Deep Model-Based Architectures for Inverse Problems Under Mismatched Priors

Shirin Shoushtari, *Graduate Student Member, IEEE*, Jiaming Liu, *Student Member, IEEE*, Yuyang Hu, *Student Member, IEEE*, and Ulugbek S. Kamilov, *Senior Member, IEEE*

***Abstract*—There is a growing interest in deep model-based architectures (DMBAs) for solving imaging inverse problems by combining physical measurement models and learned image priors specified using convolutional neural nets (CNNs). For example, well-known frameworks for systematically designing DMBAs include plug-and-play priors (PnP), deep unfolding (DU), and deep equilibrium models (DEQ). While the empirical performance and theoretical properties of DMBAs have been widely investigated, the existing work in the area has primarily focused on their performance when the desired image prior is known exactly. This work addresses the gap in the prior work by providing new theoretical and numerical insights into DMBAs under mismatched CNN priors. Mismatched priors arise naturally when there is a distribution shift between training and testing data, for example, due to test images being from a different distribution than images used for training the CNN prior. They also arise when the CNN prior used for inference is an approximation of some desired statistical estimator (MAP or MMSE). Our theoretical analysis provides explicit error bounds on the solution due to the mismatched CNN priors under a set of clearly specified assumptions. Our numerical results compare the empirical performance of DMBAs under realistic distribution shifts and approximate statistical estimators.**

***Index Terms*—Image reconstruction, inverse problems, deep learning, robustness.**

I. Introduction

ONE OF the most widely-studied problems in computational imaging is the recovery of an unknown image from a set of noisy measurements. The recovery problem is often formulated as an *inverse problem* and solved by integrating the measurement model characterizing the response of the imaging instrument with a regularizer imposing prior knowledge on the unknown image. Some well-known image priors include nonnegativity, transform-domain sparsity, and self-similarity [1], [2], [3], [4].

Deep Learning (DL) has emerged in the past decade as a powerful data-driven paradigm for solving inverse problems and has improved the state-of-the-art in a number of imaging applications (see reviews [5], [6]). A traditional DL strategy for solving inverse problems is based on training a *convolutional neural network (CNN)* to perform a regularized inversion of the forward model, thus leading to a mapping from the measurements to the unknown image. For example, U-Net [7] and DnCNN [8] are two prototypical architectures used for designing traditional DL methods for solving imaging inverse problems.

There is a growing interest in *deep model-based architectures (DMBAs)* for inverse problems that integrate physical measurement models and CNN image priors (see reviews [9], [10], [11], [12]). Well-known DMBAs that explicitly account for the measurement models include *plug-and-play priors (PnP)*, *deep unfolding (DU)*, *compressive sensing using generative models (CSGM)*, and *deep equilibrium architectures (DEQ)* [13], [14], [15], [16], [17]. DMBAs are systematically obtained from model-based iterative algorithms by parametrizing the regularization step as a CNN and training it to adapt to the empirical distribution of desired images. An important conceptual point about typical DMBAs is that they do *not* solve an optimization problem. That is, even when the original model-based algorithm solves an optimization problem, once the regularizer is replaced with a CNN, then there is no longer any corresponding function to minimize. Remarkably, the heuristic of using CNNs not associated with any explicit regularization function exhibited great empirical success and spurred much theoretical and algorithmic work [9], [10], [11], [12].

Despite the rich literature on DMBAs, the existing work in the area has primarily focused on settings where the desired image prior is known *exactly*. While this assumption has led to many useful algorithms and insights, it fails to capture the range of situations arising in imaging inverse problems. Specifically, the knowledge of the image prior is only approximate if there is a distribution shift between training and testing data, for example, due to testing images being from a different distribution than images used for training the CNN prior. Alternatively, the CNN prior used for inference within DMBA might be an approximation of some desired true statistical estimator, such as *maximum a posteriori probability (MAP)* estimator or *minimum mean squared error (MMSE)* estimator. In both of these settings, it would be valuable to gain insights on how the discrepancies in CNN priors influence the discrepancies in estimated images.

Manuscript received 26 July 2022; revised 21 October 2022; accepted 1 November 2022. Date of publication 7 November 2022; date of current version 16 March 2023. This work was supported by NSF CAREER Award under Grant CCF-2043134. *(Corresponding author: Ulugbek S. Kamilov.)*

Shirin Shoushtari, Jiaming Liu, and Yuyang Hu are with the Department of Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, MO 63130 USA (e-mail: s.shirin@wustl.edu; jiaming.liu@wustl.edu; h.yuyang@wustl.edu).

Ulugbek S. Kamilov is with the Department of Computer Science and Engineering and the Department of Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, MO 63130 USA (e-mail: kamilov@wustl.edu).

Digital Object Identifier 10.1109/JSAIT.2022.3220044

In this work, we address this gap by providing a set of new theoretical and numerical results into DMBAs under mismatched CNN priors. We focus on the architecture derived from the *steepest descent* variant of *regularization by denoising (SD-RED)* [18] by considering two types of CNN priors: (a) image denoisers trained to remove *additive white Gaussian noise (AWGN)*; (b) *artifact removal (AR)* operators learned end-to-end using DEQ. Our theoretical analysis provides explicit error bounds on the solution obtained by SD-RED due to the mismatched CNN priors under a set of explicitly specified assumptions. Our numerical results illustrate the practical influence of mismatched CNN priors on image recovery from subsampled Fourier measurements, which is a well-known problem in accelerated *magnetic resonance imaging (MRI)* [19]. Specifically, we provide numerical results on two related but distinct scenarios where: (i) CNN priors are trained on data mismatched to the testing data and (ii) CNN priors are trained to approximate an explicit image regularizer.

II. Background

Inverse Problems: We consider the problem of recovering an unknown image $\boldsymbol{x}^* \in \mathbb{R}^n$ from its measurements $\boldsymbol{y} \in \mathbb{R}^m$. The problem is traditionally formulated as an inverse problem where the solution is computed by solving an optimization problem

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}\in\mathbb{R}^n}{\arg\min} f(\boldsymbol{x}) \quad \text{with} \quad f(\boldsymbol{x}) = g(\boldsymbol{x}) + h(\boldsymbol{x}), \tag{1}$$

where g is the *data-fidelity term* enforcing consistency of the solution with $\boldsymbol{y}$ and h is the *regularizer* enforcing prior knowledge on $\boldsymbol{x}$. The formulation in eq. (1) corresponds to the MAP estimator when

$$g(\boldsymbol{x}) = -\log\left(p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{x})\right) \quad \text{and} \quad h(\boldsymbol{x}) = -\log(p_{\boldsymbol{x}}(\boldsymbol{x})) \tag{2}$$

where $p_{\boldsymbol{y}|\boldsymbol{x}}$ is the likelihood relating $\boldsymbol{x}$ to measurements $\boldsymbol{y}$ and $p_{\boldsymbol{x}}$ is the prior distribution. For example, given measurements of the form $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{A}$ is the *measurement operator* (also known as the *forward operator*) characterizing the response of the imaging instrument and $\boldsymbol{e}$ is AWGN, the data-fidelity term reduces to the quadratic function $g(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2$. On the other hand, a widely-used sparsity promoting regularizer in imaging inverse problems is *total variation (TV)* $h(\boldsymbol{x}) = \tau\|\boldsymbol{D}\boldsymbol{x}\|_1$, where $\boldsymbol{D}$ is the image gradient and $\tau > 0$ controls the strength of regularization [1], [20], [21].

Model-Based Optimization: Proximal algorithms are often used for solving problems of form (1) when h is nonsmooth (see the review [22]). Two widely used families of proximal algorithms for imaging inverse problems are the *proximal gradient method (PGM)* [21], [23], [24], [25] and the *alternating direction method of multipliers (ADMM)* [26], [27], [28], [29]. Both PGM and ADMM avoid differentiating h by using the *proximal operator*, which can be defined as

$$\text{prox}_{\sigma^2 h}(\boldsymbol{z}) \coloneqq \underset{\boldsymbol{x}\in\mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2 + \sigma^2 h(\boldsymbol{x}) \right\}, \tag{3}$$

with $\sigma > 0$, for any proper, closed, and convex function h [22]. Comparing eq. (3) and eq. (1), we see that the proximal operator can be interpreted as a MAP estimator for the AWGN denoising problem

$$\boldsymbol{z} = \boldsymbol{x}_0 + \boldsymbol{w} \quad \text{where} \quad \boldsymbol{x}_0 \sim p_{\boldsymbol{x}_0}, \quad \boldsymbol{w} \sim \mathcal{N}\left(0, \sigma^2\boldsymbol{I}\right), \tag{4}$$

by setting $h(\boldsymbol{x}) = -\log(p_{\boldsymbol{x}_0}(\boldsymbol{x}))$. It is worth noting that another less known but equally valid statistical interpretation of the proximal operator is as a MMSE estimator [30], [31].

PnP and RED: PnP [13], [14] and RED [18] are two related families of iterative algorithms that enable integrating measurement operators and CNN priors for solving imaging inverse problems (see the recent review of PnP in [12]). Since for general denoisers PnP/RED do not solve an optimization problem [32], it is common to interpret PnP/RED as fixed-point iterations of some high-dimensional operators. For example, given a denoiser $D_{\boldsymbol{\theta}} : \mathbb{R}^n \to \mathbb{R}^n$ parameterized by a CNN with weights $\boldsymbol{\theta}$, the iterations of SD-RED [18] can be written as

$$\boldsymbol{x}^k = T_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{k-1}\right) = \boldsymbol{x}^{k-1} - \gamma G_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{k-1}\right) \quad \text{with}$$
$$G_{\boldsymbol{\theta}}(\boldsymbol{x}) \coloneqq \nabla g(\boldsymbol{x}) + \tau(\boldsymbol{x} - D_{\boldsymbol{\theta}}(\boldsymbol{x})), \tag{5}$$

where g is the data-fidelity term, and $\gamma, \tau > 0$ are the step size and the regularization parameters, respectively. SD-RED seeks to compute a fixed-point $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ of the operator $T_{\boldsymbol{\theta}}$, which is equivalent to finding a zero of the operator $G_{\boldsymbol{\theta}}$

$$\bar{\boldsymbol{x}} \in \text{Fix}(T_{\boldsymbol{\theta}}) \coloneqq \left\{\boldsymbol{x} \in \mathbb{R}^n : T_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{x}\right\}$$
$$\Leftrightarrow \quad G_{\boldsymbol{\theta}}(\bar{\boldsymbol{x}}) = \nabla g(\bar{\boldsymbol{x}}) + \tau(\bar{\boldsymbol{x}} - D_{\boldsymbol{\theta}}(\bar{\boldsymbol{x}})) = \boldsymbol{0}, \tag{6}$$

The solutions of (6) balance the requirements to be both data-consistent (via ∇g) and noise-free (via $(I - D_{\boldsymbol{\theta}})$), which can be intuitively interpreted as finding an equilibrium between the physical measurement model and CNN prior. Remarkably, this heuristic of using denoisers not necessarily associated with any h within an iterative algorithm exhibited great empirical success [10], [33], [34], [35], [36], [37], [38], [39], [40], [41] and spurred a great deal of theoretical work on PnP/RED [32], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52]. Recent line of PnP work has also explored automatic selection of parameters [53] and the parameterization of the regularization functions as CNNs, thus leading to explicit loss functions [54], [55].

