Solving Inverse Problems With Deep Neural Networks – Robustness Included?

Martin Genzel, Jan Macdonald, and Maximilian März

Abstract—In the past five years, deep learning methods have become state-of-the-art in solving various inverse problems. Before such approaches can find application in safety-critical fields, a verification of their reliability appears mandatory. Recent works have pointed out instabilities of deep neural networks for several image reconstruction tasks. In analogy to adversarial attacks in classification, it was shown that slight distortions in the input domain may cause severe artifacts. The present article sheds new light on this concern, by conducting an extensive study of the robustness of deep-learning-based algorithms for solving underdetermined inverse problems. This covers compressed sensing with Gaussian measurements as well as image recovery from Fourier and Radon measurements, including a real-world scenario for magnetic resonance imaging (using the NYU-fastMRI dataset). Our main focus is on computing adversarial perturbations of the measurements that maximize the reconstruction error. A distinctive feature of our approach is the quantitative and qualitative comparison with total-variation minimization, which serves as a provably robust reference method. In contrast to previous findings, our results reveal that standard end-to-end network architectures are not only resilient against statistical noise, but also against adversarial perturbations. All considered networks are trained by common deep learning techniques, without sophisticated defense strategies.

Index Terms—Inverse problems, image reconstruction, deep neural networks, adversarial robustness, medical imaging

1 Introduction

Signal reconstruction from indirect measurements plays a central role in a variety of applications, including medical imaging [1], communication theory [2], astronomy [3], and geophysics [4]. Such tasks are typically formulated as an inverse problem, which in its prototypical, finite-dimensional form reads as follows:

$$\left\{ \begin{array}{c} \textit{Given a linear forward operator } \mathcal{A} \in \mathbb{R}^{m \times N} \\ \textit{and corrupted measurements } \boldsymbol{y} = \mathcal{A}\boldsymbol{x}_0 + \boldsymbol{e} \\ \textit{with } \|\boldsymbol{e}\|_2 \leq \eta, \textit{reconstruct the signal } \boldsymbol{x}_0. \end{array} \right\} \tag{1}$$

The ubiquitous presence of noise makes it indispensable that a reconstruction method has to be *robust* against additive perturbations $\boldsymbol{e}$. Furthermore, the measurement process is often costly and potentially harmful. Therefore, the underdetermined regime where $m \ll N$ has gained much attention during the last two decades. This restriction turns (1) into an *ill-posed inverse problem*, which does not possess a unique solution.

Under the additional assumption of sparsity, the methodology of *compressed sensing* has proven that accurate and robust reconstruction from incomplete measurements is still possible [5]. This means that a solution map $\mathrm{Rec} : \mathbb{R}^m \to \mathbb{R}^N$ for (1) satisfies an error bound of the form

$$\|\boldsymbol{x}_0 - \mathrm{Rec}(\boldsymbol{y})\|_2 \leq C \cdot \eta, \tag{2}$$

where $C > 0$ is a small constant. Although state-of-the-art in various real-world applications, the practicability of the associated algorithms is often limited by computational costs, manual parameter tuning, and a mismatch between sparsity models and data.

Building on the recent success of artificial intelligence in computer vision [6], [7], [8], there has been a considerable effort to solve the inverse problem (1) by means of *deep learning*, e.g., see [9], [10], [11], [12], [13], [14], [15], [16], [17], [18] and [19] for a recent survey. This advance is primarily based on fitting an artificial *neural network (NN)* model to a large set of data points in a supervised training procedure. It is fair to say that such data-driven approaches can significantly outperform classical methods in terms of reconstruction accuracy and speed. On the other hand, one may argue that the underlying mechanisms of NNs remain largely unclear [20]. Hence, in the absence of theoretical guarantees of the form (2), an empirical verification of their accuracy and robustness against measurement noise is crucial.

While a number of works report a remarkable resilience against noise [17], [21], [22], several alarming findings indicate that deep-learning-based reconstruction schemes are typically unstable [23], [24], [25], [26]. In particular, the recent study of Antun *et al.* [24] suggests that deep learning for inverse problems comes at the cost of instabilities, in the sense that *"[...] certain tiny, almost undetectable perturbations, both in the image and sampling domain, may result in severe artifacts in the reconstruction [...]"*. In machine learning research on classification, such a sensitivity of NNs is a well-established phenomenon. Initiated by Szegedy *et al.* [27], a

- *Martin Genzel is with Mathematical Institute, Utrecht University, 3584 CS Utrecht, The Netherlands. E-mail: m.genzel@uu.nl.*
- *Jan Macdonald and Maximilian März are with the Institute of Mathematics, Technical University of Berlin, 10623 Berlin, Germany. E-mail: {macdonald, maerz}@math.tu-berlin.de.*

Manuscript received 9 Nov. 2020; revised 3 Nov. 2021; accepted 26 Jan. 2022. Date of publication 4 Feb. 2022; date of current version 5 Dec. 2022.
Martin Genzel and Maximilian März acknowledge support by DFG Priority Programme under Grant DFG-SPP 1798.
(Corresponding author: Maximilian März.)
Recommended for acceptance by T. Pock.
Digital Object Identifier no. 10.1109/TPAMI.2022.3148324

substantial body of literature is devoted to *adversarial attacks* (and their defenses), i.e., the computation of a visually imperceptible change to the input that fools the NN. Typically, an "attacker" exploits gradient-based information in order to cross the discontinuous decision boundary of a classifier. This can be a serious issue for sensitive applications where wrong predictions impose a security risk—imagine a misclassified stop sign in autonomous driving [28], [29].

Despite these findings, it appears peculiar that solving inverse problems by deep-learning-based schemes might become unstable. Learning a reconstruction algorithm can be seen as a regression task, where measurements are mapped to a high-dimensional signal manifold (e.g., medical images). In contrast, a NN classifier maps to a low-dimensional, discrete output domain, resulting in a "vulnerable" decision boundary. Moreover, it is well known that robust and accurate algorithms exist for many inverse problems. Since these are often used as templates for NN architectures, it seems surprising that the latter should suffer from severe instabilities. Clearly, the robustness against noise is quintessential for an application of deep learning in practice, especially in sensitive fields such as biomedical imaging. Therefore, we believe that a profound study of this topic is indispensable.

1.1 Contributions

This article is dedicated to a comprehensive numerical study of the robustness of NN-based methods for solving underdetermined inverse problems. The primary objective of our experiments is to analyze how much the reconstruction error grows with the noise level η. We investigate this relationship in terms of statistical and adversarial noise: the former means that measurement noise is drawn from an appropriate probability distribution, while the latter explores worst-case perturbations that maximize the reconstruction error for fixed η. Similar to adversarial attacks in classification, computing worst-case noise is based on a non-convex formulation that is addressed by automatic differentiation and a gradient descent scheme. In the absence of an empirical certificate of robustness, a central and distinctive component of our analysis is the systematic comparison with a classical benchmark method with provable guarantees, namely total-variation (TV) minimization. In this case, evaluating the gradient is non-trivial and carried out by unrolling the underlying optimization problem.

Our experiments consider several prototypical inverse problems as use cases. This includes classical compressed sensing with Gaussian measurements as well as the reconstruction of phantom images from Radon and Fourier measurements. Furthermore, a real-world scenario for magnetic resonance imaging (MRI) is investigated, based on the NYU-fastMRI dataset [30], [31]. We examine a representative selection of learned reconstruction architectures, reaching from simple post-processing NNs to iterative schemes. In total, this work presents a robustness analysis of more than 25 NNs, each of them trained in-house with publicly available code.[1]

Our main findings may be summarized as follows:

(i) In every considered scenario, we find deep-learning-based methods that are at least as robust as TV minimization with respect to adversarial noise. This does not require sophisticated architectures or defense strategies. However, none of the trained NNs are as accurate as TV minimization for gradient-sparse signals.

(ii) All trained NNs are remarkably robust against statistical noise. Although TV minimization may yield exact recovery for noiseless measurements, it is still outperformed by learned methods in mid- to high-noise regimes.

(iii) The reconstruction performance is affected by the underlying NN architecture. For instance, promoting data consistency in iterative schemes may improve both accuracy and robustness.

(iv) One should not commit the "inverse crime" of training a NN with *noiseless* data, which may cause an unstable behavior for higher noise levels. We demonstrate that simply adding white Gaussian noise to the training measurements is an effective remedy—a regularization technique that is commonly known as *jittering* in machine learning research. This adaption has a virtually imperceptible impact on the in-distribution accuracy, but might affect out-of-distribution features (see Section 5.2). This leads to interesting trade-offs between stability and accuracy (cf. Fig. 14 bottom right panel).

Apart from these observations, our work is, to the best of our knowledge, the first to empirically characterize the performance gap between adversarial and statistical noise in the context of (1). In particular, this gap is not exclusive to deep-learning-based schemes but also appears for classical methods such as TV minimization. Our central conclusion is:

> The existence of adversarial examples in classification tasks does not always carry over to NN-based solvers for inverse problems. Such reconstruction schemes may achieve state-of-the-art accuracy and can also, in certain cases, exhibit a similar degree of robustness as classical methods. Moreover, there is an observed trade-off between stability and accuracy for both NN-based and classical methods.

1.2 Scope and Implications

The goal of our study is to show that robust solutions to ill-posed inverse problems can be obtained with data-driven methods, i.e., small perturbations of the measurements do not lead to large reconstruction errors. Clearly, the extent to what this is possible depends on the forward model and the data distribution of the underlying inverse problem. In particular, deep learning methods cannot be expected to overcome fundamental theoretical limitations. As formalized in [25], there are situations in which accuracy and stability become mutually exclusive. If the inverse problem setup is too ill-posed, i.e., well-separated signals are mapped to almost identical measurements, then any accurate method (whether learned or non-learned) must have a large local Lipschitz constant and is therefore unstable. Although this is an interesting avenue for future research, the purpose of our work is not to balance out such a trade-off.

1. Our Python implementation, based on the *PyTorch* package [32], can be found under https://github.com/jmaces/robust-nets

Since our study as it is has required massive computational resources (>2 years of GPU computation time), some other aspects have to remain unexplored, see Section 6 for a discussion. In particular, given the sheer number of NN architectures, we explicitly do *not* claim that every deep-learning-based method is stable (cf. Section 5.1). Nevertheless, our findings suggest that fairly standard workflows allow for surprisingly robust reconstruction schemes. This offers an alternative and novel perspective on the reliability of deep learning strategies in inverse problems. Therefore, we believe that the present work takes an important step towards their safe use in practice.

1.3 Organization of This Article

Section 2 is devoted to relevant previous works, followed by a conceptual overview of our approach in Section 3. The latter introduces all considered reconstruction methods, the associated NN architectures as well as our attack strategy to analyze their adversarial robustness. The main results are then presented in Section 4, complemented by several additional experiments in Section 5. We conclude with a general discussion of our findings in Section 6.

2 Related Work

Initiated by Szegedy *et al.* [27], the vulnerability of deep NNs to adversarial examples has been the subject of more than 2500 publications [33]. We refer to [34], [35] for recent surveys of the field and further references. The vast majority of existing articles is concerned with classification and related tasks, such as image segmentation [36]. On the other hand, only few works have explicitly addressed the adversarial robustness of learned solvers for inverse problems.

To the best of our knowledge, Huang *et al.* [23] have made the first effort to transfer adversarial attacks to NN-based reconstruction methods. They demonstrate that a distortion of the network's input may result in the loss of small image features. However, their initial findings are restricted to the specific problem of limited angle computed tomography, where the robust recovery of certain parts of the image is provably impossible [37]. Moreover, the proposed perturbation model is non-standard and does not correspond to noise in the measurements.

