analysis in most cases. However, they usually suffer from high computational complexity, and they are also faced with the challenges of choosing optimal transforms and tuning parameters in their solvers. Fueled by the powerful learning ability of deep networks, several deep network-based image CS re-
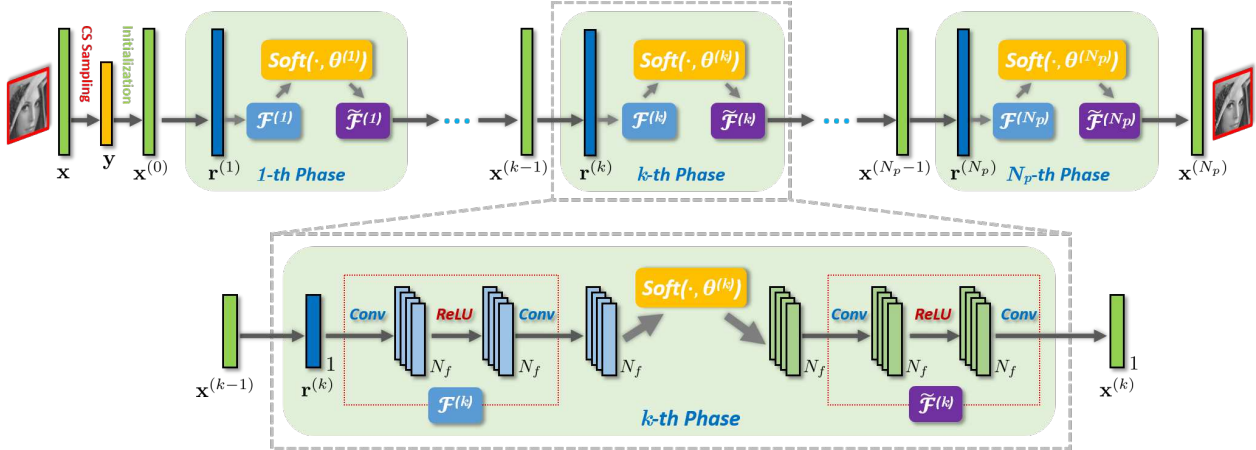
Figure 2. Illustration of our proposed ISTA-Net framework. Specifically, ISTA-Net is composed of N_p phases, and each phase strictly corresponds to one iteration in ISTA. The forward transform $\mathcal{F}^{(k)}$ is designed as a combination of two linear convolutional operators separated by a rectified linear unit (ReLU), and the backward transform $\widetilde{\mathcal{F}}^{(k)}$ is designed to exhibit a structure symmetric to that of $\mathcal{F}^{(k)}$. Note that $\mathcal{F}^{(k)}$ and $\widetilde{\mathcal{F}}^{(k)}$ satisfy the symmetry constraint $\widetilde{\mathcal{F}}^{(k)} \circ \mathcal{F}^{(k)} = \mathcal{I}$, where $\mathcal{I}$ is the identity operator. N_f denotes the number of feature maps.

construction algorithms have been recently proposed to directly learn the inverse mapping from the CS measurement domain to the original signal domain [30, 15]. Compared to optimization-based algorithms, these non-iterative algorithms dramatically reduce time complexity, while achieving impressive reconstruction performance. However, existing network-based CS algorithms are trained as a *black box*, with limited insights from the CS domain.

In this paper, we design a novel deep architecture, dubbed ISTA-Net, by mapping the traditional ISTA [4] for optimizing a general ℓ_1 norm CS reconstruction model into a deep network. In particular, ISTA-Net is composed of a fixed number of phases, each of which strictly corresponds to an ISTA-like iteration, as illustrated in Figure 2. Rather than traditional linear transforms, nonlinear learnable and sparsifying transforms are adopted in ISTA-Net, and an efficient and effective strategy to solve the proximal mapping of the nonlinear transform is developed. All the parameters involved in ISTA-Net (*e.g.* nonlinear transforms, shrinkage thresholds, step sizes, etc.) are learned end-to-end using back-propagation. Moreover, borrowing more insights from the compression realm, an enhanced version, dubbed ISTA-Net$^+$, is derived from ISTA-Net by sparsifying natural images explicitly in the residual domain. Interestingly, the skip connections introduced by ISTA-Net$^+$ further facilitate the network training. In fact, the proposed ISTA-Nets can be viewed as an attempt to bridge the gap between the two aforementioned categories of CS methods.

Contributions. The main contributions of this paper are summarized as follows: **1)** We develop a novel ISTA-Net, which adopts the structure of ISTA update steps to design a learnable deep network manifestation, where all parameters are discriminately learned instead of being hand-crafted or fixed. By learning sparsifying transforms in the residual domain, an enhanced version ISTA-Net$^+$ is derived to further

improve CS performance. As such, ISTA-Nets enjoy the advantages of fast and accurate reconstruction with well-defined interpretability. **2)** The proximal mapping problem associated to a nonlinear sparsifying transform is solved in a general and efficient way, which actually enables mapping other optimization algorithms into deep network form. **3)** Extensive experiments on natural image and MRI CS reconstruction clearly show that ISTA-Net significantly outperforms the state-of-the-art, while maintaining attractive computational complexity.

2. Related Work

We generally group existing CS reconstruction methods of images into two categories: *traditional optimization-based* CS methods and recent *network-based* CS methods. In what follows, we give a brief review of both and focus on the specific methods most relevant to our own.

Optimization-based CS reconstruction: Given the linear measurements $\mathbf{y}$, traditional image CS methods usually reconstruct the original image $\mathbf{x}$ by solving the following (generally convex) optimization problem:

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{\Phi}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{\Psi}\mathbf{x}\|_1, \qquad (1)$$

where $\mathbf{\Psi}\mathbf{x}$ denotes the transform coefficients of $\mathbf{x}$ with respect to some transform $\mathbf{\Psi}$ and the sparsity of the vector $\mathbf{\Psi}\mathbf{x}$ is encouraged by the ℓ_1 norm with λ being the (generally pre-defined) regularization parameter.

Since natural images are typically non-stationary, the classic fixed domains (*e.g.* DCT, wavelet [31], and gradient domain [22]) usually result in poor reconstruction performance. Many works incorporate additional prior knowledge about transform coefficients (*e.g.* statistical dependencies [18, 51], structure [13], etc.) into the CS reconstruction framework. Furthermore, some elaborate priors ex-

ploiting the non-local self-similarity properties of natural images have been proposed to improve CS reconstruction [48, 46, 10, 52]. Metzler *et al*. propose to integrate the well-defined BM3D denoiser into the approximate message passing (AMP) framework for CS reconstruction [28]. Quite recently, some fast and effective convolutional neural network (CNN) denoisers are trained and integrated into the half quadratic splitting (HQS) [49] and the alternating direction method of multipliers (ADMM) [7] to solve image inverse problems. However, all these traditional image CS methods require hundreds of iterations to solve Eq. (1) by means of various iterative solvers, which inevitably gives rise to high computational cost thus restricting the application of CS. In addition, the selected image prior (*e.g*. optimal transform) or the optimization parameters (*e.g*. step size and regularization parameter) are usually hand-crafted and quite challenging to pre-define.