DU and DEQ: DU (also known as *deep unrolling* or *algorithm unrolling*) is a DMBA paradigm highly related to PnP/RED with roots in sparse coding [15]. DU has gained popularity in inverse problems due to its ability to provide a systematic connection between iterative algorithms and deep neural network architectures [9], [11]. The SD-RED algorithm (5) can be turned into a DU architectures by truncating it to a fixed number of iterations $t \geq 1$, and training the corresponding architecture end-to-end in a supervised fashion by comparing the predicted image $\boldsymbol{x}^t(\boldsymbol{\theta})$ to the ground-truth $\boldsymbol{x}^*$. DEQ is a recent extension of DU that can accommodate an arbitrary number of iterations [17], [56]. DEQ can be implemented for SD-RED by replacing $\boldsymbol{x}^t(\boldsymbol{\theta})$ in the DU loss by a fixed-point $\bar{\boldsymbol{x}}(\boldsymbol{\theta})$ in eq. (6) and using implicit differentiation to update the weights $\boldsymbol{\theta}$. The benefit of DEQ over DU is that it doesn't require the storage of the intermediate variables in training, thus reducing the memory complexity. However,

DEQ requires the computation of the fixed-point $\overline{\boldsymbol{x}}(\boldsymbol{\theta})$, which can increase the computational complexity.

Other Related Work: There were a number of recent publications exploring the topic of distribution shifts in inverse problems. The use of DMBAs for adapting pre-trained CNNs to shifts in the measurement model has been discussed in several publications [39], [41], [57]. The performance gap due to distribution shifts on several well-known DL architectures has been empirically quantified for accelerated MRI in [58]. *Test-time training* was proposed as a strategy to decrease the performance gap for certain distribution shifts in [59]. The robustness of *compressive sensing (CS)* recovery using mismatched distributions with bounded Wasserstein distances was analyzed in [60]. The robustness of CSGM to changes in the ground-truth distribution and measurement operator in CS-MRI was investigated in [61]. Denoising-based approximate message passing (D-AMP) proposed in [62] coerces the signal perturbation at each iteration to be very close to AWGN. Finally, it is worth to briefly note a distinct line of work in optimization exploring the impact of *inexact* proximal operators on the convergence of the traditional proximal algorithms [63], [64], [65].

Our Contributions: Despite their conceptual differences in training, PnP, DU, and DEQ can be implemented using the same architecture during inference. For example, the SD-RED iteration in eq. (5) can be interpreted as a PnP method when the CNN prior $D_{\boldsymbol{\theta}}$ is an AWGN denoiser, as a DU architecture when $D_{\boldsymbol{\theta}}$ was trained using a fixed number of unfolded iterations, and as a DEQ architecture when $D_{\boldsymbol{\theta}}$ was trained at a fixed point. Note that the image prior in DU and DEQ is not an AWGN denoiser, but rather an artifact removal (AR) operator $D_{\boldsymbol{\theta}}$ trained by taking into account the distribution of artifacts within the iterations of SD-RED. The existing work on PnP, DU, and DEQ has primarily focused on the settings where the inference is performed assuming that $D_{\boldsymbol{\theta}}$ exactly corresponds to the desired AWGN denoiser or AR operator. However, it is clear from the discussion above, that this is an idealized scenario, in particular, when there is a distribution shift between training and testing data or when $D_{\boldsymbol{\theta}}$ is an approximation of some true statistical estimator. Our contribution is a first technical discussion on this issue in the scenario where the CNN prior used for inference in SD-RED is an approximation of some true prior. While we use the SD-RED iterations as the basis for our DMBA and corresponding mathematical analysis, the results can be extended to other DMBAs, including those based on PnP-PGM or PnP-ADMM. In short, this work presents new theoretical analysis and numerical results that are both complementary to and backward compatible with the existing literature in the area.

III. Theoretical Analysis

We focus on the DMBA based on the following modified SD-RED iteration

$$\begin{aligned} \boldsymbol{x}^k &= \boldsymbol{x}^{k-1} - \gamma \widehat{G}\left(\boldsymbol{x}^{k-1}\right) \quad \text{with} \\ \widehat{G}(\boldsymbol{x}) &\coloneqq \nabla g(\boldsymbol{x}) + \tau\left(\boldsymbol{x} - \widehat{D}(\boldsymbol{x})\right), \end{aligned} \tag{7}$$

where we refer to $\widehat{D}$ as a *mismatched prior* that approximates some *desired* or *true prior* D. We denote as $\boldsymbol{x}^* \in \operatorname{Zer}(G)$ the solution of SD-RED in (5) using the true D. We write both operators as D_σ and $\widehat{D}_\sigma$ when explicitly highlighting the *strength parameter* $\sigma > 0$ used to control the regularization strength. This parameter can account for the variance σ^2 in the proximal operator in eq. (3) and is analogous to the standard deviation parameter in the traditional PnP methods [13]. We next present a theoretical analysis of SD-RED under a mismatched prior providing: (a) error bounds on the solutions computed by (7) and (b) statistical interpretations under $\widehat{D}$ approximating a proximal operator, possibly corresponding to MAP or MMSE estimators. Our theoretical analysis builds on a set of explicitly specified assumptions that serve as sufficient conditions.

Assumption 1: The function g is convex, continuous, and has a Lipschitz continuous gradient with constant $L > 0$.

This is a standard assumption in optimization and is relatively mild in the context of imaging inverse problems. For example, it is satisfied by many traditional data-fidelity terms, including those based on the least-squares loss.

Assumption 2: The operator D is Lipschitz continuous with constant $0 < \lambda \leq 1$.

The Lipschitz continuity of CNN priors has been extensively considered in the prior work on PnP, DU, and DEQ and can be practically implemented using spectral normalization methods (see [12] for a more detailed discussion in the context of PnP). We say that D is a contractive operator when $\lambda < 1$ and it is a nonexpansive operator when $\lambda = 1$. Note also how the nonexpansiveness is *only* assumed on the desired CNN prior D rather than the mismatched one $\widehat{D}$ used for inference.

Assumption 3: The operator $\widehat{D}_\sigma$ satisfies

$$\|\widehat{D}_\sigma(\boldsymbol{x}) - D_\sigma(\boldsymbol{x})\|_2 \leq \sigma\varepsilon, \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^n,$$

where $\sigma > 0$ is the strength parameter of the prior and $\varepsilon > 0$ is some constant.

This assumption bounds the distance between the true and mismatched priors. We explicitly relate the bound to σ, since for small values of the strength parameter σ it is natural for the CNN priors to act as identity. The constant ε quantifies the approximation ability of the mismatched prior; given two approximations, the one with smaller ε is expected to be a better match. Assumption 3 can be also justified by using statistical considerations. For example, Theorem 3 below shows that when D and $\widehat{D}$ are MAP denoisers, Assumption 3 is a direct consequence of a bound on the density ratio $p_{\boldsymbol{x}}/\widehat{p}_{\boldsymbol{x}}$. A natural consequence of this argument is that when both D and $\widehat{D}$ are available, Assumption 3 can be a proxy to quantify prior distribution shifts.

We are now ready to state the first result.

Theorem 1: Run SD-RED in (7) for $t \geq 1$ iterations under Assumptions 1-3 with $\lambda < 1$ using a fixed step-size

$$0 < \gamma < \frac{(1-\lambda)\tau}{(L + (1+\lambda)\tau)^2}.$$

Then, there exists a unique $\boldsymbol{x}^* \in \operatorname{Zer}(G)$ such that

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2 \leq \eta^t R_0 + \tau\sigma\varepsilon A,$$

where $0 < \eta < 1$, $R_0 := \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2$, and $A := \gamma/(1-\eta)$ are constants independent of t.

The proof is provided in Appendix A. Theorem 1 shows that SD-RED using a mismatched CNN prior—either an AWGN denoiser or an AR operator—can approximate $\boldsymbol{x}^* \in \text{Zer}(G)$ up to an error term proportional to τ, σ, and ε. Note that it is expected that the error shrinks for small values of τ and σ since they control the influence of the CNN prior. While Theorem 1 assumes that the true prior D_σ is a contraction, our next result relaxes this condition by adopting an additional assumption.

Assumption 4: The operator G is such that $\text{Zer}(G) \neq \varnothing$. There exists a finite number $R > 0$ such that for any $\boldsymbol{x}^* \in \text{Zer}(G)$, the iterates (7) satisfy

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2 \leq R \quad \text{for all} \quad t \geq 1.$$

Assumption 4 simply states that G has a zero point in $\mathbb{R}^n$, which is equivalent to the assumption that SD-RED using the true prior has a solution. The assumption additionally states that the iterates generated via eq. (7) are bounded, which is natural for many imaging problems, since images usually have bounded pixel values in [0, 1] or [0, 255].

Theorem 2: Run SD-RED in (7) for $t \geq 1$ iterations under Assumptions 1-4 with $\lambda = 1$ using a fixed step-size $0 < \gamma < 1/(L + 2\tau)$. Then, we have that

$$\min_{1 \leq i \leq t} \|G\left(\boldsymbol{x}^{i-1}\right)\|_2^2 \leq \frac{1}{t}\sum_{i=1}^{t} \|G\left(\boldsymbol{x}^{i-1}\right)\|_2^2 \leq \frac{B_1}{t} + \tau\sigma\varepsilon B_2,$$

where

$$B_1 := \left((L + 2\tau)R^2\right)/\gamma,$$
$$B_2 := (L + 2\tau)(2R + \gamma\tau\sigma\varepsilon)$$

are constants independent of t.

The proof is provided in Appendix B. Theorem 2 is more general since it relaxes the assumption that D is a contraction to it being a nonexpansive operator. This comes with a cost of a slower sublinear convergence of the first term, compared to the linear convergence of the corresponding term under a contractive prior. It is worth highlighting that Theorems 1 and 2 generalize the prior work on DMBAs in the sense that for $\varepsilon = 0$ one recovers the traditional convergence results in the literature [17], [38], [45], [46]. The convergence rate and the error terms in the theorems are also compatible with the traditional results on inexact first-order methods [63], [64], [65].

Theorems 1 and 2 establish general error bounds on the solutions of SD-RED using a mismatched operator $\widehat{D}$, under the assumption that for the same input the distance between the outputs of $\widehat{D}$ and D are bounded. The following result provides a statistical interpretation to this assumption by considering denoisers that perform MAP estimation.

Theorem 3: Let $p_{\boldsymbol{x}}$ and $\widehat{p}_{\boldsymbol{x}}$ denote two log-concave continuous probability density functions supported over $\mathbb{R}^n$, and D_σ and $\widehat{D}_\sigma$ denote corresponding MAP estimators for the AWGN problem in eq. (4). Let $r := p_{\boldsymbol{x}}/\widehat{p}_{\boldsymbol{x}}$ denote the density ratio of p and $\widehat{p}$. If $\exp(-\varepsilon^2/2) \leq r(\boldsymbol{x}) \leq \exp(\varepsilon^2/2)$ for all $\boldsymbol{x} \in \mathbb{R}^n$, then we have that $\|D_\sigma(\boldsymbol{x}) - \widehat{D}_\sigma(\boldsymbol{x})\|_2 \leq \sigma\varepsilon$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

The proof is provided in Appendix C. Theorem 3 shows that if two prior distributions $p_{\boldsymbol{x}}$ and $\widehat{p}_{\boldsymbol{x}}$ are close to each other, then the distance between corresponding MAP denoisers is small, finally leading to a small error terms in Theorems 1 and 2. It is worth mentioning that the density ratio is a common tool for quantifying the distances between probability densities and is used, for example, in the Kullback–Leibler divergence $\mathbb{E}_{p_{\boldsymbol{x}}}[\log(r(\boldsymbol{x}))]$.

