

# Optimization Learning: Perspective, Method, and Applications

Risheng Liu

International School of Information Science & Engineering  
Dalian University of Technology  
rslu@dlut.edu.cn

## Abstract

Numerous tasks at the core of statistics, learning, and vision areas are specific cases of ill-posed inverse problems. Recently, learning-based (e.g., deep) iterative methods have been empirically shown to be useful for these problems. Nevertheless, integrating learnable structures into iterations is still a laborious process, which can only be guided by intuitions or empirical insights. Moreover, there is a lack of rigorous analysis of the convergence behaviors of these reimplemented iterations, and thus the significance of such methods is a little bit vague. We move beyond these limits and propose a theoretically guaranteed optimization learning paradigm, a generic and provable paradigm for nonconvex inverse problems, and develop a series of convergent deep models. Our theoretical analysis reveals that the proposed optimization learning paradigm allows us to generate globally convergent trajectories for learning-based iterative methods. Thanks to the superiority of our framework, we achieve state-of-the-art performance on different real applications.

## 1 Introduction

In applications throughout statistics, machine learning and computer vision, one is often faced with the challenge of solving ill-posed inverse problems. In general, the basic inverse problem leads to a discrete linear system of the form  $\mathcal{T}(\mathbf{x}) = \mathbf{y} + \mathbf{n}$ , where  $\mathbf{x} \in \mathbb{R}^D$  is the latent variable to be estimated,  $\mathcal{T}$  denotes some given linear operations on  $\mathbf{x}$ , and  $\mathbf{y}, \mathbf{n} \in \mathbb{R}^D$  are the observation and an unknown error term, respectively. Typically, these inverse problems can be addressed by solving the composite minimization model:

$$\min_{\mathbf{x}} \Psi(\mathbf{x}) := f(\mathbf{x}; \mathcal{T}, \mathbf{y}) + g(\mathbf{x}), \quad (1)$$

where  $f$  is the fidelity that captures the loss of data fitting, and  $g$  refers to the prior that promotes desired distribution on the solution. Recent studies illustrate that many problems (e.g., image deconvolution, matrix factorization and dictionary learning) naturally require to be solved in the nonconvex scenario. This trend motivates us to investigate Nonconvex Inverse Problems (NIPs) in the form of Eq. (1) and with the

practical configuration that  $f$  is continuously differentiable,  $g$  is nonsmooth, and both  $f$  and  $g$  are possibly nonconvex.

Over the past decades, a broad class of first-order methods have been developed to solve special instances of Eq. (1). Many learning and vision tasks have been formulated as the problems of optimizing hand-designed mathematical convex models. Unfortunately, abstractly designed general optimization models may be lack of flexibility and robustness, especially in real-world scenarios.

In recent years, various learning-based strategies have been proposed to address practical inverse problems in the form of Eq. (1). A variety of deep neural networks (DNNs) have been established and trained for different learning and vision problems (e.g., [LeCun *et al.*, 2015][He *et al.*, 2016]). Though with relatively good practical performance on specific applications, the theoretical investigations (e.g., the interpretability and guarantees) are the most important missing footstones for existing experience-based DNNs.

Some preliminary works have been developed to design DNNs based on optimization process (e.g., [Chen *et al.*, 2017]). However, due to the naive combination strategies (e.g., directly replace iterations by networks), it is still challenging to strictly analyze their propagation behaviors.

To break the limits of prevalent approaches, we establish an optimization learning paradigm [Liu *et al.*, 2019a], a *generic* and *convergent* algorithmic framework that combines together the learnable architecture (e.g., mainstream deep networks) with principled knowledges (formulated by mathematical models), illustrated in Fig. 1. Based on the optimization learning paradigm, we develop a series of deep models [Liu *et al.*, 2018c][Liu *et al.*, 2018b][Liu *et al.*, 2018d][Liu *et al.*, 2019c][Liu *et al.*, 2019b][Liu *et al.*, 2018a][Liu *et al.*, 2019d] to incorporate optimization process and trainable architectures to design knowledge-driven deep learning models.