More recently, the topic was brought to attention by the inspiring article of Antun *et al.* [24]. Their numerical experiments show instabilities of existing deep NNs with respect to adversarial noise, out-of-distribution features, and changes in the number of measurements. An important difference to our work is that adversarial noise is only computed for learned schemes. We believe that a comparative "attack" of a classical benchmark method is crucial for a fair assessment of robustness. Furthermore, the results of [24] are reported qualitatively by visualizing reconstructed images, as it is common in adversarial machine learning. We argue that the mathematical setup of the inverse problem (1) calls for a quantitative error analysis that is in line with the bound of (2). Finally, the training stage of the networks in [24] does not seem to account for noise, which we have identified as a potential source of instability, see Section 5.1. Note that our study also analyzes the FBPConvNet architecture [13], a relative of AUTOMAP [17], and an iterative scheme similar to DeepMRI [38]. Nevertheless, a one-to-one comparison to [24] is subtle due to task-specific architectures and data processing. A follow-up work of [24] presents a theoretical characterization of instabilities in terms of the kernel of the forward operator [25]. Our results provide empirical evidence that the considered deep-learning-based schemes could be kernel aware (cf. Section 5.4).

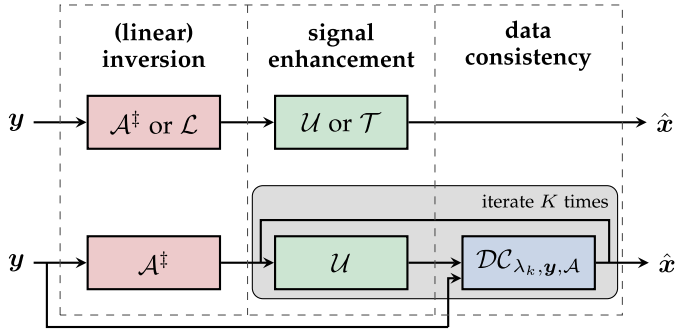

Fig. 1. Schematic network reconstruction pipelines of UNet, TiraFL (top), and ItNet (bottom).

As a countermeasure to the outcome of [24], Raj *et al.* [26] suggest a sophisticated defense strategy resulting in robust networks. This work also addresses shortcomings of the attack strategy in [24], see Section 3.4 for details. In line with our findings, Kobler *et al.* [39] propose the data-driven *total deep variation* regularizer and demonstrate its adversarial robustness for image denoising.

Finally, in another line of research, [40] conducts a theoretical error analysis of a family of mappings (referred to as *RegNets*) that post-process classical regularization methods. Under the assumption of Lipschitz-continuous networks, convergence rates in the spirit of (2) are derived for the limit $\eta \to 0$. However, due to their asymptotic nature, such results do not directly address the adversarial perturbation scenarios of the present work, where a whole range of noise levels η is analyzed. Apart from that, the convergence rates in [40] linearly depend on the global Lipschitz constants, which are hard to compute and control in practice [27], [41], [42]. Our simulations reveal that *pointwise* Lipschitz constants for common reconstruction NNs are well-behaved, regardless of possibly large global constants.

3 Methods and Preliminaries

In this section, we briefly introduce the considered reconstruction schemes for solving the inverse problem (1). This includes a representative selection of NN-based methods and total-variation minimization as a classical benchmark. Furthermore, our attack strategy to analyze their adversarial robustness is presented.

3.1 Neural Network Architectures

In the past five years, numerous deep-learning-based approaches for solving inverse problems have been developed; see [19], [43] for overviews. The present work focuses on a selection of widely used *end-to-end network schemes* that define an explicit reconstruction map from $\mathbb{R}^m$ to $\mathbb{R}^N$, see also Fig. 1.

The first considered method is a *post-processing network*

$$\mathrm{UNet} : \mathbb{R}^m \to \mathbb{R}^N,\ \boldsymbol{y} \mapsto [\mathcal{U} \circ \mathcal{A}^\dagger](\boldsymbol{y}).$$

It employs the U-Net architecture $\mathcal{U} : \mathbb{R}^N \to \mathbb{R}^N$ [44] as a residual network [45] to enhance an initial, model-based reconstruction $\mathcal{A}^\dagger(\boldsymbol{y})$. Here, $\mathcal{A}^\dagger : \mathbb{R}^m \to \mathbb{R}^N$ is an approximate inversion of the forward operator $\mathcal{A}$, e.g., the filtered back-projection for Radon measurements. Despite its simplicity, it has been demonstrated in [13] that UNet is an effective solution method for (1); see also [12], [15], [21], [46], [47] for related approaches.

Our second reconstruction scheme is a *fully-learned network*

$$\mathrm{TiraFL} : \mathbb{R}^m \to \mathbb{R}^N,\ \boldsymbol{y} \mapsto [\mathcal{T} \circ \mathcal{L}](\boldsymbol{y}),$$

which is closely related to UNet, but differs in two aspects: It is based on the Tiramisu architecture $\mathcal{T} : \mathbb{R}^N \to \mathbb{R}^N$ [48] as a residual network, which can be seen as a refinement of the U-Net. While $\mathcal{T}$ shares the same multi-level structure, it is built from fully-convolutional dense-blocks [49] instead of standard convolutional blocks. More importantly, the fixed inversion $\mathcal{A}^\dagger$ is replaced by a learnable linear layer $\mathcal{L} \in \mathbb{R}^{N \times m}$, so that TiraFL does not contain fixed model-based components anymore. The approach of TiraFL is similar to [17], [50], which makes use of a fully-learned reconstruction map for MRI. For the sake of completeness, we have also conducted experiments for Tira, a Tiramisu-based post-processing network, as well as for UNetFL, a U-Net-based fully-learned network, see Sections S1– S3 in the supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2022.3148324, for results.

Finally, we also analyze an *iterative network*

$$\mathrm{ItNet} : \mathbb{R}^m \to \mathbb{R}^N,\ \boldsymbol{y} \mapsto \left[\left(\bigcirc_{k=1}^K [\mathcal{DC}_{\lambda_k,\boldsymbol{y},\mathcal{A}} \circ \mathcal{U}]\right) \circ \mathcal{A}^\dagger\right](\boldsymbol{y}),$$

where

$$\mathcal{DC}_{\lambda_k,\boldsymbol{y},\mathcal{A}} : \mathbb{R}^N \to \mathbb{R}^N,\ \boldsymbol{x} \mapsto \boldsymbol{x} - \lambda_k \cdot \mathcal{A}^*(\mathcal{A}\boldsymbol{x} - \boldsymbol{y}).$$

The scalar parameters λ_k are learnable and $\mathcal{A}^*$ denotes the adjoint of $\mathcal{A}$. Mathematically, $\mathcal{DC}_{\lambda_k,\boldsymbol{y},\mathcal{A}}$ performs a gradient step on the loss $\boldsymbol{x} \mapsto \frac{\lambda_k}{2}\|\mathcal{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2$, promoting *data consistent* solutions. Therefore, the alternating cascade of ItNet can be seen as a proximal gradient descent scheme, where the proximal operator is replaced by a trainable enhancement network. Here, the U-Net architecture is used again, due to its omnipresence in image-to-image processing tasks. Unrolled methods in the spirit of ItNet are frequently used to solve inverse problems, e.g., see [9], [10], [14], [38], [51], [52], [53], [54].

3.2 Neural Network Training

The learnable parameters of the networks are trained from sample data pairs $\{(\boldsymbol{y}^i = \mathcal{A}\boldsymbol{x}_0^i + \boldsymbol{e}^i, \boldsymbol{x}_0^i)\}_{i=1}^M$ by minimizing an empirical loss function. Depending on the use case, the signals $\boldsymbol{x}_0^i$ are either drawn from a fixed publicly available training dataset or according to a synthetic probability distribution. If $\mathrm{Net}[\boldsymbol{\theta}] : \mathbb{R}^m \to \mathbb{R}^N$ denotes a reconstruction network with all learnable parameters collected in $\boldsymbol{\theta}$, then the training amounts to (approximately) solving

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^M \ell\big(\mathrm{Net}[\boldsymbol{\theta}](\boldsymbol{y}^i), \boldsymbol{x}_0^i\big) + \mu \cdot \|\boldsymbol{\theta}\|_2^2, \tag{3}$$

for some cost function $\ell : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}_{\geq 0}$, which is the squared distance unless stated otherwise. Overfitting is addressed by ℓ_2-regularization with a hyper-parameter $\mu \geq 0$.[2] In order to solve (3), we utilize mini-batch stochastic gradient descent and the Adam optimizer [55]. We found that larger mini-batches were beneficial for the training performance during later epochs. Technically, this is achieved by gradient accumulation, i.e., the gradient is cumulatively summed over several mini-batches before executing a descent step. For each network, the hyper-parameters were selected manually until satisfactory precision was achieved, see Tables S9 and S10, available online.

Due to the ubiquitous presence of noise in inverse problems, it is natural to account for it in the training data. In many applications, measurement noise is modeled as an independent random variable, for instance, following a Gaussian distribution. Therefore, the perturbation $\boldsymbol{e}^i$ is treated as statistical noise during the training phase, i.e., a fresh realization is randomly drawn in each epoch. This technique is well known as *jittering* in machine learning research, where it is primarily used to avoid overfitting [56], [57], [58]; see also [59]. In Section 5.1, we relate jittering to the phenomenon of inverse crimes and demonstrate its importance for the robustness of learned reconstruction schemes. Due to varying noise levels in the evaluation of our models, we design $\boldsymbol{e}^i$ as a centered Gaussian vector with random variance, such that its expected norm $\mathbb{E}[\|\boldsymbol{e}^i\|_2]$ is distributed uniformly in a range $[0, \widetilde{\eta}]$ for a fixed $\widetilde{\eta} \geq 0$. This means that a particular NN is trained via (3) to handle multiple levels of (adversarial and statistical) noise at same time.

3.3 Total-Variation Minimization

Dating back to the seminal work of Rudin *et al.* [60], *total-variation (TV) minimization* has become a standard tool for solving signal and image reconstruction tasks [61], [62]. We apply it to the problem (1) in the following form:

$$\mathrm{TV}[\eta] : \mathbb{R}^m \to \mathbb{R}^N,\quad \boldsymbol{y} \mapsto \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathbb{R}^N} \|\nabla \boldsymbol{x}\|_1 \quad \text{s.t.} \quad \|\mathcal{A}\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \eta, \tag{4}$$

where ∇ denotes a discrete gradient operator. Crucial to the above optimization problem is the use of the ℓ_1-norm, which is known to promote gradient-sparse solutions. Indeed, under suitable assumptions on $\mathcal{A}$, compressed sensing theory suggests an error bound of the form (2) for a gradient-sparse signal $\boldsymbol{x}_0$ and $\mathrm{Rec} = \mathrm{TV}[\eta]$, e.g., see [63], [64], [65], [66]. In other words, TV minimization is provably robust with a near-optimal dependence on η. This particularly justifies its use as a reference method, allowing us to empirically characterize the robustness of learned reconstruction schemes.

2. This is often referred to as *weight decay* in deep learning, since the ℓ_2-term corresponds to a shrinkage of the weights $\boldsymbol{\theta}$ by a constant factor when performing the gradient update, e.g., see [8, Section 7.1.1].