Network-based CS reconstruction: Inspired by the powerful learning capability of deep networks [42] and its success in computer vision tasks [20, 25], several deep network-based image CS reconstruction algorithms have recently been proposed [30, 15, 21, 29]. Mousavi *et al*. first propose to apply a stacked denoising auto-encoder (SDA) to learn the representation from training data and to reconstruct test data from their CS measurements [30]. Adler *et al*. and Iliadis *et al*. separately propose to utilize fully-connected neural networks for image and video CS reconstruction [2, 15]. Kulkarni *et al*. further develop a CNN-based CS algorithm, dubbed ReconNet, which learns to regress an image block (output) from its CS measurement (input) [21]. Mousavi and Baraniuk recently propose an all-convolutional network for image CS reconstruction [29]. A main feature of network-based image CS methods is that they are non-iterative, which dramatically reduces time complexity as compared with their optimization-based counterparts. However, this is done with either fully-connected or repetitive convolutional layers. We believe that their lack of structural diversity, which originates from the absence of CS domain specific insights inherent to optimization-based methods, is the bottleneck for further performance improvement.

The tremendous success of deep learning for many image processing applications has also led researchers to consider relating iterative optimization methods to neural networks [16, 39, 43, 32]. For instance, in the context of sparse coding, Grefor and LeCun propose a fast algorithm to calculate good approximations of optimal sparse codes by introducing the Learned ISTA (LISTA), in which two matrices in classical ISTA are learned instead of using pre-computed ones [12]. Mark *et al*. extend approximate message passing (AMP) algorithms to so-called Learned AMP networks for solving sparse linear inverse problems [5]. Relying on LISTA, some sparse-coding based networks for image

super-resolution and deblocking are proposed [41, 40, 23]. For image denoising and deconvolution, Schmidt and Roth propose to learn the linear filters and shrinkage functions under the framework of half-quadratic optimization [35]. Chen *et al*. propose a trainable reaction diffusion model by learning several parameterized linear filters and influence functions for image denoising and deblocking [8]. In the context of CS for sparse signals, Kamilov and Mansour propose to learn the optimal thresholding functions for ISTA based on a B-spline decomposition [17].

Recently, Yang *et al*. propose a so-called ADMM-Net architecture by reformulating ADMM for CS magnetic resonance imaging (CS-MRI) using deep networks [44]. Although both ADMM-Net and our proposed ISTA-Net have similar inspirations, they are quite different. In fact, there are two main differences between both methods. First, ADMM-Net is specifically designed and developed for CS-MRI based on ADMM, while our ISTA-Net is much more general, since it works well for both general CS and CS-MRI based on ISTA. Second, ADMM-Net only utilizes several linear filters, while ISTA-Net goes beyond that to adopt nonlinear transforms to more effectively sparsify natural images and develops an efficient strategy for solving their proximal mapping problems. The detailed comparison with ADMM-Net for CS-MRI can be found in Section 5.3.

In a nutshell, the proposed ISTA-Net can be essentially viewed as a significant extension of LISTA [12], from the sparse coding problem to general CS reconstruction. Compared with traditional optimization-based CS methods, ISTA-Net is able to learn its optimal parameters, *i.e*. thresholds, step sizes as well as nonlinear transforms, without hand-crafted settings. In addition, ISTA-Net has the same computational complexity as several iterations of traditional ISTA, which is more than 100 times faster than existing methods of this category. Compared with network-based CS methods, ISTA-Net borrows insights from traditional optimization methods to allow for interpretability in its network design and it utilizes the structural diversity originating from the CS domain. Extensive experiments demonstrate that ISTA-Net significantly outperforms the existing optimization-based and network-based CS methods, even when compared against methods that are designed for a specific domain (*e.g*. CS-MRI).

3. Proposed ISTA-Net for Compressive Sensing

In this section, we first briefly review traditional ISTA optimization for image CS reconstruction, and then elaborate on the design of our proposed ISTA-Net.

3.1. Traditional ISTA for CS

The iterative shrinkage-thresholding algorithm (ISTA) is a popular first order proximal method, which is well suited for solving many large-scale linear inverse problems.

Specifically, ISTA solves the CS reconstruction problem in Eq. (1) by iterating between the following update steps:

$$\mathbf{r}^{(k)} = \mathbf{x}^{(k-1)} - \rho\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\mathbf{x}^{(k-1)} - \mathbf{y}), \qquad (2)$$

$$\mathbf{x}^{(k)} = \arg\min_{\mathbf{x}} \ \frac{1}{2}\|\mathbf{x} - \mathbf{r}^{(k)}\|_2^2 + \lambda\|\boldsymbol{\Psi}\mathbf{x}\|_1. \qquad (3)$$

Here, k denotes the ISTA iteration index, and ρ is the step size. Eq. (2) is trivial, while Eq. (3) is actually a special case of the so-called proximal mapping, *i.e.* $\mathrm{prox}_{\lambda\phi}(\mathbf{r}^{(k)})$, when $\phi(\mathbf{x}) = \|\boldsymbol{\Psi}\mathbf{x}\|_1$. Formally, the proximal mapping of regularizer ϕ denoted by $\mathrm{prox}_{\lambda\phi}(\mathbf{r})$ is defined as

$$\mathrm{prox}_{\lambda\phi}(\mathbf{r}) = \arg\min_{\mathbf{x}} \ \frac{1}{2}\|\mathbf{x} - \mathbf{r}\|_2^2 + \lambda\phi(\mathbf{x}). \qquad (4)$$

Solving $\mathrm{prox}_{\lambda\phi}(\mathbf{r})$ in an efficient and effective way is critical for ISTA [47], as well as for other optimization methods, such as ADMM [3] and AMP [28]. For example, when $\phi(\mathbf{x}) = \|\mathbf{W}\mathbf{x}\|_1$ ($\mathbf{W}$ is wavelet transform matrix), we have $\mathrm{prox}_{\lambda\phi}(\mathbf{r}) = \mathbf{W}^\top soft(\mathbf{W}\mathbf{r}, \lambda)$ due to the orthogonality of $\mathbf{W}$. However, it remains non-trivial to obtain $\mathbf{x}^{(k)}$ in Eq. (3) for a more complex non-orthogonal (or even nonlinear) transform $\boldsymbol{\Psi}$. In addition, ISTA usually requires many iterations to obtain a satisfactory result, suffering from extensive computation. The optimal transform $\boldsymbol{\Psi}$ and all the parameters such as ρ and λ are pre-defined (do not change with k), and very challenging to tune apriori.

3.2. ISTA-Net Framework

By taking full advantage of the merits of ISTA-based and network-based CS methods, the basic idea of ISTA-Net is to map the previous ISTA update steps to a deep network architecture that consists of a fixed number of phases, each of which corresponds to one iteration in traditional ISTA.