On the one hand, our framework actually provides a way to learn data-dependent numerical solvers [Liu *et al.*, 2018c][Liu *et al.*, 2018b][Liu *et al.*, 2018a][Liu *et al.*, 2019b][Liu *et al.*, 2019a] for the instances of Eq. (1). Compared with these naive unrolling-based approaches (e.g., [Chen *et al.*, 2017]), which lack theoretical guarantees, our main advantage is that we can obtain iterative process, which strictly converge to the critical point of the considered optimization model, even in the complex nonconvex and nonsmooth scenarios. On the other hand, by slightly relaxing the exact optimality constraints

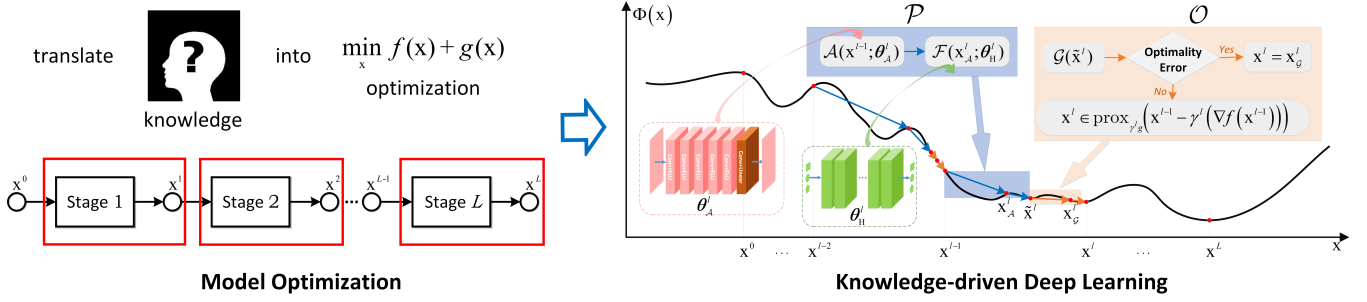


Figure 1: Illustration of our knowledge-driven deep learning paradigm.

during propagations, we can obtain an interpretable framework to integrate mathematical principles (*i.e.*, formulated by model-based building-block) and experience of the tasks (*i.e.*, network structures designed in heuristic manners) for collaborative end-to-end learning [Liu *et al.*, 2018b][Liu *et al.*, 2018a][Liu *et al.*, 2019b][Liu *et al.*, 2019a].

In summary, the contributions of our works mainly include:

- We provided a model-inspired paradigm to establish building-block modules for deep model design. Different from existing trainable iteration methods, in which the architectures are built either from specific prior formulations or completely in heuristic manners, we develop a flexible framework to integrate both data (investigated from training set) and knowledge (incorporated into principled priors) for deep propagations construction.
- By introducing an optimality error checking condition together with a proximal feedback mechanism, we prove in theory that our generated deep propagation is globally convergent to the critical point of the given optimization model. Such strict convergent guarantee is just the main advantage against these existing deep iterations designed in heuristic manner.
- We provide a practical and effective ensemble of domain knowledge and sophisticated learned data distributions and design a series of algorithms to address practical vision applications. We bring the expressive power of knowledge-based and data-driven methodologies to yield state-of-the-art performance on real vision tasks.

## 2 Our Convergent Optimization Paradigm

This section develops our general and convergent optimization learning paradigm for nonconvex inverse problems in Eq. (1). The convergence behaviors are also investigated accordingly. Hereafter, some fairly loose assumptions are enforced on Eq. (1):  $f$  is proper and Lipschitz smooth (with modulus  $L$ ) on a bounded set,  $g$  is proper, lower semi-continuous and proximal<sup>1</sup> and  $\Psi$  is coercive.

### 2.1 Abstract Iterative Modularization

A large amount of first-order methods can be summarized as forward-backward-type iterations. This motivates us to

<sup>1</sup>The function  $g$  is proximal if  $\min_{\mathbf{x}} g(\mathbf{x}) + \frac{\gamma}{2} \|\mathbf{x} - \mathbf{y}\|^2$  can be easily solved by the given  $\mathbf{y}$  and  $\gamma > 0$ .

### Algorithm 1 General Optimization Learning Paradigm

**Require:**  $\mathbf{x}^0$ ,  $\mathcal{A} = \{\mathcal{A}_g, \mathcal{A}_f\}$ ,  $\{0 < 2C^k < \mu^k < \infty\}$ , and  $\{0 < \gamma^k < 1/L\}$ .

