

Task-Oriented Convex Bilevel Optimization with Latent Feasibility

Risheng Liu, *Member, IEEE*, Long Ma, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang

Abstract—This paper firstly proposes a convex bilevel optimization paradigm to formulate and optimize popular learning and vision problems in real-world scenarios. Different from conventional approaches, which directly design their iteration schemes based on given problem formulation, we introduce a task-oriented energy as our latent constraint which integrates richer task information. By explicitly re-characterizing the feasibility, we establish an efficient and flexible algorithmic framework to tackle convex models with both shrunken solution space and powerful auxiliary (based on domain knowledge and data distribution of the task). In theory, we present the convergence analysis of our latent feasibility re-characterization based numerical strategy. We also analyze the stability of the theoretical convergence under computational error perturbation. Extensive numerical experiments are conducted to verify our theoretical findings and evaluate the practical performance of our method on different applications.

Index Terms—Convex optimization, latent constraint, global convergence, image processing.

I. INTRODUCTION

Over the past decades, convex optimization techniques have been widely used to address machine learning and computer vision problems [1], [2], [3], [4]. The main idea behind these approaches is to approximate the implicit task energy (possibly non-convex) by a convex surrogate. Then numerical solvers can be adopted to obtain desirable global solutions. However, due to the complexity of tasks and data distributions, it is usually challenging to exactly obtain the task-desired optimal solutions only based on these simple convex optimization formulations. For example, as in Fig. 1 (the red dashed rectangle),

This work is partially supported by the National Key R&D Program of China (2020YFB1313503), the National Natural Science Foundation of China (Nos. 61922019 and 11971220), Shenzhen Science and Technology Program (No. RCYX20200714114700072), the Stable Support Plan Program of Shenzhen Natural Science Fund (No. 20200925152128002), Guangdong Basic and Applied Basic Research Foundation (2019A1515011152), the Pacific Institute for the Mathematical Sciences (PIMS), and the Fundamental Research Funds for the Central Universities.

R. Liu is with DUT-RU International School of Information Science & Engineering, Dalian University of Technology, Dalian, 116024, China. E-mail: rslu@dlut.edu.cn.

L. Ma is with the School of Software Technology, Dalian University of Technology, Dalian, 116024, China. He is also with the Peng Cheng Laboratory, Shenzhen, 518052, China. E-mail: malone94319@gmail.com.

X. Yuan is with the Department of Mathematics, The University of Hong Kong, Hong Kong. E-mail: xmyuan@hku.hk.

S. Zeng is with the Department of Mathematics and Statistics, University of Victoria, Canada. E-mail: zengshangzhi@uvic.ca.

J. Zhang is with the Department of Mathematics, SUSTech International Center for Mathematics, Southern University of Science and Technology, National Center for Applied Mathematics Shenzhen, Shenzhen, 518055, China. (Corresponding author, E-mail: zhangj9@sustech.edu.cn.)

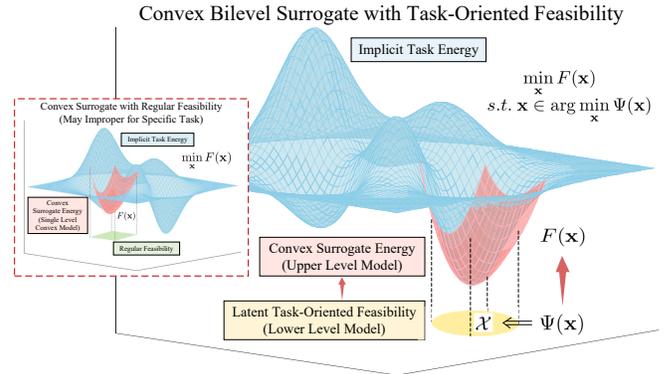


Fig. 1. Illustrating the mechanism of TOLF. The blue and red surfaces represent the implicit nonconvex energy and convex surrogate, respectively. It can be seen that with improper feasibility (e.g., the green region in the red dashed rectangle), standard convex optimization methods (e.g., [1], [10], [11]) may not obtain the desired solutions for the specific task. In contrast, TOLF aims to introduce latent feasibility to collect rich task information to narrow down the solution space (illustrated as the yellow region) and then solve a convex bilevel formulation to obtain the task-specific optimal solution.

a convex optimization model with improper feasibility may directly lead to incorrect solutions for the given task.

Non-convex optimization techniques are usually suggested to encode complex prior information for the task solution. However, in theory, finding global minimizers to non-convex problems is too ambitious. Even worse, according to prevailing non-convex optimization theories, iterative sequences generated by non-convex optimization solvers converge to superficial critical points, which might be saddle points or local maximizers [5]. In recent years, a variety of plug-and-play iterative modules have been introduced to perform task-specific optimization. The idea is to unroll an existing optimization process and replace the explicit iterative updating rule with hand-designed operators and/or learned architectures [6], [7], [8]. Unfortunately, due to the uncontrolled inexact computational modules, it is hard to theoretically guarantee the convergence of these methods. The work in [9] tried to introduce error control rules to correct improper modules and thus ensure the convergence of their trained iterations. However, these additional error checking process will slow down the particular computation when handling challenging problems. Moreover, solid theoretical investigations (e.g., the stability of the iterations) are still missing.

In this work, we shall propose a new framework, named Task-Oriented Latent Feasibility (TOLF), to construct convex bilevel optimization models with the support of task information (e.g., principled knowledge and training data)

for complex learning and vision problems. Specifically, as illustrated in Fig. 1, instead of directly optimizing the convex surrogate on the entire feature space or with explicitly defined constraints, we introduce a task-oriented energy (e.g., the lower-level model in Fig. 1) to narrow down the solution space. The convex objective (e.g., the upper-level model in Fig. 1) is then optimized upon the optimal solution set of the lower-level model¹, resulting in a convex bilevel optimization problem [12]. Along this direction, we actually suggest a convex optimization model with hierarchies. Two subproblems with hierarchical structures are incorporated to simultaneously formulate the objective and constraint of the given tasks (possibly with non-convex implicit energy). Since in TOLF we consider general convex composite (not necessarily smooth) energies in both upper and lower subproblems, to our best knowledge, no efficient bilevel optimization algorithms can handle the resulted model. Fortunately, by re-characterizing the latent feasibility as explicit set constraints, an efficient bilevel iteration scheme is developed to tackle TOLF. Theoretically, we prove that our iteration scheme can strictly converge to the global optimal solution of the established convex bilevel model. We also show that the computational errors caused by lower-level solution set re-characterization can be successfully dominated, and thus the proposed method is sufficiently stable. Extensive experiments on real-world image processing task demonstrate the efficiency and effectiveness of the proposed framework. In summary, our contributions mainly include:

- TOLF provides a flexible framework from a new bilevel perspective. In particular, TOLF incorporates task information to narrow down the solution space and improve the iteration behaviors, leading to a powerful convex bilevel scheme for challenging problems (possibly with unknown nonconvex energies, illustrated in Fig. 1).
- By investigating the convergence as well as its stability, our latent feasibility re-characterization based solution strategy for solving nested energy model in TOLF is strictly convinced. Thus we obtain solid theoretical guarantees for the proposed task-oriented convex bilevel optimization paradigm.
- We develop a practical method based on TOLF to exploit both the domain knowledge and data-driven engines for real-world applications (e.g., image processing). Our method successfully addresses the issues in standard numerical schemes (hard to leverage data information) and plug-and-play iterations (lack of theoretical guarantees).

II. RELATED WORKS

In this section, we briefly review popular convex, non-convex and plug-and-play optimization techniques in learning and vision fields.

Convex Optimization. Recognizing or formulating a given problem as a convex optimization model has been a prevailing manner in a wide range of application areas, e.g., LASSO [13], matrix completion [14]. Algorithms for solving convex optimization models are rich, efficient and reliable. They can

be easily embedded in analysis tools or control systems [15]. More importantly, stationary solutions to convex models are globally optimal with solid theoretical guarantee [10]. Unfortunately, the real-world tasks [16], [11] are with large complexity and diversity. The advantageous convexity is usually missing, thus it is too ambitious to accurately realize the ideal task-desired optimal solution for a complex task in terms of convex surrogates.

Non-convex Optimization. To ameliorate better modeling power, various non-convex prior regularizations [17], [18], [19] have been derived from given tasks. Unfortunately, the non-convexity creates difficulties in designing effective solution schemes [20], [21]. A popular workaround is the relaxation from non-convexity to convexity. However, this step involves losses of accuracy, which sometimes is fateful in applications [22]. Therefore, solving the non-convex optimizations directly surpasses relaxation-based techniques in some senses, and tremendous successes have been witnessed [23]. However, general non-convex optimization strategies can hardly guarantee desirable solution qualities.

Plug-and-Play Optimization. Recently proposed plug-and-play strategies [24], [25], [7] incorporate some task-related computational modules (e.g., handcrafted operations or trained network architecture) into certain optimization procedures. Substituting subproblems in an optimization process with data-driven networks gains popularity among the plug-and-play literature [6], [26]. The essential idea behind the plug-and-play schemes is the derivation of learned architectures from optimization models to incorporate data priors. Unfortunately, despite the observable high-quality performance in practice, the lack of theoretical convergence limits its scientific contributions. To address this issue, the work in [7] has presented the convergence analysis after introducing an error control rule with a loop body, but this step brings extra computational loads and the final theoretical results which just reach the critical points are still not fully satisfied. In [27], by designing spectral normalization to train a denoiser, the theoretical convergence analysis of the plug-and-play method is explored based on the forward-backward splitting algorithm. Moreover, a series of works [28], [29], [30], [31] proved the convergence under a set of explicit assumptions, e.g., strongly convex, Lipschitz continuous gradient. There was another approach to prove the convergence which introduced the additional assumptions on the plug-and-play denoisers (e.g., non-expansive, symmetric) [32], [33], [34], [35], [36]. Different from them, our TOLF is established on a new bilevel optimization paradigm, which allows us to narrow down the solution space by regarding task information as the lower-level subproblem, rather than focusing on defining the regularization like works mentioned. On the other hand, the iterative process of TOLF we designed is derived based on re-characterization of the latent feasibility, which can bring solid theoretical guarantees and don't need for any additional assumptions. More importantly, our work is flexible enough with a newly-designed bilevel model and we demonstrates the superiority in multiple challenging image-processing applications (e.g., low-light image enhancement) but some have not been completed in other works mentioned.

¹Since we always consider the convex (but not strongly convex) lower-level energy, the feasibly solution set of our model may not be a singleton.

III. TASK-ORIENTED CONVEX BILEVEL OPTIMIZATION

Similar to standard convex optimization paradigms, we also consider the following composite optimization model:

$$\min_{\mathbf{x}} F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}), \quad (1)$$

where both f, g are extended-valued convex functions $\mathbb{R}^n \rightarrow (-\infty, \infty]$ and g is possibly nonsmooth. But instead of directly optimizing Eq. (1) on the entire feature space or enforcing explicit constraints in existing approaches [1], we aim to establish a new method, named Task-Oriented Latent Feasibility (TOLF), to collect task information to assist in solving Eq. (1) in challenging real-world scenarios.

A. Energy-based Latent Feasibility

The latent feasible set of Eq. (1) can be formulated as the minimizers of another optimization model with the task-oriented composite energy:

$$\min_{\mathbf{x}} \Psi(\mathbf{x}) = \psi(\mathbf{x}) + \varphi(\mathbf{x}), \quad (2)$$

where ψ, φ are also extended-valued convex functions $\mathbb{R}^n \rightarrow (-\infty, \infty]$ and φ is possibly nonsmooth. We will demonstrate in Sec. V how to define ψ and φ based on the principled knowledge and/or collected training data for complex tasks in learning and vision communities. After this process, we amount to solving the following convex bilevel optimization model (i.e., TOLF):

$$\min_{\mathbf{x}} F(\mathbf{x}) \quad s.t. \quad \mathbf{x} \in \arg \min_{\mathbf{x}} \Psi(\mathbf{x}), \quad (3)$$

which implicitly integrate two different hierarchies of task information (i.e., F and Ψ). From an optimization perspective, we actually utilize the lower-level subproblem Eq. (2) to characterize the feasible region of Eq. (3). The main benefit of such strategy is that we can take full advantage of our domain knowledge of the task. Indeed, Eq. (3) can be recognized as a specific convex bilevel model. However, both upper and lower levels are in general in lack of smoothness and strong convexity, and thus the existing solution schemes [37], [38], [39] no longer admit any theoretical validity. Specifically, when the upper-level subproblem is in the absence of strong convexity, directly solving Eq. (3) with such latent feasibility is extremely challenging.