In our numerical simulations, the problem of (4) is solved by the *alternating direction method of multipliers (ADMM)* [67], [68]. For 1D signals, $\nabla \in \mathbb{R}^{N\times N}$ is chosen as a forward finite difference operator with Neumann boundary conditions, extended by a constant row vector to capture the mean of the signal. For image signals, $\nabla \in \mathbb{R}^{2N\times N}$ corresponds to a forward finite difference operator with periodic boundary conditions. Due to the non-separability of $\|\nabla(\cdot)\|_1$ in 2D, the formulation of $\mathrm{TV}[\eta]$ in (4) becomes computationally infeasible for finding adversarial noise. In imaging scenarios, we therefore solve the unconstrained version of $\mathrm{TV}[\eta]$ instead, i.e., the objective function is changed to $\boldsymbol{x} \mapsto \lambda \cdot \|\nabla \boldsymbol{x}\|_1 + \|\mathcal{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2$. Note that this strategy is theoretically equivalent [5 Appx. B] , but requires an appropriate choice of the regularization parameter $\lambda > 0$. A near-optimal selection with respect to the relative ℓ_2-error is determined by grid searches over the test set and a densely sampled range of noise levels η. In contrast to the NNs, $\mathrm{TV}[\eta]$ is explicitly adapted to the amount of perturbation of the measurements.

3.4 Adversarial Noise

In the setup of (1), adversarial noise for a given reconstruction method $\mathrm{Rec} : \mathbb{R}^m \to \mathbb{R}^N$ can be computed by solving an optimization problem: for a fixed signal $\boldsymbol{x}_0 \in \mathbb{R}^N$ and noise level $\eta \geq 0$, find an additive perturbation $\boldsymbol{e}_{\mathrm{adv}} \in \mathbb{R}^m$ of the noiseless measurements $\boldsymbol{y}_0 = \mathcal{A}\boldsymbol{x}_0$ that maximizes the reconstruction error, i.e.,

$$\boldsymbol{e}_{\mathrm{adv}} = \operatorname*{arg\,max}_{\boldsymbol{e}\in\mathbb{R}^m} \|\mathrm{Rec}(\boldsymbol{y}_0 + \boldsymbol{e}) - \boldsymbol{x}_0\|_2 \quad \text{s.t.} \quad \|\boldsymbol{e}\|_2 \leq \eta. \quad (5)$$

Such an attack strategy is a straightforward adaption of a common approach in adversarial machine learning [34]. In contrast to [24], we consider a constrained optimization problem that avoids shortcomings of an unconstrained formulation; in particular, this allows for precise control over the noise level. Moreover, (5) explores a natural perturbation model, operating directly in the measurement domain, cf. the discussion in [26].

In order to solve the problem (5), we use the projected gradient descent algorithm in conjunction with the Adam optimizer, which was found to be most effective (cf. [69]). The non-convexity of (5) is accounted for by choosing the worst perturbation out of multiple runs with random initialization. Assuming a whitebox model (i.e., Rec is fully accessible), we use PyTorch's automatic differentiation [32] to compute gradients of the considered NN schemes.

A central aspect of our work is that the above perturbation strategy is also applied to $\mathrm{TV}[\eta]$. This is non-trivial, since the gradient of the implicit map $\boldsymbol{y} \mapsto \mathrm{TV}[\eta](\boldsymbol{y})$ has to be computed. The large-scale nature of imaging problems prevents us from using the recent concept of differentiable convex optimization layers [70]. Instead, we calculate the gradient of the unrolled ADMM scheme for TV minimization by automatic differentiation. In general, a large number of iterations might be required for the convergence of ADMM, which in turn ensures an accurate gradient approximation for $\mathrm{TV}[\eta]$. This leads to numerical difficulties in automatic differentiation, due to memory & time constraints and error accumulation. We address this issue by decreasing the number of ADMM iterations for the gradient computation (denoted by k_{grad}). To compensate for a loss of accuracy, we use a pre-initialization of the primal and dual variables by the output of a fully converged ADMM scheme with input $\boldsymbol{y}_0$. Such a warm start is beneficial, since the gradient is only evaluated in an η-ball around $\boldsymbol{y}_0$ during the attack. Note that the actual TV reconstructions are always computed by a fully converged ADMM algorithm (with $k_{\mathrm{rec}} \gg k_{\mathrm{grad}}$ iterations).

4 Main Results

This section studies the robustness of NN-based solution methods for three different instances of the inverse problem (1). The goal of our experiments is to assess the loss of reconstruction accuracy caused by noise. To that end, we rely on two types of visualization:

- *Noise-to-error curves* are generated by plotting the relative noise level $\eta/\|\mathcal{A}\boldsymbol{x}_0\|_2$ against the relative reconstruction error $\|\boldsymbol{x}_0 - \mathrm{Rec}(\mathcal{A}\boldsymbol{x}_0 + \boldsymbol{e})\|_2/\|\boldsymbol{x}_0\|_2$.
- *Individual reconstruction results* are shown for different relative noise levels and a randomly selected signal from the test set.

In both cases, the perturbation vector $\boldsymbol{e}$ is either of *statistical* or *adversarial* type. The former means that $\boldsymbol{e}$ is a random vector such that $\mathbb{E}[\|\boldsymbol{e}\|_2^2] = \eta^2$, whereas the latter is found by (5). While noise-to-error curves are of quantitative nature, individual reconstructions facilitate a qualitative judgment of robustness. Note that the sensitivity to noise is different in each considered scenario. Therefore, we have selected the maximal level of adversarial noise such that the benchmark of TV minimization does not yield a (subjectively) acceptable performance anymore. A specification of all empirically selected hyper-parameters can be found in the supplementary material (see Tables S9–S11), available online.

4.1 Case Study A: Compressed Sensing With Gaussian Measurements

Our first study is devoted to sparse recovery of 1D signals from Gaussian measurements, which is a standard benchmark setup in the field of compressed sensing (CS) theory [5]. This means that the entries of the forward operator $\mathcal{A}$ in (1) are independent Gaussian random variables with zero mean and variance $1/m$. We consider two different scenarios based on (approximately) gradient-sparse signals; note that such a model is canonical for TV minimization and compatible with the local connectivity of our convolutional NN schemes.

Scenario A1. We draw $\boldsymbol{x}_0$ from a synthetic distribution of *piecewise constant signals* with zero boundaries and well-controlled random jumps, see Fig. 3 for an example. In this scenario, we choose $m = 100$, $N = 256$, and use $M = 200\mathrm{k}$ training samples.

Scenario A2. We sample $\boldsymbol{x}_0 \in [0,1]^{28\times 28}$ from the widely used *MNIST database* [71] with $M = 60\mathrm{k}$ training images of handwritten digits. In the context of (1), the images are treated as 1D signals[3] of dimension $N = 28^2 = 784$. The number of Gaussian measurements is $m = 300$.

3. We have decided for a vectorized data processing (i.e., $\mathrm{TV}[\eta]$ and the NNs operate on vectorized images), since Scenario A2 is regarded as a direct continuation of the idealistic situation in A1. However, for visual purposes, all reconstructions are displayed as images, see Fig. 5.

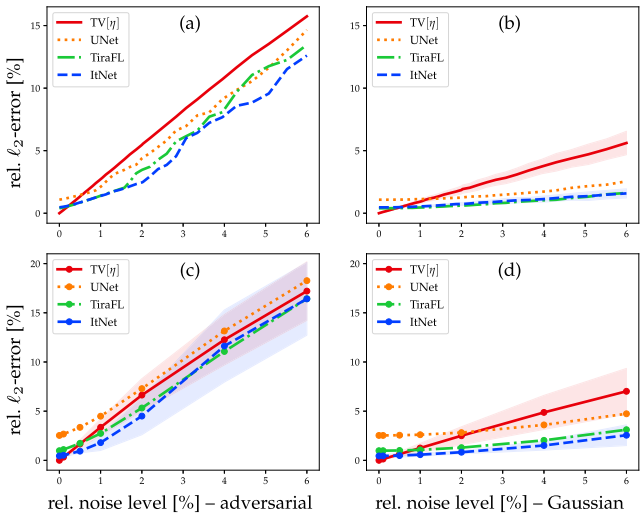

Fig. 2. *Scenario A1 – CS with 1D signals.* (a) shows the adversarial noise-to-error curve for the randomly selected signal of Fig. 3. (b) shows the corresponding Gaussian noise-to-error curve, where the mean and standard deviation are computed over 200 draws of e. (c) and (d) display the respective curves averaged over 50 signals from the test set. For the sake of clarity, we have omitted the standard deviations for UNet and TiraFL, which behave similarly.

In both scenarios, we chose the model-based, linear inversion layer of the networks as a generalized Tikhonov matrix, i.e., $\mathcal{A}^{\ddagger} = (\mathcal{A}^T\mathcal{A} + \alpha \cdot \nabla^T\nabla)^{-1}\mathcal{A}^T \in \mathbb{R}^{N\times m}$ with the empirically chosen regularization parameter $\alpha = 0.02$. We were not able to train the NNs to a comparable reconstruction accuracy with other natural choices, such as $\mathcal{A}^{\ddagger} = \mathcal{A}^T$. The above matrix is also used to initialize the inversion layer $\mathcal{L} \in \mathbb{R}^{N\times m}$ of the fully-learned schemes.

Fig. 2 shows the noise-to-error curves for *Scenario A1 (CS with 1D signals)*; see also Tables S1 and S2, available online. The associated individual reconstructions for adversarial noise are displayed in Fig. 3; see Fig. S2, available online, for the corresponding results with Gaussian noise. Fig. S1 supplements, available online, the simulation of Figs. 2b and 2d by two additional types of random noise, drawn from the uniform and Bernoulli distribution. Both exhibit results that are virtually indistinguishable from the Gaussian case. Fig. 4 shows the noise-to-error curves for *Scenario A2 (CS with MNIST)*; see also Tables S3 and S4, available online. The associated individual reconstructions for adversarial noise are displayed in Fig. 5; see Fig. S3, available online, for two additional digits and Fig. S4, available online, for the corresponding results with Gaussian noise.

Conclusions. The above results confirm that the considered NN-based schemes are as least as robust to adversarial perturbations as the benchmark of TV minimization. Although TV[η] is perfectly tuned to each noise level η, it is clearly outperformed in the case of statistical noise. The gap between statistical and adversarial perturbations is comparable for all methods.