In order to improve reconstruction performance and increase network capacity and instead of the hand-crafted transform $\boldsymbol{\Psi}$ in Eq. (1), ISTA-Net adopts a general nonlinear transform function to sparsify natural images, denoted by $\mathcal{F}(\cdot)$, whose parameters are learnable. In particular and inspired by the powerful representation power of CNN [9] and its universal approximation property [14], we propose to design $\mathcal{F}(\cdot)$ as a combination of two linear convolutional operators (without bias terms) separated by a rectified linear unit (ReLU). As illustrated in Figure 2, the first convolutional operator in $\mathcal{F}(\cdot)$ corresponds to N_f filters (each of size 3×3 in our experiments) and the second convolutional operator corresponds to another set of N_f filters (each of size $3 \times 3 \times N_f$ in our experiments). In our implementation, we set $N_f = 32$ by default. Obviously, $\mathcal{F}(\cdot)$ can also be equivalently formulated in matrix form as $\mathcal{F}(\mathbf{x}) = \mathbf{B}ReLU(\mathbf{A}\mathbf{x})$, where $\mathbf{A}$ and $\mathbf{B}$ correspond to the above two convolutional operators, respectively. With its learnable and nonlinear characteristics, $\mathcal{F}(\cdot)$ is expected to be able to achieve a richer representation for natural images.

Replacing $\boldsymbol{\Psi}$ in Eq. (1) with $\mathcal{F}(\cdot)$, we obtain the following sparsity-inducing regularization problem with a nonlinear transform:

$$\min_{\mathbf{x}} \ \frac{1}{2}\|\boldsymbol{\Phi}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathcal{F}(\mathbf{x})\|_1. \qquad (5)$$

By solving Eq. (5) using ISTA, Eq. (2) is unchanged while Eq. (3) becomes

$$\mathbf{x}^{(k)} = \arg\min_{\mathbf{x}} \ \frac{1}{2}\|\mathbf{x} - \mathbf{r}^{(k)}\|_2^2 + \lambda\|\mathcal{F}(\mathbf{x})\|_1. \qquad (6)$$

In the following, we argue that the above two steps Eq. (2) and Eq. (6) in the k-*th* ISTA iteration both admit efficient solutions, and we cast them into two separate modules in the k-*th* phase of ISTA-Net, namely the $\mathbf{r}^{(k)}$ **module** and the $\mathbf{x}^{(k)}$ **module**, as illustrated in Figure 2.

• $\mathbf{r}^{(k)}$ **Module:** It corresponds to Eq. (2) and is used to generate the immediate reconstruction result $\mathbf{r}^{(k)}$. Note that $\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\mathbf{x}^{(k-1)} - \mathbf{y})$ is essentially the gradient of the data-fidelity term $\frac{1}{2}\|\boldsymbol{\Phi}\mathbf{x} - \mathbf{y}\|_2^2$, computed at $\mathbf{x}^{(k-1)}$. To preserve the ISTA structure while increasing network flexibility, we allow the step size ρ to vary across iterations (while it is fixed in traditional ISTA), so the output of this module with input $\mathbf{x}^{(k-1)}$ is finally defined as:

$$\mathbf{r}^{(k)} = \mathbf{x}^{(k-1)} - \rho^{(k)}\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\mathbf{x}^{(k-1)} - \mathbf{y}). \qquad (7)$$

• $\mathbf{x}^{(k)}$ **Module:** It aims to compute $\mathbf{x}^{(k)}$ according to Eq. (6) with input $\mathbf{r}^{(k)}$. Note that Eq. (6) is actually the proximal mapping $\mathrm{prox}_{\lambda\mathcal{F}}(\mathbf{r}^{(k)})$ associated with the nonlinear transform $\mathcal{F}(\cdot)$. In this paper, we propose to solve $\mathrm{prox}_{\lambda\mathcal{F}}(\mathbf{r}^{(k)})$ efficiently in two steps, which is also one of our main contributions.

First, note that $\mathbf{r}^{(k)}$ is the immediate reconstruction result of $\mathbf{x}^{(k)}$ at the k-*th* iteration. In the context of image inverse problems, one general and reasonable assumption is that each element of $(\mathbf{x}^{(k)} - \mathbf{r}^{(k)})$ follows an independent normal distribution with common zero mean and variance σ^2 [46]. Here, we also make this assumption, and then we further prove the following theorem:

Theorem 1 *Let* $X_1, ..., X_n$ *be independent normal random variables with common zero mean and variance* σ^2. *If* $\vec{X} = [X_1, ..., X_n]^\top$ *and given any matrices* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *and* $\mathbf{B} \in \mathbb{R}^{s \times m}$, *define a new random variable* $\vec{Y} = \mathbf{B}ReLU(\mathbf{A}\vec{X}) = \mathbf{B}\max(\mathbf{0}, \mathbf{A}\vec{X})$. *Then,* $\mathbb{E}[\|\vec{Y} - \mathbb{E}[\vec{Y}]\|_2^2]$ *and* $\mathbb{E}[\|\vec{X} - \mathbb{E}[\vec{X}]\|_2^2]$ *are linearly related,* i.e. $\mathbb{E}[\|\vec{Y} - \mathbb{E}[\vec{Y}]\|_2^2] = \alpha\mathbb{E}[\|\vec{X} - \mathbb{E}[\vec{X}]\|_2^2]$, *where* α *is only a function of* $\mathbf{A}$ *and* $\mathbf{B}$. *(Please refer to the **supplementary material** for the proof and more details.)*

Theorem 1 can be easily extended to a normal distribution. Suppose that $\mathbf{r}^{(k)}$ and $\mathcal{F}(\mathbf{r}^{(k)})$ are the mean values of $\mathbf{x}$ and $\mathcal{F}(\mathbf{x})$ respectively, then we can make the following
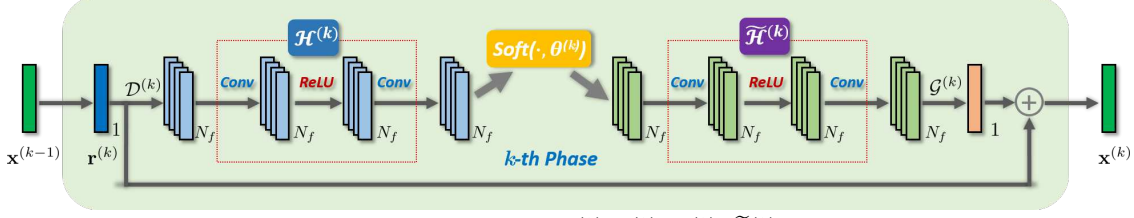
Figure 3. Illustration of the *k-th* phase of the proposed ISTA-Net$^+$. $\mathcal{D}^{(k)}, \mathcal{G}^{(k)}, \mathcal{H}^{(k)}, \widetilde{\mathcal{H}}^{(k)}$ are learnable linear convolutional operators.