The next result enables a statistical interpretation of Theorems 1 and 2. This is due to the fact that both MAP and MMSE denoisers for the AWGN problem in eq. (4) can be expressed as proximal operators [31], [49], [66]. Using this result one can obtain a novel interpretation of PnP, DU, and DEQ algorithms under mismatched priors as using CNNs approximating true priors corresponding to some statistical estimators. For the rest of this section, we set $\tau = 1/\sigma^2$ to simplify the mathematical analysis and consider the true prior of form $D_\sigma = \text{prox}_{\sigma^2 h}$, where the function h satisfies the following assumption.

Assumption 5: The function h is closed, proper, convex, and Lipschitz continuous with constant $S > 0$.

This assumption is commonly adopted in nonsmooth optimization and implies the existence of a global upper bound on subgradients [67], [68], [69]. It is satisfied by a large number of functions, including the ℓ_1-norm and TV regularizers. We are now ready to state the final result, which can be seen as an extension of the analysis in [70] to mismatched CNN priors.

Theorem 4: Run SD-RED in (7) for $t \geq 1$ iterations under Assumptions 1-5 using a fixed step-size $0 < \gamma < 1/(L+2\tau)$. Then, we have that

$$\min_{1 \leq i \leq t}\left(f\left(\boldsymbol{x}^{i-1}\right) - f^*\right) \leq \frac{1}{t}\sum_{i=1}^{t}\left(f\left(\boldsymbol{x}^{i-1}\right) - f^*\right) \leq \frac{2(L+2\tau)R^3}{\gamma t} + \frac{\epsilon^2 R}{\sigma^2} + \frac{S^2\sigma^2}{2},$$

where $f = g + h$, $D_\sigma = \text{prox}_{\sigma^2 h}$, and $\tau = 1/\sigma^2$.

The proof is given in Appendix D. It states that SD-RED using a mismatched CNN prior $\widehat{D}_\sigma$, which approximates some proximal operator (possibly corresponding to MAP or MMSE statistical estimators), can approximate the minimum f^* up to an error term that is a function of ϵ and σ^2. The error term is minimized to $\epsilon S\sqrt{2R}$ when $\sigma^2 = \sqrt{2\epsilon^2 R/S^2}$.

To conclude this section, we theoretically analyzed the SD-RED iteration in eq. (7), where a mismatched prior $\widehat{D}$ is used instead of the desired or true prior D. Our analysis shows that one can get explicit error bounds on the solution of DMBAs that depend on the level of mismatch. Our analysis also reveals in the context of MAP estimation that the obtained error bounds can be related to the prior distribution shifts. In the next section, we provide a set of numerical results illustrating the empirical performance of SD-RED under distribution shifts and approximate proximal operators.

IV. Numerical Evaluation

There has been significant interest in understanding the performance of CNN priors for image recovery from noisy

and sub-sampled measurements. The recent work on DMBAs has shown that CNN priors can lead to significant improvements over traditional image priors such as TV in a wide variety of inverse problems. The results presented in this section evaluate the performance under mismatched CNN priors obtained due to shifts in the data distribution (for example, training on natural images but testing on MRI images) and due to approximate proximal operators (for example, training a CNN as an empirical MAP or MMSE estimator). It is worth mentioning that our focus is to highlight the impact of mismatched CNN priors, not to justify SD-RED as a DL architecture (such justifications can be found elsewhere, for example, see [12], [18], [32]).

We present three distinct sets of simulations on image recovery from subsampled Fourier measurements: *(a) using severely mismatched CNN priors corresponding to AWGN denoisers and AR operators trained on MRI, CT, and natural images*; *(b) using CNNs trained to approximate the traditional TV proximal operator*; and *(c) using moderately mismatched CNN priors for MRI trained on a different anatomical regions or MRI modalities*. While our simulations focus on subsampled measurements without noise, we expect that the reported relative performances will be preserved for moderate levels of noise. In all simulations, we use the classical TV regularization as a representative of the traditional model-based image reconstruction [21].

A. Mismatched Priors From Training on CT, Natural, and MR Images

We consider image recovery from subsampled Fourier measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \in \mathbb{C}^m$, where $\boldsymbol{A}$ performs radial Fourier sampling [19]. The measurement operator can be written as $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{F}$, where $\boldsymbol{F}$ is the Fourier transform and $\boldsymbol{P}$ is a diagonal sampling matrix. We train three classes of CNN priors modeling different empirical data distributions: (a) natural grayscale images from [71], (b) brain images from [72], and (c) CT images from the *low dose CT grand challenge* of Mayo Clinic [73]. Ten 180×180 images from Set11 were randomly chosen as natural test images. From 50 slices of 256×256 images provided for testing in [72], ten random images were chosen as MRI test images. Ten random CT images of 512×512 were chosen as CT test images. All images are rescaled to the range $[0, 255]$. The recovery performance of SD-RED using all three classes of CNN priors, as well as the traditional TV regularizer [21], is quantified using *peak-signal-to-noise ratio (PSNR)* in dB and *structural similarity index measure (SSIM)*.

Beyond the data distribution considerations, we also consider two types of CNN priors: (i) AWGN denoisers and (ii) AR operators. AWGN denoisers are extensively used within PnP due to their simplicity and effectiveness, while AR operators have been widely reported to achieve state-of-the-art imaging performance [12], [74]. We train one AWGN denoiser for each noise level $\sigma \in \{1, 2, 3, 5, 7, 8, 10, 12, 15\}$ and image distribution (natural, MRI, and CT) using the DnCNN architecture with batch normalization layers removed (as was done in [74]). The architecture of the AR operators is identical to that of the AWGN denoisers. The AR operators are trained via DEQ using pre-trained AWGN denoisers for initialization [17]. We train one AR operator for each sampling ratio considered in the simulations (10%, 20%, and 30%) and image distribution (natural, MRI, and CT). Nesterov [75] and Anderson et al.[76] acceleration techniques are used in forward and backward DEQ iterations for faster convergence. Spectral normalization is used for controlling the Lipschitz constants of all our CNNs [74], [77]. For each experiment, we select σ and τ achieving the best PSNR performance. We use `fminbound` in the `scipy.optimize` toolbox to identify the optimal regularization parameters for TV and all the SD-RED methods for each image at the inference time. We set the number of iterations for SD-RED to 500 and TV to 300, which we observed to be enough for convergence.

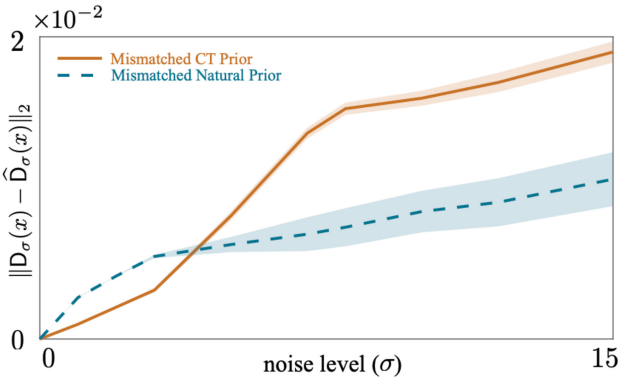

Fig. 1. Numerical evaluation of the mismatch between CT and natural image CNN priors with respect to the true MRI prior on MRI test images at various noise levels σ. As supposed in Assumption 3 the distance between the mismatched and true CNN priors is bounded and proportional to the noise level σ.

Figure 1 plots the distances between the outputs of the true AWGN denoisers trained on MRI images and mismatched ones trained on CT and natural images. The plot is generated using MRI test images. Average distance of the denoisers is plotted against noise level σ. Shaded area illustrates the range of values obtained across all test images. Our theoretical analysis assumes that the distance between the outputs of CNN priors is bounded and proportional to σ. Figure 1 numerically shows that the distance between the CNN priors is indeed bounded with a constant proportional to the noise level.

Theorem 2 states that SD-RED with shifted priors converges to an element of $\mathrm{Zer}(G)$ up to an error term that depends on τ, σ, and ϵ. Figure 2 illustrates the convergence of SD-RED with CT, natural, and MRI AWGN priors on MRI test data under 10% subsampling ratio. The average value of $\|G(\boldsymbol{x}^k)\|_2^2/\|G(\boldsymbol{x}^0)\|_2^2$ and PSNR (dB) are plotted against iterations of the algorithm. The distance to the zero-set quantifies the convergence behavior of the algorithm and is expected to be smaller for matched priors compared to mismatched priors. For reference, we also provide the evolution of the PSNR values obtained using TV reconstruction. Figure 4 shows the influence of τ and σ on the convergence of SD-RED using a mismatched natural AWGN prior to $\mathrm{Zer}(G)$, where G uses the true MRI AWGN prior at 10% subsampling. The average value

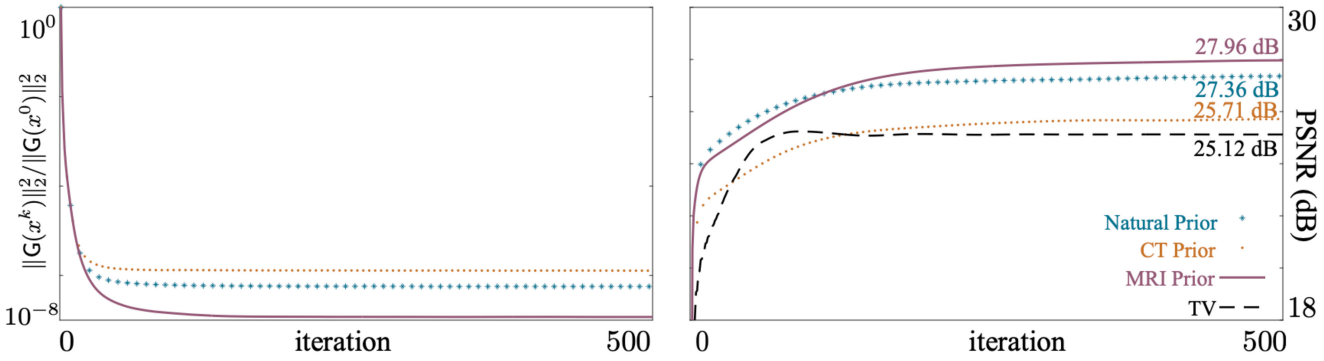

Fig. 2. **Left:** Empirical evaluation of convergence of SD-RED using the true and mismatched AWGN priors for brain MRI reconstruction at 10% sampling. Average normalized distance to Zer(G) is plotted against the iterations. **Right:** PSNR (dB) is plotted against iterations for TV and three AWGN priors. Note the effect of mismatched priors on the convergence of SD-RED to Zer(G) and the superior performance of mismatched CNN priors over TV.