B. Feasibility Re-characterization and Optimization

Motivated by the observation that in applications, the latent feasibility in Eq. (2) usually possesses some underlying structures, in this paper, we develop a new optimization strategy with solid theoretical guarantees. In particular, we re-characterize the latent feasibility in Eq. (2), and by doing this, we shall reformulate Eq. (3) in terms of the re-characterization of Eq. (2) into a standard optimization problem which is numerically tractable. To this end, we first list some structural assumptions, which are necessary for our following analysis

and are easy to be satisfied in applications of practical interests. Specifically, throughout this paper, we suppose that ψ has the following structural properties.²

Assumption 1. $\psi(\mathbf{x}) = h(\mathcal{A}(\mathbf{x}))$, where \mathcal{A} is some given linear operator and function h is closed, proper, convex and admits the properties that (i) h is continuously differentiable on $\text{dom}h$, assumed to be open, and (ii) h is local strongly convex on $\text{dom}h$.

We are now ready to state the following theorem to investigate the feasibility of our problem.

Theorem 1. (Latent feasibility re-characterization) *Let \mathcal{X} be the solution set of Eq. (2) (i.e., $\mathcal{X} := \arg \min_{\mathbf{x}} \Psi(\mathbf{x})$), then \mathcal{A} is invariant on \mathcal{X} . That is, given any $\bar{\mathbf{x}} \in \mathcal{X}$, \mathcal{X} can be explicitly characterized as*

$$\mathcal{X} = \{\mathbf{x} | \mathcal{A}(\mathbf{x}) = \mathcal{A}(\bar{\mathbf{x}}), \varphi(\mathbf{x}) \leq \varphi(\bar{\mathbf{x}})\}. \quad (4)$$

Following Theorem 1, we define $\mathcal{X}_\varphi := \{\mathbf{x} | \varphi(\mathbf{x}) \leq \varphi(\bar{\mathbf{x}})\}$ and $\bar{\mathbf{y}} = \mathcal{A}(\bar{\mathbf{x}})$. Then the original bilevel model in Eq. (3) can be equivalently reformulated as the following single-level constrained optimization problem:

$$\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x}) + \iota_{\mathcal{X}_\varphi}(\mathbf{x}), \quad s.t. \quad \mathcal{A}(\mathbf{x}) = \bar{\mathbf{y}}, \quad (5)$$

where $\iota_{\mathcal{X}_\varphi}$ denotes the indicator of \mathcal{X}_φ . Now our model (i.e., Eq. (3)) is reformulated to the single level standard optimization Eq. (5). To solve the single level reformulation, we introduce the following augmented Lagrangian function with auxiliary variables \mathbf{s}, \mathbf{z}

$$\begin{aligned} \mathcal{L}_\beta(\mathbf{x}, \mathbf{z}, \mathbf{s}, \{\boldsymbol{\lambda}_i\}_{i=1}^3) &:= f(\mathbf{x}) + g(\mathbf{z}) + \iota_{\mathcal{X}_\varphi}(\mathbf{s}) \\ &+ \langle \boldsymbol{\lambda}_1, \mathbf{z} - \mathbf{x} \rangle + \langle \boldsymbol{\lambda}_2, \mathbf{s} - \mathbf{x} \rangle + \langle \boldsymbol{\lambda}_3, \mathcal{A}(\mathbf{x}) - \bar{\mathbf{y}} \rangle \\ &+ \frac{\beta}{2} (\|\mathbf{z} - \mathbf{x}\|^2 + \|\mathbf{s} - \mathbf{x}\|^2 + \|\mathcal{A}(\mathbf{x}) - \bar{\mathbf{y}}\|^2), \end{aligned}$$

where $\{\boldsymbol{\lambda}_i\}_{i=1}^3$ denote the dual multipliers and $\beta > 0$ denotes the penalty parameter. Then the proximal ADMM (with $\tau > 0$) reads as follows:

$$\left\{ \begin{array}{l} \mathbf{z}^{k+1} \in \arg \min_{\mathbf{z}} \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{z}, \mathbf{s}^k, \{\boldsymbol{\lambda}_i^k\}_{i=1}^3) \\ \quad + \frac{\tau}{2} \|\mathbf{z} - \mathbf{z}^k\|^2, \\ \mathbf{s}^{k+1} \in \arg \min_{\mathbf{s}} \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{z}^k, \mathbf{s}, \{\boldsymbol{\lambda}_i^k\}_{i=1}^3) \\ \quad + \frac{\tau}{2} \|\mathbf{s} - \mathbf{s}^k\|^2, \\ \mathbf{x}^{k+1} \in \arg \min_{\mathbf{x}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{z}^{k+1}, \mathbf{s}^{k+1}, \{\boldsymbol{\lambda}_i^k\}_{i=1}^3) \\ \quad + \frac{\tau}{2} \|\mathbf{x} - \mathbf{x}^k\|^2, \\ \boldsymbol{\lambda}_1^{k+1} = \boldsymbol{\lambda}_1^k + \beta(\mathbf{z}^{k+1} - \mathbf{x}^{k+1}), \\ \boldsymbol{\lambda}_2^{k+1} = \boldsymbol{\lambda}_2^k + \beta(\mathbf{s}^{k+1} - \mathbf{x}^{k+1}), \\ \boldsymbol{\lambda}_3^{k+1} = \boldsymbol{\lambda}_3^k + \beta(\mathcal{A}(\mathbf{x}^{k+1}) - \bar{\mathbf{y}}). \end{array} \right. \quad (6)$$

Thanks to Theorem 1, we directly have a corollary to guarantee the convergence of Eq. (6) toward the global optimal solutions of Eq. (3).

Corollary 1. *Suppose that the problem in Eq. (5) has KKT solutions. Let $\{(\mathbf{x}^k, \mathbf{z}^k, \mathbf{s}^k, \boldsymbol{\lambda}^k)\}$ be the sequence generated by Eq. (6) on problem (5), then $\{(\mathbf{x}^k, \mathbf{z}^k, \mathbf{s}^k, \boldsymbol{\lambda}^k)\}$ converges to*

²Some commonly used functions in learning and vision (e.g., linear regression $h(\mathbf{z}) = \frac{1}{2} \|\mathbf{z} - \mathbf{b}_1\|^2$, logistic regression $h(\mathbf{z}) = \sum_{i=1}^m \log(1 + e^{\mathbf{z}^i}) - \langle \mathbf{b}_2, \mathbf{z} \rangle$ and likelihood estimation under Poisson noise $h(\mathbf{z}) = -\sum_{i=1}^m \log(\mathbf{z}_i) + \langle \mathbf{b}_3, \mathbf{z} \rangle$) automatically satisfy these assumptions, where $\mathbf{b}_1, \mathbf{b}_2$ and \mathbf{b}_3 are parameters.

the KKT point set of Eq. (5). In particular, the sequence $\{\mathbf{x}^k\}$ converges to the global optimal solutions of Eq. (3).

Remark 1. *Indeed, we can further estimate a nice linear convergence rate of Eq. (6) for particular models. That is, if f takes the form that $f(\mathbf{x}) = h(\tilde{A}\mathbf{x})$ where \tilde{A} is some given linear operator, h satisfies Assumption 1, g represents (1) convex polyhedral regularizer; (2) group-lasso regularizer; (3) sparse group lasso regularizer, and ψ is a convex polyhedral function, then $\{(\mathbf{x}^k, \mathbf{z}^k, \mathbf{s}^k, \boldsymbol{\lambda}^k)\}$ converges linearly to the KKT point set of the problem in Eq. (5). In particular, the sequence $\{\mathbf{x}^k\}$ converges linearly to the global optimal solutions of Eq. (3).*

IV. THEORETICAL INVESTIGATIONS

Thanks to Theorem 1, the solutions returned by Eq. (5) can exactly optimize the bilevel problem in Eq. (3). Our single-level reformulation based optimization scheme relies on the re-characterization of the solution set. To be specific, we require one solution of the constraint subproblem (i.e., $\bar{\mathbf{x}} \in \arg \min_{\mathbf{x}} \psi(\mathbf{x}) + \varphi(\mathbf{x})$) to construct the solution set \mathcal{X} .

In general, obtaining such a solution exactly is intractable. That is, we in practice can only calculate a solution with computation errors for the constraint subproblem in Eq. (2), i.e., obtain a point $\bar{\mathbf{x}}_\delta$ satisfying $d(\bar{\mathbf{x}}_\delta, \mathcal{X}) \leq \delta$, where d is the distance mapping, and $\delta \geq 0$ measures the computational errors. As a consequence, we consider the practical optimization process of Eq. (3) as solving an approximation of Eq. (5), which can be formulated as follows³:

$$\min_{\mathbf{x}} F(\mathbf{x}), \text{ s.t. } \mathcal{A}(\mathbf{x}) = \mathcal{A}(\bar{\mathbf{x}}_\delta), \varphi(\mathbf{x}) \leq \varphi(\bar{\mathbf{x}}_\delta). \quad (7)$$

In the following, we shall analyze the convergence behaviors and stability properties of our practical computation (can be abstractly formulated as Eq. (7)) from the perturbation analysis perspective. Specifically, we consider the errors for solving $\bar{\mathbf{x}}$ as the perturbation of optimizing Eq. (5) and obtain the following constructive results:

- Convergence (Theorem 2): As the error δ decreases to 0 in Eq. (7), the solution sequence strictly converges to our desired solution of the bilevel problem in Eq. (3).
- Stability (Theorem 3): The proximity from the optimal solution of Eq. (7) to the solution set of the bilevel problem in Eq. (3) can be strictly dominated in terms of δ .

A. Convergence Analysis

Before proving our formal convergence result, we first introduce some necessary notations. By respectively considering $\mathcal{A}(\bar{\mathbf{x}}_\delta)$ and $\varphi(\bar{\mathbf{x}}_\delta)$ in Eq. (7) as perturbed $\bar{\mathbf{y}}$ and $\bar{\mathbf{s}}$, we are now aiming to investigate the stability of the following parameterized optimization problem

$$(P_{\mathbf{p}}) \quad \min_{\mathbf{x}} F(\mathbf{x}), \text{ s.t. } \begin{cases} \mathcal{A}(\mathbf{x}) - \bar{\mathbf{y}} = \mathbf{p}_1, \\ \varphi(\mathbf{x}) \leq \bar{\mathbf{s}} + \mathbf{p}_2, \end{cases} \quad (8)$$

³Please notice that Eq. (7) is only used for our theoretical analysis, but not practical computation.

where $\mathbf{p} = \{\mathbf{p}_1, \mathbf{p}_2\}$, $\bar{\mathbf{y}} = \mathcal{A}(\bar{\mathbf{x}})$ and $\bar{\mathbf{s}} = \varphi(\bar{\mathbf{x}})$ for any given $\bar{\mathbf{x}} \in \mathcal{X}$. Moreover, we shall need the following notations.

- The feasible set mapping of $P_{\mathbf{p}}$: $\mathcal{S}_{feas}(\mathbf{p}) := \{\mathbf{x} | \mathcal{A}(\mathbf{x}) - \bar{\mathbf{y}} = \mathbf{p}_1, \varphi(\mathbf{x}) \leq \bar{\mathbf{s}} + \mathbf{p}_2\}$.
- The optimal value mapping of $P_{\mathbf{p}}$: $\mathcal{S}_{val}(\mathbf{p}) := \inf_{\mathbf{x}} \{F(\mathbf{x}) | \mathcal{A}(\mathbf{x}) - \bar{\mathbf{y}} = \mathbf{p}_1, \varphi(\mathbf{x}) \leq \bar{\mathbf{s}} + \mathbf{p}_2\}$.
- The solution set mapping of $P_{\mathbf{p}}$: $\mathcal{S}_{sol}(\mathbf{p}) := \{\mathbf{x} \in \mathcal{S}_{feas}(\mathbf{p}) | F(\mathbf{x}) = \mathcal{S}_{val}(\mathbf{p})\}$.