approximation based on **Theorem** 1:

$$\|\mathcal{F}(\mathbf{x}) - \mathcal{F}(\mathbf{r}^{(k)})\|_2^2 \approx \alpha\|\mathbf{x} - \mathbf{r}^{(k)}\|_2^2, \qquad (8)$$

where α is a scalar that is only related to the parameters of $\mathcal{F}(\cdot)$. By incorporating this linear relationship into Eq. (6), we obtain the following optimization:

$$\mathbf{x}^{(k)} = \arg\min_{\mathbf{x}} \ \frac{1}{2}\|\mathcal{F}(\mathbf{x}) - \mathcal{F}(\mathbf{r}^{(k)})\|_2^2 + \theta\|\mathcal{F}(\mathbf{x})\|_1, \quad (9)$$

where λ and α are merged into one parameter θ, *i.e.* $\theta = \lambda\alpha$. Therefore, we get a closed-form version of $\mathcal{F}(\mathbf{x}^{(k)})$:

$$\mathcal{F}(\mathbf{x}^{(k)}) = soft(\mathcal{F}(\mathbf{r}^{(k)}), \theta). \qquad (10)$$

Second, motivated by the invertible characteristics of the wavelet transform that leads to a closed-form solution for Eq. (3), we introduce the left inverse of $\mathcal{F}(\cdot)$, denoted by $\widetilde{\mathcal{F}}(\cdot)$ such that $\widetilde{\mathcal{F}} \circ \mathcal{F} = \mathcal{I}$, where $\mathcal{I}$ is the identity operator. Specifically, $\widetilde{\mathcal{F}}(\cdot)$ is designed to exhibit a structure symmetric to that of $\mathcal{F}(\cdot)$, so it is also modeled as two linear convolutional operators separated by a ReLU operator, as shown in Figure 2. Because $\mathcal{F}(\cdot)$ and $\widetilde{\mathcal{F}}(\cdot)$ are both learnable, we will enforce the ***symmetry constraint*** $\widetilde{\mathcal{F}} \circ \mathcal{F} = \mathcal{I}$ by incorporating it into the loss function during network training. Therefore, $\mathbf{x}^{(k)}$ can be efficiently computed in closed-form as:

$$\mathbf{x}^{(k)} = \widetilde{\mathcal{F}}(soft(\mathcal{F}(\mathbf{r}^{(k)}), \theta)). \qquad (11)$$

It is worth emphasizing that θ, as a shrinkage threshold, is a learnable parameter in this module. Similarly, to increase network capacity, we do not constrain that $\mathcal{F}(\cdot)$, $\widetilde{\mathcal{F}}(\cdot)$, and θ be the same at each phase. That is, each phase of ISTA-Net has its own $\{\mathcal{F}^{(k)}(\cdot), \widetilde{\mathcal{F}}^{(k)}(\cdot), \theta^{(k)}\}$, as illustrated in Figure 2. Therefore, with all the learnable parameters, the output $\mathbf{x}^{(k)}$ in this module should be updated as:

$$\mathbf{x}^{(k)} = \widetilde{\mathcal{F}}^{(k)}(soft(\mathcal{F}^{(k)}(\mathbf{r}^{(k)}), \theta^{(k)})). \qquad (12)$$

Figure 2 clearly illustrates how Eq. (6) with the closed-form solution in Eq. (12) is mapped into a deep network in the *k-th* phase of ISTA-Net.

Parameters in ISTA-Net: Each module in each phase of ISTA-Net strictly corresponds to the updates steps in an ISTA iteration. The learnable parameter set in ISTA-Net, denoted by $\boldsymbol{\Theta}$, includes the step size $\rho^{(k)}$ in the $\mathbf{r}^{(k)}$ module, the parameters of the forward and backward transforms $\mathcal{F}^{(k)}(\cdot)$ and $\widetilde{\mathcal{F}}^{(k)}(\cdot)$, and the shrinkage threshold $\theta^{(k)}$ in the $\mathbf{x}^{(k)}$ module. As such, $\boldsymbol{\Theta} = \{\rho^{(k)}, \theta^{(k)}, \mathcal{F}^{(k)}, \widetilde{\mathcal{F}}^{(k)}\}_{k=1}^{N_p}$,

where N_p is the total number of ISTA-Net phases. All these parameters will be learned as neural network parameters.

Initialization: Like traditional ISTA, ISTA-Net also requires an initialization denoted by $\mathbf{x}^{(0)}$ in Figure 2. Instead of random values, we propose to directly use a linear mapping to compute the initialization. Specifically, given the training data pairs that include the image blocks and their corresponding CS measurements, *i.e.* $\{(\mathbf{y}_i, \mathbf{x}_i)\}_{i=1}^{N_b}$ with $\mathbf{x}_i \in \mathbb{R}^N, \mathbf{y}_i \in \mathbb{R}^M$, the linear mapping matrix, denoted by $\mathbf{Q}_{init}$, can be computed by solving a least squares problem: $\mathbf{Q}_{init} = \arg\min_{\mathbf{Q}} \|\mathbf{QY} - \mathbf{X}\|_F^2 = \mathbf{XY}^\top(\mathbf{YY}^\top)^{-1}$. Here, $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_{N_b}]$, and $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_{N_b}]$. Hence, given any input CS measurement $\mathbf{y}$, its corresponding ISTA-Net initialization $\mathbf{x}^{(0)}$ is computed as: $\mathbf{x}^{(0)} = \mathbf{Q}_{init}\mathbf{y}$.

3.3. Loss Function Design

Given the training data pairs $\{(\mathbf{y}_i, \mathbf{x}_i)\}_{i=1}^{N_b}$, ISTA-Net first takes the CS measurement $\mathbf{y}_i$ as input and generates the reconstruction result, denoted by $\mathbf{x}_i^{(N_p)}$, as output. We seek to reduce the discrepancy between $\mathbf{x}_i$ and $\mathbf{x}_i^{(N_p)}$ while satisfying the symmetry constraint $\widetilde{\mathcal{F}}^{(k)} \circ \mathcal{F}^{(k)} = \mathcal{I} \ \forall k = 1, \ldots, N_p$. Therefore, we design the end-to-end loss function for ISTA-Net as follows:

$$\mathcal{L}_{total}(\boldsymbol{\Theta}) = \mathcal{L}_{discrepancy} + \gamma\mathcal{L}_{constraint}, \qquad (13)$$

with: $\begin{cases} \mathcal{L}_{discrepancy} = \frac{1}{N_b N}\sum_{i=1}^{N_b}\|\mathbf{x}_i^{(N_p)} - \mathbf{x}_i\|_2^2 \\ \mathcal{L}_{constraint} = \frac{1}{N_b N}\sum_{i=1}^{N_b}\sum_{k=1}^{N_p}\|\widetilde{\mathcal{F}}^{(k)}(\mathcal{F}^{(k)}(\mathbf{x}_i)) - \mathbf{x}_i\|_2^2, \end{cases}$

where N_p, N_b, N, and γ are the total number of ISTA-Net phases, the total number of training blocks, the size of each block $\mathbf{x}_i$, and the regularization parameter, respectively. In our experiments, γ is set to 0.01.