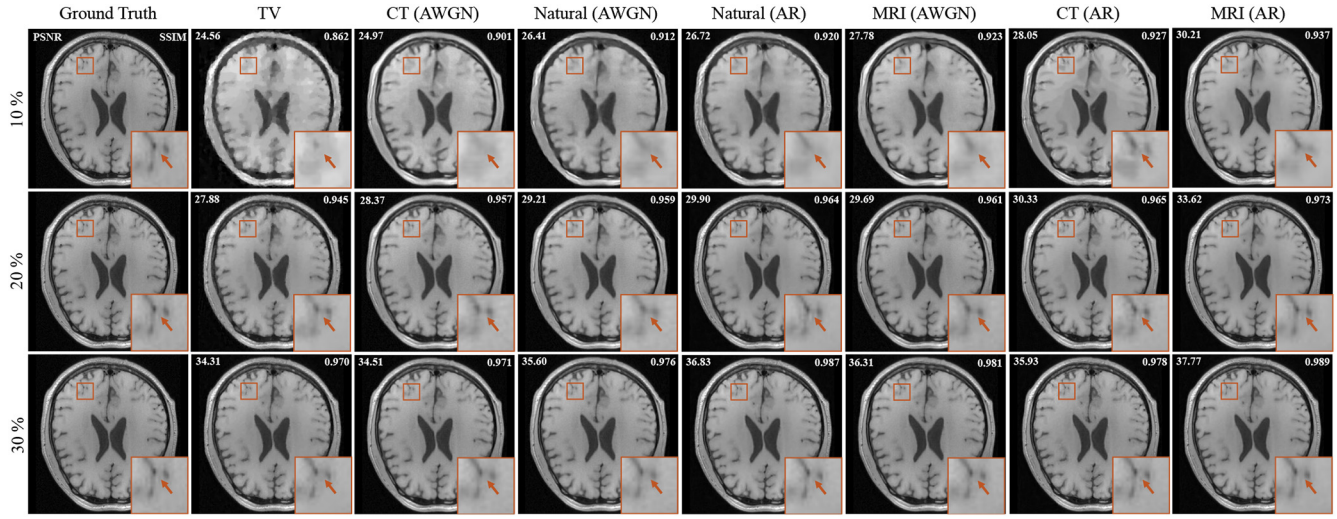

Fig. 3. Visual evaluation of several AWGN and AR priors used within SD-RED for brain MRI reconstruction from radially subsampled Fourier measurements at 10%, 20%, and 30%. Note how all the learned priors (both true and mismatched) outperform the traditional TV prior. Additionally, note the improvement in terms of imaging performance due to the mismatched AR priors compared to the true and mismatched AWGN priors.

of $\|G(\boldsymbol{x}^k)\|_2^2/\|G(\boldsymbol{x}^0)\|_2^2$ is plotted against iteration number for $\sigma \in \{5, 10, 30\}$ and $\tau \in \{0.001, 0.01, 0.1\}$. The results are consistent with the theoretical analysis and show that increase in both τ and σ increases the error term.

Table I reports the recovery PSNR for natural, CT, and MRI test images from subsampled Fourier measurements using the corresponding true and mismatched CNN priors. The best PSNR values for AWGN and AR priors are highlighted separately in bold. Figure 3 presents corresponding visual comparisons on a MRI image with 10%, 20%, and 30% sampling. As expected, matched priors lead to better performance in all the experiments, with matched AR priors achieving the best performance. While mismatched priors result in performance drops for both AWGN and AR priors, we observe better overall results for AR priors. Finally, note the inferior performance of TV compared to SD-RED under both true and mismatched CNN priors.

B. Approximating the TV Proximal Operator Using a CNN

In this section, we numerically evaluate CNN priors trained to approximate proximal operators. To that end, we train DnCNN to approximate the TV proximal operator and use the trained network within SD-RED as a mismatched CNN prior. We will refer to the mismatched prior as *TV Approx* and the true prior as *TV Exact*. We generate the training dataset by adding AWGN to natural grayscale images from [71]. We pre-train TV Approx as a CNN approximation of TV proximal by adding AWGN with $\sigma \in \{1, 2, 3, 5, 7, 8, 10, 12, 15\}$ to the training data and using the true TV solution as a training label. As a result, we have 9 pre-trained TV Approx CNNs. During the inference time, we select the TV Approx that yields the best reconstruction performance in terms of PSNR. We consider the same recovery problem as in the previous section, where the goal is to recover an image from its subsampled Fourier measurements at 10%, 20%, and 30% sampling rates. Theorem 4 shows that SD-RED using *TV Approx* can approximate the solution of (1) up to an error term. This behaviour is illustrated in Figure 6 (Left) for natural images reconstructed at 10% subsampling. The average value of $f(\boldsymbol{x}^k)/f(\boldsymbol{x}^0)$ is plotted against iterations. Figure 6 (Right) shows that the approximate TV prior can indeed achieve performance similar to the true TV prior. These plots highlight that despite the

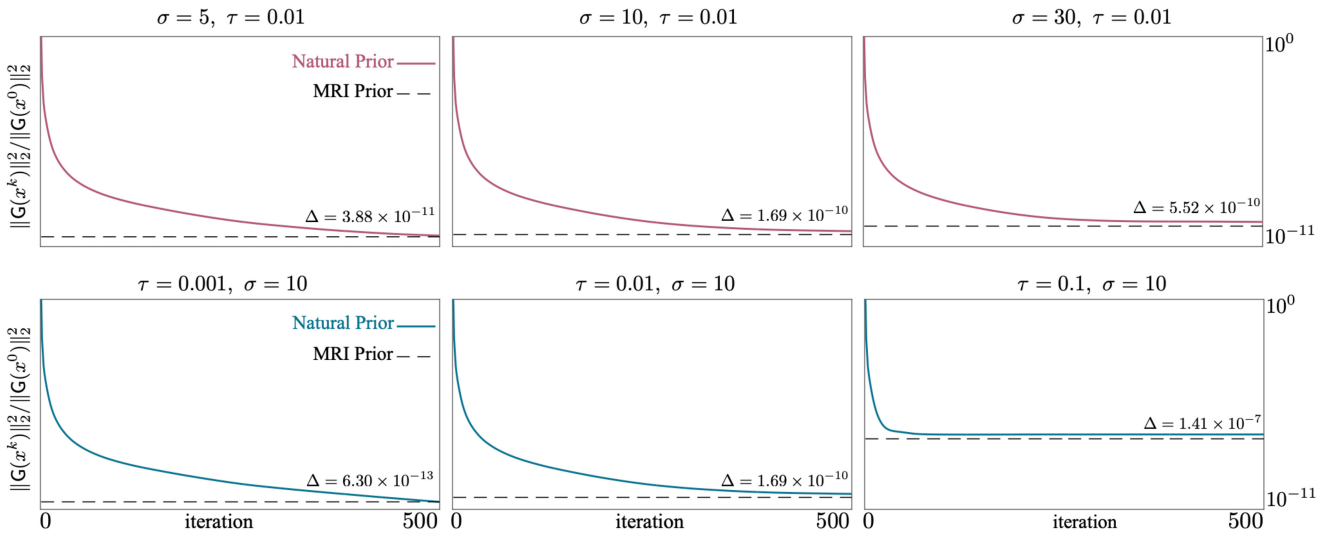

Fig. 4. Illustration of influence of noise level (σ) and regularization parameter (τ) on the error term in the convergence analysis of SD-RED under mismatched priors. Average normalized distance to Zer(G) is plotted against iterations for MRI and natural AWGN priors for MRI reconstruction at 10% sampling. **Top:** τ is set to a constant to evaluate the effect of σ. **Bottom:** σ is set to a constant to evaluate the effect of τ. The gap illustrates the error term due to the use of mismatched CNN priors. Note how the gap increases with the increase of both parameters τ and σ.

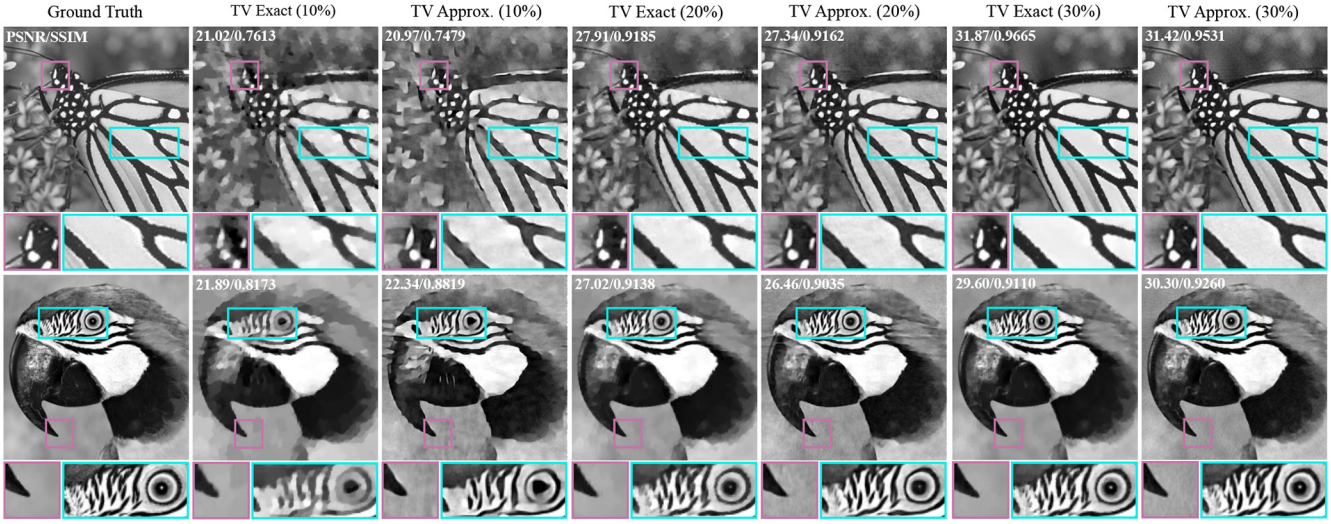

Fig. 5. Recovery of *Butterfly* and *Parrot* from 10%, 20%, and 30% Fourier samples using SD-RED under DnCNN trained as an approximate TV prior. Results of the traditional TV are also provided. Note the visual and quantitative similarities between the exact and approximate TV results at all sampling ratios.

constant error term in the objective function due to the prior mismatch, the approximation can still lead to nearly identical PSNR and SSIM values. Figure 5 illustrates the recovery of two test images at three sampling rates, highlighting the ability of TV Approx to match the performance of the true TV prior.