- 1: **while** not converged **do**
- 2:  $\mathbf{u}^k = \mathcal{A}_g \circ \mathcal{A}_f(\mathbf{x}^k)$ .
- 3:  $\tilde{\mathbf{u}}^k \in \text{prox}_{\gamma^k g}(\mathbf{u}^k - \gamma^k (\nabla f(\mathbf{u}^k) + \mu^k(\mathbf{u}^k - \mathbf{x}^k)))$ .
- 4: **if**  $\|\mathbf{d}_{\Psi^k}^{\tilde{\mathbf{u}}^k}\| \leq C^k \|\tilde{\mathbf{u}}^k - \mathbf{x}^k\|$  **then**
- 5:  $\mathbf{v}^k = \tilde{\mathbf{u}}^k$ .
- 6: **else**
- 7:  $\mathbf{v}^k = \mathbf{x}^k$ .
- 8: **end if**
- 9:  $\mathbf{x}^{k+1} \in \text{prox}_{\gamma^k g}(\mathbf{v}^k - \gamma^k \nabla f(\mathbf{v}^k))$ .
- 10: **end while**

consider the following even more abstract updating principle:

$$\mathbf{x}^{k+1} = \mathcal{A}_g \circ \mathcal{A}_f(\mathbf{x}^k), \quad (2)$$

where  $\mathcal{A}_f$  and  $\mathcal{A}_g$  respectively stand for the user-specified modules for  $f$  and  $g$ , and  $\circ$  denotes operation composition. Building upon this formulation, it is easy to see that designing a learning-based iterative method reduces to the problem of iteratively specifying and learning  $\mathcal{A}_f$  and  $\mathcal{A}_g$ .

It is straightforward that most prevalent approaches naturally fall into this general formulation. Nevertheless, currently it is still impossible to provide any strict theoretical results for practical trajectories of Eq. (2). This is mainly due to the lack of efficient mechanisms to control the propagations generated by these handcrafted operations. Fortunately, in the following, we will introduce our policies to automatically guide the iterations in Eq. (2), resulting in a series of *theoretically convergent learning-based iterative methods*.

### 2.2 Implicit Momentum via Error Control

In this subsection, we show how to address this issue by controlling the first-order optimality error during iterations. Specifically, we consider the auxiliary of  $\Psi$  at  $\mathbf{x}^k$  (denoted as  $\Psi^k$ ) and denote its sub-differential (denoted as  $\mathbf{d}_{\Psi^k}^{\mathbf{x}}$ )<sup>2</sup> as

$$\begin{aligned} \Psi^k(\mathbf{x}) &= f(\mathbf{x}) + g(\mathbf{x}) + \frac{\mu^k}{2} \|\mathbf{x} - \mathbf{x}^k\|^2, \\ \mathbf{d}_{\Psi^k}^{\mathbf{x}} &= \mathbf{d}_g^{\mathbf{x}} + \nabla f(\mathbf{x}) + \mu^k(\mathbf{x} - \mathbf{x}^k) \in \partial \Psi^k(\mathbf{x}), \end{aligned} \quad (3)$$

<sup>2</sup>Strictly speaking,  $\partial \Psi^k(\mathbf{x})$  is the so-called limiting Frechét sub-differential. We state its formal definition and propose a practical computation scheme for  $\mathbf{d}_{\Psi^k}^{\mathbf{x}}$  in Appendix of [Liu *et al.*, 2019a].