4. Enhanced Version: ISTA-Net$^+$

Motivated by the fact that the residuals of natural images and videos are more compressible [38, 37], an enhanced version, dubbed ISTA-Net$^+$, is derived from ISTA-Net to further improve CS performance. Starting from Eq. (6), we assume that $\mathbf{x}^{(k)} = \mathbf{r}^{(k)} + \mathbf{w}^{(k)} + \mathbf{e}^{(k)}$, where $\mathbf{e}^{(k)}$ stands for some noise and $\mathbf{w}^{(k)}$ represents some missing high-frequency component in $\mathbf{r}^{(k)}$, which can be extracted by a linear operator $\mathcal{R}(\cdot)$ from $\mathbf{x}^{(k)}$, *i.e.* $\mathbf{w}^{(k)} = \mathcal{R}(\mathbf{x}^{(k)})$. Furthermore, $\mathcal{R}(\cdot)$ is defined as $\mathcal{R} = \mathcal{G} \circ \mathcal{D}$, where $\mathcal{D}$ corresponds to N_f filters (each of size 3×3 in our experiments)
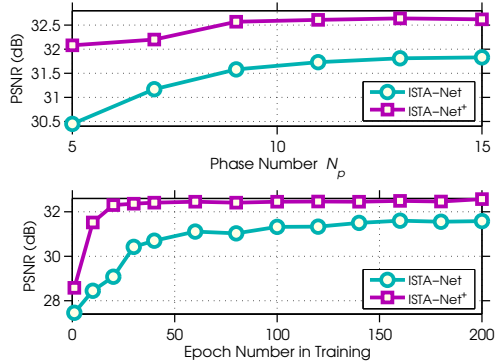
Figure 4. PSNR comparison between ISTA-Net and ISTA-Net$^+$ with various numbers of phases and epochs during training.

and $\mathcal{G}$ corresponds to 1 filter (with size $3 \times 3 \times N_f$). By modeling $\mathcal{F} = \mathcal{H} \circ \mathcal{D}$, where $\mathcal{H}$ consists of two linear convolutional operators and one ReLU, as illustrated in Figure 3, we can replace $\mathcal{F}$ in Eq. (9) with $\mathcal{H} \circ \mathcal{D}$ to obtain:

$$\min_{\mathbf{x}} \quad \frac{1}{2} ||\mathcal{H}(\mathcal{D}(\mathbf{x})) - \mathcal{H}(\mathcal{D}(\mathbf{r}^{(k)}))||_2^2 + \theta ||\mathcal{H}(\mathcal{D}(\mathbf{x}))||_1. \quad (14)$$

By exploiting the approximation used in Eq. (9) and following the same strategy as in ISTA-Net, we define the left inverse of $\mathcal{H}$ as $\widetilde{\mathcal{H}}$, which has a structure symmetric to that of $\mathcal{H}$ and satisfies the symmetry constraint $\widetilde{\mathcal{H}} \circ \mathcal{H} = \mathcal{I}$. Thus, the closed form of the ISTA-Net$^+$ update for $\mathbf{x}^{(k)}$ is:

$$\mathbf{x}^{(k)} = \mathbf{r}^{(k)} + \mathcal{G}(\widetilde{\mathcal{H}}(soft(\mathcal{H}(\mathcal{D}(\mathbf{r}^{(k)})), \theta))). \quad (15)$$

Similar to ISTA-Net, each phase of ISTA-Net$^+$ also has its own learnable parameters, and the k-th phase of ISTA-Net$^+$ is illustrated in Figure 3. Hence, the learnable parameter set Θ^+ of ISTA-Net$^+$ is $\Theta^+ = \{\rho^{(k)}, \theta^{(k)}, \mathcal{D}^{(k)}, \mathcal{G}^{(k)}, \mathcal{H}^{(k)}, \widetilde{\mathcal{H}}^{(k)}\}_{k=1}^{N_p}$. The loss function of ISTA-Net$^+$ is analogously designed by incorporating the constraints $\widetilde{\mathcal{H}}^{(k)} \circ \mathcal{H}^{(k)} = \mathcal{I}$ into Eq. (13).

5. Experimental Results

For fair comparison, we use the same set of 91 images as in [21] for training. Following common practices in previous CS work, we generate the training data pairs $\{(\mathbf{y}_i, \mathbf{x}_i)\}_{i=1}^{N_b}$ by first extracting the luminance component of 88,912 randomly cropped image blocks (each of size 33×33), *i.e.* N_b=88,912 and N=1,089. Then, for a given CS ratio, the corresponding measurement matrix $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$ is constructed by generating a random Gaussian matrix and then orthogonalizing its rows, *i.e.* $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Applying $\mathbf{y}_i = \boldsymbol{\Phi}\mathbf{x}_i$ yields the set of CS measurements, where $\mathbf{x}_i$ is the vectorized version of an image block. We use TensorFlow [1] to implement and train the ISTA-Nets separately for a range of CS ratios $\{1\%, 4\%, 10\%, 25\%, 30\%, 40\%, 50\%\}$. To train the networks, we use Adam optimization [19] with a learning

rate of 0.0001 (200 epochs), and a batch size of 64. All the experiments are performed on a workstation with Intel Core i7-6820 CPU and GTX1060 GPU. Training ISTA-Nets with phase number N_p=9 roughly takes 10 hours. For testing, we utilize two widely used benchmark datasets: Set11 [21] and BSD68 [27], which have 11 and 68 gray images, respectively. The reconstruction results are reported as the average Peak Signal-to-Noise Ratio (PSNR) over the test images.

5.1. ISTA-Net vs. ISTA-Net$^+$

To demonstrate the superiority of ISTA-Net$^+$ over ISTA-Net, we compare them in two aspects: performance and convergence. Figure 4 (top) shows the average PSNR curves for the testing set (Set11) with respect to different phase numbers, when the CS ratio is 25%. We observe that both PSNR curves increase as phase number N_p increases; however, the curves are almost flat when $N_p \geq 9$. Thus, considering the tradeoff between network complexity and reconstruction performance, we set the default phase number N_p=9 for both ISTA-Net and ISTA-Net$^+$ in the rest of the experiments. Clearly, ISTA-Net$^+$ achieves about 1 dB gain over ISTA-Net when N_p=9. Furthermore, Figure 4 (bottom) plots the average PSNR curves for Set11 with respect to different numbers of epochs during training, when the CS ratio is 25% and N_p=9. Both ISTA-Nets get higher PSNR when trained for a larger number of epochs, but ISTA-Net$^+$ achieves faster training convergence and better reconstruction performance on the test set (Set11). Due to limited space, please refer to **supplementary material** for the filters that are learned by ISTA-Nets.