Continuity properties of set-valued mapping $\mathcal{S} : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ is developed in terms of outer and inner limits:

$$\begin{aligned} \limsup_{\mathbf{p} \rightarrow \bar{\mathbf{p}}} \mathcal{S}(\mathbf{p}) &:= \{\mathbf{x} | \exists \mathbf{p}^\nu \rightarrow \bar{\mathbf{p}}, \exists \mathbf{x}^\nu \rightarrow \mathbf{x} \text{ with } \mathbf{x}^\nu \in \mathcal{S}(\mathbf{p}^\nu)\}, \\ \liminf_{\mathbf{p} \rightarrow \bar{\mathbf{p}}} \mathcal{S}(\mathbf{p}) &:= \{\mathbf{x} | \forall \mathbf{p}^\nu \rightarrow \bar{\mathbf{p}}, \exists \mathbf{x}^\nu \rightarrow \mathbf{x} \text{ with } \mathbf{x}^\nu \in \mathcal{S}(\mathbf{p}^\nu)\}. \end{aligned}$$

Definition 1. *A set-valued mapping $\mathcal{S} : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ is outer semicontinuous (OSC) at $\bar{\mathbf{p}}$ when $\limsup_{\mathbf{p} \rightarrow \bar{\mathbf{p}}} \mathcal{S}(\mathbf{p}) \subseteq \mathcal{S}(\bar{\mathbf{p}})$ and inner semicontinuous (ISC) at $\bar{\mathbf{p}}$ when $\liminf_{\mathbf{p} \rightarrow \bar{\mathbf{p}}} \mathcal{S}(\mathbf{p}) \supseteq \mathcal{S}(\bar{\mathbf{p}})$. It is called continuous at $\bar{\mathbf{p}}$ when it is both OSC and ISC at $\bar{\mathbf{p}}$, as expressed by $\lim_{\mathbf{p} \rightarrow \bar{\mathbf{p}}} \mathcal{S}(\mathbf{p}) = \mathcal{S}(\bar{\mathbf{p}})$.*

Lemma 1. *Suppose that F is a continuous function. If $\mathcal{S}_{feas}(\mathbf{p})$ is continuous at 0 and $\mathcal{S}_{sol}(0) \neq \emptyset$, then $\mathcal{S}_{sol}(\mathbf{p})$ is outer semicontinuous at 0.*

Remark 2. *We shall clarify the continuity assumption regarding \mathcal{S}_{feas} in Lemma 1. In fact, when φ is a convex polyhedral function, \mathcal{S}_{feas} is a closed polyhedral convex mapping. Then according to Theorem 3C.3 in [40], we know that \mathcal{S}_{feas} is Lipschitz continuous, i.e. there exists $\kappa \geq 0$ such that for all $\mathbf{p}_1, \mathbf{p}_2 \in \text{dom} \mathcal{S}_{feas}$,*

$$h(\mathcal{S}_{feas}(\mathbf{p}_1), \mathcal{S}_{feas}(\mathbf{p}_2)) \leq \kappa \|\mathbf{p}_1 - \mathbf{p}_2\|,$$

where for any nonempty sets \mathcal{E} and \mathcal{F} , $h(\mathcal{E}, \mathcal{F})$ is given by $h(\mathcal{E}, \mathcal{F}) = \max\{e(\mathcal{E}, \mathcal{F}), e(\mathcal{F}, \mathcal{E})\}$, and $e(\mathcal{E}, \mathcal{F}) = \sup_{\mathbf{x} \in \mathcal{E}} d(\mathbf{x}, \mathcal{F})$. Therefore, when φ is a convex polyhedral function, all the assumptions about \mathcal{S}_{feas} in the lemma above are satisfied.

Now we are ready to induce the main result to guarantee the convergence of our proposed optimization scheme.

Theorem 2. *Suppose that φ is a convex polyhedral function and let $\{\mathbf{x}_{\delta_k}^*\}$ be the solution returned by solving Eq. (7) with errors $\{\delta_k\}$. If $\delta_k \rightarrow 0$, then*

- 1) *For any accumulation point \mathbf{x}^* of the sequence $\{\mathbf{x}_{\delta_k}^*\}$, we have that $\mathbf{x}^* \in \mathcal{S}_{sol}(0)$. That is, \mathbf{x}^* solves bilevel problem Eq. (3).*
- 2) *If F is coercive, then the sequence $\{\mathbf{x}_{\delta_k}^*\}$ is bounded and hence admits at least one accumulation point.*

B. Stability Analysis

Before we establish the desired stability result for Eq. (7), we need the stability analysis as preliminaries.

Proposition 1. *Suppose that there exists neighborhood \mathcal{N} of some point $\bar{\mathbf{x}} \in \mathcal{S}_{sol}(0)$ such that $F(\mathbf{x}) \geq F(\bar{\mathbf{x}}) + \frac{\epsilon}{2}d(\mathbf{x}, \mathcal{S}_{sol}(0))^2$, $\forall \mathbf{x} \in \mathcal{N} \cap \mathcal{S}_{feas}(0)$, where F is Lipschitz continuous with modulus L on \mathcal{N} , and there exist κ_1, κ_2 such that $\mathcal{S}_{feas}(\mathbf{p}) \cap \mathcal{N} \subseteq \mathcal{S}_{feas}(0) + \kappa_1 \|\mathbf{p}\| \mathbb{B}$, and $\mathcal{S}_{feas}(0) \cap \mathcal{N} \subseteq$*

TABLE I
THE SPECIFIC FORMS OF THE LINEAR OPERATOR \mathbf{A} IN DIFFERENT APPLICATIONS.

Application	Linear operator \mathbf{A}	Remarks
IR	\mathbf{K}	\mathbf{K} denotes the blur matrix
CS-MRI	\mathbf{PH}	\mathbf{P} is the under-sampling matrix \mathbf{H} is the Fourier transform
LLIE	\mathbf{T}	\mathbf{T} denotes the illumination matrix

$\mathcal{S}_{feas}(\mathbf{p}) + \kappa_2 \|\mathbf{p}\|_{\mathbb{B}}$. Then for any $\mathbf{x}_p \in \mathcal{S}_{sol}(\mathbf{p}) \cap \mathcal{N}$, we have $d(\mathbf{x}_p, \mathcal{S}_{sol}(0)) \leq \kappa_1 \|\mathbf{p}\| + \left(\frac{(\kappa_1 + \kappa_2)L}{c}\right)^{\frac{1}{2}} \|\mathbf{p}\|^{\frac{1}{2}}$.

According to the results obtained above, together with the arguments given in the proof of Theorem 2, we also have following stability guarantees as follows.

Theorem 3. For given $\delta > 0$, let \mathbf{x}_δ^* represents the solution returned by solving Eq. (7). Suppose that φ is a convex polyhedral function. If there exists a neighborhood \mathcal{N} of some point $\bar{\mathbf{x}} \in \mathcal{S}_{sol}(0)$ such that $F(\mathbf{x}) \geq F(\bar{\mathbf{x}}) + \frac{\epsilon}{2} d(\mathbf{x}, \mathcal{S}_{opt}(0))^2$, $\forall \mathbf{x} \in \mathcal{N} \cap \mathcal{S}_{feas}(0)$, and F is Lipschitz continuous on \mathcal{N} , then there exist $c_1, c_2 > 0$ such that for any $\mathbf{x}_\delta^* \in \mathcal{N}$, we have $d(\mathbf{x}_\delta^*, \mathcal{S}_{sol}(0)) \leq c_1 \delta + c_2 \delta^{1/2}$.

V. APPLICATIONS

This section shows how to apply TOLF to integrate task-oriented information to solve the classical ℓ_1 regularized convex optimization model for a variety of challenging applications, such as Image Restoration (IR), Compressed Sensing MRI (CS-MRI) and Low-Light Image Enhancement (LLIE). Please notice that the last two tasks actually have never been addressed by such a simple convex model in previous works. Specifically, given the observed image \mathbf{y} , by representing the latent image as $\mathbf{F}\mathbf{x}$ (i.e., \mathbf{x} represents the sparse code of the latent image on the inverse wavelet transform \mathbf{F}), we consider the upper-level objective in Eq. (3) as the following ℓ_1 -regularized energy:

$$\min_{\mathbf{x}} F(\mathbf{x}) = \frac{1}{2} \|\mathbf{A}\mathbf{F}\mathbf{x} - \mathbf{y}\|^2 + \eta \|\mathbf{x}\|_1, \quad (9)$$

where \mathbf{A} denotes a task-specific matrix, and $\eta > 0$ is a trade-off parameter. The model in Eq. (9) is actually a classical convex formulation for image restoration [1], in which we consider \mathbf{A} as the blur matrix. In Table I, we summarize the specific forms of \mathbf{A} for three image processing applications including Image Restoration (IR), Compressed Sensing MRI (CS-MRI), and Low-Light Image Enhancement (LLIE).

A. Latent Constraints

Within TOLF, we introduce an energy-based latent constraint on \mathbf{x} for Eq. (9) as follows:

$$\mathbf{x} \in \mathcal{X} := \arg \min_{\mathbf{x}} \Psi(\mathbf{x}) = \frac{1}{2} \|\mathbf{D}(\mathbf{F}\mathbf{x} - \mathbf{u})\|^2 + \gamma \|\mathbf{D}\mathbf{F}\mathbf{x}\|_1, \quad (10)$$

where \mathbf{D} denotes the gradient operator, $\mathbf{u} = \mathcal{T}(\mathbf{y})$, and γ is a balancing parameter. As for the first term in Eq. (10),

it is introduced to ensure the prominent structural similarity (extracted by the gradient map) between the warm-start \mathbf{u} and the desired output $\mathbf{F}\mathbf{x}$. The second term is to enforce the sparsity of image gradients. In fact, Ψ embeds the task information from two different perspectives on the image gradient domain. On the one hand, we utilize $\|\mathbf{D}\mathbf{F}\mathbf{x}\|_1$ to enforce the sparsity of image gradients (i.e., total variation prior). On the other hand, we incorporate a task-specific operation \mathcal{T}^4 to generate warm-start to guide the optimization process.

Here, We would like to clarify that we actually have not made such an assumption that the problem in Eq. (10) has multiple solutions. Indeed, the multiple solutions property is completely due to the underlying structure of the latent constraint defined in Eq. (10). This also justifies our motivation to study the bilevel optimization paradigm. Particularly, the non-emptiness of the latent constraint defined by the solution set of the optimization problem in Eq. (10) can be explicitly shown as following. We may first consider an auxiliary problem defined by

$$\min_{\mathbf{y}} \tilde{\Psi}(\mathbf{y}) = \frac{1}{2} \|\mathbf{y} - \mathbf{D}\mathbf{u}\|^2 + \gamma \|\mathbf{y}\|_1.$$

As $\tilde{\Psi}$ is strongly convex, the above optimization has a unique solution, that is there exists $\bar{\mathbf{y}}$ such that

$$\bar{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmin}} \tilde{\Psi}(\mathbf{y}).$$

Then the solution set of the optimization problem in Eq. (10) can be characterized by

$$\{\mathbf{x} \mid \mathbf{D}\mathbf{F}\mathbf{x} = \bar{\mathbf{y}}\} = \underset{\mathbf{x}}{\operatorname{argmin}} \Psi(\mathbf{x}).$$

According to the definition of \mathbf{D} and \mathbf{F} , the linear operator $\mathbf{D}\mathbf{F}$ is not injective, and thus the solution set of the optimization problem in Eq. (10) has multiple solutions.

We also make detailed explanation about why using proximal ADMM. Applying the convergence result of primal variables for vanilla ADMM established in [41] to the optimization problem in Eq. (5) requires the linear operator \mathcal{A} to be injective. However, when \mathcal{A} is injective, the latent feasible set of Eq. (1) (i.e., the solution set of the lower-level problem in Eq. (2)) is unique, this is not the case that we are interested in. Furthermore, for the latent feasible set in Eq. (10), the linear operator \mathcal{A} is chosen as $\mathbf{D}\mathbf{F}$ and it is not injective. The added proximal term can make the subproblems in the proximal ADMM scheme be strongly convex, specially the \mathbf{x} -update subproblem. This can help the subproblems be more stable and easier to be solved.

We emphasize that \mathcal{T} actually implements the mechanism similar to plug-and-play architecture [24], [7], but we only need to calculate it once at the initial stage, and thus it reduces the computational burden than that in existing plug-and-play methods. More importantly, TOLF obtains much better theoretical properties than plug-and-play approaches [24], [7], [25] in terms of both convergence and stability.

⁴The specific form and analysis can be found in Sec. VI-A and VI-B.

TABLE II
QUANTITATIVE RESULTS OF DIFFERENT OPTIMIZATION MECHANISMS (I.E., FISTA, PP-ADMM AND TOLF) ON SOLVING EQ. (9) FOR THE IMAGE RESTORATION TASK.

Method	FISTA	PP-ADMM		TOLF	
		(DC)	(RF)	(DC)	(RF)
PSNR	29.82	27.69	30.06	31.64	30.95
SSIM	0.8818	0.8037	0.8735	0.8997	0.8871

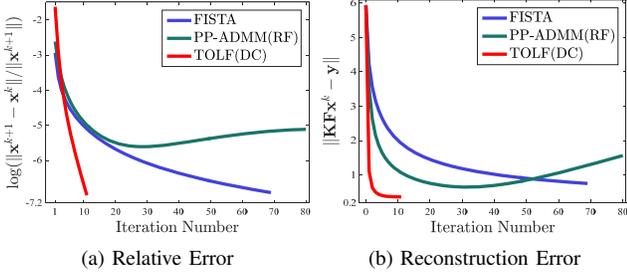


Fig. 2. Iteration behaviors of different optimization mechanisms (i.e., FISTA, PP-ADMM and TOLF) on Eq. (9). As for PP-ADMM and TOLF categories, we just adopt the settings with the best performance in Table II (i.e., PP-ADMM (RF) and TOLF (DC)).