TV minimization is a perfect match for Scenario A1. In particular, exact recovery from noiseless measurements is guaranteed by CS theory [66], [72]. Although this cannot be

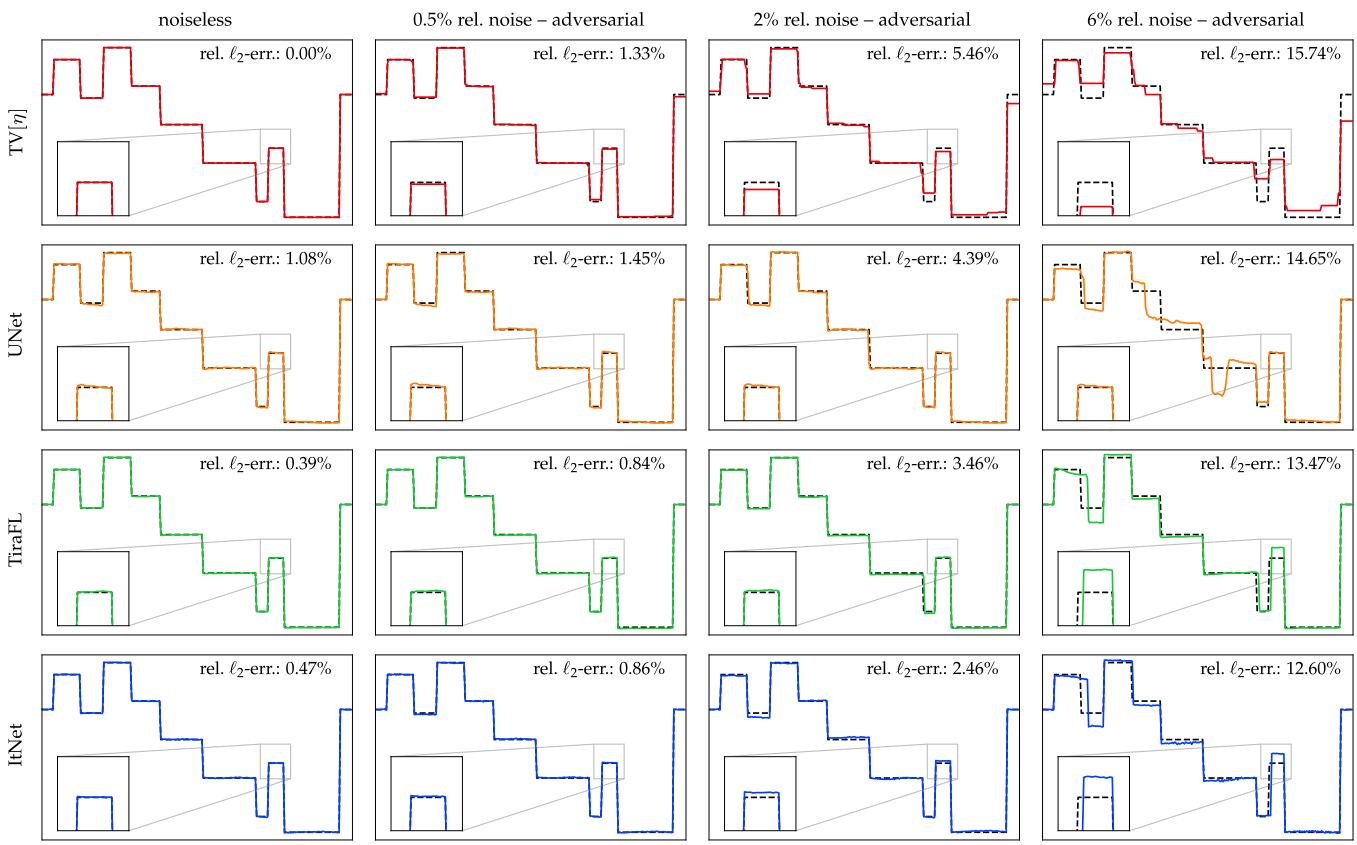

Fig. 3. *Scenario A1 – CS with 1D signals.* Individual reconstructions of a randomly selected signal from the test set for different levels of adversarial noise. The ground truth signal is visualized by a dashed line.

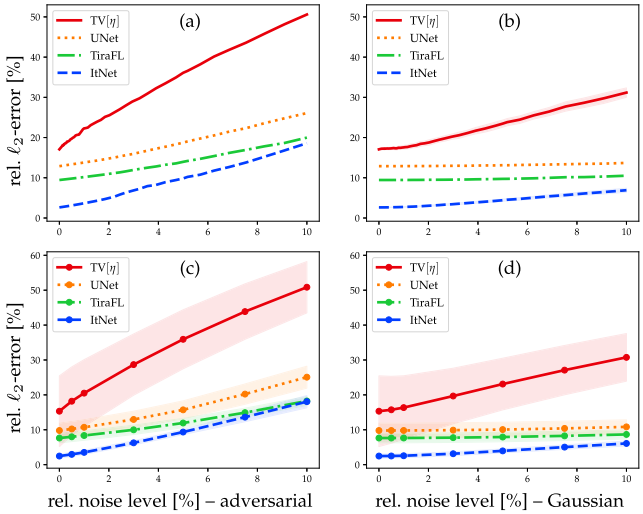

Fig. 4. *Scenario A2 – CS with MNIST.* (a) shows the adversarial noise-to-error curve for the randomly selected digit 3 of Fig. 5. (b) shows the corresponding Gaussian noise-to-error curve, where the mean and standard deviation are computed over 200 draws of e. (c) and (d) display the respective curves averaged over 50 signals from the test set.

expected for NN-based solvers, they still come with an overall superior robustness against noise. The situation is even more striking in Scenario A2. Here, TV minimization performs worse, since the signals are only approximately gradient-sparse. In contrast, the NN-based reconstruction schemes adapt well to the simple MNIST database, leading to significantly better outcomes in every regard. Hence, the increase in accuracy by learned methods does not necessarily imply a loss of robustness.

The performance ranking of the considered deep NNs is as one might expect: First, data consistency as encouraged by the ItNet-architecture is beneficial. Furthermore, Tables S1–S4, available online, reveal that the Tiramisu architecture is superior to a simple U-Net, and that a learnable inversion layer improves the recovery. The latter observation is not surprising, since Thikonov regularization is known to work poorly in conjunction with subsampled Gaussian measurements.

4.2 Case Study B: Image Recovery of Phantom Ellipses

Our second set of experiments concerns the recovery of phantom ellipses from Fourier or Radon measurements. These tasks correspond to popular simulation studies for biomedical imaging, e.g., see [13], [18], [52], [73]. We sample $\boldsymbol{x}_0 \in [0,1]^{256\times 256}$ from a distribution of superimposed random ellipses with mild linear intensity gradients and well-controlled geometric properties, see Fig. 7 for an example. The training is performed on $M = 25\text{k}$ images. We consider the following two measurement scenarios for (1), associated with the problems of *compressed sensing MRI* [1] and *low-dose computed tomography (CT)* [13], [74], respectively:

Scenario B1. The forward operator takes the form $\mathcal{A} = P\mathcal{F} \in \mathbb{C}^{m\times N}$, where $\mathcal{F} \in \mathbb{C}^{N\times N}$ is the *2D discrete Fourier transform* and $P \in \{0,1\}^{m\times N}$ is a subsampling operator defined by a golden-angle radial mask with 40 lines ($m = 10941$ and $N = 256^2 = 65536$). Note that the entire data processing is complex-valued, while the actual reconstructions are computed as real-valued magnitude images, as common in MRI. We use the canonical inversion layer $\mathcal{A}^\ddagger = \mathcal{A}^* = \mathcal{F}^{-1}P^* \in \mathbb{C}^{N\times m}$.

Scenario B2. The forward operator $\mathcal{A} \in \mathbb{R}^{m\times N}$ is given by a sparse-angle *Radon transform* with 60 views ($m = 21780$ and $N = 65536$). The non-linear inversion layer $\mathcal{A}^\ddagger : \mathbb{R}^m \to \mathbb{R}^N$ is chosen as the filtered back-projection algorithm (FBP) with a Hann filter.

In contrast to Case Study A, the aforementioned problems are of significantly higher dimensionality. Therefore, fully-learned schemes are difficult to realize, since the size of the inversion layer scales multiplicatively in the image dimensions. In the Fourier case, the number of free parameters can be reduced by enforcing a Kronecker product structure on $\mathcal{L} \in \mathbb{C}^{N\times m}$; this exploits the fact that $\mathcal{F}$ is a tensor product of two 1D Fourier transforms, cf. [50].

Fig. 6 shows the noise-to-error curves for *Scenario B1 (Fourier meas. with ellipses)*; see also Tables S5 and S6, available online. The associated individual reconstructions for $\text{TV}[\eta]$ and ItNet with adversarial noise are displayed in Fig. 7; see Fig. S5, available online, for the remaining networks and Fig. S6, available online, for the corresponding results with Gaussian noise. In the tables and individual reconstructions, we have also reported the *peak signal-to-noise ratio (PSNR)* and *structural similarity index measure (SSIM)* [75]. In the case of *Scenario B2 (Radon meas. with ellipses)*, we only present individual reconstructions based on $\text{TV}[\eta]$ and UNet; see Fig. 8 for adversarial noise and Fig. S7, available online, for the common Poisson noise model. This restriction is due to the more complicated nature of the Radon transform, and in particular, the need for automatic differentiation. The used implementation [76] requires significantly more computational effort, compared to the fast Fourier transform.

Conclusions. The main findings of Case Study A remain valid: (i) the adversarial robustness of NN-based methods and TV minimization is similar with respect to the ℓ_2-error; (ii) NNs are more resilient against statistical perturbations in mid- to high-noise regimes (see also the individual reconstructions in Figs. S6 and S7, available online); (iii) there is a clear gap between adversarial and statistical noise that is comparable for model-based and learned schemes.

The individual reconstruction results in Figs. 7 and 8 allow for further insights. First, the effect of adversarial noise for $\text{TV}[\eta]$ manifests itself in the well-known staircasing phenomenon, a considerable loss of resolution as well as point-like artifacts (see the zoomed region in Fig. 7). In contrast, NN-based methods always produce sharp images, with almost imperceptible visual errors up to 3% relative noise in the case of Fourier measurements (1% noise in the case of Radon measurements). For the highest noise level, on the other hand, they exhibit unnatural ellipsoidal artifacts.

At first sight, this observation might indicate a vulnerability to adversarial noise. However, a simple *transferability test* refutes this conclusion (cf. [77]): plugging the perturbed measurements for ItNet into $\text{TV}[\eta]$ leads to the same ellipsoidal artifacts; see Figs. 7 and S8, available online. Furthermore, Fig. 8 reveals that the corresponding artifacts are already present in the FBP inversion and are not caused by the post-processing network. This shows that the learned solvers do not suffer from undesired instabilities, but the observed artifacts are due to actual features in the corrupted measurements. Interestingly, adversarial perturbations

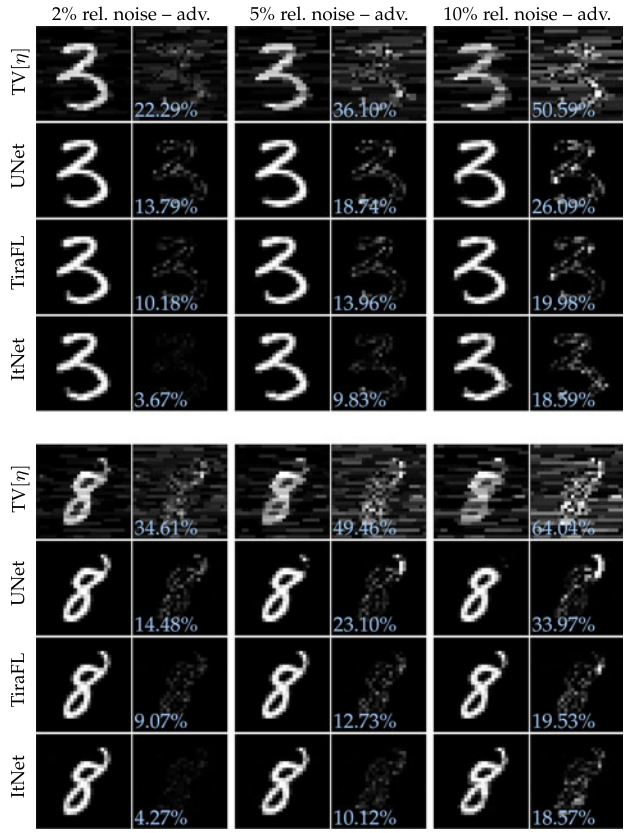

Fig. 5. *Scenario A2 – CS with MNIST.* Individual reconstructions of two randomly selected digits from the test set for different levels of adversarial noise. The reconstructed digits and their error plots (with relative ℓ_2-error) are displayed in the windows [0,1] and $[0, 0.6]$, respectively. The horizontal line artifacts in the $\mathrm{TV}[\eta]$-solutions are due to the fact that the MNIST images are treated as vectorized 1D signals. Remarkably, although relying on 1D convolutional filters, the NN-based reconstructions do not suffer from these artifacts.

found for $\mathrm{TV}[\eta]$ do not transfer to NN-based methods, see Fig. S8, available online. Overall, the attack strategy of (5) has different qualitative effects on each reconstruction paradigm: while known flaws of TV minimization are amplified, the NNs are perturbed by adding "real" ellipsoidal features to the measurements.