C. Using Mismatched MRI Priors for Accelerated Parallel MRI

In this section, we consider only MRI images to evaluate the impact of CNN priors trained on data with moderate distribution shifts. Our measurement model uses multi-coil Cartesian Fourier sampling mask from the fastMRI challenge [78] with 2× and 4× accelerated acquisitions. The measurement operator for each coil can be written as $\boldsymbol{A}_i = \boldsymbol{PFS}_i$, where $\boldsymbol{P}$ is a subsampling mask, $\boldsymbol{F}$ is the Fourier transform, and $\boldsymbol{S}_i$ is a coil sensitivity map. We use datasets in [72] and [78] to pre-train four separate AWGN denoisers on brain, knee, AXT1PRE, and AXT2 images. The denoisers are trained on noise levels corresponding to $\sigma \in \{1, 2, 3, 5, 8, 10, 12, 15\}$. We select the denoiser yielding the best performance in terms of PSNR as the prior for SD-RED. Thirty images are randomly chosen from each dataset for testing. In each experiment, 128 coil sensitivity maps are synthesized using SigPy [79]. Table II reports PSNR and SSIM for 2× and 4× accelerated MRI reconstruction using SD-RED. Note the superior performance of Knee prior compared to other mismatched brain priors in the reconstruction of MRI AXT2. Figure 7 presents a visual comparison on one test image from the fastMRI AXT2 dataset

TABLE I
THE RECOVERY OF NATURAL, MRI, AND CT IMAGES IN TERMS OF PSNR (DB) USING THE TRUE AND MISMATCHED AR AND AWGN PRIORS

Method	MRI			Natural			CT		
	10%	20%	30%	10%	20%	30%	10%	20%	30%
TV	25.12	27.92	31.59	20.88	25.03	27.89	26.21	31.76	35.52
MRI (AWGN)	**27.96**	**33.41**	**37.13**	21.57	25.94	28.16	26.37	32.75	35.89
Natural (AWGN)	27.36	31.57	36.00	**25.71**	**28.26**	**31.65**	29.09	32.61	35.67
CT (AWGN)	25.71	28.26	31.65	21.02	25.11	28.50	**30.28**	**33.62**	**36.57**
MRI (AR)	**30.09**	**36.10**	**39.47**	22.30	27.56	31.58	34.11	41.69	46.15
Natural (AR)	28.18	33.82	37.95	**25.11**	**29.71**	**32.89**	33.52	41.37	45.46
CT (AR)	28.37	32.38	37.48	21.32	27.18	28.76	**37.77**	**41.81**	**46.73**

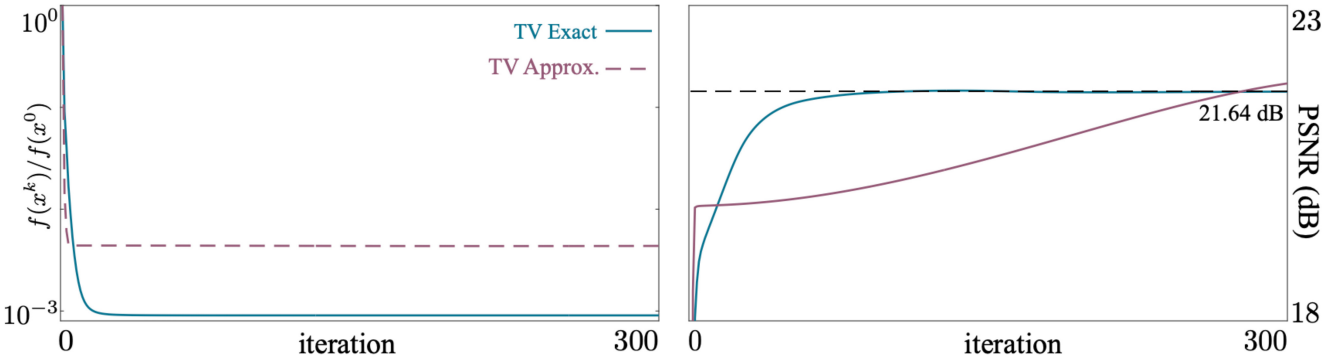

Fig. 6. **Left:** Evolution of the average TV objective using the true TV prior and its DnCNN approximation over the test set of natural images. **Right:** Evolution of average PSNR values on the true and approximate TV priors over the same test set of natural images. Note that despite the constant error term predicted in the objective due to the use of mismatched priors, the approximate TV prior achieves nearly identical PSNR compared to the true TV prior.

TABLE II
RECOVERY OF MRI AXT2 AND AXT1PRE TEST IMAGES FROM ACCELERATED PARALLEL MRI MEASUREMENTS USING SEVERAL IMAGE PRIORS

Method	MRI AXT2 [78]				MRI AXT1PRE [78]			
	2x		4x		2x		4x	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
TV	39.96	0.984	33.35	0.937	43.20	0.988	35.68	0.931
Knee Prior	41.84	0.988	34.63	0.944	43.81	0.991	37.80	0.958
AXT2 Prior	**43.64**	**0.991**	**36.46**	**0.954**	45.85	0.993	36.71	0.955
AXT1PRE Prior	41.96	0.989	34.76	0.945	**46.64**	**0.994**	**37.81**	**0.959**
Brain MRI Prior	41.63	0.988	34.42	0.937	45.95	0.993	36.52	0.954

using several CNN priors. The results indicate that mismatched CNN priors can be useful when true CNN priors are not available. They also suggest that under the settings considered in our simulations, mismatched CNN priors can outperform the traditional TV reconstruction in terms of PSNR/SSIM.

V. CONCLUSION

In this work, we explored the topic of mismatched CNN priors within DMBAs by theoretically analyzing the error bounds due to the mismatch and numerically illustrating the impact of the mismatch on the imaging performance. While our focus has been on the SD-RED architecture, similar results can certainly be carried out using other DMBAs. Our results show how the severity of the mismatch on the CNN prior translates to that of the final recovered images, relate the mismatch in CNN priors to the distribution shifts in statistical priors, and highlight the potential of DMBAs using mismatched CNN priors to outperform the traditional TV regularizer. In the future work, we would like to explore alternative characterizations of the mismatch and develop methods to reduce the influence of mismatch on the final recovery performance. Another direction for future research would be the derivation of lower bounds analogous the upper bounds in Theorems 1 and 2 (lower-bounds for traditional first-order methods have been investigated before [80]).

APPENDIX

We adopt *monotone operator theory* [81], [82] as a framework for a unified analysis of SD-RED under mismatched priors. In Appendix A, we analyze SD-RED under the assumption that the true CNN prior D is a contraction. In Appendix B, we extend the analysis to nonexpansive operators D. In Appendix C, we show how the bound on the density ratio $r = p_{\boldsymbol{x}}/\widehat{p}_{\boldsymbol{x}}$ leads to the bound in Assumption 3. In Appendix D, we analyze SD-RED under approximate proximal operators. In Appendices E and F, we present some key results from monotone operator theory and traditional optimization that are useful for our analysis.

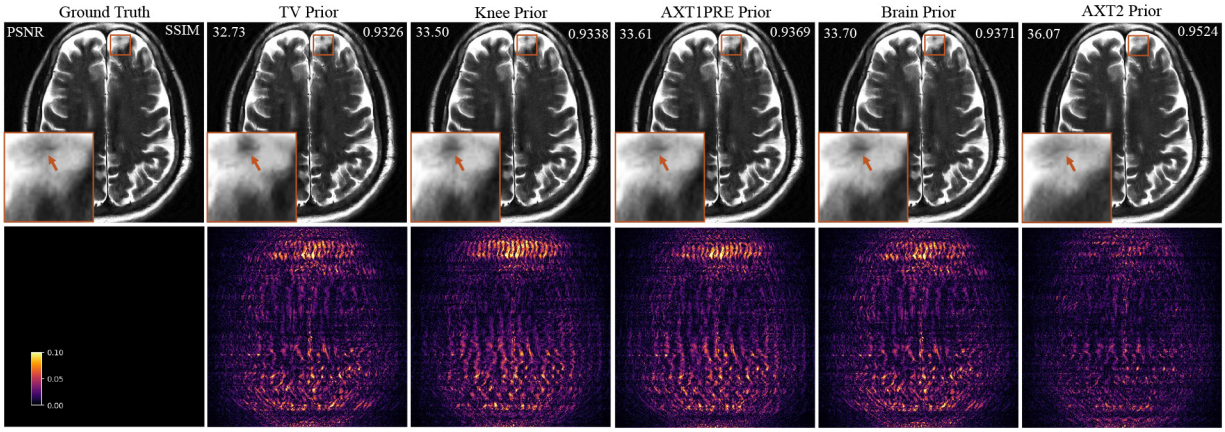

Fig. 7. Image recovery from 4× accelerated parallel MRI measurements of an image from AXT2 fastMRI dataset using several CNN priors trained as AWGN denoisers. Note how despite the performance drop due to the distribution mismatched, all CNN priors significantly outperform the traditional TV regularizer.

A. Proof of Theorem 1

First note that

$$\|G(\boldsymbol{x}) - \widehat{G}(\boldsymbol{x})\|_2 = \tau \|(I - D)(\boldsymbol{x}) - \big(I - \widehat{D}\big)(\boldsymbol{x})\|_2 \leq \tau\sigma\epsilon.$$

Consider a single iteration $\boldsymbol{x}^+ = \boldsymbol{x} - \gamma\widehat{G}(\boldsymbol{x})$ and $\boldsymbol{x}^* \in \mathrm{Zer}(G)$

$$\begin{aligned}\|\boldsymbol{x}^+ - \boldsymbol{x}^*\|_2 &= \|\boldsymbol{x} - \gamma\widehat{G}(\boldsymbol{x}) - \boldsymbol{x}^*\|_2 \\ &\leq \|\boldsymbol{x} - \gamma G(\boldsymbol{x}) - \boldsymbol{x}^*\|_2 + \gamma\|G(\boldsymbol{x}) - \widehat{G}(\boldsymbol{x})\|_2 \\ &\leq \eta\|\boldsymbol{x} - \boldsymbol{x}^*\|_2 + \gamma\tau\sigma\epsilon,\end{aligned}$$

where in the first inequality we used the triangular inequality and in the second Proposition 4 in Appendix E. By iterating this inequality for $t \geq 1$ iterations, we obtain

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2 \leq \eta^t\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2 + \tau\sigma\epsilon A,$$

where $\eta^2 \coloneqq 1 - 2\gamma[\tau(1-\lambda)] + \gamma^2[L + (1+\lambda)\tau]^2 \in (0, 1)$ and $A \coloneqq \gamma/(1-\eta)$ are fixed constants.

B. Proof of Theorem 2

Consider a single iteration $\boldsymbol{x}^+ = \boldsymbol{x} - \gamma\widehat{G}(\boldsymbol{x})$ and $\boldsymbol{x}^* \in \mathrm{Zer}(G)$

$$\begin{aligned}&\|\boldsymbol{x}^+ - \boldsymbol{x}^*\|_2^2 \\ &= \|\boldsymbol{x} - \gamma\widehat{G}(\boldsymbol{x}) - \boldsymbol{x}^*\|_2^2 = \|\boldsymbol{x} - \gamma G(\boldsymbol{x}) - \boldsymbol{x}^*\|_2^2 \\ &\quad + 2\gamma\big(G(\boldsymbol{x}) - \widehat{G}(\boldsymbol{x})\big)^{\mathsf{T}}\big(\boldsymbol{x} - \gamma G(\boldsymbol{x}) - \boldsymbol{x}^*\big) \\ &\quad + \gamma^2\|G(\boldsymbol{x}) - \widehat{G}(\boldsymbol{x})\|_2^2 \\ &\leq \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 - \frac{\gamma}{L+2\tau}\|G(\boldsymbol{x})\|_2^2 \\ &\quad + 2\gamma\|G(\boldsymbol{x}) - \widehat{G}(\boldsymbol{x})\|_2\|\boldsymbol{x} - \gamma G(\boldsymbol{x}) - \boldsymbol{x}^*\|_2 \\ &\quad + \gamma^2\|G(\boldsymbol{x}) - \widehat{G}(\boldsymbol{x})\|_2^2 \\ &\leq \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 - \frac{\gamma}{L+2\tau}\|G(\boldsymbol{x})\|_2^2 + 2\gamma\tau\sigma\epsilon R + \gamma^2\tau^2\sigma^2\epsilon^2.\end{aligned}$$

By rearranging the terms, we get

$$\begin{aligned}\|G(\boldsymbol{x})\|_2^2 \leq \left(\frac{L+2\tau}{\gamma}\right)\Big[\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}^+ - \boldsymbol{x}^*\|_2^2\Big] \\ + (L+2\tau)\Big(2\tau\sigma\epsilon R + \gamma\tau^2\sigma^2\epsilon^2\Big).\end{aligned}$$

By averaging over $t \geq 1$ iterations, we get

$$\frac{1}{t}\sum_{i=1}^{t}\|G\Big(\boldsymbol{x}^{i-1}\Big)\|_2^2 \leq \frac{B_1}{t} + \tau\sigma\epsilon B_2,$$

where $B_1 \coloneqq ((L+2\tau)R^2)/\gamma$ and $B_2 \coloneqq (L+2\tau)(2R+\gamma\tau\sigma\epsilon)$.