B. Iteration Scheme

With $\Psi(\mathbf{x})$ defined in (10), we have $\varphi(\mathbf{x}) = \gamma \|\mathbf{D}\mathbf{F}\mathbf{x}\|_1$ and $\psi(\mathbf{x}) = \frac{1}{2} \|\mathbf{D}(\mathbf{F}\mathbf{x} - \mathbf{u})\|^2$ satisfying Assumption 1 with $\mathcal{A}(\mathbf{x}) = \mathbf{D}\mathbf{F}\mathbf{x}$ and $h(\cdot) = \frac{1}{2} \|\cdot - \mathbf{D}\mathbf{u}\|^2$. With an (approximation) lower-level solution $\hat{\mathbf{x}}$ obtained from solving the optimization problem in Eq. (10), applying Theorem 1 gives us the following (approximation) formula of the feasible set:

$$\hat{\mathcal{X}} = \{\mathbf{x} | \mathbf{D}\mathbf{F}\mathbf{x} = \mathbf{D}\mathbf{o}, \|\mathbf{D}\mathbf{F}\mathbf{x}\|_1 \leq \|\mathbf{D}\mathbf{F}\hat{\mathbf{x}}\|_1\},$$

where $\mathbf{o} = \mathbf{F}\hat{\mathbf{x}}$. By introducing auxiliary variables $\mathbf{s}_1, \mathbf{s}_2, r, \mathbf{w}$, we can rewrite the feasible set as

$$\hat{\mathcal{X}} = \{\mathbf{x} | \mathbf{D}\mathbf{F}\mathbf{x} = \mathbf{D}\mathbf{o}, \mathbf{D}\mathbf{F}\mathbf{x} + \mathbf{s}_1 - \mathbf{w} = 0, \mathbf{D}\mathbf{F}\mathbf{x} - \mathbf{s}_2 + \mathbf{w} = 0, \langle \mathbf{e}, \mathbf{w} \rangle + r - \hat{t} = 0, \mathbf{s}_1, \mathbf{s}_2, r \geq 0\},$$

where $\hat{t} = \|\mathbf{D}\mathbf{F}\hat{\mathbf{x}}\|_1$ and \mathbf{e} denotes the all one vector.

Here we make the clarification about the reason of rewriting feasible set. Since, if \mathbf{x} satisfies $\|\mathbf{D}\mathbf{F}\mathbf{x}\|_1 \leq \hat{t}$, by setting $\mathbf{w}_i = |(\mathbf{D}\mathbf{F}\mathbf{x})_i|$, we have $\mathbf{D}\mathbf{F}\mathbf{x} \leq \mathbf{w}$, $-\mathbf{D}\mathbf{F}\mathbf{x} \leq \mathbf{w}$ and $\langle \mathbf{e}, \mathbf{w} \rangle \leq \hat{t}$, which yields the existences of $\mathbf{s}_1, \mathbf{s}_2, r \geq 0$ such that $\mathbf{D}\mathbf{F}\mathbf{x} + \mathbf{s}_1 - \mathbf{w} = 0$, $\mathbf{D}\mathbf{F}\mathbf{x} - \mathbf{s}_2 + \mathbf{w} = 0$, and $\langle \mathbf{e}, \mathbf{w} \rangle = \hat{t}$. On the other hand, for any given \mathbf{x} , if there exist $\mathbf{w}, \mathbf{s}_1, \mathbf{s}_2, r$ satisfying $\mathbf{D}\mathbf{F}\mathbf{x} + \mathbf{s}_1 - \mathbf{w} = 0$, $\mathbf{D}\mathbf{F}\mathbf{x} - \mathbf{s}_2 + \mathbf{w} = 0$, $\langle \mathbf{e}, \mathbf{w} \rangle + r - \hat{t} = 0$ and $\mathbf{s}_1, \mathbf{s}_2, r \geq 0$, we can obtain from $\mathbf{D}\mathbf{F}\mathbf{x} \leq \mathbf{w}$, $-\mathbf{D}\mathbf{F}\mathbf{x} \leq \mathbf{w}$ that $|(\mathbf{D}\mathbf{F}\mathbf{x})_i| \leq \mathbf{w}_i$ and thus $\langle \mathbf{e}, \mathbf{w} \rangle \leq \hat{t}$ implies that $\|\mathbf{D}\mathbf{F}\mathbf{x}\|_1 \leq \hat{t}$.

With an additional auxiliary variable \mathbf{z} and equality constraint $\mathbf{x} - \mathbf{z} = 0$, the convex bilevel optimization model can be reformulated into a single level constrained optimization

problem with objective $\frac{1}{2} \|\mathbf{A}\mathbf{F}\mathbf{x} - \mathbf{y}\|^2 + \eta \|\mathbf{z}\|_1 + \delta_{\geq 0}(\mathbf{s}_1, \mathbf{s}_2, r)$ and linear equality constraints

$$\begin{cases} \mathbf{x} - \mathbf{z} = 0, \\ \mathbf{D}\mathbf{F}\mathbf{x} = \mathbf{D}\mathbf{o}, \\ \mathbf{D}\mathbf{F}\mathbf{x} + \mathbf{s}_1 - \mathbf{w} = 0, \\ \mathbf{D}\mathbf{F}\mathbf{x} - \mathbf{s}_2 + \mathbf{w} = 0, \\ \langle \mathbf{e}, \mathbf{w} \rangle + r - \hat{t} = 0. \end{cases}$$

Based on this reformulation, the proximal ADMM [42] scheme with $(\mathbf{z}, \mathbf{s}_1, \mathbf{s}_2, r)$ as the first block and (\mathbf{w}, \mathbf{x}) as the second block gives following updating rule:

$$\begin{cases} [\mathbf{z}^{k+1}]_i = \text{sign}([\mathbf{v}^k]_i) \max\{0, |[\mathbf{v}^k]_i| - \zeta\}, \\ [\mathbf{s}_1^{k+1}]_i = \max\left\{0, \frac{\beta([\mathbf{w}^k]_i - [\mathbf{D}\mathbf{F}\mathbf{x}^k]_i) - [\lambda_{\mathbf{s}_1}^k]_i + \tau[\mathbf{s}_1^k]_i}{\beta + \tau}\right\}, \\ [\mathbf{s}_2^{k+1}]_i = \max\left\{0, \frac{\beta([\mathbf{w}^k]_i + [\mathbf{D}\mathbf{F}\mathbf{x}^k]_i) + [\lambda_{\mathbf{s}_2}^k]_i + \tau[\mathbf{s}_2^k]_i}{\beta + \tau}\right\}, \\ r^{k+1} = \max\left\{0, \frac{\beta(\hat{t} - \langle \mathbf{e}, \mathbf{w}^k \rangle) - \lambda_r^k + \tau r^k}{\beta + \tau}\right\}, \\ \mathbf{w}^{k+1} = ((2\beta + \tau)\mathbf{I} + \beta\mathbf{e}\mathbf{e}^\top)^{-1}\mathbf{b}^{k+1}, \\ \mathbf{x}^{k+1} = (\mathbf{c}^{k+1})^{-1}\mathbf{d}^{k+1}, \\ \lambda_{\mathbf{x}}^{k+1} = \lambda_{\mathbf{x}}^k + \beta(\mathbf{D}\mathbf{F}\mathbf{x}^{k+1} - \mathbf{D}\mathbf{o}), \\ \lambda_{\mathbf{z}}^{k+1} = \lambda_{\mathbf{z}}^k + \beta(\mathbf{z}^{k+1} - \mathbf{x}^{k+1}), \\ \lambda_{\mathbf{s}_1}^{k+1} = \lambda_{\mathbf{s}_1}^k + \beta(\mathbf{D}\mathbf{F}\mathbf{x}^{k+1} + \mathbf{s}_1^{k+1} - \mathbf{w}^{k+1}), \\ \lambda_{\mathbf{s}_2}^{k+1} = \lambda_{\mathbf{s}_2}^k + \beta(\mathbf{D}\mathbf{F}\mathbf{x}^{k+1} - \mathbf{s}_2^{k+1} + \mathbf{w}^{k+1}), \\ \lambda_r^{k+1} = \lambda_r^k + \beta(\langle \mathbf{e}, \mathbf{w}^{k+1} \rangle + r^{k+1} - \hat{t}), \end{cases}$$

where $\beta, \tau, \zeta = \frac{\eta}{\beta + \tau} > 0$, are respectively the penalty and regularization parameters, $[\cdot]_i$ denotes the i -th element of the given vector, and \mathbf{I} is the identity matrix. The symbol $\text{sign}(\cdot)$ denotes the sign function. The other variables are presented as

$$\begin{cases} \mathbf{v}^k = \frac{\beta\mathbf{x}^k - \lambda_{\mathbf{z}}^k + \tau\mathbf{z}^k}{\beta + \tau}, \\ \mathbf{b}^{k+1} = \beta\mathbf{s}_1^{k+1} + \lambda_{\mathbf{s}_1}^k + \beta\mathbf{s}_2^{k+1} - \lambda_{\mathbf{s}_2}^k + \varrho\mathbf{e} + \tau\mathbf{w}^k, \\ \mathbf{c}^{k+1} = \mathbf{F}^\top \mathbf{A}^\top \mathbf{A} \mathbf{F} + 3\beta\mathbf{F}^\top \mathbf{D}^\top \mathbf{D} \mathbf{F} + (\beta + \tau)\mathbf{I}, \\ \mathbf{d}^{k+1} = \mathbf{F}^\top \mathbf{A}^\top \mathbf{y} + \beta\mathbf{F}^\top \mathbf{D}^\top \mathbf{D} \mathbf{o} - \mathbf{F}^\top \mathbf{D}^\top \lambda_{\mathbf{x}}^k + \beta\mathbf{z}^{k+1} + \lambda_{\mathbf{z}}^k \\ + \mathbf{F}^\top \mathbf{D}^\top (\beta(\mathbf{s}_2^{k+1} - \lambda_{\mathbf{s}_2}^k) - \beta(\mathbf{s}_1^{k+1} + \lambda_{\mathbf{s}_1}^k)) + \tau\mathbf{x}^k, \\ \varrho = \beta\hat{t} - \beta r^{k+1} - \lambda_r^k. \end{cases}$$

VI. EXPERIMENTAL RESULTS

This section first explored the iteration behaviors of TOLF to verify our theoretical results of TOLF and then researched parameter and network architecture analysis. Finally, we compared our proposed algorithm with state-of-the-art approaches on three real-world image processing applications. All the experiments were conducted on a PC with an Intel Core i7 CPU at 3.7GHz, 32GB RAM, and an NVIDIA GeForce GTX 1080Ti 11GB GPU.

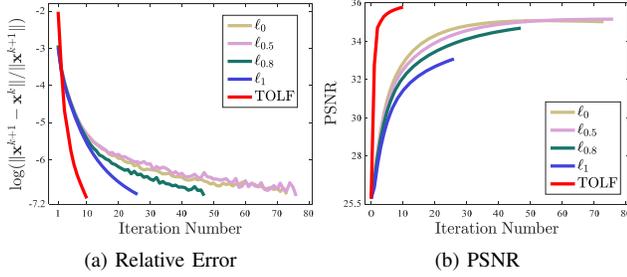


Fig. 3. Iteration behaviors of convex and non-convex optimization formulations. These models are based on Eq. (9) but with different regularization, i.e., convex ℓ_1 -norm, solved by FISTA (denoted as ℓ_1) and TOLF, and non-convex ℓ_p -norm ($p = 0, 0.5, 0.8$), solved by [20].

TABLE III
QUANTITATIVE COMPARISONS BETWEEN COMPOUND REGULARIZATION AND OUR TOLF ON DIFFERENT SETTINGS. “*” REPRESENTS THAT DC IS PERFORMED BEFORE DENOISING.

Method	Compound Regularization			TOLF		
	(DC)	(RF)	(DnCNN)	(DC)	(RF*)	(DnCNN*)
PSNR	30.27	31.73	32.54	31.64	32.22	33.13
SSIM	0.8684	0.9001	0.9197	0.8997	0.9103	0.9324

A. Iteration Behaviors Analysis

In this part, we compared TOLF to classical convex, non-convex and plug-and-play optimization techniques on the image restoration task. Specifically, we first considered to adopt different optimization mechanisms, including standard convex scheme (e.g., FISTA [1]), plug-and-play strategy (i.e., Plug-and-Play ADMM [24], PP-ADMM for short) and our TOLF, to solve the classical convex model in Eq. (9). As for PP-ADMM and TOLF, we introduced two different task-specific operations including the classical image filter, i.e., the recursive filter [24] (RF for short), and the task-based deconvolution process [7], i.e., $\arg \min_{\mathbf{x}} \|\mathbf{K}\mathbf{F}\mathbf{x} - \mathbf{y}\|^2 + \alpha \|\mathbf{F}\mathbf{x} - \mathbf{y}\|^2$ with a trade-off $\alpha > 0$ (fix it as 10^{-4} for all the experiments) (DC for short). The lower-level subproblem was solved by APG [43] with the relative errors $\|\mathbf{x}^{k+1} - \mathbf{x}^k\| / \|\mathbf{x}^{k+1}\| \leq 10^{-3}$, to find an approximate solution to $\bar{\mathbf{x}}$, which works as an initialization to the problem in Eq. (4). In Table II, we reported quantitative performances of these compared methods on an example image from Levin et al.’s benchmark [44]. Since we cannot obtain convergence sequences for PP-ADMM even after 80 iterations, we had to report the best results during their iterations, i.e., the 37th and 29th steps for PP-ADMM (RF) and PP-ADMM (DC), respectively. It can be seen that in most cases introducing task-specific operations improved the performance of Eq. (9) for image restoration. That is, the results of PP-ADMM (RF), TOLF (DC) and TOLF (RF) were all better than the classical FISTA method. Meanwhile, we also observed that PP-ADMM (DC) was even worse than FISTA. This is mainly because the DC operation is repeatedly performed within the plug-and-play mechanism, which may overly smooth the image details.