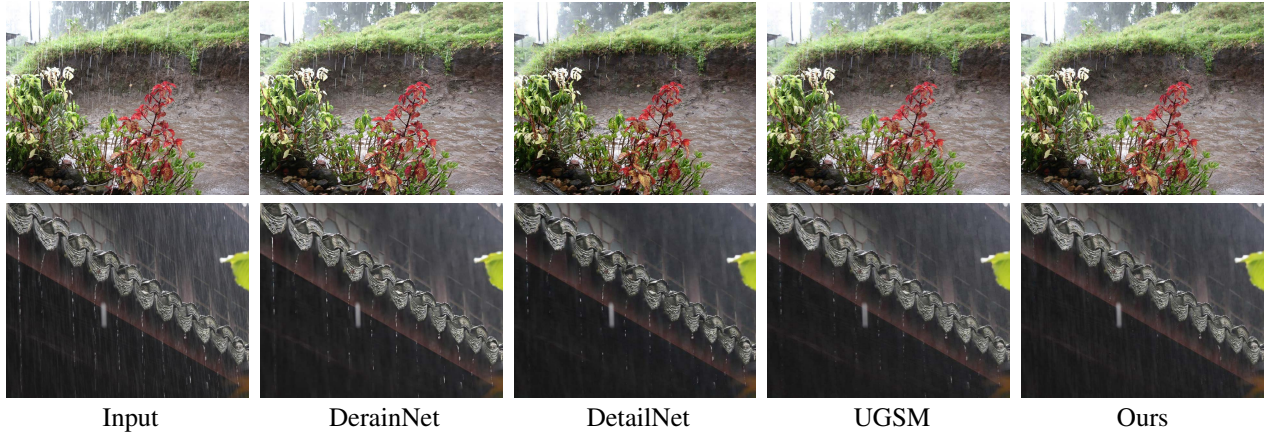


Figure 2: Rain streaks removal results of our algorithm with comparisons to the state-of-the-art approaches on real-world rainy images.

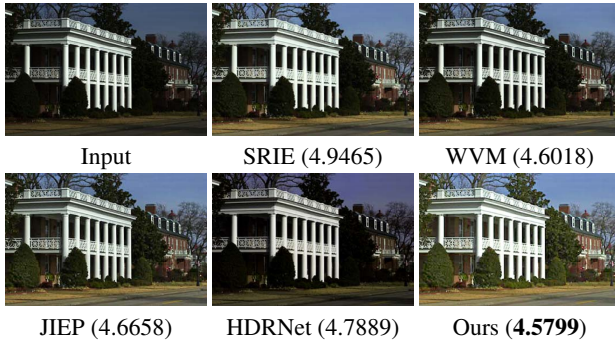


Figure 3: Low-light image enhancement results on an example selected from NASA benchmark. The scores in the brackets are the NIQE values.



Figure 4: Non-blind deblurring results. The PSNR / SSIM scores are reported below each image.

where  $\mu^k > 0$  is the penalty parameter and  $\mathbf{d}_g^x \in \partial g(\mathbf{x})$ .

As shown in Alg. 1, at stage  $k$ , a variable  $\tilde{\mathbf{u}}^k$  is obtained by proximally minimizing  $\Psi^k$  at  $\mathbf{u}^k$  (i.e., Step 3 of Alg. 1). Roughly, this new variable is just an ensemble of the last updated  $\mathbf{x}^k$  and the output  $\mathbf{u}^k$  of user-specified  $\mathcal{A}$  following the specific proximal structure in Eq. (1). Then the monitor is obtained by checking the boundedness of  $\mathbf{d}_{\Psi^k}^{\tilde{\mathbf{u}}^k}$ . Notice that the constant  $C^k$  actually reveals our tolerance to the inexactness of  $\mathcal{A}$  at  $k$ -th iteration.

**Proposition 1.** *Let  $\{\mathbf{x}^k, \tilde{\mathbf{u}}^k, \mathbf{v}^k\}_{k \in \mathbb{N}}$  be the sequences generated by Alg. 1. Then there exist two sequences  $\{\alpha^k | \alpha^k > 0\}_{k \in \mathbb{N}}$  and  $\{\beta^k | \beta^k > 0\}_{k \in \mathbb{N}}$  and  $\Psi(\tilde{\mathbf{u}}^k) \leq \Psi(\mathbf{x}^k) - \beta^k \|\tilde{\mathbf{u}}^k - \mathbf{x}^k\|^2$  are respectively satisfied.*

Equipped with Proposition 1, it will be straightforward to guarantee that the objective values generated by Alg. 1 (i.e.,  $\{\Psi(\mathbf{x}^k)\}_{k \in \mathbb{N}}$ ) also has sufficient descent. Then the global convergence of our flexible iterative modularization algorithm is proved as follows. Please refer to [Liu *et al.*, 2019a] for more details.

**Theorem 1.** *Let  $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$  be the sequence generated by iFI-MA. Then  $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$  is bounded and any of its accumulation points are the critical points of  $\Psi$ . If  $\Psi$  is semi-algebraic, we further have that  $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$  is a Cauchy sequence, thus globally converges to a critical point of  $\Psi(\mathbf{x})$  in Eq. (1).*

### 3 Applications

This section illustrates how to apply the proposed general optimization learning paradigm to tackle practical inverse problems in different practical computer vision applications, such as low-light image enhancement, rain streaks removal, image restoration problems, and medical image analysis [Liu *et al.*, 2018d][Liu *et al.*, 2019c][Liu *et al.*, 2019d] and so on. This section also compares the performance of our algorithm with other state-of-the-art learning-based iterative methods on real-world inverse problems.