C. Proof of Theorem 3

The MAP solution to (4) can be expressed as the proximal operator (3). Consider two density functions $p_{\boldsymbol{x}}$ and $\widehat{p}_{\boldsymbol{x}}$ and corresponding MAP regularizers $h(\boldsymbol{x}) = -\log(p_{\boldsymbol{x}}(\boldsymbol{x}))$ and $\widehat{h}(\boldsymbol{x}) = -\log(\widehat{p}_{\boldsymbol{x}}(\boldsymbol{x}))$. The log-concavity and continuity of $p_{\boldsymbol{x}}$ and $\widehat{p}_{\boldsymbol{x}}$ imply that $\mathrm{prox}_{\sigma^2 h}$ and $\mathrm{prox}_{\sigma^2\widehat{h}}$ are unique minimizers of 1-strongly convex functions $\phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2 + \sigma^2 h(\boldsymbol{x})$ and $\widehat{\phi}(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2 + \sigma^2\widehat{h}(\boldsymbol{x})$, respectively. From the definition of the proximal operator, ϕ and $\widehat{\phi}$ are minimized at $\boldsymbol{x}^* = D_\sigma(\boldsymbol{z}) = \mathrm{prox}_{\sigma^2 h}(\boldsymbol{z})$ and $\widehat{\boldsymbol{x}} = \widehat{D}_\sigma(\boldsymbol{z}) = \mathrm{prox}_{\sigma^2\widehat{h}}(\boldsymbol{z})$, respectively, where $\boldsymbol{z} \in \mathbb{R}^n$ is any vector. From strong convexity

$$\begin{aligned}&\phi(\widehat{\boldsymbol{x}}) \geq \phi(\boldsymbol{x}^*) + \tfrac{1}{2}\|\boldsymbol{x}^* - \widehat{\boldsymbol{x}}\|_2^2 \\ &\widehat{\phi}(\boldsymbol{x}^*) \geq \widehat{\phi}(\widehat{\boldsymbol{x}}) + \tfrac{1}{2}\|\boldsymbol{x}^* - \widehat{\boldsymbol{x}}\|_2^2 \quad \Rightarrow \\ &\|\boldsymbol{x}^* - \widehat{\boldsymbol{x}}\|_2^2 \leq \sigma^2\big(\widehat{h}\big(\boldsymbol{x}^*\big) - h\big(\boldsymbol{x}^*\big) + h(\widehat{\boldsymbol{x}}) - \widehat{h}(\widehat{\boldsymbol{x}})\big). \qquad (8)\end{aligned}$$

We can re-write the bound on the density ratio as

$$\begin{aligned}&\mathrm{e}^{-\varepsilon^2/2} \leq p_{\boldsymbol{x}}(\boldsymbol{x})/\widehat{p}_{\boldsymbol{x}}(\boldsymbol{x}) \leq \mathrm{e}^{\varepsilon^2/2} \quad \Rightarrow \\ &\quad -\varepsilon^2/2 \leq \log(p_{\boldsymbol{x}}(\boldsymbol{x})) - \log(\widehat{p}_{\boldsymbol{x}}(\boldsymbol{x})) \leq \varepsilon^2/2 \quad \Rightarrow \\ &|h(\boldsymbol{x}) - \widehat{h}(\boldsymbol{x})| \leq \varepsilon^2/2.\end{aligned}$$

By combining this inequality with (8), we get the desired conclusion

$$\|D_\sigma(\boldsymbol{z}) - \widehat{D}_\sigma(\boldsymbol{z})\|_2^2 \leq \sigma^2\epsilon^2,$$

which is true for any $\boldsymbol{z} \in \mathbb{R}^n$.

D. Proof of Theorem 4

Moreau smoothing is well-known in the optimization literature and has been extensively used to design and analyze non-smooth algorithms (see, for example, [68]). It has been

previously used in [70, Th. 2] to analyze the block-coordinate RED (called BC-RED). The contribution of the following analysis is to extend [70] the prior result to a mismatched CNN prior $\widehat{D}$. Following [70], we fix $\tau = 1/\sigma^2$ for convenience, which enables us to have only one regularization parameter, namely σ^2.

We define the following loss function to approximate $f = g + h$

$$f_{\sigma^2}(\boldsymbol{x}) = g(\boldsymbol{x}) + \tau h_{\sigma^2}(\boldsymbol{x}),$$

where we set $\tau = (1/\sigma^2)$ and h_{σ^2} is known as the Moreau envelope of h

$$h_{\sigma^2}(\boldsymbol{x}) \coloneqq \min_{\boldsymbol{v}\in\mathbb{R}^n}\left\{\frac{1}{2}\|\boldsymbol{v}-\boldsymbol{x}\|_2^2 + \sigma^2 h(\boldsymbol{v})\right\}. \tag{9}$$

Lemma 5 and Assumption 5 imply that

$$\begin{aligned} 0 \le h(\boldsymbol{x}) - \tau h_{\sigma^2}(\boldsymbol{x}) \le \frac{S\sigma^2}{2} \quad &\Leftrightarrow \\ 0 \le f(\boldsymbol{x}) - f_{\sigma^2}(\boldsymbol{x}) \le \frac{S\sigma^2}{2}. \end{aligned} \tag{10}$$

Additionally, Lemma 4 implies that

$$G_\sigma(\boldsymbol{x}) = \nabla g(\boldsymbol{x}) + \tau\big(\boldsymbol{x} - \operatorname{prox}_{\sigma^2 h}(\boldsymbol{x})\big) = \nabla f_{\sigma^2}(\boldsymbol{x}), \tag{11}$$

where ∇f_σ^2 is $(L+2\tau)$-Lipschitz continuous. Eq. (11) implies that a single iteration of SD-RED using the true CNN prior is a gradient step on the smoothed version f_{σ^2} of f. From (9) and the convexity of the Moreau envelope, we have for all $\boldsymbol{x} \in \mathbb{R}^n$

$$f_{\sigma^2}^* = f_{\sigma^2}\big(\boldsymbol{x}^*\big) \le f_{\sigma^2}(\boldsymbol{x}) \le f(\boldsymbol{x}), \tag{12}$$

where $\boldsymbol{x}^* \in \operatorname{Zer}(G)$. Hence, there exists a finite f^* such that $f(\boldsymbol{x}) \ge f^*$ with $f^* \ge f_{\sigma^2}^*$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

Consider a single SD-RED update using a mismatched prior $\boldsymbol{x}^+ = \boldsymbol{x} - \gamma\widehat{G}(\boldsymbol{x})$

$$\begin{aligned} f_{\sigma^2}\big(\boldsymbol{x}^+\big) &\le f_{\sigma^2}(\boldsymbol{x}) + \nabla f_{\sigma^2}^{\mathsf{T}}\big(\boldsymbol{x}^+ - \boldsymbol{x}\big) + \frac{(L+2\tau)}{2}\|\boldsymbol{x}^+ - \boldsymbol{x}\|_2^2 \\ &= f_{\sigma^2}(\boldsymbol{x}) - \gamma\nabla f_{\sigma^2}(\boldsymbol{x})^{\mathsf{T}}\widehat{G}(\boldsymbol{x}) + \frac{\gamma^2(L+2\tau)}{2}\|\widehat{G}(\boldsymbol{x})\|_2^2 \\ &\le f_{\sigma^2}(\boldsymbol{x}) + \frac{\gamma}{2}\Big[\|\widehat{G}(\boldsymbol{x})\|_2^2 - 2\nabla f_{\sigma^2}(\boldsymbol{x})^{\mathsf{T}}\widehat{G}(\boldsymbol{x})\Big] \\ &\le f_{\sigma^2}(\boldsymbol{x}) - \frac{\gamma}{2}\|\nabla f_{\sigma^2}(\boldsymbol{x})\|_2^2 + \frac{\gamma\tau^2\sigma^2\epsilon^2}{2}, \end{aligned} \tag{13}$$

where in first inequality we used the Lipschitz continuity of ∇f_{σ^2}, in the second the fact that $\gamma \le 1/(L+2\tau)$, and the third

$$\begin{aligned} &\|\widehat{G}(\boldsymbol{x}) - \nabla f_{\sigma^2}(\boldsymbol{x})\|_2 \le \tau\sigma\epsilon \quad \Leftrightarrow \\ &\|\widehat{G}(\boldsymbol{x})\|_2^2 - 2\nabla f_{\sigma^2}(\boldsymbol{x})^{\mathsf{T}}\widehat{G}(\boldsymbol{x}) \le (\tau\sigma\epsilon)^2 - \|\nabla f_{\sigma^2}(\boldsymbol{x})\|_2^2. \end{aligned}$$

Consider the iteration $t \ge 1$, then we have that

$$\begin{aligned} \min_{1\le i\le t}\ \Big(f\Big(\boldsymbol{x}^{i-1}\Big) - f^*\Big) &\le \frac{1}{t}\sum_{i=1}^{t}\Big(f\Big(\boldsymbol{x}^{i-1}\Big) - f^*\Big) \\ &\le \frac{1}{t}\sum_{i=1}^{t}\Big(f_{\sigma^2}\Big(\boldsymbol{x}^{i-1}\Big) - f_{\sigma^2}^*\Big) + \frac{S^2\sigma^2}{2} \\ &\le \frac{R}{t}\sum_{i=1}^{t}\|\nabla f_{\sigma^2}\Big(\boldsymbol{x}^{i-1}\Big)\|_2^2 + \frac{S^2\sigma^2}{2} \\ &\le \frac{2R}{\gamma t}\Big(f_{\sigma^2}\Big(\boldsymbol{x}^0\Big) - f_{\sigma^2}^*\Big) + \frac{\epsilon^2 R}{\sigma^2} + \frac{S^2\sigma^2}{2} \\ &\le \frac{2R^3(L+2\tau)}{\gamma t} + \frac{\epsilon^2 R}{\sigma^2} + \frac{S^2\sigma^2}{2}, \end{aligned}$$

where in the second inequality we used eq. (10) and (12), in the third the convexity of f_{σ^2} and Assumption 4, in the forth eq. (13) and that $\tau = 1/\sigma^2$, and in the final the convexity of f_{σ^2} and Assumption 4.

E. Useful Results for the Main Theorems

Proposition 1: Suppose Assumptions 1-2 are true. Then, $G = \nabla g + \tau R$ is $(1/(L+2\tau))$-cocoercive.

Proof: Since ∇g is L-Lipschitz continuous, from Lemma 1 is also $(1/L)$-cocoercive. Then, from Lemma 2, $\mathsf{I} - (2/L)\nabla g$ is nonexpansive.

Since $D = I - R$ is nonexpansive and any convex combination of nonexpansive operators is nonexpansive, we have that the following operator is also nonexpansive

$$\begin{aligned} &I - \frac{2}{L+2\tau}G \\ &= \left[\frac{2}{L+2\tau}\cdot\frac{2\tau}{2}\right](I-R) + \left[\frac{2}{L+2\tau}\cdot\frac{L}{2}\right](I - (2/L)\nabla g) \\ &= \left[1 - \frac{2}{L+2\tau}\cdot\frac{L}{2}\right](I-R) + \left[\frac{2}{L+2\tau}\cdot\frac{L}{2}\right](I - (2/L)\nabla g). \end{aligned}$$

Thus, Lemma 2 implies that G is $1/(L+2\tau)$-cocoercive. ■

Proposition 2: Suppose Assumptions 1-2 are true. Then, for any $0 < \gamma \le 1/(L+2\tau)$, we have

$$\|\boldsymbol{x} - \gamma G(\boldsymbol{x}) - \boldsymbol{x}^*\|_2^2 \le \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 - \left(\frac{\gamma}{L+2\tau}\right)\|G(\boldsymbol{x})\|_2^2,$$

where $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{x}^* \in \operatorname{Zer}(G)$.