On a final note, we confirm the ranking of architectures as pointed out in Case Study A. Nevertheless, there is no clear superiority of the fully learned schemes as in case of Gaussian measurements, since the inverse Fourier transform appears to be a near-optimal choice of model-based inversion layer.

4.3 Case Study C: MRI on Real-World Data (fastMRI)

The third case study of this article is devoted to a real-world MRI scenario. To this end, we use the publicly available *fastMRI* knee dataset, which consists of 1594 multi-coil diagnostic knee MRI scans.[4] Our experiments are based on the subset of 796 coronal proton-density weighted scans

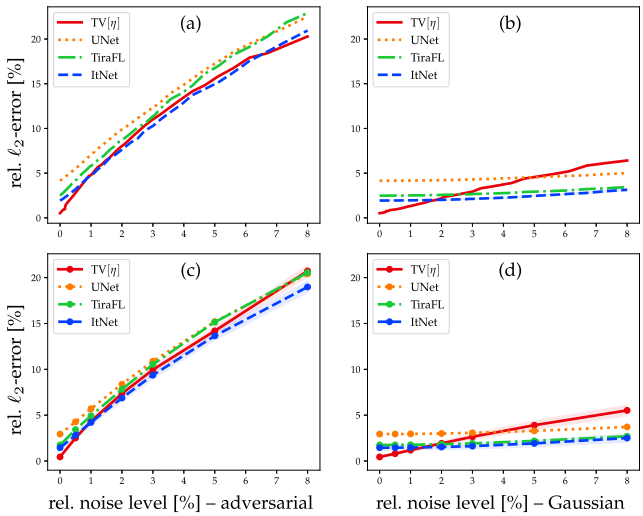

Fig. 6. *Scenario B1 – Fourier meas. with ellipses.* (a) shows the adversarial noise-to-error curve for the randomly selected image of Fig. 7. (b) shows the corresponding Gaussian noise-to-error curve, where the mean and (almost imperceptible) standard deviation are computed over 50 draws of e. (c) and (d) display the respective curves averaged over 50 images from the test set. For the sake of clarity, we have omitted the standard deviations for UNet and TiraFL, which behave similarly.

without fat-suppression, resulting in $M \approx 17\mathrm{k}$ training images. We draw magnitude images $\boldsymbol{x}_0 \in \mathbb{R}^{320\times320}$, obtained from fully-sampled multi-coil[5] data, and consider subsampled Fourier measurements as in Scenario B1 with 50 radial lines ($m = 17178$ and $N = 320^2 = 102400$). As before, the data processing is complex-valued, while the actual reconstructions are computed as real-valued magnitude images. The model-based and learned inversion layers are realized as in Scenario B1. As common in the fastMRI challenge, we have trained all networks with a cost function based on a combination of the ℓ_1- and SSIM-distance, see also [78]. TV minimization is solved in the unconstrained formulation, with the regularization parameter determined by a grid search over a subset of the validation set.

Fig. 9 shows the noise-to-error curves; see also Tables S7 and S8, available online. The associated individual reconstructions for $\mathrm{TV}[\eta]$ and TiraFL with adversarial noise are displayed in Fig. 10; see Fig. S9, available online, for the remaining networks and Fig. S10, available online, for the corresponding results with Gaussian noise.

Conclusions. Our experimental results show that the main findings of Case Study A and B carry over to real-world data. The noise-to-error curves in Fig. 9 reveal a superior robustness of the learned reconstruction schemes over TV minimization, even for noiseless measurements (cf. Scenario A2). Fig. 10 underpins this observation from a qualitative

4. Data used in the preparation of this article were obtained from the NYU fastMRI Initiative database [30], [31] (https://fastmri.med.nyu.edu). As such, NYU fastMRI investigators provided data but did not participate in analysis or writing of this article. The primary goal of fastMRI is to test whether machine learning can aid in the reconstruction of medical images.

5. Note that our measurement model actually corresponds to the simpler modality of subsampled single-coil MRI. While the fastMRI challenge also provides single-coil data, it is based on retrospective masking of *emulated* Fourier measurements. The subsampling is done by omitting k-space lines in the phase-encoding direction, which we found less suitable for our robustness analysis; see Section 5.4 for an experiment with the original setup. Since emulating single-coil measurements is unavoidable, we have decided to sample from the multi-coil magnitude reconstructions in favor of higher image quality. This was found to be particularly important to ensure that TV minimization can serve as a competitive benchmark method, at least for noiseless measurements.

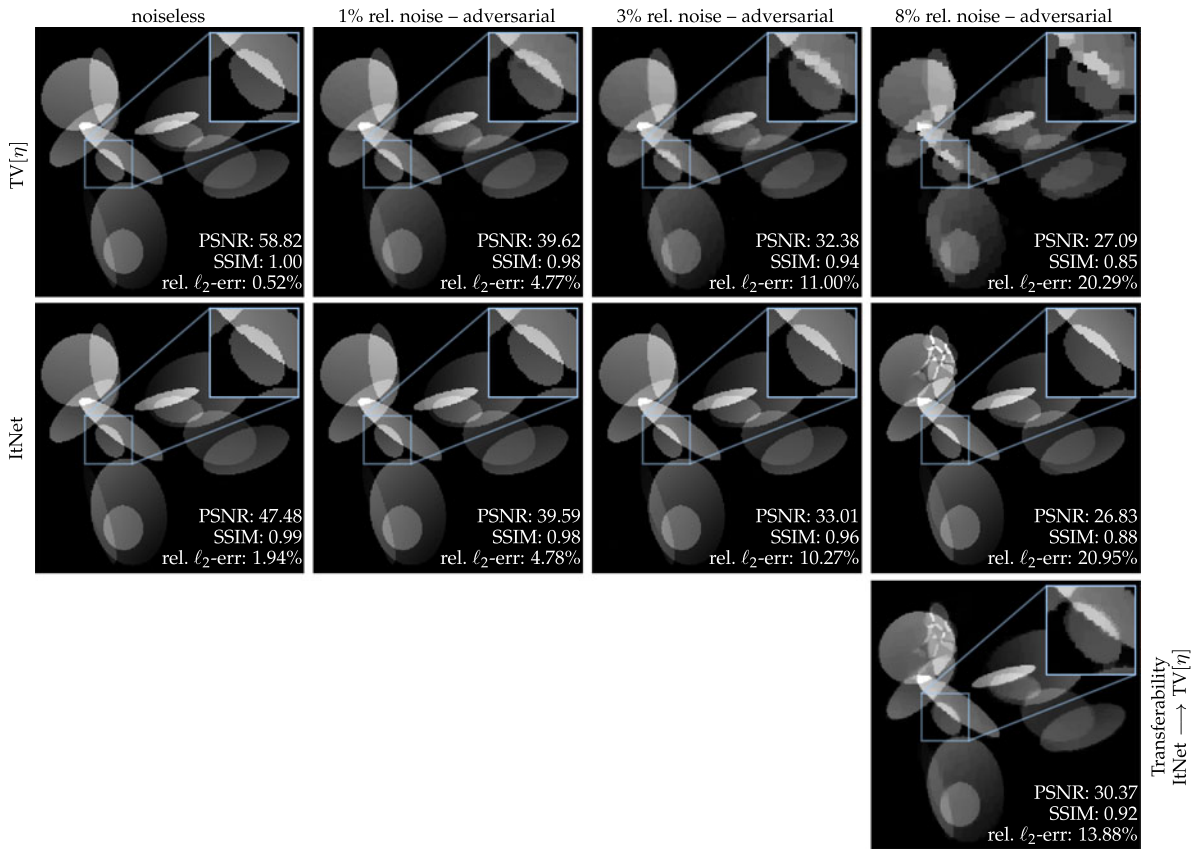

Fig. 7. *Scenario B1 – Fourier meas. with ellipses.* Individual reconstructions of a randomly selected image from the test set for different levels of adversarial noise. The reconstructed images are displayed in the window [0,0.9], which is also used for the computation of the PSNR and SSIM. For error plots and the results of UNet and TiraFL, we refer to Fig. S5, available online. The bottom right figure concerns the *transferability* of adversarial noise: it shows the reconstruction $\mathrm{TV}[\eta](\boldsymbol{y}_{\mathrm{adv}})$, where $\boldsymbol{y}_{\mathrm{adv}}$ is the perturbation found for ItNet with 8% noise; see Fig. S8, available online, for additional experiments. The ground truth image $\boldsymbol{x}_0$ has been omitted, as it is visually indistinguishable from the noiseless reconstruction by $\mathrm{TV}[\eta]$.

viewpoint: the model-based prior of $\mathrm{TV}[\eta]$ tends to blur fine details in the reconstructed images—this "oil painting" effect becomes stronger with larger perturbations. In contrast, the NN-based reconstructions always yield high resolution images. Despite adversarial noise, the central image region—which is of main medical interest—remains largely unaffected, whereas tiny vessel structures appear in the outside (fat) region. Such an amplification of existing patterns is comparable to the ellipsoidal artifacts in Case Study B. We emphasize that this phenomenon only occurs for large adversarial perturbations, where the benchmark of TV minimization already suffers from severe distortions. In particular, the performance of the learned methods is not impaired by the same amount of Gaussian noise (see Fig. S10, available online).

5 Further Aspects of Robustness

This section presents several additional experiments that allow for further insights into the robustness of learned methods.

5.1 Training Without Noise – An Inverse Crime?

In this section, the importance of *jittering* for the stability of deep-learning-based reconstruction schemes is discussed (see Section 3.2). We have found that this technique can be beneficial for promoting adversarial robustness, in particular, for iterative architectures. The previous claim is verified by an ablation study, comparing two versions of ItNet for Scenario A2, one trained with jittering and the other without. The resulting noise-to-error curves in Fig. 11 reveal that noiseless training data can have drastic consequences. Indeed, the relative recovery error blows up at $\sim$15% adversarial noise if jittering is not used. In a similar experiment, we analyze the adversarial robustness of image recovery from Radon measurements as in Scenario B2. The results of Fig. 12 show a clear superiority of the UNet that was subjected to noise during training (see also Fig. S7, available online, for the effect of Poisson noise). Without jittering, almost imperceptible distortions in the FBP inversions are intensified by the post-processing network (see blue arrows).