We attribute the superiority of ISTA-Net$^+$ over ISTA-Net to two factors. First, ISTA-Net$^+$ explicitly sparsifies the images in the residual domain, leading to a sparser representation as compared to ISTA-Net. Second, the skip connections introduced by ISTA-Net$^+$ coincide with the central idea of the popular ResNet [13] architecture, which facilitates the training of deeper networks.

5.2. Comparison with State-of-the-Art Methods

We compare our proposed ISTA-Net and ISTA-Net$^+$ with five recent and state-of-the-art image CS methods, namely TVAL3 [22], D-AMP [28], IRCNN [49], SDA [30], and ReconNet [21][1]. The first three are optimization-based methods, while the last two are network-based methods. The average PSNR reconstruction performance on Set11 with respect to seven CS ratios are summarized in Table 1. For fair comparison and following the evaluation strategy of [21], all the competing methods reconstruct each image block from its CS measurement independently. From Table 1, we observe that SDA and ReconNet work better at

---

[1]We use the source code made publicly available by the authors of T-VAL3 [22], D-AMP [28], IRCNN [49], and ReconNet [21] and implement SDA [30] ourselves, since its source code is not available.
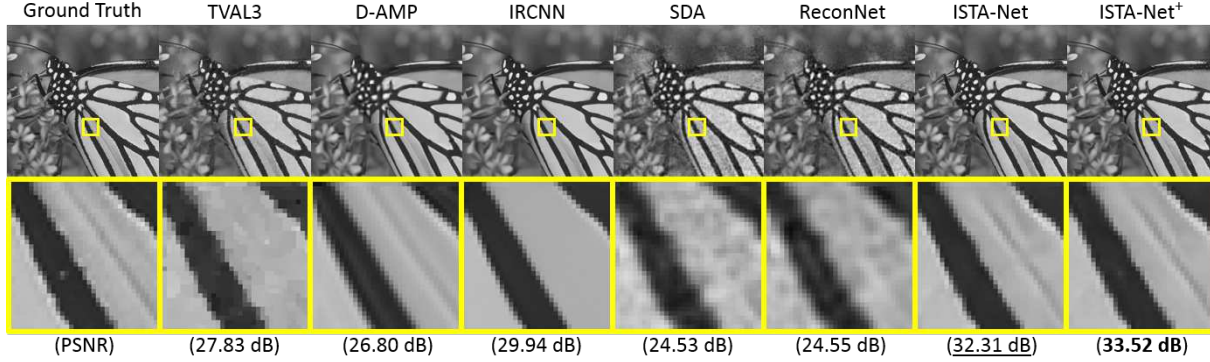
Figure 5. Comparison of seven CS reconstruction methods (including our ISTA-Net and ISTA-Net$^+$), when applied to the *Butterfly* image in Set11 (CS ratio is 25%).

Table 1. Average PSNR (dB) performance comparisons on Set11 with different CS ratios. The best performance is labeled in bold and the second best is underlined. Note that the last two columns is a run-time analysis of all the competing methods, showing the average time to reconstruct a 256×256 image and the corresponding frames-per-second (FPS).

| Algorithm | CS Ratio | | | | | | | Time | FPS |
	50%	40%	30%	25%	10%	4%	1%	CPU/GPU	CPU/GPU
TVAL3 [22]	33.55	31.46	29.23	27.92	22.99	18.75	16.43	3.135s/——	0.32/——
D-AMP [28]	35.92	33.56	30.39	28.46	22.64	18.40	5.21	51.21s/——	0.02/——
IRCNN [49]	36.23	34.06	31.18	30.07	24.02	17.56	7.70	——/68.42s	——/0.015
SDA [30]	28.95	27.79	26.63	25.34	22.65	20.12	17.29	——/0.0032s	——/312.5
ReconNet [21]	31.50	30.58	28.74	25.60	24.28	20.63	17.27	——/0.016s	——/62.5
ISTA-Net	<u>37.43</u>	<u>35.36</u>	<u>32.91</u>	<u>31.53</u>	<u>25.80</u>	<u>21.23</u>	<u>17.30</u>	0.923s/0.039s	1.08/25.6
ISTA-Net$^+$	**38.07**	**36.06**	**33.82**	**32.57**	**26.64**	**21.31**	**17.34**	1.375s/0.047s	0.73/21.3

extremely low CS ratios of 1% and 4%, while traditional optimization-based methods perform better at higher CS ratios. However, ISTA-Net and ISTA-Net$^+$ outperform all the existing methods by a large margin across all the CS ratios. This clearly demonstrates that they combine the merits of both categories of CS methods. As expected, ISTA-Net$^+$ performs better than ISTA-Net. The last two columns in Table 1 is a run-time analysis of all the competing methods. These results indicate that the proposed ISTA-Nets produce consistently better reconstruction results, while remaining computationally attractive. In Figure 5, we show the reconstructions of all seven methods of the *Butterfly* image when the CS ratio is 25%. The proposed ISTA-Net$^+$ is able to reconstruct more details and sharper edges.

To further validate the generalizability of our ISTA-Nets, we also compare them to network-based methods SDA and ReconNet on the larger BSD68 dataset. As shown in Table 2, ISTA-Net$^+$ achieves the best performance, while ISTA-Net registers second best among all five CS ratios. ISTA-Nets outperform SDA and ReconNet, especially at higher CS ratios. In addition, it is worth emphasizing that a pre-trained ISTA-Net or ISTA-Net$^+$ using one CS measurement matrix Φ can be directly used for any new measurement matrix with the same CS ratio as Φ, avoiding training new network from scratch. The only difference is that we need to recalculate the initialization matrix $\mathbf{Q}_{init}$ for the new mea-

Table 2. Average PSNR (dB) performance comparison of various network-based algorithms on the BSD68 dataset.

| Algorithm | CS Ratio | | | | |
	50%	40%	30%	10%	4%
SDA [30]	28.35	27.41	26.38	23.12	21.32
ReconNet [21]	29.86	29.08	27.53	24.15	21.66
ISTA-Net	<u>33.60</u>	<u>31.85</u>	<u>29.93</u>	<u>25.02</u>	<u>22.12</u>
ISTA-Net$^+$	**34.01**	**32.21**	**30.34**	**25.33**	**22.17**

surement matrix, which usually takes less than 1 second. Please refer to **supplementary material** for more results.