Notably, our TOLF (DC) is superior to TOLF (RF), but PP-ADMM (RF) is superior to PP-ADMM (DC). It is because DC was only performed at the first stage of TOLF, which actually

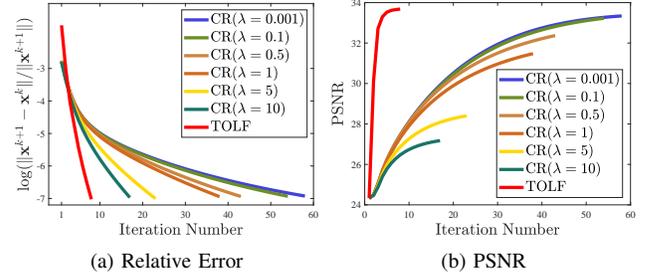


Fig. 4. Iteration behavior of Compound Regularization (CR) with different regularization parameters λ and our TOLF. We just adopt the settings with the best performance in Table III (i.e., Compound Regularization (DnCNN) and TOLF (DnCNN)).

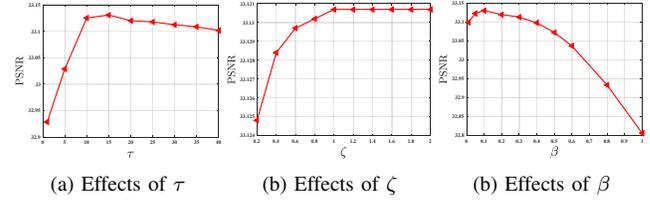


Fig. 5. Sensitivity analysis with respect to parameters τ , ζ , and β .

provided a task-specific warm start to optimize the blurry input for our iterations, but RF is task-independent which just presents a denoising operation for the plugged step. In other words, different from PP-ADMM, the task-specific operation is a crucial component for the warm start procedure in our TOLF.

We also compared the iteration behaviors of FISTA, PP-ADMM and TOLF in Fig. 2. For the last two methods, we only chose the inner operator \mathcal{T} with better performances in Table II (i.e., RF for PP-ADMM and DC for TOLF) to provide clearer illustrations. We observed that the speed of FISTA was slow and the iterations converged after 70 steps. It also can be seen that PP-ADMM (RF) did not converge even after 80 steps, but it obtained lower reconstruction errors than FISTA near the 30th step (see the right subfigure). In contrast, TOLF (DC) converged only after 10 iterations and achieved the lowest reconstruction errors.

Next, we adopted some recently proposed non-convex regularization techniques [17], [19] (e.g., replace ℓ_1 -norm by ℓ_p -norm, $p = 0, 0.5, 0.8$) to reformulate Eq. (9) for image restoration. The same non-convex accelerated proximal gradient scheme [20] was performed to solve these ℓ_p -regularized models. Their iteration behaviors are visualized in Fig. 3. We also plotted the iteration curves of FISTA and our TOLF on the original ℓ_1 -norm regularized model in this figure. It can be seen that these non-convex regularization models improved the practical performance (i.e., PSNR in the right subfigure) of the original convex model in Eq. (9) (solved by FISTA, denoted as ℓ_1), but their trajectories fluctuated after several iterations (see the left subfigure). In contrast, TOLF obtained the fastest convergence speed and the best final performance among all the compared methods, even by only solving a convex model.

Finally, we presented numerical results and iterative behaviors in terms of the compound regularization (i.e., $\min_{\mathbf{x}} F(\mathbf{x}) + \lambda \Psi(\mathbf{x})$, where λ is the positive balancing parame-

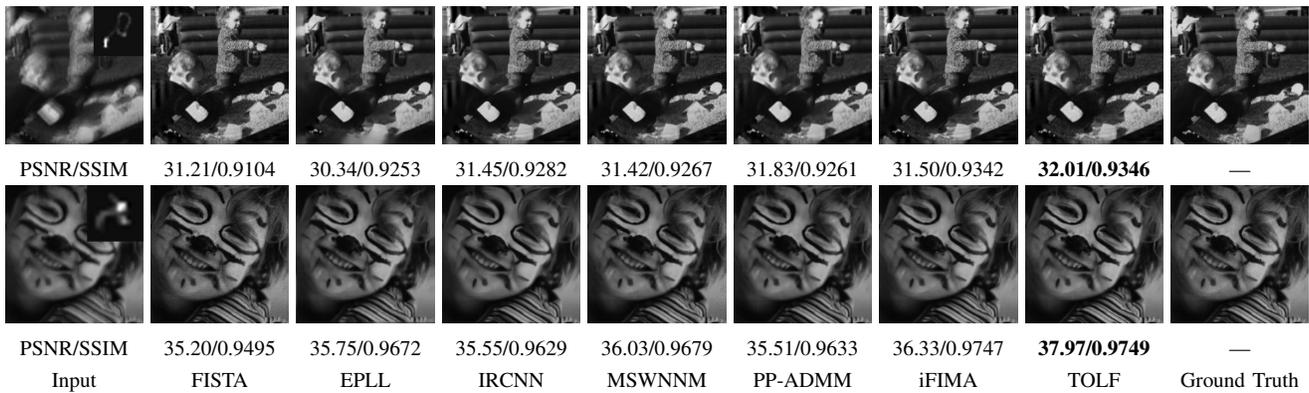


Fig. 6. Image restoration results on two examples in Levin et al.’ dataset. We chose top six performance methods from Table. V.



Fig. 7. Image restoration results on a challenging color image with a large-sized kernel (75×75). The PSNR score is reported below each subfigure.

TABLE IV
QUANTITATIVE RESULTS OF DIFFERENT NETWORKS (I.E., TNRD, DNCNN, AND IRCNN). “*” REPRESENTS THAT DC IS PERFORMED BEFORE DENOISING.

Method	PP-ADMM			TOLF		
	(TNRD)	(IRCNN)	(DnCNN)	(TNRD*)	(IRCNN*)	(DnCNN*)
PSNR	31.63	32.58	32.60	32.33	32.97	33.13
SSIM	0.8902	0.9240	0.9296	0.9112	0.9295	0.9324

ter) and our model (i.e., $\min_{\mathbf{x}} F(\mathbf{x}) \text{ s.t. } \mathbf{x} \in \arg \min_{\mathbf{x}} \Psi(\mathbf{x})$). To ensure fairness, we adopted the proximal ADMM (used for our TOLF) to solve the model of compound regularization. In addition, we considered the same \mathcal{T} to define \mathbf{u} in $\Psi(\mathbf{x})$. As presented in Table II, we recognized that DC (deconvolution process) is a crucial step for obtaining \mathbf{u} , so we would like to emphasize that we performed DC before denoising operation (e.g., RF, and DnCNN [45]) in the following experiments. Table III reported quantitative results of these two models by adopting different settings. Note that $\lambda = 0.1$ in the compound regularization. Thanks to our newly-introduced latent feasibility, it can be easily seen that our results were consistently superior to compound regularization in all cases. The fact that the quantitative scores of TOLF (RF*) (32.22/0.9103) obtained a significant improvement than TOLF (RF) (30.95/0.8871) (* represents that DC is performed before denoising), also justified the necessity of the task-specific operation for T. Fig. 4 displayed iteration behaviors of compound regularization and our TOLF. To make a rigorous evaluation, we change settings of compound regularization by adopting different λ . We can easily observe that with the value of λ decreasing, the compound regularization converged more slowly but obtained better performance. By contrast, our TOLF realized a higher value at a faster speed than all naive compound regularizations with different regularization param-

eters. This experiment indicates that compound regularization cannot promote numerical improvement and accelerate the convergence speed simultaneously, while TOLF we proposed can obtain a satisfying outcome.

B. Parameter and Network Architecture Analysis

The proposed algorithm involves many parameters as described in Sec. V-B, some of which are iteration variables introduced based on the proximal ADMM scheme, i.e., $\lambda_{\mathbf{x}}, \lambda_{\mathbf{z}}, \lambda_{\mathbf{s}_1}, \lambda_{\mathbf{s}_2}, \lambda_r$. As for them, we followed the commonly-used setting (set as zero) to initialize them. Here we mainly explored the effects of some algorithmic parameters including τ, ζ , and β . As shown in Fig. 5, τ and ζ were insensitive to different settings to some extent, while the parameter β was sensitive when it increased. Large β caused poor performance. Additionally, based on these results, we defined $\tau = 15, \zeta = 1, \beta = 0.1$ as our default settings for solving image restoration.

Table IV reported quantitative scores of using different networks for PP-ADMM and our method. Among them, TNRD [47] is a famous image denoising framework based on a nonlinear reaction-diffusion model; DnCNN [45] is a well-known CNN-based network for image denoising; IRCNN is a recently-proposed plug-and-play framework for image restoration, and here we just utilized its denoising architecture. It can be easily seen that DnCNN reached the highest scores both in PP-ADMM and our method because it owned stronger denoising ability than IRCNN and TNRD. Thus we chose DnCNN as \mathcal{T} for image restoration in the experiments mentioned above and in the following experiments. Moreover, the results of our method were consistently better than PP-ADMM under different networks. It showed the superiority of our designed computational framework against the classical PP-ADMM after plugging the learnable architecture.

TABLE V
AVERAGED IMAGE RESTORATION PERFORMANCE ON LEVIN ET AL.' BENCHMARK [44].

	FTVd	FISTA	HL	IDDBM3D	EPLL	MLP	IRCNN	MSWNNM	FDN	GLRA	PP-ADMM	MEDAEP	eFIMA	iFIMA	TOLF
PSNR	29.38	31.81	30.12	31.53	31.65	31.32	32.28	32.50	32.04	19.75	32.60	19.77	32.68	32.72	33.13
SSIM	0.8819	0.9010	0.8961	0.9043	0.9258	0.8991	0.9200	0.9247	0.9277	0.5134	0.9296	0.4639	0.9291	0.9296	0.9324

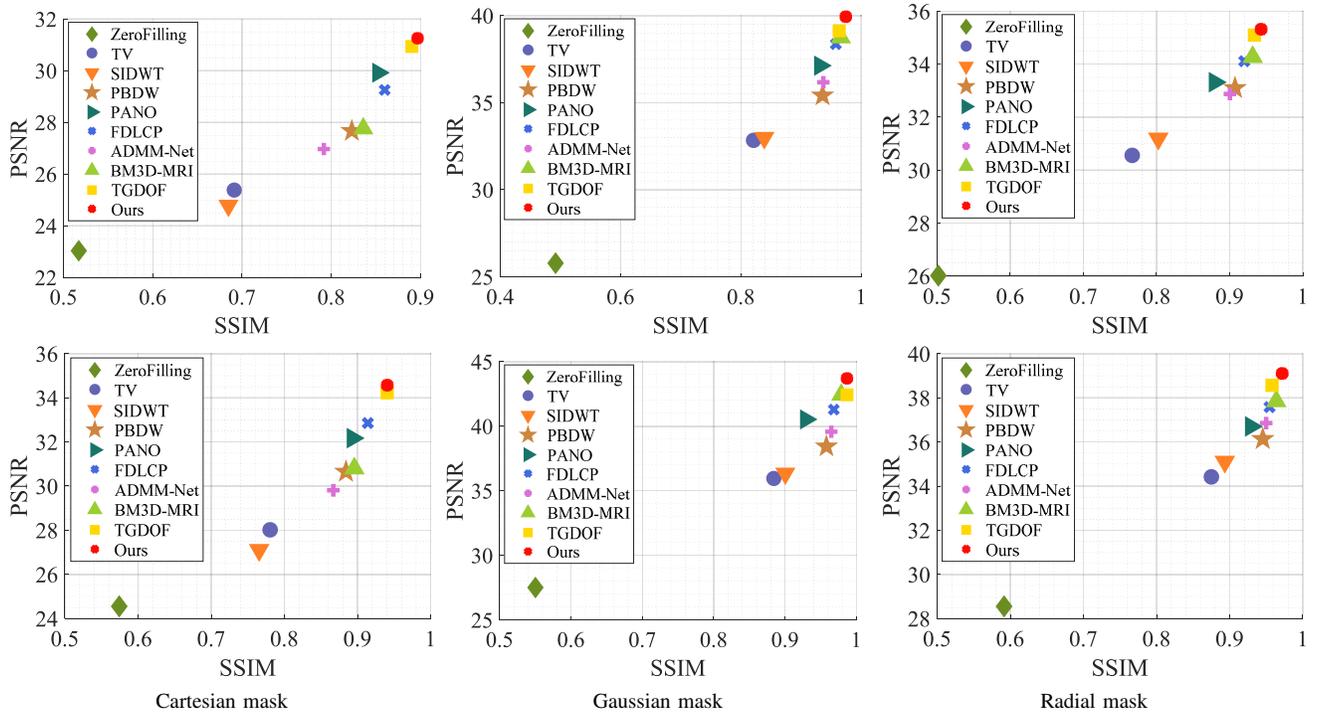


Fig. 8. Averaged CS-MRI results on IXI dataset among state-of-the-art methods. Top row: 20% sampling rate. Bottom row: 30% sampling rate. In each subfigure, the upper right is the best.