The above observations can be related to the notion of *inverse crimes* in the literature on inverse problems, e.g., see [79], [80]. This term is commonly used to explain the phenomenon of exact, but highly unstable, recovery from noiseless, simulated measurements. In a similar way, networks seem to learn accurate, but unstable, reconstruction rules if they are trained with noiseless data. We note that this does not only concern simulated phantom data but also real-world scenarios. Indeed, in medical imaging applications, one often acquires fully sampled (noisy) reference scans $\{\tilde{\boldsymbol{y}}^i\}_{i=1}^M$, which are used to generate the ground truth training images $\boldsymbol{x}_0^i = \mathcal{A}_{\mathrm{full}}^{-1}\tilde{\boldsymbol{y}}^i$. The measurements are usually subsampled retrospectively by $\boldsymbol{y}^i = P\tilde{\boldsymbol{y}}^i$, where P denotes an appropriate selection operator. NN-based solution methods

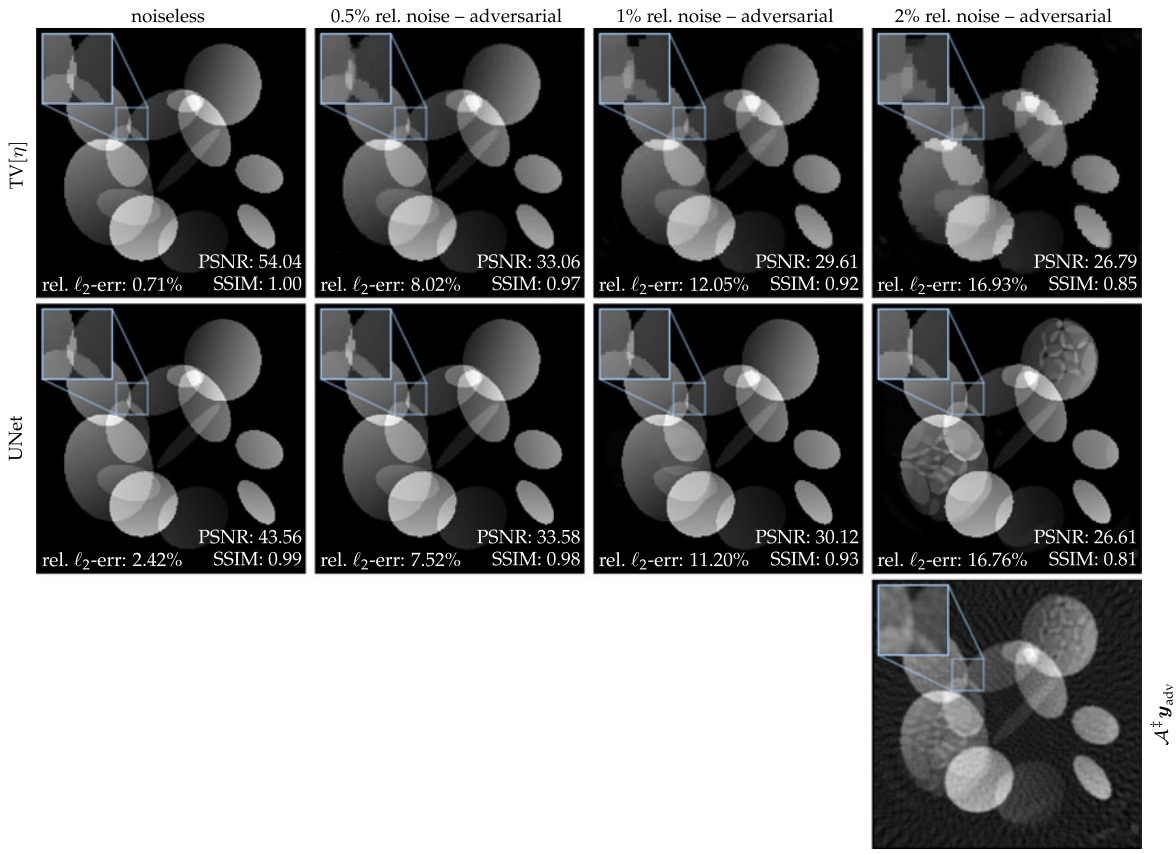

Fig. 8. *Scenario B2 – Radon meas. with ellipses.* Individual reconstructions of a randomly selected image from the test set for different levels of adversarial noise. The reconstructed images are displayed in the window [0,1], which is also used for the computation of the PSNR and SSIM. The bottom right figure shows the FBP inversion of the 2%-adversarial perturbation found for UNet. The ground truth image x_0 has been omitted, as it is visually indistinguishable from the noiseless reconstruction by $\mathrm{TV}[\eta]$.

for the limited data problem (1) with $\mathcal{A} = P\mathcal{A}_{\mathtt{full}}$ are then obtained by training on $\{(\boldsymbol{y}^i, \boldsymbol{x}_0^i)\}_{i=1}^M$. Importantly, such data pairs also "commit" an inverse crime, since they follow the noiseless forward model $\mathcal{A}\boldsymbol{x}_0^i = P\mathcal{A}_{\mathtt{full}}\boldsymbol{x}_0^i = \boldsymbol{y}^i$. Hence, we believe that simulating additional noise might be helpful in the situation of real-world measurements as well. Jittering is a simple and natural remedy in that regard that can additionally reduce overfitting [56]. The exploration of further regularization techniques or more sophisticated ways of injecting noise during training is left to future research.

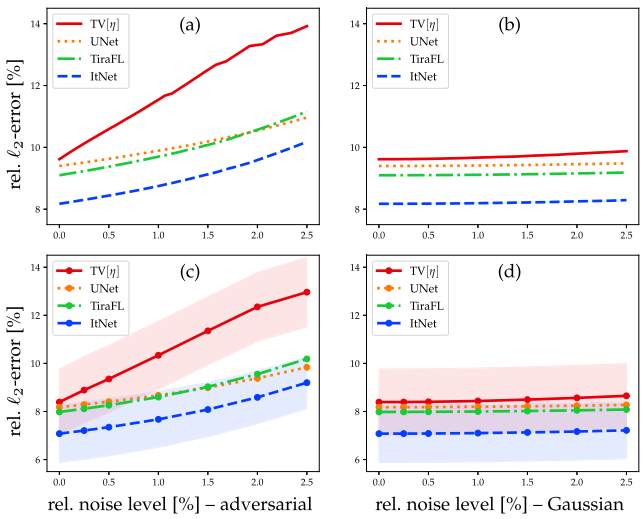

Fig. 9. *Case Study C – fastMRI.* (a) shows the adversarial noise-to-error curve for the randomly selected image of Fig. 10. (b) shows the corresponding Gaussian noise-to-error curve, where the mean and (almost imperceptible) standard deviation are computed over 50 draws of e. (c) and (d) display the respective curves averaged over 30 images from the validation set. For the sake of clarity, we have omitted the standard deviations for UNet and TiraFL, which behave similarly.

5.2 Training With Noise – Losing Accuracy?

One might wonder whether the aforementioned robustification via jittering has a detrimental effect on the resulting reconstruction scheme for unperturbed inputs. Indeed, if a method—not necessarily learned—is too insensitive to small changes in the input, it might become incapable of reconstructing fine details. In the context of our study, it is useful to distinguish between the recovery accuracy with respect to *in-distribution* and *out-of-distribution* (*OOD*) features. The former simply corresponds to a task evaluation on regular images from the test set. Regarding this aspect, we have observed only a marginal impact of jittering: across all considered scenarios, no significant performance loss was found when training with noise (e.g., see left column of Fig. 13), and occasionally, the accuracy even improved slightly (e.g., see Fig. 11 and left column of Fig. S7, available online).

The behavior might be different for OOD attributes. Following [24], we address this situation by exposing an

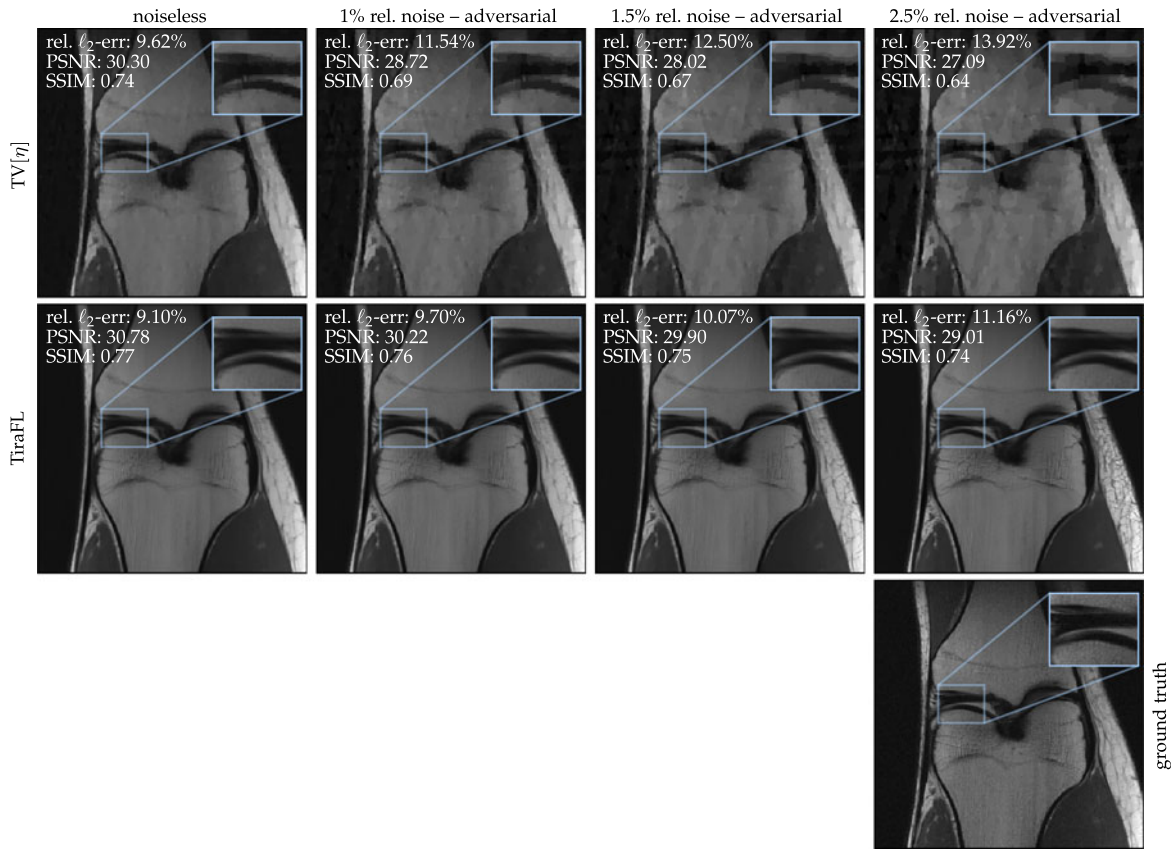

Fig. 10. *Case Study C – fastMRI.* Individual reconstructions of a central slice of a randomly selected volume from the validation set for different levels of adversarial noise. The reconstructed images are displayed in the window [0.05,4.50], which is also used for the computation of the PSNR and SSIM. For error plots and the results of UNet and ItNet, we refer to Fig. S9, available online. The ground truth image x_0 is shown at the bottom right.

NN-based solver to structural details that do not belong to the data distribution. Fig. 13 shows that inserted text and a 3×3-square are recognizable with and without jittering. While the feature contrast is higher in the latter case, no essential information was missed due to training with noise.

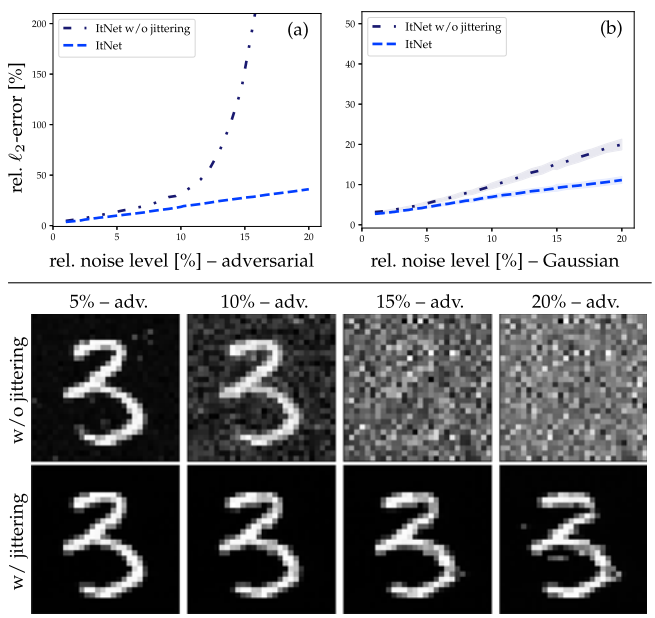

Fig. 11. *An inverse crime?* A comparison between ItNet trained with and without jittering. The above noise-to-error curves are generated for the MNIST-digit 3 from Fig. 5 with (a) adversarial and (b) Gaussian noise. Individual reconstructions for adversarial noise are shown below (the intermediate steps performed by ItNet are visualized in Fig. S11, available online).