5.3. Comparison with ADMM-Net for CS-MRI

To demonstrate the generality of ISTA-Net$^+$, we directly extend ISTA-Net$^+$ to the specific problem of CS MRI reconstruction, which aims at reconstructing MR images from a small number of under-sampled data in *k*-space. In this application and following common practices, we set the sampling matrix Φ in Eq. (1) to $\Phi = \mathbf{PF}$, where $\mathbf{P}$ is an under-sampling matrix and $\mathbf{F}$ is the discrete Fourier transform. In this case, we compare against ADMM-Net [44][2], which is a network-based method inspired by ADMM and specifically designed for the CS-MRI domain. It is worthwhile to note that ADMM-Net cannot be trivially extended to other CS domains, since it imposes a specific structure to the sampling matrix Φ. Utilizing the same training and
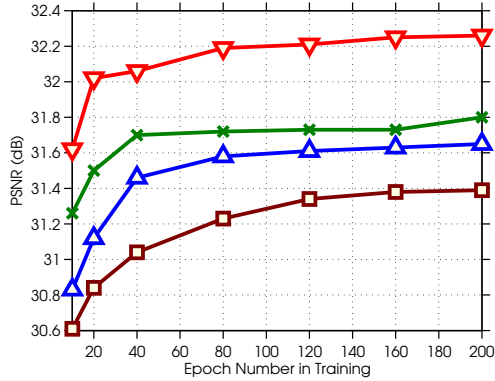
---

[2]https://github.com/yangyan92/Deep-ADMM-Net

Figure 6. PSNR (dB) comparison between two versions of ISTA-Net$^+$: with and without ReLU (when N_f=8 and N_f=16).

Table 3. Average PSNR (dB) comparison between ADMM-Net [44] and our proposed ISTA-Nets for CS-MRI.

Algorithm	CS Ratio				Time
	20%	30%	40%	50%	GPU
ADMM-Net	37.17	39.84	41.56	43.00	0.9535s
ISTA-Net	38.30	40.52	42.12	43.60	0.1246s
ISTA-Net$^+$	**38.73**	**40.89**	**42.52**	**44.09**	0.1437s

Table 4. Average PSNR (dB) performance with different shared types of ISTA-Net$^+$.

Shared Type	Number of Parameters	PSNR
Shared $\rho^{(k)}, \theta^{(k)}, \mathcal{T}^{(k)}$	(37440+1+1)=**37,442**	31.53
Shared $\rho^{(k)}, \mathcal{T}^{(k)}$	(37440+1)+1*9=37,450	32.28
Shared $\theta^{(k)}, \mathcal{T}^{(k)}$	(37440+1)+1*9=37,450	32.08
Shared $\mathcal{T}^{(k)}$	37440+(1+1)*9=37,458	32.36
Unshared (default)	(37440+1+1)*9=336,978	**32.57**

testing brain medical images as ADMM-Net, the CS-MRI results of ISTA-Nets with N_p=11 phases are summarized in Table 3 for CS ratios of 20%, 30%, 40% and 50%. It is clear that ISTA-Nets outperform ADMM-Net not only in terms of reconstruction but also in terms of runtime.

5.4. Ablation Studies and Discussions

This section mainly focuses on the nonlinearity and flexibility of the proposed ISTA-Nets. In what follows, we analyze ISTA-Net$^+$ with N_p=9 phases.

• **Linear vs. Nonlinear Transforms:** The nonlinearity of ISTA-Net$^+$ is introduced by the ReLU operator in $\mathcal{H}^{(k)}$ and $\widetilde{\mathcal{H}}^{(k)}$, as shown in Figure 3. To evaluate the impact of the nonlinearity, we train ISTA-Net$^+$ models with ReLU (nonlinear transforms) and without ReLU (linear transforms). Figure 6 plots the average PSNR curves for each ISTA-Net$^+$ variant on Set11 throughout training. Note that parameter N_f, the number of feature maps in $\mathcal{H}^{(k)}$ and $\widetilde{\mathcal{H}}^{(k)}$, is set to 8 or 16 in this experiment. It is clear that the nonlinearity introduced by the ReLU is critical for high fidelity CS reconstruction performance. In addition, when $N_f > 30$, experiments indicate that ISTA-Net$^+$ without ReLU is sig-

nificantly less stable in training than ISTA-Net$^+$ with ReLU, which still performs well. We conclude that the nonlinearity plays an important role in facilitating satisfaction of the symmetry constraint, improving network stability, and learning a suitable transform possible for CS.

• **Shared vs. Unshared:** As described previously, each phase of ISTA-Net$^+$ (N_f=32) has three types of parameters with their dimensionality listed in parentheses: step size $\rho^{(k)}$ (1), threshold $\theta^{(k)}$ (1), and transform $\mathcal{T}^{(k)} = \{\mathcal{D}^{(k)}, \mathcal{G}^{(k)}, \mathcal{H}^{(k)}, \widetilde{\mathcal{H}}^{(k)}\}$ ($32 \times 3 \times 3 + 32 \times 3 \times 3 \times 32 \times 2 + 32 \times 3 \times 3 \times 32 \times 2 + 1 \times 3 \times 3 \times 32 = 37440$). The flexibility of ISTA-Net$^+$ indicates that the same type of parameters in different phases do not need to be the same. To demonstrate the impact of this flexibility, we train several variants of ISTA-Net$^+$, where we vary the parameters that are shared among the phases. A summary of the average PSNR results on Set11 at a 25% CS ratio is reported in Table 4. Obviously, the default *unshared* ISTA-Net$^+$ (most flexible with largest number of parameters) achieves the best performance, while the variant of ISTA-Net$^+$ that shares all parameters $(\rho^{(k)}, \theta^{(k)}, \mathcal{T}^{(k)})$ in all its phases (least flexible with smallest number of parameters) obtains the worst performance. When only $(\rho^{(k)}, \mathcal{T}^{(k)})$ or $(\theta^{(k)}, \mathcal{T}^{(k)})$ are shared, these ISTA-Net$^+$ variants register 0.75dB and 0.55dB gains over he variant with all shared parameters. Interestingly, the ISTA-Net$^+$ variant with only shared transforms $\mathcal{T}^{(k)}$ obtains very competitive PSNR results compared to the unshared variant. This indicates that further compression in ISTA-Net$^+$ parameters is possible, with limited affect on reconstruction performance.

6. Conclusion and Future Work

Inspired by the Iterative Shrinkage-Thresholding Algorithm (ISTA), we propose a novel structured deep network for image compressive sensing (CS) reconstruction, dubbed ISTA-Net, as well as, its enhanced version ISTA-Net$^+$. The proposed ISTA-Nets have well-defined interpretability, and make full use of the merits of both optimization-based and network-based CS methods. All the parameters in ISTA-Nets are discriminately learned end-to-end. Extensive experiments show that ISTA-Nets greatly improve upon the results of state-of-the-art CS methods, while maintaining a fast runtime. Since the developed strategy to solve the proximal mapping problem associated to a nonlinear sparsifying transform is quite general and efficient, one direction of interest is to design deep networks based on other optimization inspirations, such as FISTA [4]. The other direction of our future work is to extend ISTA-Nets for other image inverse problems, such as deconvolution and inpainting.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] A. Adler, D. Boublil, M. Elad, and M. Zibulevsky. A deep learning approach to block-based compressed sensing of images. *arXiv preprint arXiv:1606.01519*, 2016.

[3] M. V. Afonso, J. M. Bioucas-Dias, and M. A. Figueiredo. An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems. *IEEE Transactions on Image Processing*, 20(3):681–695, 2011.

[4] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[5] M. Borgerding, P. Schniter, and S. Rangan. AMP-inspired deep networks for sparse linear inverse problems. *IEEE Transactions on Signal Processing*, 65(16):4293–4308, 2017.