Proof: We have the following set of relations

$$\begin{aligned} \|\boldsymbol{x} - \gamma G(\boldsymbol{x}) - \boldsymbol{x}^*\|_2^2 &= \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 - 2\gamma G(\boldsymbol{x})^{\mathsf{T}}\big(\boldsymbol{x} - \boldsymbol{x}^*\big) \\ &\quad + \gamma^2\|G(\boldsymbol{x})\|_2^2 \\ &\le \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 - \left(\frac{2\gamma}{L+2\tau}\right)\|G(\boldsymbol{x})\|_2^2 \\ &\quad + \left(\frac{\gamma}{L+2\tau}\right)\|G(\boldsymbol{x})\|_2^2 \\ &= \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 - \left(\frac{\gamma}{L+2\tau}\right)\|G(\boldsymbol{x})\|_2^2, \end{aligned}$$

where in the second row we used Proposition 1 and the fact that $\gamma \le 1/(L+2\tau)$. ■

Proposition 3: Suppose Assumptions 1-2 are true with $\lambda < 1$. Then, $G = \nabla g + \tau R$ is $[\tau(1-\lambda)]$-strongly monotone and $[L + (1+\lambda)\tau]$-Lipschitz continuous.

Proof: From the convexity and smoothness of g, we know that ∇g is monotone. Due to the contractiveness of D, Lemma 3 implies that R is $(1-\lambda)$-strongly monotone. Thus, we have that

$$\begin{aligned} (G(\boldsymbol{x}) - G(\boldsymbol{z}))^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{z}) &= (\nabla g(\boldsymbol{x}) - \nabla g(\boldsymbol{z}))^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{z}) \\ &\quad + \tau(R(\boldsymbol{x}) - R(\boldsymbol{z}))^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{z}) \ge \tau(1-\lambda)\|\boldsymbol{x} - \boldsymbol{z}\|_2^2. \end{aligned}$$

To see the Lipschitz continuity, note that

$$\|G(\boldsymbol{x}) - G(\boldsymbol{y})\|_2 \leq \|\nabla g(\boldsymbol{x}) - \nabla g(\boldsymbol{y})\|_2 + \tau\|R(\boldsymbol{x}) - R(\boldsymbol{y})\|_2 \leq L + \tau(1+\lambda).$$ ■

Proposition 4: Suppose Assumptions 1-2 are true with $\lambda < 1$. Suppose that the step-size parameter is selected to satisfy

$$0 < \gamma < \frac{(1-\lambda)\tau}{(L+(1+\lambda)\tau)^2}.$$

Then, for any $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{x}^* \in \mathrm{Zer}(G)$, we have

$$\|\boldsymbol{x} - \gamma G(\boldsymbol{x}) - \boldsymbol{x}^*\|_2^2 \leq \eta^2 \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2,$$

where $\eta^2 = 1 - 2\gamma[\tau(1-\lambda)] + \gamma^2[L+(1+\lambda)\tau]^2 \in (0,1)$.

Proof: Let $\ell = L + (1+\lambda)\tau$ and $\mu = (1-\lambda)\tau$.

$$\begin{aligned}\|\boldsymbol{x} - \gamma G(\boldsymbol{x}) - \boldsymbol{x}^*\|_2^2 &= \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 - 2\gamma G(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^*) \\ &\quad + \gamma^2\|G(\boldsymbol{x})\|_2^2 \\ &\leq \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 - 2\gamma\mu\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 \\ &\quad + \gamma^2\ell^2\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 \\ &= \eta^2\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2,\end{aligned}$$

where $\eta^2 = (1 - 2\gamma\mu + \gamma^2\ell^2)$. Thus, for any $0 < \gamma < 2\mu/\ell^2$, we have that $0 < \eta^2 < 1$. ■

Proposition 5: Suppose Assumptions 1-5 the update

$$\boldsymbol{x}^+ = \boldsymbol{x} - \gamma\widehat{G}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^n,$$

under, where $D_\sigma = \mathrm{prox}_{\sigma^2 h}$ and $0 < \gamma \leq 1/(L+2\tau)$. Then, for any, we have

$$\|\nabla f_{(1/\tau)}(\boldsymbol{x})\|_2^2 \leq \frac{2}{\gamma}\big(f_{(1/\tau)}(\boldsymbol{x}) - f_{(1/\tau)}(\boldsymbol{x}^+)\big) + \sigma^2\varepsilon^2,$$

F. Background Material

The results in this section are well-known in the optimization literature and can be found in different forms in standard textbooks [75], [81], [83], [84]. We summarize the results useful for our analysis by restating them in a convenient form.

Lemma 1: For convex and continuously differentiable function g, we have

$$\nabla g \text{ is } L\text{-Lipschitz continuous} \quad \Leftrightarrow \quad \nabla g \text{ is } (1/L)\text{-cocoercive.}$$

Proof: See [75, Th. 2.1.5 and Sec. 2.1] ■

Lemma 2: Consider an operator $\mathsf{T} : \mathbb{R}^n \to \mathbb{R}^n$ and $\beta > 0$. The following properties are equivalent

1) T is β-cocoercive;
2) $\beta\mathsf{T}$ is firmly nonexpansive;
3) $\mathsf{I} - \beta\mathsf{T}$ is firmly nonexpansive;
4) $\beta\mathsf{T}$ is $(1/2)$-averaged;
5) $\mathsf{I} - 2\beta\mathsf{T}$ is nonexpansive.

Proof: See Proposition 4 in the Supplementary Material[1] of [51]. ■

Lemma 3: Consider an operator $\mathsf{R} = \mathsf{I} - \mathsf{D}$ where $\mathsf{D} : \mathbb{R}^n \to \mathbb{R}^n$.

$$\mathsf{D} \text{ is Lipschitz continuous with constant } \lambda < 1 \quad \Rightarrow$$
$$\mathsf{R} \text{ is } (1-\lambda)\text{-strongly monotone.}$$

[1]It can also be found in the pre-print https://arxiv.org/pdf/2006.03224.pdf.

Proof: See Proposition 2 in the Supplementary Material of [51]. ■

Definition 1: Consider a proper, closed, and convex function h and a constant $\mu > 0$. We define the *proximal operator*

$$\mathrm{prox}_{\mu h}(\boldsymbol{x}) = \underset{z \in \mathbb{R}^n}{\arg\min}\left\{\frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2 + \mu h(\boldsymbol{z})\right\}$$

and the *Moreau envelope*

$$h_\mu(\boldsymbol{x}) = \min_{z \in \mathbb{R}^n}\left\{\frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2 + \mu h(\boldsymbol{z})\right\}.$$

The following two lemmas provide useful results on the Moreau envelope. The Moreau envelope is convex and smooth.

Lemma 4: The function h_μ is convex and continuously differentiable with a 1-Lipschitz gradient

$$\nabla h_\mu(\boldsymbol{x}) = \boldsymbol{x} - \mathrm{prox}_{\mu h}(\boldsymbol{x}).$$

Proof: See [38, Proposition 8]. ■

The Moreau envelope can also serve as a smooth approximation of a nonsmooth function.

Lemma 5: Consider $h \in \mathbb{R}^n$ and its Moreau envelope $h_\mu(\boldsymbol{x})$ for $\mu > 0$. Then,

$$0 \leq h(\boldsymbol{x}) - \tfrac{1}{\mu}h_\mu(\boldsymbol{x}) \leq \tfrac{\mu}{2}G_{\boldsymbol{x}}^2 \quad \text{with}$$
$$G_{\boldsymbol{x}}^2 := \min_{\boldsymbol{g} \in \partial h(\boldsymbol{x})}\|\boldsymbol{g}\|^2, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

Proof: See [38, Proposition 9]. ■

REFERENCES

[1] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D Nonlinear Phenomena*, vol. 60, nos. 1–4, pp. 259–268, Nov. 1992.

[2] M. A. T. Figueiredo and R. D. Nowak, "Wavelet-based image estimation: An empirical Bayes approach using Jeffreys' noninformative prior," *IEEE Trans. Image Process.*, vol. 10, no. 9, pp. 1322–1331, Sep. 2001.

[3] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[4] A. Danielyan, "Block-based collaborative 3-D transform domain modeing in inverse imaging," Tampereen teknillinen yliopisto, Tampere Univ. Technol., Tampere, Finland, Rep. 1145, May 2013.

[5] M. T. McCann, K. H. Jin, and M. Unser, "Convolutional neural networks for inverse problems in imaging: A review," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 85–95, Nov. 2017.

[6] A. Lucas, M. Iliadis, R. Molina, and A. K. Katsaggelos, "Using deep neural networks for inverse problems in imaging: Beyond analytical methods," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 20–36, Jan. 2018.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image. Comput. Comput. Assist. Intervent.*, 2015, pp. 234–241.

[8] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[9] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett, "Deep learning techniques for inverse problems in imaging," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 39–56, May 2020.

[10] R. Ahmad et al., "Plug-and-play methods for magnetic resonance imaging: Using denoisers for image recovery," *IEEE Signal Process. Mag.*, vol. 37, no. 1, pp. 105–116, Jan. 2020.

[11] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, Mar. 2021.

[12] U. S. Kamilov, C. A. Bouman, G. T. Buzzard, and B. Wohlberg, "Plug-and-play methods for integrating physical and learned models in computational imaging," *IEEE Signal Process. Mag.*, vol. 40, no. 1, pp. 85–97, Jan. 2023, doi: 10.1109/MSP.2022.3199595.

[13] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," in *Proc. IEEE Global Conf. Signal Process. Inf. Process.*, Austin, TX, USA, 2013, pp. 945–948.

[14] S. Sreehari et al., "Plug-and-play priors for bright field electron tomography and sparse interpolation," *IEEE Trans. Comput. Imag.*, vol. 2, no. 4, pp. 408–423, Dec. 2016.

[15] K. Gregor and Y. LeCun, "Learning fast approximation of sparse coding," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, Jun. 2010, pp. 399–406.

[16] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative priors," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, Sydney, NSW, Australia, Aug. 2017, pp. 537–546.

[17] D. Gilton, G. Ongie, and R. Willett, "Deep equilibrium architectures for inverse problems in imaging," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 1123–1133, Oct. 2021.

[18] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (RED)," *SIAM J. Imag. Sci.*, vol. 10, no. 4, pp. 1804–1844, 2017.

[19] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing MRI," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 72–82, Mar. 2008.

[20] J. M. Bioucas-Dias and M. A. T. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2992–3004, Dec. 2007.

[21] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[22] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 123–231, 2014.

[23] M. A. T. Figueiredo and R. D. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, Aug. 2003.

[24] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, Nov. 2004.

[25] J. Bect, L. Blanc-Feraud, G. Aubert, and A. Chambolle, "A ℓ_1-unified variational framework for image restoration," in *Proc. Eur. Conf. Comp. Vis.*, vol. 3024, New York, NY, USA, 2004, pp. 1–13.

[26] J. Eckstein and D. P. Bertsekas, "On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 55, pp. 293–318, Apr. 1992.