Nevertheless, the OOD generalization of learned methods is still poorly understood in general. Among other factors, the outcome may also depend on the "richness" of the training data. Fig. 14 presents a similar experiment as in Fig. 13 with phantom ellipses (Case Study B). In this case, the ItNet with jittering is not able to recover the added text feature. We also give some evidence that this limitation is a consequence of our naive (non-adaptive) jittering strategy. In fact, a slightly modified training procedure provides a simple remedy.[6] On the other hand, the noise-to-error curve in Fig. 14 indicates a trade-off between robustness and accuracy, since the modified ItNet exhibits a larger reconstruction error for higher noise levels.

5.3 Adversarial Examples for Classification From Compressed Measurements

In medical healthcare, image recovery is merely one component of the entire data-processing chain. Indeed, machine learning techniques are particularly suitable for automated diagnosis or personalized treatment recommendations. As argued in the introduction of this article, the study of

6. This modification consists in training the ItNet with jittering first (as before) and then performing a second training phase with a much lower jittering level.

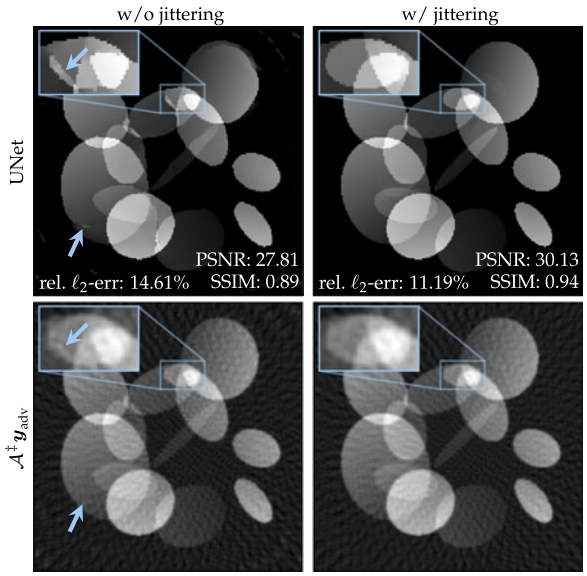

Fig. 12. *An inverse crime?* A comparison between UNet trained with and without jittering for image recovery from sparse-angle Radon measurements, see Fig. 8 in Scenario B2. The reconstructions are obtained for 1% adversarial noise. The bottom figures show the FBP inversions of the found perturbations, respectively. The blue arrows highlight tiny distortions that are amplified by the post-processing network.

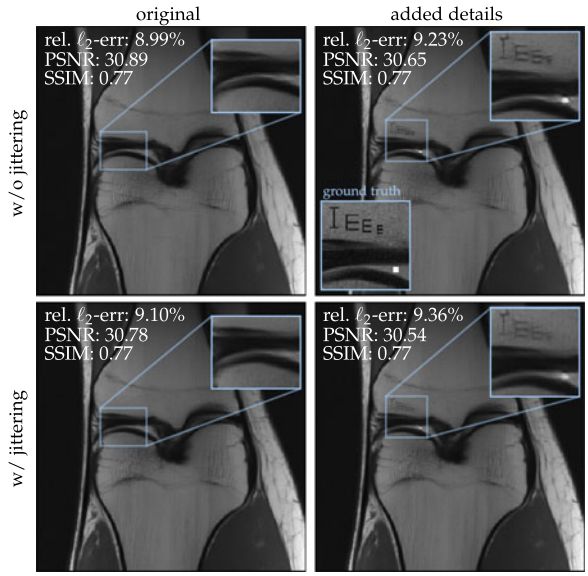

Fig. 13. *Losing accuracy?* The left column compares the image of Fig. 10, when reconstructed by TiraFL with and without jittering, respectively. The right column displays analogous results after adding two out-of-distribution features (text and a 3×3-square) to the ground truth; note that the smallest letter 'E' has the lowest possible resolution. The reconstructed images are displayed in the window [0.05,4.50], which is also used for the computation of the PSNR and SSIM.

adversarial examples for such classification tasks differs from the robustness analysis of reconstruction methods. In this section, we shed further light on this subject by analyzing classification from compressed measurements—think of detecting a tumor from a subsampled MRI scan.

To this end, we revisit the benchmark model of Scenario A2, with the goal to predict MNIST digits from their Gaussian measurements. This is realized by training a basic convolutional NN classifier $\text{ConvNet} : \mathbb{R}^N \to [0,1]^{10}$, mapping images to class probabilities for each of the 10 digits. The concatenation with a reconstruction method $\text{Rec} : \mathbb{R}^m \to \mathbb{R}^N$ then yields the following classification map:

$$\text{CC} : \mathbb{R}^m \to [0,1]^{10}, \ \boldsymbol{y} \mapsto [\text{ConvNet} \circ \text{Rec}](\boldsymbol{y}). \tag{6}$$

The approach of CC can be seen as a simplified model for the automated diagnosis from subsampled measurements; see also [82] and the references therein for the related problem of *compressed classification*.

Inspired by [69], we adapt the attack strategy (5) to the classification setting by (approximately) solving

$$\boldsymbol{e}_{\text{adv}} = \underset{\|\boldsymbol{e}\|_2 \leq \eta}{\arg\max} \ \max_{k \neq c} \ [\text{CC}(\boldsymbol{y}_0 + \boldsymbol{e})]_k - [\text{CC}(\boldsymbol{y}_0 + \boldsymbol{e})]_c,$$

where $c \in \{0, 1, \ldots, 9\}$ is the true class label of $\boldsymbol{x}_0$. Fig. 15 shows a noise-to-*accuracy* curve visualizing the relative amount of correct classifications for different choices of Rec. The corresponding image reconstructions $\text{Rec}(\boldsymbol{y}_0 + \boldsymbol{e}_{\text{adv}})$ as well as the predicted classes $\arg\max_k [\text{CC}(\boldsymbol{y}_0 + \boldsymbol{e}_{\text{adv}})]_k$ for an example digit are presented below.

All classifiers exhibit a transition behavior: the success rate is almost perfect for small perturbations and then drops to zero at some point. The associated images show that we have found adversarial examples in the ordinary sense of machine learning. Indeed, every visualized reconstruction is still recognizable as the digit 9. In other words, although being stable, each of the recovery methods is capable of producing slightly perturbed images that fool the ConvNet-part. Remarkably, this phenomenon occurs independently of using a model-based or learned solver for (1). We conclude that deep-learning-based data-processing pipelines (as in medical healthcare) remain vulnerable to adversarial attacks, even if provably robust reconstruction schemes are employed.

5.4 The Original fastMRI Challenge Setup

This section demonstrates that the original fastMRI challenge data for single-coil MRI is more susceptible to adversarial noise. In contrast to Case Study C, the challenge measurement setup is based on omitting k-space lines in the phase-encoding direction (corresponding to 4-fold acceleration), i.e., the subsampling mask is defined by vertical lines. The resulting undersampling ratio of ∼23% is higher than in Case Study C (∼17%). Fig. 16 shows individual image reconstructions for TV[η] and Tira.[7] Compared to Fig. 10, the outcomes indicate a loss of adversarial robustness, as the reconstructed images exhibit undesired line-shaped artifacts (see blue box in Fig. 16). This phenomenon occurs regardless of using a model-based (TV[η]) or learned method (Tira). Nevertheless, a noteworthy pitfall is that such defects are not easily detectable for learned schemes, since they still produce realistic images.

The observed artifacts are a consequence of the underlying measurement system: the anisotropic mask pattern implies that vertical image features become more "aligned"

7. Since the fastMRI challenge setup does not rely on a fixed subsampling mask, the fully-learned approach for Tiramisu is not available here. Our Tira-net performs competitively in the fastMRI public leaderboard: We have achieved an SSIM of 0.765, whereas the leading method has 0.783 (https://fastmri.org/leaderboards/, teamname `AnItalianDessert`, accessed on 2020-11-08).

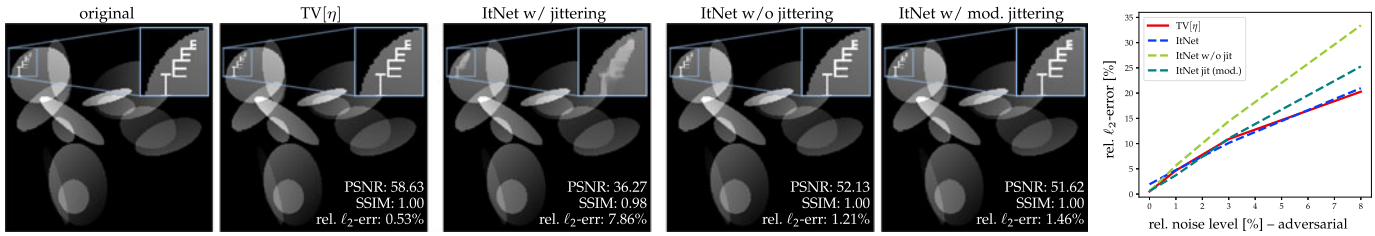

Fig. 14. *Losing accuracy?* The top row compares different methods when reconstructing the image of Fig. 7 with an added OOD text feature. As the underlying signal is sufficiently gradient-sparse, $\mathrm{TV}[\eta]$ provides a highly accurate result. The ItNet with jittering fails to recover the text, as opposed to its non-jittering counterpart; see also [81, Fig. 5] for a similar experiment. A simple remedy is shown in the right column: a slight modification of the jittering approach (training with jittering first, then with a much lower jittering level in later epochs) produces a network that is similarly robust to $\mathrm{TV}[\eta]$ for low- and mid-noise levels (see noise-to-error curve), while it successfully recovers the text. Remarkably, none of the networks has ever seen text features in the training stage. We anticipate that further improvements are possible with a more advanced approach, for instance, by learning noise-aware NNs for multiple values of η. This is comparable to tuning $\mathrm{TV}[\eta]$ with respect to the given noise level. Moreover, it is notable that compared to the other methods, ItNet does not achieve near-zero reconstruction error in the noiseless limit. The reconstructed images are displayed in the window [0,1], which is also used for the computation of the PSNR and SSIM.

with the kernel of the forward operator. Hence, clearly visible distortions may be caused by relatively small perturbations of the measurements (cf. [25]). This confirms that the design of sampling patterns does not only influence the accuracy of a reconstruction method (e.g., see [83]), but also its adversarial robustness.

6 Discussion

In an extensive series of experiments, this work has analyzed the robustness of deep-learning-based solution methods for inverse problems. Central to our approach was to study the effect of adversarial noise, i.e., worst-case perturbations of the measurements that maximize the reconstruction error. A systematic comparison with a model-based reference method has shown that standard deep NN schemes are remarkably resilient against statistical and adversarial distortions. On the other hand, we have demonstrated that instabilities might be caused by the "inverse crime" of training with noiseless data. A simple remedy in that regard is jittering—a standard regularization and robustification technique in deep learning [8]. However, it is well known that this does not cure the adversarial vulnerability of deep NN classifiers, which requires more sophisticated defense strategies [84]. While such defenses may also improve the robustness in the context of image recovery [26], our results allow for a surprising conclusion: Injecting Gaussian random noise in the training phase seems sufficient to obtain solution methods for inverse problems that are resistant to other types of noise, including adversarial perturbations.