[6] E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

[7] J. H. R. Chang, C.-L. Li, B. Póczos, B. V. K. V. Kumar, and A. C. Sankaranarayanan. One network to solve them all — solving linear inverse problems using deep projection models. *ICCV*, 2017.

[8] Y. Chen, W. Yu, and T. Pock. On learning optimized reaction diffusion processes for effective image restoration. In *CVPR*, 2015.

[9] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199. Springer, 2014.

[10] W. Dong, G. Shi, X. Li, Y. Ma, and F. Huang. Compressive sensing via nonlocal low-rank regularization. *IEEE Transactions on Image Processing*, 23(8):3618–3632, 2014.

[11] M. F. Duarte, M. A. Davenport, D. Takbar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):83–91, 2008.

[12] K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. In *International Conference on Machine Learning*, pages 399–406, 2010.

[13] L. He and L. Carin. Exploiting structure in wavelet-based bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 57(9):3488–3497, 2009.

[14] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[15] M. Iliadis, L. Spinoulas, and A. K. Katsaggelos. Deep fully-connected networks for video compressive sensing. *Digital Signal Processing*, 72:9–18, 2018.

[16] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[17] U. S. Kamilov and H. Mansour. Learning optimal nonlinearities for iterative thresholding algorithms. *IEEE Signal Processing Letters*, 23(5):747–751, 2016.

[18] Y. Kim, M. S. Nadar, and A. Bilgin. Compressed sensing using a gaussian scale mixtures model in wavelet domain. In *ICIP*, pages 3365–3368. IEEE, 2010.

[19] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[21] K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, and A. Ashok. ReconNet: non-iterative reconstruction of images from compressively sensed measurements. In *CVPR*, pages 449–458, 2016.

[22] C. Li, W. Yin, H. Jiang, and Y. Zhang. An efficient augmented lagrangian method with applications to total variation minimization. *Computational Optimization and Applications*, 56(3):507–530, 2013.

[23] D. Liu, Z. Wang, B. Wen, J. Yang, W. Han, and T. S. Huang. Robust single image super-resolution via deep networks with sparse prior. *IEEE Transactions on Image Processing*, 25(7):3194–3207, 2016.

[24] A. Liutkus, D. Martina, S. Popoff, G. Chardon, O. Katz, G. Lerosey, S. Gigan, L. Daudet, and I. Carron. Imaging with nature: Compressive imaging using a multiply scattering medium. *Scientific reports*, 4, 2014.

[25] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[26] M. Lustig, D. Donoho, and J. M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[27] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*. IEEE, 2001.

[28] C. A. Metzler, A. Maleki, and R. G. Baraniuk. From denoising to compressed sensing. *IEEE Transactions on Information Theory*, 62(9):5117–5144, 2016.

[29] A. Mousavi and R. G. Baraniuk. Learning to invert: Signal recovery via deep convolutional networks. *ICASSP*, 2017.

[30] A. Mousavi, A. B. Patel, and R. G. Baraniuk. A deep learning approach to structured signal recovery. In *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1336–1343. IEEE, 2015.

[31] S. Mun and J. E. Fowler. Block compressed sensing of images using directional transforms. In *ICIP*, pages 3021–3024. IEEE, 2009.

[32] G. Riegler, M. Rüther, and H. Bischof. Atgv-net: accurate depth super-resolution. In *ECCV*, pages 268–284, 2016.

[33] F. Rousset, N. Ducros, A. Farina, G. Valentini, C. DAndrea, and F. Peyrin. Adaptive basis scan by wavelet prediction for single-pixel imaging. *IEEE Transactions on Computational Imaging*, 3(1):36–46, 2017.

[34] A. C. Sankaranarayanan, C. Studer, and R. G. Baraniuk. Csmuvi: Video compressive sensing for spatial-multiplexing cameras. In *ICCP*, pages 1–10. IEEE, 2012.

[35] U. Schmidt and S. Roth. Shrinkage fields for effective image restoration. In *CVPR*, pages 2774–2781, 2014.

[36] S. K. Sharma, E. Lagunas, S. Chatzinotas, and B. Ottersten. Application of compressive sensing in cognitive radio communications: A survey. *IEEE Communication Surveys & Tutorials*, 18(3):1838–1860, 2016.

[37] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1649–1668, 2012.

[38] G. K. Wallace. The JPEG still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992.

[39] S. Wang, S. Fidler, and R. Urtasun. Proximal deep structured models. In *NIPS*, pages 865–873, 2016.

[40] Z. Wang, D. Liu, S. Chang, Q. Ling, Y. Yang, and T. S. Huang. D3: Deep dual-domain based fast restoration of JPEG-compressed images. In *CVPR*, pages 2764–2772, 2016.

[41] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang. Deep networks for image super-resolution with sparse prior. In *ICCV*, pages 370–378, 2015.

[42] J. Xie, L. Xu, and E. Chen. Image denoising and inpainting with deep neural networks. In *NIPS*, pages 341–349, 2012.

[43] B. Xin, Y. Wang, W. Gao, D. Wipf, and B. Wang. Maximal sparsity with deep networks? In *NIPS*, pages 4340–4348, 2016.

[44] Y. Yang, J. Sun, H. Li, and Z. Xu. Deep ADMM-Net for compressive sensing MRI. In *NIPS*, pages 10–18, 2016.

[45] J. Zhang, C. Zhao, D. Zhao, and W. Gao. Image compressive sensing recovery using adaptively learned sparsifying basis via l0 minimization. *Signal Processing*, 103:114–126, 2014.

[46] J. Zhang, D. Zhao, and W. Gao. Group-based sparse representation for image restoration. *IEEE Transactions on Image Processing*, 23(8):3336–3351, 2014.

[47] J. Zhang, D. Zhao, F. Jiang, and W. Gao. Structural group sparse representation for image compressive sensing recovery. In *Data Compression Conference (DCC)*, pages 331–340, 2013.

[48] J. Zhang, D. Zhao, C. Zhao, R. Xiong, S. Ma, and W. Gao. Image compressive sensing recovery via collaborative sparsity. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2(3):380–391, 2012.

[49] K. Zhang, W. Zuo, S. Gu, and L. Zhang. Learning deep CNN denoiser prior for image restoration. *CVPR*, 2017.

[50] Z. Zhang, T.-P. Jung, S. Makeig, and B. D. Rao. Compressed sensing for energy-efficient wireless telemonitoring of noninvasive fetal ECG via block sparse bayesian learning. *IEEE Transactions on Biomedical Engineering*, 60(2):300–309, 2013.

[51] C. Zhao, S. Ma, J. Zhang, R. Xiong, and W. Gao. Video compressive sensing reconstruction via reweighted residual sparsity. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(6):1182–1195, 2017.

[52] C. Zhao, J. Zhang, S. Ma, and W. Gao. Nonconvex lp nuclear norm based admm framework for compressed sensing. In *Data Compression Conference (DCC)*, pages 161–170, 2016.