#### 3.1 Low-level Vision Problems

We conduct experiments on various low-level vision tasks (i.e., rain streaks removal and low-light image enhancement)



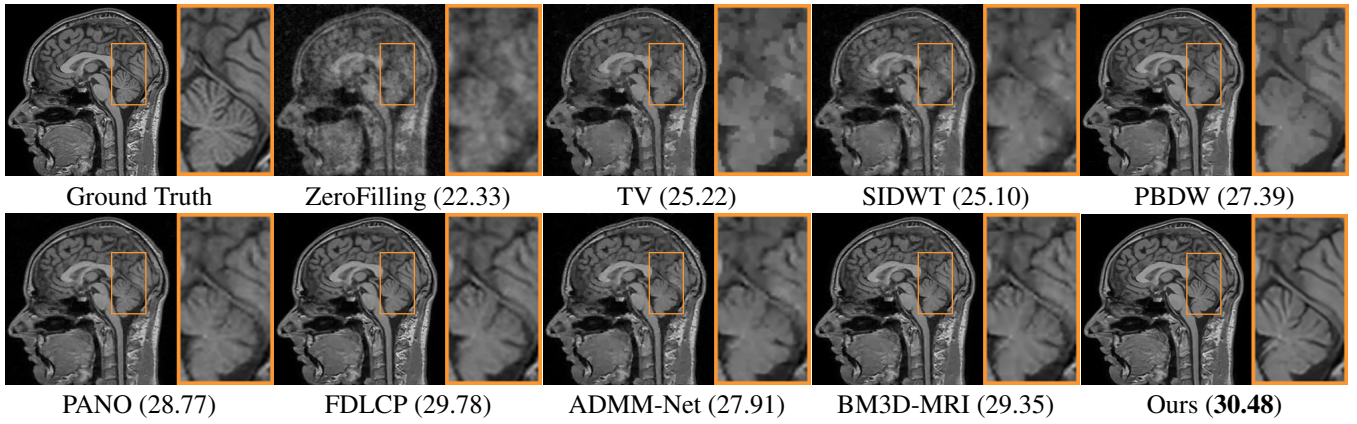


Figure 6: Qualitative comparisons on  $T_1$ -weighted brain MRI data using Gaussian mask at a sampling ratio of 10% .

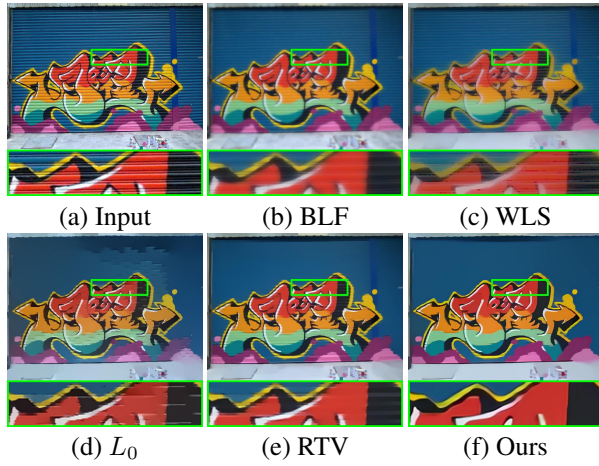


Figure 5: Edge-preserved smoothing results on an example with abundant textures.

and demonstrate the superiority of our proposed framework, compared with existing state-of-the-art approaches in [Liu *et al.*, 2019a][Liu *et al.*, 2019b]. Fig. 2 shows the rain streaks removal results of our paradigm on real-world rainy images. As can be observed, our proposed method can remove more rain streaks and preserve the more detail textures than others. Fig. 3 shows quantitative results on low-light image enhancement. As it shows, our method recover more details in the dark and presents a high contrast with a lower NIQE score.