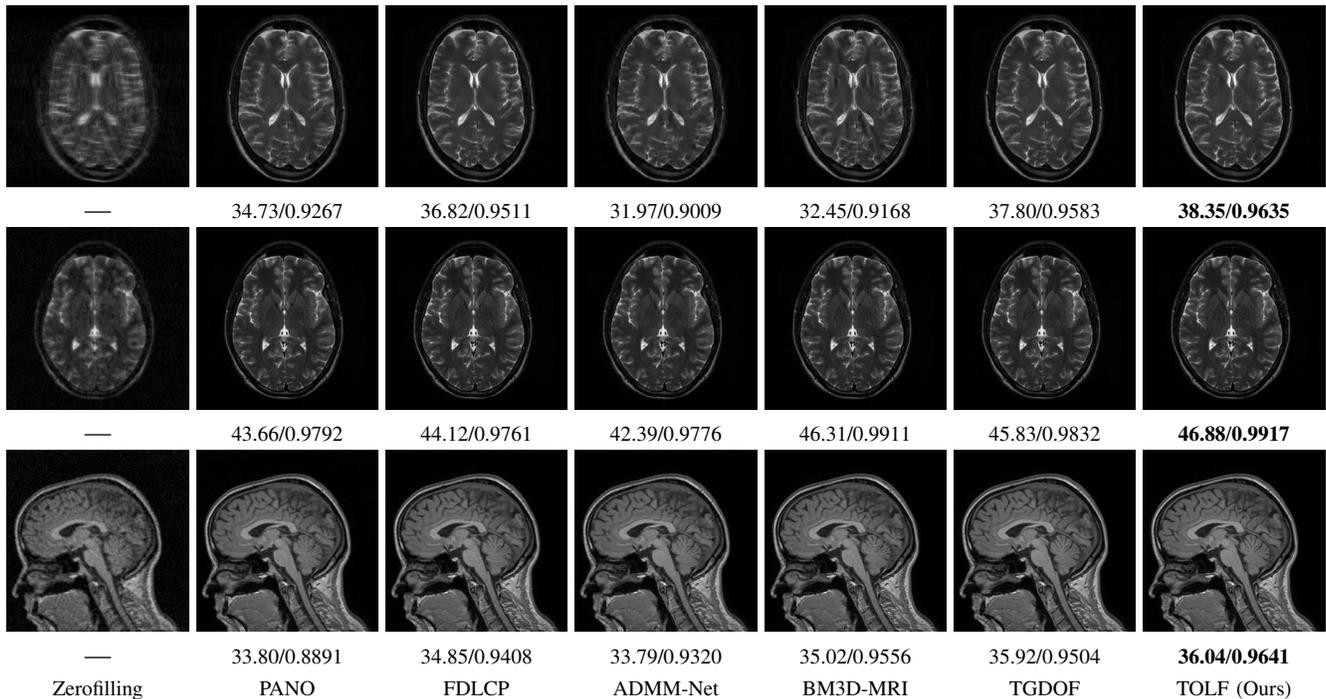


Fig. 9. Visual comparison among state-of-the-art methods of Compressive Sensing MRI at the sparse k -space data with different undersampling patterns and at a 30% sampling rates (Top row: Cartesian mask. Middle row: Gaussian mask. Bottom row: Radial mask).

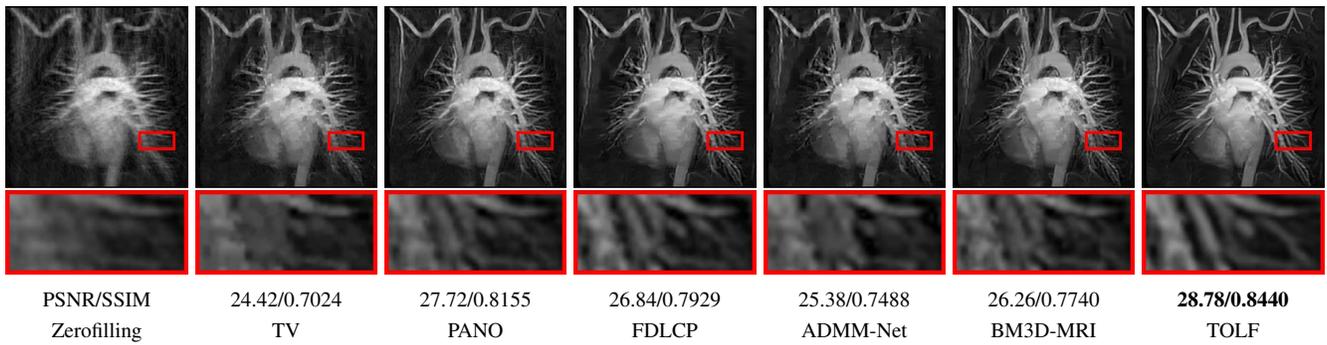


Fig. 10. CS-MRI results on chest data with Cartesian mask (30% sampling rate).

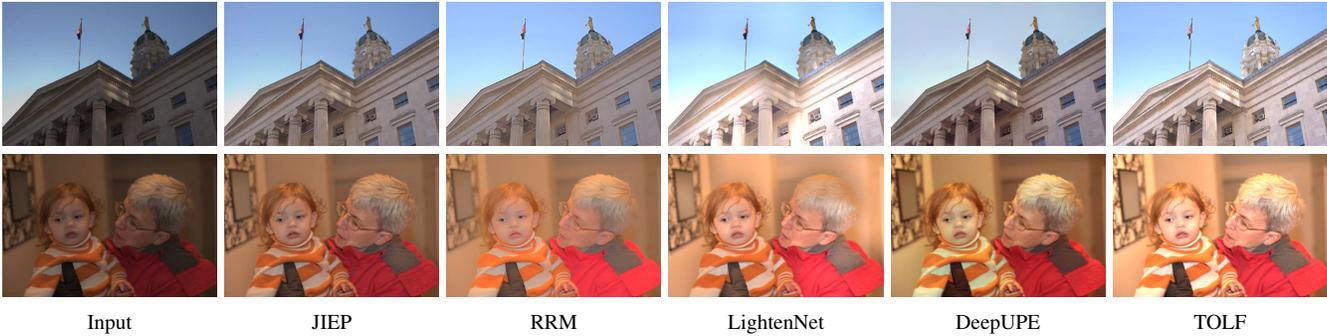


Fig. 11. Illustrating visual results on testing images collected by [46].

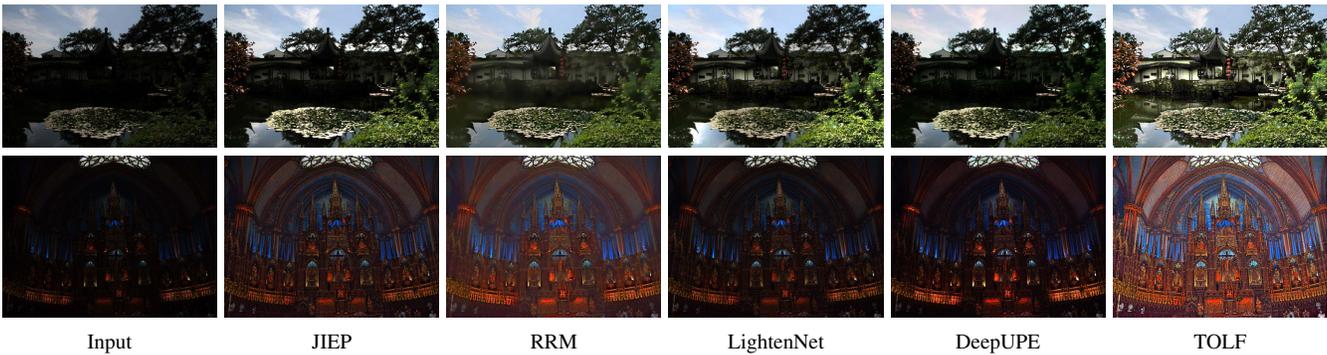


Fig. 12. Illustrating visual results on challenging low-light images.

C. Image Processing Applications

Image Restoration. We compared TOLF with many approaches, including traditional optimization methods (i.e., FTVd [2], FISTA [1], HL [17], IDDBM3D [48], EPLL [16]), and learning-based methods (i.e., MLP [49], IRCNN [6], MSWNNM [50], FDN [51], GLRA [52], PP-ADMM [24], MEDAEP [53], FIMA [7] (contains eFIMA and iFIMA)). As for TOLF, we introduced a denoising CNN [45] architecture as \mathcal{T} (according to Table IV). We conducted experiments on the Levin et al.’ benchmark [44], which includes 32 images of the size 255×255 and blurred by 8 different kernels of the size ranging from 13×13 to 27×27 . We reported the quantitative scores in Table V. The visual comparisons on an example image from this benchmark were plotted in Fig. 6. Obviously, deep-learning-based IRCNN achieved much better performance than other traditional optimization methods. The recently-proposed FIMA (includes eFIMA and

iFIMA) considered integrating the data and knowledge by an optimization unrolling strategy, thus its results were even better. Nevertheless, thanks to the novel modeling mechanism, TOLF obtained the best quantitative and qualitative results. In addition, a color image (612×342) corrupted by a very large kernel (75×75) was used to further evaluate the performance, shown in Fig. VI-A. Again, TOLF recovered richer textures and details, and thus performed the best.

Compressed Sensing MRI (CS-MRI). We then evaluated TOLF on the CS-MRI task. Here we defined the task-specific operation $\arg \min_{\mathbf{x}} \|\mathbf{P}\mathbf{H}\mathbf{F}\mathbf{x} - \mathbf{y}\|^2 + \alpha \|\mathbf{F}\mathbf{x} - \mathbf{y}\|^2$ with a trade-off $\alpha > 0$ (fix it as 10^{-3}) and the pre-trained denoising model described in [6] together as \mathcal{T} . Specifically, we conducted experiments on 55 images from the widely-used IXI MRI benchmark⁵. Our experiments contained three types of undersampling patterns (i.e., Cartesian, Gaussian, and Radial

⁵<http://brain-development.org/ixi-dataset/>.

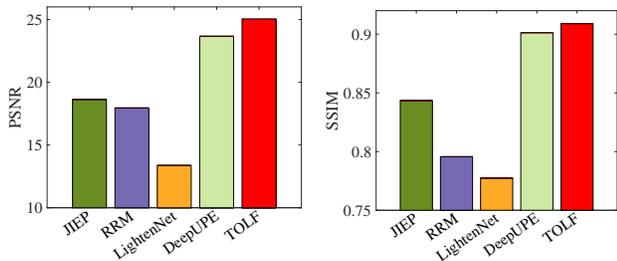


Fig. 13. Averaged LLIE results on 500 test images collected by [46].

mask) and two sampling rates (i.e., 20%, 30%) to generate the sparse k -space data. We compared TOLF with many state-of-the-art methods including TV [54], SIDWT [55], PBDW [56], PANO [57], FDLCP [58], ADMM-Net [59], TGDOF [9], and BM3D-MRI [60]. Fig. 8 illustrated the quantitative results of these approaches on the test dataset. Our TOLF achieved the best results according to both PSNR and SSIM scores (i.e., the upper right red points). At a sampling rate of 30%, Fig. 10 demonstrated CS-MRI results on a challenging chest data [9] with Cartesian mask. Obviously, TOLF obtained the best quantitative and qualitative performance.

Low-Light Image Enhancement (LLIE). Lastly, we applied TOLF to solve LLIE. Here we defined the task-specific operation as $\mathbf{x} = \mathbf{y} \oslash \mathcal{T}(\mathbf{y})$, where \oslash represents the element-wise division. We compared TOLF with two optimization methods (i.e., JIEP [61] and RRM [11]) and two recent deep learning techniques (i.e., LightenNet [62] and DeepUPE [46]). Fig. 12 illustrated the visual results of these methods on two challenging images. Quantitative results were reported in Fig. 13. By comparison, TOLF obtained the best visual quality and the highest scores. In Fig. 14, we compared the performance of TOLF with different warm-start operations \mathcal{T} , including the naive low-light input (denoted as \mathbf{y}), the result of the gamma correction (denoted as $\mathbf{y}^{1/a}$ with $a = 2.2$), a simplified relative total variation filter [18] (denoted as $\text{RTV}(\mathbf{y})$), and the simple denoising CNN architecture used in the above image restoration experiment (denoted as $\text{CNN}(\mathbf{y})$). We first observed that TOLF with RTV and CNN performed better than the other two choices. On the other hand, we emphasize that even with different warm starts, the TOLF process consistently improved the overall performance.