[27] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2345–2356, Sep. 2010.

[28] M. K. Ng, P. Weiss, and X. Yuan, "Solving constrained total-variation image restoration and reconstruction problems via alternating direction methods," *SIAM J. Sci. Comput.*, vol. 32, no. 5, pp. 2710–2736, Aug. 2010.

[29] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends® Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jul. 2011.

[30] R. Gribonval, G. Chardon, and L. Daudet, "Blind calibration for compressed sensing by convex optimization," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 2713–2716.

[31] R. Gribonval and P. Machart, "Reconciling "priors" & "priors" without prejudice?" in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2013, pp. 2193–2201.

[32] E. T. Reehorst and P. Schniter, "Regularization by denoising: Clarifications and new interpretations," *IEEE Trans. Comput. Imag.*, vol. 5, no. 1, pp. 52–67, Mar. 2019.

[33] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3929–3938.

[34] C. Metzler, P. Schniter, A. Veeraraghavan, and R. Baraniuk, "prDeep: Robust phase retrieval with a flexible deep network," in *Proc. 36th Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018, pp. 3501–3510.

[35] W. Dong, P. Wang, W. Yin, G. Shi, F. Wu, and X. Lu, "Denoising prior driven deep neural network for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2305–2318, Oct. 2019.

[36] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1671–1681.

[37] G. Mataev, P. Milanfar, and M. Elad, "DeepRED: Deep image prior powered by RED," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Oct. 2019, pp. 1–10.

[38] Y. Sun, S. Xu, Y. Li, L. Tian, B. Wohlberg, and U. S. Kamilov, "Regularized Fourier ptychography using an Online plug-and-play algorithm," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 7665–7669.

[39] J. Liu, Y. Sun, C. Eldeniz, W. Gan, H. An, and U. S. Kamilov, "RARE: Image reconstruction using deep priors learned without ground truth," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 6, pp. 1088–1099, Oct. 2020.

[40] K. Wei, A. Aviles-Rivero, J. Liang, Y. Fu, C.-B. Schönlieb, and H. Huang, "Tuning-free plug-and-play proximal algorithm for inverse imaging problems," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 10158–10169.

[41] M. Xie, J. Liu, Y. Sun, W. Gan, B. Wohlberg, and U. S. Kamilov, "Joint reconstruction and calibration using regularization by denoising with application to computed tomography," in *Proc. IEEE Int. Conf. Comp. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 4028–4037.

[42] S. H. Chan, X. Wang, and O. A. Elgendy, "Plug-and-play ADMM for image restoration: Fixed-point convergence and applications," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 84–98, Mar. 2017.

[43] T. Meinhardt, M. Moeller, C. Hazirbas, and D. Cremers, "Learning proximal operators: Using denoising networks for regularizing inverse imaging problems," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 1799–1808.

[44] G. T. Buzzard, S. H. Chan, S. Sreehari, and C. A. Bouman, "Plug-and-play unplugged: Optimization free reconstruction using consensus equilibrium," *SIAM J. Imag. Sci.*, vol. 11, no. 3, pp. 2001–2020, Sep. 2018.

[45] E. K. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin, "Plug-and-play methods provably converge with properly trained denoisers," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97. Long Beach, CA, USA, Jun. 2019, pp. 5546–5557.

[46] Y. Sun, B. Wohlberg, and U. S. Kamilov, "An Online plug-and-play algorithm for regularized image reconstruction," *IEEE Trans. Comput. Imag.*, vol. 5, no. 3, pp. 395–408, Sep. 2019.

[47] T. Tirer and R. Giryes, "Image restoration by iterative denoising and backward projections," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1220–1234, Mar. 2019.

[48] A. M. Teodoro, J. M. Bioucas-Dias, and M. Figueiredo, "A convergent image fusion algorithm using scene-adapted Gaussian-mixture-based denoising," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 451–463, Jan. 2019.

[49] X. Xu, Y. Sun, J. Liu, B. Wohlberg, and U. S. Kamilov, "Provable convergence of plug-and-play priors with MMSE denoisers," *IEEE Signal Process. Lett.*, vol. 27, pp. 1280–1284, Jul. 2020.

[50] J. Tang and M. Davies, "A fast stochastic plug-and-play ADMM for imaging inverse problems," 2020, *arXiv:2006.11630*.

[51] Y. Sun, Z. Wu, X. Xu, B. Wohlberg, and U. S. Kamilov, "Scalable plug-and-play ADMM with convergence guarantees," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 849–863, Jul. 2021.

[52] R. Cohen, M. Elad, and P. Milanfar, "Regularization by denoising via fixed-point projection (red-pro)," *SIAM J. Imag. Sci.*, vol. 14, no. 3, pp. 1374–1406, 2021.

[53] X. Wang and S. H. Chan, "Parameter-free plug-and-play ADMM for image restoration," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2017, pp. 1323–1327.

[54] R. Cohen, Y. Blau, D. Freedman, and E. Rivlin, "It has potential: Gradient-driven denoisers for convergent solutions to inverse problems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 18152–18164.

[55] S. Hurault, A. Leclaire, and N. Papadakis, "Gradient step denoiser for convergent plug-and-play," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Kigali, Rwanda, May 2022.

[56] S. Bai, J. Z. Kolter, and V. Koltun, "Deep equilibrium models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 688–699.

[57] D. Gilton, G. Ongie, and R. Willett, "Model adaptation for inverse problems in imaging," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 661–674, Jul. 2021.

[58] M. Z. Darestani, A. S. Chaudhari, and R. Heckel, "Measuring robustness in deep learning based compressive sensing," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 2433–2444.

[59] M. Z. Darestani, J. Liu, and R. Heckel, "Test-time training can close the natural distribution shift performance gap in deep learning based compressed sensing," in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, Baltimore, MD, USA, Jul. 2022, pp. 4754–4776.

[60] A. Jalal, S. Karmalkar, A. Dimakis, and E. Price, "Instance-optimal compressed sensing via posterior sampling," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 4709–4720.
[61] A. Jalal, M. Arvinte, G. Daras, E. Price, A. G. Dimakis, and J. Tamir, "Robust compressed sensing MRI with deep generative priors," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2021, pp. 14938–14954.
[62] C. A. Metzler, A. Maleki, and R. G. Baraniuk, "From denoising to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 62, no. 9, pp. 5117–5144, Sep. 2016.
[63] D. P. Bertsekas, "Incremental proximal methods for large scale convex optimization," *Math. Program.*, vol. 129, pp. 163–195, Jun. 2011.
[64] M. Schmidt, N. Le Roux, and F. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, Granada, Spain, Dec. 2011, pp. 1458–1466.
[65] O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," *Math. Program.*, vol. 146, nos. 1–2, pp. 37–75, 2013.
[66] R. Gribonval, "Should penalized least squares regression be interpreted as maximum a posteriori estimation?" *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2405–2410, May 2011.
[67] H. Ouyang, N. He, L. Tran, and A. Gray, "Stochastic alternating direction method of multipliers," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 80–88.
[68] Y. Yu, "Better approximation and faster algorithm using the proximal average," in *Proc. Neural Inf. Process. Syst. (NIPS)*, Lake Tahoe, CA, USA, Dec. 2013, pp. 458–466.
[69] S. Boyd and L. Vandenberghe, "Subgradients," in *Convex Optimization II*. Stanford, CA, USA: Stanford Univ., Apr. 2008. [Online]. Available: http://see.stanford.edu/materials/lsocoee364b/01-subgradients_notes.pdf
[70] Y. Sun, J. Liu, and U. S. Kamilov, "Block coordinate regularization by denoising," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2019, pp. 382–392.
[71] 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 *Proc. IEEE Int. Conf. Comp. Vis. (ICCV)*, Vancouver, BC, Canada, Jul. 2001, pp. 416–423.
[72] J. Zhang and B. Ghanem, "ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 1828–1837.
[73] C. McCollough, "TU-FG-207A-04: Overview of the low dose CT grand challenge," *Med. Phys.*, vol. 43, no. 6, pp. 3759–3760, 2016.
[74] J. Liu, S. Asif, B. Wohlberg, and U. S. Kamilov, "Recovery analysis for plug-and-play priors using the restricted eigenvalue condition," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2021, pp. 5921–5933.
[75] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. New York, NY, USA: Kluwer Acad. Publ., 2004.
[76] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An Introduction to Random Matrices*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[77] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, Apr. 2018.
[78] F. Knoll et al., "FastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning," *Radiol. Artif. Intell.*, vol. 2, no. 1, 2020, Art. no. e190007.
[79] F. Ong and M. Lustig, "SigPy: A python package for high performance iterative reconstruction," in *Proc. ISMRM 27th Annu. Meeting*, Montreal, QC, Canada, vol. 4819, 2019.
[80] Y. E. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$," *Dokl. Akad. Nauk SSSR*, vol. 269, no. 3, pp. 543–547, 1983.
[81] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed. New York, NY, USA: Springer, 2017.
[82] E. K. Ryu and S. Boyd, "A primer on monotone operator methods," *Appl. Comput. Math.*, vol. 15, no. 1, pp. 3–43, 2016.
[83] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1970.
[84] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K: Cambridge Univ. Press, 2004.

Shirin Shoushtari (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from the Sharif University of Technology, Tehran, Iran, respectively, in 2017 and 2020, respectively. She is currently pursuing the Ph.D. degree with the Computational Imaging Group, Washington University in St. Louis, St. Louis, MO, USA. Her research interests include computational imaging, deep learning, image processing, and optimization.

Jiaming Liu (Student Member, IEEE) received the B.Sc. degree in electronic and information engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2017, and the M.S. degree in electrical and engineering from Washington University in St. Louis, St. Louis, MO, USA, in 2018, where he is currently pursuing the Ph.D. degree with Computational Imaging Group. His current research interests include machine learning, image processing, and optimization.

Yuyang Hu (Student Member, IEEE) received the B.Sc. degree in electronic and information engineering from Nanjing Tech University, Nanjing, China, in 2020, and the M.S. degree in electrical engineering from Washington University in St. Louis, St. Louis, MO, USA, in 2022, where he is currently pursuing the Ph.D. degree with Computational Imaging Group. His current research interests include machine learning, signal and image processing, and optimization.

Ulugbek S. Kamilov (Senior Member, IEEE) received the B.Sc./M.Sc. degree in communication systems and the Ph.D. degree in electrical engineering from EPFL, Switzerland, in 2011 and 2015, respectively. He is the Director of Computational Imaging Group and an Assistant Professor of Electrical and Systems Engineering and Computer Science and Engineering with Washington University in St. Louis. From 2015 to 2017, he was a Research Scientist with Mitsubishi Electric Research Laboratories, Cambridge, MA, USA. He was a recipient of the NSF CAREER Award and the IEEE Signal Processing Society's 2017 Best Paper Award. He was among 55 early-career researchers in the USA selected as a Fellow for the Scialog initiative on "Advancing Bioimaging" in 2021. His Ph.D. thesis was selected as a finalist for the EPFL Doctorate Award in 2016. He has served as a Senior Member of the Editorial Board of *IEEE Signal Processing Magazine* and as an Associate Editor of IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING. He has served on IEEE Signal Processing Society's Computational Imaging Technical Committee and Bioimaging and Signal Processing Technical Committee. He was a Plenary Speaker at iTWIST 2018 and is the Program Co-Chair for the International Biomedical and Astronomical Signal Processing Frontiers Conference for 2023.