### 3.2 Image Restoration Problems

Non-blind Deconvolution aims to restore the latent image  $z$  from corrupted observation  $y$  with known blur kernel  $b$ . We propose a collaborative learning framework in [Liu *et al.*, 2018d] to address the blind image deblurring issues. We design two modules, named Generator and Corrector, to extract the intrinsic image structures from the data-driven and knowledge-based perspectives, respectively. By introducing a collaborative methodology to cascade these modules, we can strictly prove the convergence of our image propagations to a deblurring-related optimal solution. As a nontrivial byproduct,

we also apply the proposed method to address other related tasks, such as edge-preserved smoothing. Edge-preserved image smoothing is a fundamental tool for image editing and processing, such as pencil sketch rendering and cartoon artifact removal. This task aims to increase the steepness of transition while eliminating a manageable degree of low-amplitude structures in the image. Fig. 4 illustrate the non-blind deblurring results. As shown, our algorithm maintains the advantage to recover more details and textures. Fig. 5 illustrates the edge-preserved smoothing results on an example image. It can be seen that our method removes most of the horizontal shutter door textures, while there still exists some horizontal lines in the results of other methods.

### 3.3 Medical Image Analysis Problems

Compressed Sensing Magnetic Resonance Imaging (CS-MRI) enables fast acquisition at sampling rate much lower than Nyquist rate with performance guarantee. However, existing CS-MRI deep optimization techniques have limitation in providing a theoretically converged scheme to restore MR images utilizing both domain knowledge and learnable architecture. To address this issue, we propose a deep optimization framework in [Liu *et al.*, 2019d], which can integrate domain knowledge and learning-based architectures by checking the first-order optimal condition. By the automatic checking mechanism from optimal condition, we can obtain a converged sequence which trends to the critical point of specific model. We further extend our paradigm to address real-world CS-MRI with Rician noise. Fig.6 offers the qualitative comparison. Visualized results demonstrate our method has better performance on both artifacts removing and details restoration.

## 4 Conclusions

We establish a theoretically guaranteed paradigm to design deep models for different learning tasks. We propose our novel mechanisms to adaptively guide the trajectories of learning-based iterations and proved their strict convergence. We also showed how to apply our guaranteed paradigm for different real-world applications, such as low-level vision problems, image restoration problems, and medical image analysis.

## References

- [Chen *et al.*, 2017] Yutian Chen, Matthew W. Hoffman, Sergio Gomez Colmenarejo, Misha Denil, Timothy P. Lillicrap, Matthew Botvinick, and Nando de Freitas. Learning to learn without gradient descent by gradient descent. In *ICML*, pages 748–756, 2017.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [Liu *et al.*, 2018a] Risheng Liu, Shichao Cheng, Yi He, Xin Fan, and Zhongxuan Luo. Toward designing convergent deep operator splitting methods for task-specific nonconvex optimization. In *IJCAI*, pages 2468–2474, 2018.
- [Liu *et al.*, 2018b] Risheng Liu, Shichao Cheng, Xiaokun Liu, Long Ma, Xin Fan, and Zhongxuan Luo. A bridging framework for model optimization and deep propagation. In *NeurIPS*, pages 4323–4332, 2018.
- [Liu *et al.*, 2018c] Risheng Liu, Xin Fan, Shichao Cheng, Xiangyu Wang, and Zhongxuan Luo. Proximal alternating direction network: A globally converged deep unrolling framework. In *AAAI*, pages 1371–1378, 2018.
- [Liu *et al.*, 2018d] Risheng Liu, Yi He, Shichao Cheng, Xin Fan, and Zhongxuan Luo. Learning collaborative generation correction modules for blind image deblurring and beyond. In *ACM Multimedia*, pages 1921–1929, 2018.
- [Liu *et al.*, 2019a] Risheng Liu, Shichao Cheng, Yi He, Xin Fan, Zhouchen Lin, and Zhongxuan Luo. On the convergence of learning-based iterative methods for nonconvex inverse problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [Liu *et al.*, 2019b] Risheng Liu, Shichao Cheng, Long Ma, Xin Fan, and Zhongxuan Luo. Deep proximal unrolling: Algorithmic framework, convergence analysis and applications. *IEEE Transactions on Image Processing*, 28(10):5013–5026, 2019.
- [Liu *et al.*, 2019c] Risheng Liu, Long Ma, Yiyang Wang, and Lei Zhang. Learning converged propagations with deep prior ensemble for image enhancement. *IEEE Transactions on Image Processing*, 28(3):1528–1543, 2019.
- [Liu *et al.*, 2019d] Risheng Liu, Yuxi Zhang, Shichao Cheng, Xin Fan, and Zhongxuan Luo. A theoretically guaranteed deep optimization framework for robust compressive sensing MRI. In *AAAI*, pages 4368–4375, 2019.