VII. CONCLUSION AND FUTURE WORK

This paper developed a task-oriented convex bilevel optimization with latent feasibility for handling complex problems. The convergence and stability were strictly proved to realize our solid theoretical guarantee. Experiments on iteration behaviors verified the properties of TOLF. Extensive comparisons on three real-world applications demonstrated our outstanding performance against existing advanced methods.

Actually, our TOLF is designed towards general learning and vision models. In this work, we mainly focus on low-level vision tasks to evaluate the performance. In the future, we will consider extending our designed method for more challenging vision tasks, e.g., weakly supervised learning. Here we provide two possible research directions for related readers. The one

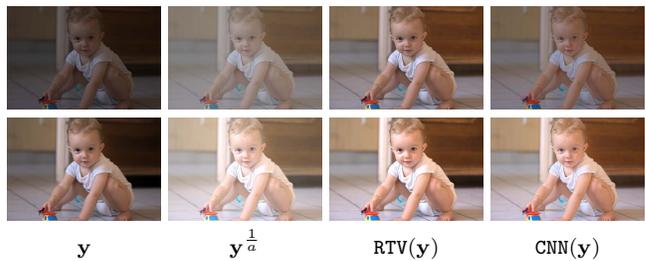


Fig. 14. Illustrating the LLIE results of TOLF with different \mathcal{T} . Top row: results of warm start from \mathcal{T} . Bottom row: final enhanced results of TOLF.

is to follow the existing deep unrolling schemes to unroll the TOLF and introduce the task-specific architecture into the iteration step to further establish an end-to-end network. The other is to extend the TOLF to generate a gradient-based propagation algorithm for improving the training efficiency towards general learning issues.

APPENDIX

In this part, we present the detailed proofs for all the theoretical results in our algorithm.

A. Proof of Theorem 1

Proof. Given a solution $\bar{\mathbf{x}} \in \mathcal{X}$. First, for any $\mathbf{x} \in \{\mathbf{x} | \mathcal{A}(\mathbf{x}) = \mathcal{A}(\bar{\mathbf{x}}), \varphi(\mathbf{x}) \leq \varphi(\bar{\mathbf{x}})\}$, we have $\Psi(\mathbf{x}) = h(\mathcal{A}(\bar{\mathbf{x}})) + \varphi(\mathbf{x}) \leq h(\mathcal{A}(\bar{\mathbf{x}})) + \varphi(\bar{\mathbf{x}}) = \min_{\mathbf{x}} \Psi(\mathbf{x})$, and thus

$$\{\mathbf{x} | \mathcal{A}(\mathbf{x}) = \mathcal{A}(\bar{\mathbf{x}}), \varphi(\mathbf{x}) \leq \varphi(\bar{\mathbf{x}})\} \subseteq \mathcal{X}. \quad (11)$$

For any $\mathbf{x} \in \mathcal{X}$, if $\mathcal{A}(\mathbf{x}) \neq \mathcal{A}(\bar{\mathbf{x}})$, let $\mathbf{x}_\alpha = (1 - \alpha)\bar{\mathbf{x}} + \alpha\mathbf{x}$, and then $\mathbf{x}_\alpha \in \mathcal{X}$ because \mathcal{X} is convex. As h is locally strongly convex around $\bar{\mathbf{x}}$, there exists neighborhood \mathcal{N} of $\mathcal{A}(\bar{\mathbf{x}})$ such that h is strongly convex on \mathcal{N} . There exists sufficiently small $\alpha > 0$ such that $\mathcal{A}(\mathbf{x}_\alpha) \in \mathcal{N}$ and $\mathcal{A}(\mathbf{x}_\alpha) \neq \mathcal{A}(\bar{\mathbf{x}})$. Then, there exists $\sigma > 0$ such that

$$h(\mathcal{A}(\mathbf{x}_\alpha)) \geq h(\mathcal{A}(\bar{\mathbf{x}})) + \alpha \langle \mathbf{D}h(\mathcal{A}(\bar{\mathbf{x}})), \mathcal{A}(\mathbf{x}) - \mathcal{A}(\bar{\mathbf{x}}) \rangle + \frac{\sigma}{2} \alpha^2 \|\mathcal{A}(\mathbf{x}) - \mathcal{A}(\bar{\mathbf{x}})\|^2.$$

And since $0 \in \mathcal{A}^T \mathbf{D}h(\mathcal{A}(\bar{\mathbf{x}})) + \partial\varphi(\bar{\mathbf{x}})$, by the convexity of φ , we have

$$\varphi(\mathbf{x}_\alpha) \geq \varphi(\bar{\mathbf{x}}) + \alpha \langle -\mathcal{A}^T \mathbf{D}h(\mathcal{A}(\bar{\mathbf{x}})), \mathbf{x} - \bar{\mathbf{x}} \rangle.$$

Combining the two inequalities given above, we obtain

$$\Psi(\mathbf{x}_\alpha) \geq \Psi(\bar{\mathbf{x}}) + \frac{\sigma}{2} \alpha^2 \|\mathcal{A}(\mathbf{x}) - \mathcal{A}(\bar{\mathbf{x}})\|^2 > \Psi(\bar{\mathbf{x}}),$$

which contradicts to the fact that $\mathbf{x}_\alpha \in \mathcal{X}$. Next, since $\mathcal{A}(\mathbf{x}) = \mathcal{A}(\bar{\mathbf{x}})$, and $\Psi(\mathbf{x}) = \Psi(\bar{\mathbf{x}})$, we have $\varphi(\mathbf{x}) = \varphi(\bar{\mathbf{x}})$, and thus

$$\mathcal{X} \subseteq \{\mathbf{x} | \mathcal{A}(\mathbf{x}) = \mathcal{A}(\bar{\mathbf{x}}), \varphi(\mathbf{x}) \leq \varphi(\bar{\mathbf{x}})\}. \quad (12)$$

Upon combining Eq. (11) and Eq. (12), we reach the re-characterization of \mathcal{X} as Eq. (4). \square

B. Proof of Corollary 1

Proof. By Theorem 1, we directly have the equivalence between Eq. (3) and Eq. (5). And since Eq. (6) is the iterations of two block proximal ADMM applied on Eq. (5). Thus, the convergence of the iterations in solving problem Eq. (5) can be directly guaranteed via standard convergence results of proximal ADMM [63]. Moreover, by applying the investigations in [64], we can further obtain a linear convergence rate estimation for our iterations in solving problem Eq. (5). \square

C. Proof of Lemma 1

Proof. First, we show that $\limsup_{\mathbf{p} \rightarrow 0} \mathcal{S}_{val}(\mathbf{p}) \leq \mathcal{S}_{val}(0)$. Let $\bar{\mathbf{x}} \in \mathcal{S}_{sol}(0)$. Since $\mathcal{S}_{feas}(\mathbf{p})$ is continuous at 0, it is inner semicontinuous at 0. Thus for any sequence $\mathbf{p}^k \rightarrow 0$, we get the existence of a sequence of points $\mathbf{x}^k \in \mathcal{S}_{feas}(\mathbf{p}^k)$ such that $\mathbf{x}^k \rightarrow \bar{\mathbf{x}}$ as $k \rightarrow \infty$. Then for any $\epsilon > 0$ there exists $N > 0$ such that

$$\mathcal{S}_{val}(\mathbf{p}^k) \leq F(\mathbf{x}^k) \leq F(\bar{\mathbf{x}}) + \epsilon = \mathcal{S}_{val}(0) + \epsilon, \quad \forall k \geq N,$$

which implies

$$\limsup_{\mathbf{p} \rightarrow 0} \mathcal{S}_{val}(\mathbf{p}) \leq \mathcal{S}_{val}(0).$$

We next show the outer semicontinuity of \mathcal{S}_{sol} at 0. For any $\mathbf{p}^k \rightarrow 0$ with $\mathbf{x}^k \in \mathcal{S}_{opt}(\mathbf{p}^k)$ such that $\mathbf{x}^k \rightarrow \bar{\mathbf{x}}$, since \mathcal{S}_{feas} is outer semicontinuous at 0, we have $\bar{\mathbf{x}} \in \mathcal{S}_{feas}(0)$. By the continuity of F and upper semicontinuity of \mathcal{S}_{val} at 0, we have

$$\mathcal{S}_{val}(0) \leq F(\bar{\mathbf{x}}) = \lim_{k \rightarrow \infty} F(\mathbf{x}^k) = \limsup_{k \rightarrow \infty} \mathcal{S}_{val}(\mathbf{p}^k) \leq \mathcal{S}_{val}(0),$$

which implies

$$\bar{\mathbf{x}} \in \mathcal{S}_{sol}(0).$$

That is, $\mathcal{S}_{sol}(\mathbf{p})$ is outer semicontinuous at 0 according to Definition 1 in the manuscript. \square

D. Proof of Theorem 2

Proof. For any \mathbf{x}_δ^* , we have $\mathbf{x}_\delta^* \in \mathcal{S}_{sol}(\mathbf{p})$ with $\mathbf{p}_1 = \mathcal{A}(\bar{\mathbf{x}}_\delta) - \mathcal{A}(\bar{\mathbf{x}})$ and $\mathbf{p}_2 = \varphi(\bar{\mathbf{x}}_\delta) - \varphi(\bar{\mathbf{x}})$. Therefore, $\|\mathbf{p}\| \leq \|\mathcal{A}(\bar{\mathbf{x}}_\delta) - \mathcal{A}(\bar{\mathbf{x}})\| + \|\varphi(\bar{\mathbf{x}}_\delta) - \varphi(\bar{\mathbf{x}})\| \leq (\|\mathcal{A}\| + L_\varphi)d(\bar{\mathbf{x}}_\delta, \mathcal{X}) \leq (\|\mathcal{A}\| + L_\varphi)\delta$, where L_φ is the Lipschitz continuity modulus of φ . Note that the Lipschitz continuity modulus is guaranteed to exist because φ is a convex polyhedral function. Then the first argument follows from Lemma 1 directly. The second argument actually follows from the fact that $\limsup_{k \rightarrow \infty} F(\mathbf{x}_{\delta_k}^*) \leq \mathcal{S}_{val}(0)$ and F is coercive. \square

E. Proof of Proposition 1

Proof. For any $\mathbf{x}_p \in \mathcal{S}_{opt}(\mathbf{p}) \cap \mathcal{N}$, let $\mathbf{z} := \text{Proj}_{\mathcal{S}_{feas}(0)}(\mathbf{x}_p)$ and $\mathbf{x}_0 := \text{Proj}_{\mathcal{S}_{sol}(0)}(\mathbf{z})$, and since $\mathcal{S}_{feas}(0)$ and $\mathcal{S}_{sol}(0)$ are both closed convex sets, \mathbf{z} and \mathbf{x}_0 are well defined. Because $\bar{\mathbf{x}} \in \mathcal{S}_{sol}(0)$, we have $\|\mathbf{x}_0 - \bar{\mathbf{x}}\| \leq \|\mathbf{z} - \bar{\mathbf{x}}\| \leq \|\mathbf{x}_p - \bar{\mathbf{x}}\|$, and thus $\mathbf{z}, \mathbf{x}_0 \in \mathcal{N}$ and $\|\mathbf{x}_p - \mathbf{z}\| \leq \kappa_1 \|\mathbf{p}\|$.

Since $\mathbf{x}_p \in \mathcal{S}_{sol}(\mathbf{p})$, for any point $\mathbf{y} \in \mathcal{S}_{feas}(\mathbf{p})$, we have

$$F(\mathbf{x}_p) - F(\mathbf{x}_0) = F(\mathbf{x}_p) - F(\mathbf{y}) + F(\mathbf{y}) - F(\mathbf{x}_0) \leq L\|\mathbf{y} - \mathbf{x}_0\|.$$

Since \mathbf{y} can be any point in $\mathcal{S}_{feas}(\mathbf{p})$, we have

$$F(\mathbf{x}_p) - F(\mathbf{x}_0) \leq \kappa_2 L \|\mathbf{p}\|. \quad (13)$$

Next, we have

$$\begin{aligned} F(\mathbf{x}_p) - F(\mathbf{x}_0) &\geq F(\mathbf{z}) - F(\mathbf{x}_0) - |F(\mathbf{x}_p) - F(\mathbf{z})| \\ &\geq c\|\mathbf{z} - \mathbf{x}_0\|^2 - L\|\mathbf{x}_p - \mathbf{z}\| \\ &\geq c(\|\mathbf{x}_p - \mathbf{x}_0\| - \|\mathbf{z} - \mathbf{x}_p\|)^2 - \kappa_1 L \|\mathbf{p}\| \\ &\geq c(\|\mathbf{x}_p - \mathbf{x}_0\| - \kappa_1 \|\mathbf{p}\|)^2 - \kappa_1 L \|\mathbf{p}\|. \end{aligned}$$

Combining with Eq. (13), we get

$$\kappa_2 L \|\mathbf{p}\| \geq c(\|\mathbf{x}_p - \mathbf{x}_0\| - \kappa_1 \|\mathbf{p}\|)^2 - \kappa_1 L \|\mathbf{p}\|,$$

and thus

$$d(\mathbf{x}_p, \mathcal{S}_{sol}(0)) \leq \|\mathbf{x}_p - \mathbf{x}_0\| \leq \kappa_1 \|\mathbf{p}\| + \sqrt{\frac{(\kappa_1 + \kappa_2)L}{c}} \|\mathbf{p}\|. \quad \square$$

F. Proof of Theorem 3

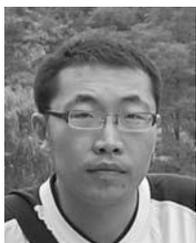
Proof. As stated in our manuscript, according to the results proved in Proposition 1, together with the arguments given in the proof of Theorem 2, we can directly have the stability guarantees in this theorem. \square

REFERENCES

- [1] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.
- [2] C. Li, W. Yin, H. Jiang, and Y. Zhang, "An efficient augmented lagrangian method with applications to total variation minimization," *Computational Optimization and Applications*, vol. 56, no. 3, pp. 507–530, 2013.
- [3] X. Peng, J. Feng, S. Xiao, W.-Y. Yau, J. T. Zhou, and S. Yang, "Structured autoencoders for subspace clustering," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5076–5086, 2018.
- [4] Z. Huang, P. Hu, J. T. Zhou, J. Lv, and X. Peng, "Partially view-aligned clustering," *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [5] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.
- [6] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep cnn denoiser prior for image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3929–3938.
- [7] R. Liu, S. Cheng, Y. He, X. Fan, Z. Lin, and Z. Luo, "On the convergence of learning-based iterative methods for nonconvex inverse problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 12, pp. 3027–3039, 2019.
- [8] R. Liu, P. Mu, J. Chen, X. Fan, and Z. Luo, "Investigating task-driven latent feasibility for nonconvex image modeling," *IEEE Transactions on Image Processing*, vol. 29, pp. 7629–7640, 2020.
- [9] R. Liu, Y. Zhang, S. Cheng, X. Fan, and Z. Luo, "A theoretically guaranteed deep optimization framework for robust compressive sensing mri," in *Association for the Advancement of Artificial Intelligence*, 2019.
- [10] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.
- [11] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust retinex model," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2828–2841, 2018.
- [12] S. Dempe, V. Kalashnikov, G. A. Prez-Valds, and N. Kalashnykova, *Bilevel Programming Problems: Theory, Algorithms and Applications to Energy Networks*. Springer Publishing Company, Incorporated, 2015.
- [13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

- [14] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.
- [15] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
- [16] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 479–486.
- [17] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-laplacian priors," in *Neural Information Processing Systems*, 2009, pp. 1033–1041.
- [18] L. Xu, Q. Yan, Y. Xia, and J. Jia, "Structure extraction from texture via relative total variation," *ACM Transactions on Graphics*, vol. 31, no. 6, 2012.
- [19] J. Pan, Z. Hu, Z. Su, and M.-H. Yang, " l_0 -regularized intensity and gradient prior for deblurring text images and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 342–355, 2016.
- [20] H. Li and Z. Lin, "Accelerated proximal gradient methods for nonconvex programming," in *Neural Information Processing Systems*, 2015, pp. 379–387.
- [21] P. Jain, P. Kar *et al.*, "Non-convex optimization for machine learning," *Foundations and Trends® in Machine Learning*, vol. 10, no. 3–4, pp. 142–336, 2017.
- [22] S. N. Negahban, P. Ravikumar, M. J. Wainwright, B. Yu *et al.*, "A unified framework for high-dimensional analysis of m -estimators with decomposable regularizers," *Statistical Science*, vol. 27, no. 4, pp. 538–557, 2012.
- [23] Y. Chen and M. J. Wainwright, "Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees," *arXiv preprint arXiv:1509.03025*, 2015.
- [24] S. H. Chan, X. Wang, and O. A. Elgandy, "Plug-and-play admm for image restoration: Fixed-point convergence and applications," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 84–98, 2016.
- [25] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising," *SIAM Journal on Imaging Sciences*, vol. 10, no. 4, pp. 1804–1844, 2017.
- [26] W. Dong, P. Wang, W. Yin, G. Shi, F. Wu, and X. Lu, "Denoising prior driven deep neural network for image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 10, pp. 2305–2318, 2018.
- [27] E. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin, "Plug-and-play methods provably converge with properly trained denoisers," in *International Conference on Machine Learning*, 2019, pp. 5546–5557.
- [28] S. Sreehari, S. V. Venkatakrishnan, B. Wohlberg, G. T. Buzzard, L. F. Drummy, J. P. Simmons, and C. A. Bouman, "Plug-and-play priors for bright field electron tomography and sparse interpolation," *IEEE Transactions on Computational Imaging*, vol. 2, no. 4, pp. 408–423, 2016.
- [29] A. M. Teodoro, J. M. Bioucas-Dias, and M. A. Figueiredo, "A convergent image fusion algorithm using scene-adapted gaussian-mixture-based denoising," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 451–463, 2018.
- [30] Y. Sun, B. Wohlberg, and U. S. Kamilov, "An online plug-and-play algorithm for regularized image reconstruction," *IEEE Transactions on Computational Imaging*, vol. 5, no. 3, pp. 395–408, 2019.
- [31] Y. Sun, Z. Wu, X. Xu, B. Wohlberg, and U. S. Kamilov, "Scalable plug-and-play admm with convergence guarantees," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 849–863, 2021.
- [32] E. T. Reehorst and P. Schniter, "Regularization by denoising: Clarifications and new interpretations," *IEEE Transactions on Computational Imaging*, vol. 5, no. 1, pp. 52–67, 2018.
- [33] P. Nair, R. G. Gavaskar, and K. N. Chaudhury, "Fixed-point and objective convergence of plug-and-play algorithms," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 337–348, 2021.
- [34] R. Cohen, M. Elad, and P. Milanfar, "Regularization by denoising via fixed-point projection (red-pro)," *SIAM Journal on Imaging Sciences*, vol. 14, no. 3, pp. 1374–1406, 2021.
- [35] R. G. Gavaskar, C. D. Athalye, and K. N. Chaudhury, "On plug-and-play regularization using linear denoisers," *IEEE Transactions on Image Processing*, vol. 30, pp. 4802–4813, 2021.
- [36] X. Xu, Y. Sun, J. Liu, B. Wohlberg, and U. S. Kamilov, "Provable convergence of plug-and-play priors with mmse denoisers," *IEEE Signal Processing Letters*, vol. 27, pp. 1280–1284, 2020.
- [37] M. Solodov, "An explicit descent method for bilevel convex optimization," *Journal of Convex Analysis*, vol. 14, no. 2, p. 227, 2007.
- [38] A. Beck and S. Sabach, "A first order method for finding minimal norm-like solutions of convex optimization problems," *Mathematical Programming*, vol. 147, no. 1–2, pp. 25–46, 2014.
- [39] S. Sabach and S. Shtern, "A first order method for solving convex bilevel optimization problems," *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 640–660, 2017.
- [40] A. L. Dontchev and R. T. Rockafellar, "Implicit functions and solution mappings," *Springer Monographs in Mathematics*. Springer, vol. 208, 2009.
- [41] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1, pp. 293–318, 1992.
- [42] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.
- [43] J. M. Bioucas-Dias and M. A. Figueiredo, "A new twist: two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2992–3004, 2007.
- [44] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Understanding and evaluating blind deconvolution algorithms," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1964–1971.
- [45] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.
- [46] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6849–6857.
- [47] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1256–1272, 2016.
- [48] A. Danielyan, V. Katkovnik, and K. Egiazarian, "Bm3d frames and variational image deblurring," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1715–1728, 2012.
- [49] C. J. Schuler, H. C. Burger, S. Harmeling, and B. Scholkopf, "A machine learning approach for non-blind image deconvolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1067–1074.
- [50] N. Yair and T. Michaeli, "Multi-scale weighted nuclear norm image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3165–3174.
- [51] J. Kruse, C. Rother, and U. Schmidt, "Learning to push the limits of efficient fft-based image deconvolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4586–4594.
- [52] W. Ren, J. Zhang, L. Ma, J. Pan, X. Cao, W. Zuo, W. Liu, and M.-H. Yang, "Deep non-blind deconvolution via generalized low-rank approximation," in *Advances in Neural Information Processing Systems*, 2018, pp. 297–307.
- [53] S. Li, B. Qin, J. Xiao, Q. Liu, Y. Wang, and D. Liang, "Multi-channel and multi-model-based autoencoding prior for grayscale image restoration," *IEEE Transactions on Image Processing*, vol. 29, pp. 142–156, 2019.
- [54] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing mri," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 72–82, 2008.
- [55] R. G. Baraniuk, "Compressive sensing [lecture notes]," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–121, 2007.
- [56] X. Qu, D. Guo, B. Ning, Y. Hou, Y. Lin, S. Cai, and Z. Chen, "Undersampled mri reconstruction with patch-based directional wavelets," *Magnetic Resonance Imaging*, vol. 30, no. 7, pp. 964–97, 2012.
- [57] X. Qu, Y. Hou, F. Lam, D. Guo, J. Zhong, and Z. Chen, "Magnetic resonance image reconstruction from undersampled measurements using a patch-based nonlocal operator," *Medical Image Analysis*, vol. 18, no. 6, pp. 843–856, 2014.
- [58] Z. Zhan, J.-F. Cai, D. Guo, Y. Liu, Z. Chen, and X. Qu, "Fast multiclass dictionaries learning with geometrical directions in mri reconstruction," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 9, pp. 1850–1861, 2016.
- [59] Y. Yang, J. Sun, H. Li, and Z. Xu, "Admm-csnet: A deep learning approach for image compressive sensing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
- [60] E. M. Eksioğlu, "Decoupled algorithm for mri reconstruction using nonlocal block matching model: Bm3d-mri," *Journal of Mathematical Imaging and Vision*, vol. 56, no. 3, pp. 430–440, 2016.

- [61] B. Cai, X. Xu, K. Guo, K. Jia, B. Hu, and D. Tao, "A joint intrinsic-extrinsic prior model for retinex," in *International Conference on Computer Vision*, 2017.
- [62] C. Li, J. Guo, F. Porikli, and Y. Pang, "Lightnet: a convolutional neural network for weakly illuminated image enhancement," *Pattern Recognition Letters*, vol. 104, pp. 15–22, 2018.
- [63] B. He, L.-Z. Liao, D. Han, and H. Yang, "A new inexact alternating directions method for monotone variational inequalities," *Mathematical Programming*, vol. 92, no. 1, pp. 103–118, 2002.
- [64] X. Yuan, S. Zeng, and J. Zhang, "Discerning the linear convergence of admm for structured convex optimization through the lens of variational analysis," *Journal of Machine Learning Research*, vol. 21, pp. 83–1, 2020.



Risheng Liu (M'12-) received the BSc and PhD degrees both in mathematics from the Dalian University of Technology in 2007 and 2012, respectively. He was a visiting scholar in the Robotic Institute of Carnegie Mellon University from 2010 to 2012. He served as Hong Kong Scholar Research Fellow at the Hong Kong Polytechnic University from 2016 to 2017. He is currently a professor with the International School of Information Science & Engineering, Dalian University of Technology. His research interests include machine learning, optimization, computer vision and multimedia.

He was a co-recipient of the IEEE ICME Best Student Paper Award in both 2014 and 2015. Two papers were also selected as Finalist of the Best Paper Award in ICME 2017. He is a member of the IEEE and ACM.



Long Ma received the M.S. degree in software engineering at Dalian University of Technology, Dalian, China, in 2019. He is currently pursuing the Ph. D. degree in software engineering at Dalian University of Technology, Dalian, China. His research interests include computer vision, image enhancement and machine learning. He is a reviewer for CVPR, ICCV, AAAI, ACCV, IEEE TCSVT, and Neurocomputing.



Xiaoming Yuan is Professor at Department of Mathematics, The University of Hong Kong. His main research interests include numerical optimization, scientific computing and optimal control. Recently, he is particularly interested in optimization problems in various AI and cloud computing areas.



Shangzhi Zeng received the B.Sc. degree in Mathematics and Applied Mathematics from Wuhan University in 2015, the M.Phil. degree from Hong Kong Baptist University in 2017, and the Ph.D. degree from the University of Hong Kong in 2021. He is currently a PIMS postdoctoral fellow in the Department of Mathematics and Statistics at University of Victoria. His current research interests include variational analysis and bilevel optimization.



Jin Zhang received the B.A. degree in Journalism from the Dalian University of Technology in 2007. He pursued a degree in mathematics and received the M.S. degree in Operational Research and Cybernetics from the Dalian University of Technology, China, in 2010, and the Ph. D. degree in Applied Mathematics from University of Victoria, Canada, in 2015. After working in Hong Kong Baptist University for 3 years, he joined Southern University of Science and Technology as a tenure-track assistant professor in the department of mathematics. His broad research

area is comprised of optimization, variational analysis and their applications in economics, engineering and data science.