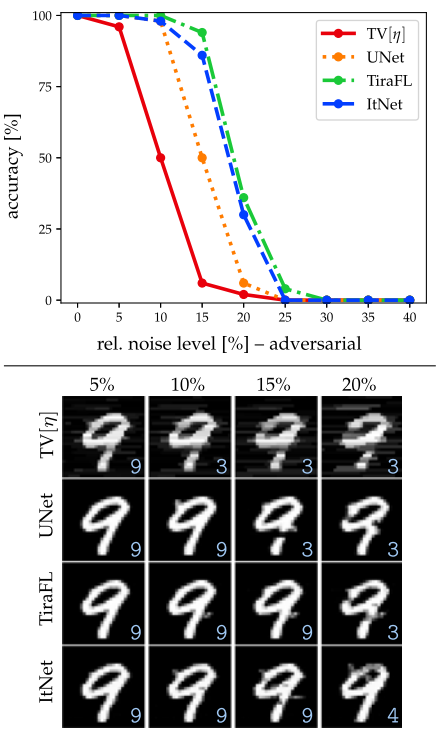

Fig. 15. *Classification from compressed measurements.* The above curve plots the relative adversarial noise level against the prediction accuracy of the classfier (6) for different recovery methods (averged over 50 digits from the test set). The intermediate reconstructions of a randomly selected digit are shown below for different noise levels. Their predicted class labels are displayed in the bottom right corner.

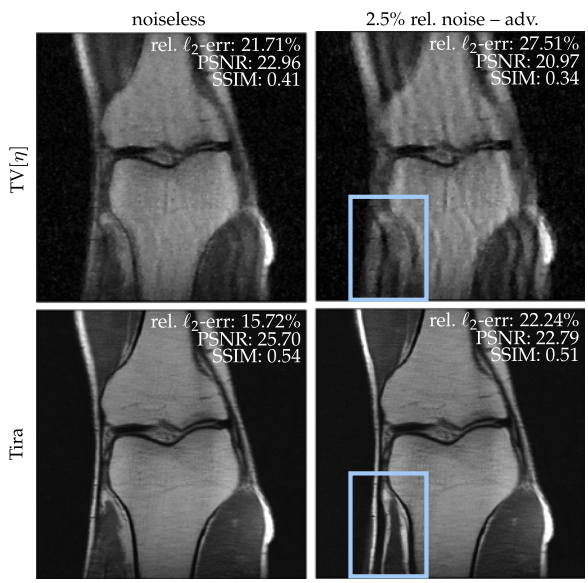

Fig. 16. *The original fastMRI challenge setup.* Reconstructions of a randomly selected image from the validation set. Compared to the analogous experiment in Fig. 10, the Fourier subsampling operator is based on vertical lines in the k-space instead of a radial mask. The reconstructed images are displayed in the window [0.05,4.50], which is also used for the computation of the PSNR and SSIM. Note that the data are given as emulated single-coil (ESC) measurements, whereas the reconstructions in Fig. 10 are based on multi-coil images. Hence, the signal-to-noise ratios are not directly comparable.

Admittedly, there are several aspects that go beyond the scope of our study: (i) We are restricted to a selection of end-to-end NN architectures, excluding other approaches, such as generative models [16], the deep image prior [85], or learned regularizers [86]. However, since these algorithms typically involve more model-based components, we expect their robustness to be comparable to the schemes considered in the present work. (ii) Due to the non-convexity of (5), a theoretical optimality certificate for our attack strategy is lacking. Nevertheless, our results provide empirical evidence that we have solved the problem adequately: The gap between worst-case and statistical perturbations appears consistent across all considered scenarios. More importantly, we have verified the ability to detect an error blowup caused by adversarial noise (see Fig. 11). (iii) Our analysis takes a mathematical perspective on robustness, thereby relying on standard similarity measures, in particular, the euclidean norm. It is well known that such quantitative metrics are insensitive to several types of visual distortions. For example, a characteristic feature of data-driven methods is that they tend to generate realistic images, even when compromised (cf. Section 5.4). This can hinder the detection of failure modes and possibly lead to false-positives/negatives [25]. (iv) The reliability of NN-based reconstructions may suffer from other shortcomings that are not directly linked to a lack of adversarial robustness. For instance, even the winning networks of the 2019 fastMRI challenge were occasionally unable to capture certain tiny pathological features that rarely appear in the data [87]. This issue was specifically addressed in the 2020 fastMRI challenge, which focuses on pathology depiction instead of an overall image quality assessment [88]. However, this time, hallucinations were noted, i.e., non-physical features created by the reconstruction networks. An investigation of causes and remedies seems to be a promising direction for future research, e.g., see [89].

The relevance of artificial intelligence for future healthcare is undeniable. Reliable reconstruction methods are indispensable in this field, since errors caused by instabilities can be fatal. In light of the threat of intentional manipulations in medical imaging [90], it is reassuring to know the limits of what could go wrong in principle. Of similar practical interest is the robustness against random perturbations, which is the standard noise model for common imaging modalities. We believe that our work makes progress in both regards, by showing optimistic results on the use of deep NNs for inverse problems in imaging.

Acknowledgments

We express our gratitude to the Institute of Mathematics of the Technical University of Berlin for providing us hardware resources to realize the numerical experiments presented in this work. Moreover, the authors would like to thank the anonymous referees for their useful comments and suggestions. Martin Genzel, Jan Macdonald, and Maximilian März have contributed equally.

References

[1] 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.

[2] J. Haupt, W. U. Bajwa, M. Rabbat, and R. Nowak, "Compressed sensing for networked data," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 92–101, Mar. 2008.

[3] J. L. Starck, E. Pantin, and F. Murtagh, "Deconvolution in astronomy: A review," *Pub. Astronomical Soc. Pacific*, vol. 114, no. 800, pp. 1051–1069, 2002.

[4] A. Tarantola and B. Valetta, "Inverse problems = quest for information," *J. Geophys.*, vol. 50, no. 1, pp. 159–170, 1981.

[5] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Basel, Switzerland: Birkhäuser, 2013.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

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

[10] Y. Yang, J. Sun, H. Li, and Z. Xu, "Deep ADMM-net for compressive sensing MRI," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 10–18.

[11] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[12] E. Kang, J. Min, and J. C. Ye, "A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction," *Med. Phys.*, vol. 44, no. 10, pp. e360–e375, 2017.

[13] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4509–4522, Sep. 2017.

[14] K. Hammernik *et al.*, "Learning a variational network for reconstruction of accelerated MRI data," *Magn. Reson. Med.*, vol. 79, no. 6, pp. 3055–3071, 2018.

[15] H. Chen *et al.*, "Low-dose CT via convolutional neural network," *Biomed. Opt. Exp.*, vol. 8, no. 2, pp. 679–694, 2017.

[16] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 537–546.

[17] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen, "Image reconstruction by domain-transform manifold learning," *Nature*, vol. 555, no. 7697, pp. 487–492, 2018.

[18] T. A. Bubba *et al.* "Learning the invisible: A hybrid deep learning-shearlet framework for limited angle computed tomography," *Inverse Problems*, vol. 35, no. 6, 2019, Art. no. 064002.

[19] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb, "Solving inverse problems using data-driven models," *Acta Numerica*, vol. 28, pp. 1–174, 2019.

[20] M. Elad, "Deep, deep trouble: Deep learning's impact on image processing, mathematics, and humanity," SIAM News, 2017. [Online]. Available: https://sinews.siam.org/Details-Page/deep-deep-trouble

[21] H. Chen *et al.*, "Low-dose CT with a residual encoder-decoder convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 36, no. 12, pp. 2524–2535, Dec. 2017.

[22] A. Hauptmann, J. Adler, S. Arridge, and O. Öktem, "Multi-scale learned iterative reconstruction," *IEEE Trans. Comput. Imag.*, vol. 6, no. 1, pp. 843–856, Apr. 2020.

[23] Y. Huang, T. Würfl, K. Breininger, L. Liu, G. Lauritsch, and A. Maier, "Some investigations on robustness of deep learning in limited angle tomography," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2018, pp. 145–153.

[24] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen, "On instabilities of deep learning in image reconstruction and the potential costs of AI," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 48, pp. 30088–30095, 2020.

[25] N. M. Gottschling, V. Antun, B. Adcock, and A. C. Hansen, "The troublesome kernel: Why deep learning for inverse problems is typically unstable," 2020, *arXiv:2001.01258*.

[26] A. Raj, Y. Bresler, and B. Li, "Improving robustness of deep-learning-based image reconstruction," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 7932–7942.

[27] C. Szegedy *et al.*, "Intriguing properties of neural networks," 2014, *arXiv:1312.6199*.

[28] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2017, *arXiv:1607.02533*.

[29] K. Eykholt *et al.*, "Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1625–1634.
[30] J. Zbontar *et al.*, "fastMRI: An open dataset and benchmarks for accelerated MRI," 2018, *arXiv:1811.08839*.
[31] 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.
[32] A. Paszke *et al.*, "Automatic differentiation in PyTorch," Contribution to the NIPS 2017 Autodiff Workshop, 2017. [Online]. Available: https://openreview.net/forum?id=BJJsrmfCZ
[33] N. Carlini, "A complete list of all (arXiv) adversarial example papers," 2020. Accessed: Nov. 02, 2020. [Online]. Available: https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html
[34] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, Sep. 2019.
[35] G. Ortiz-Jimenez, A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard, "Optimism in the face of adversity: Understanding and improving deep learning through adversarial robustness," 2020, *arXiv:2010.09624*.
[36] A. Arnab, O. Miksik, and P. H. S. Torr, "On the robustness of semantic segmentation models to adversarial attacks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 12, pp. 3040–3053, Dec. 2020.
[37] E. T. Quinto, "Singularities of the X-Ray transform and limited data tomography in $\mathbb{R}^2$ and $\mathbb{R}^3$," *SIAM J. Math. Anal.*, vol. 24, no. 5, pp. 1215–1225, 1993.
[38] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for dynamic MR image reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 491–503, Feb. 2018.
[39] E. Kobler, A. Effland, K. Kunisch, and T. Pock, "Total deep variation: A stable regularizer for inverse problems," 2020, *arXiv:2006.08789*.
[40] J. Schwab, S. Antholzer, and M. Haltmeier, "Big in Japan: Regularizing networks for solving inverse problems," *J. Math. Imag. Vis.*, vol. 62, pp. 445–455, 2020.
[41] A. Virmaux and K. Scaman, "Lipschitz regularity of deep neural networks: Analysis and efficient estimation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 3839–3848.
[42] T. Huster, C.-Y. J. Chiang, and R. Chadha, "Limitations of the Lipschitz constant as a defense against adversarial examples," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2018, pp. 16–29.
[43] G. Ongie, A. Jalal, R. G. Baraniuk, C. A. Metzler, 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.
[44] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[46] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Generative adversarial networks for noise reduction in low-dose CT," *IEEE Trans. Med. Imag.*, vol. 36, no. 12, pp. 2536–2545, Dec. 2017.
[47] Q. Yang *et al.*, "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1348–1357, Jun. 2018.
[48] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers Tiramisu: Fully convolutional DenseNets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1175–1183.
[49] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.
[50] J. Schlemper *et al.*, "dAUTOMAP: Decomposing AUTOMAP to achieve scalability and enhance performance," 2019, *arXiv:1909.10995*.
[51] H. K. Aggarwal, M. P. Mani, and M. Jacob, "MoDL: Model-based deep learning architecture for inverse problems," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 394–405, Feb. 2019.
[52] J. Adler and O. Öktem, "Learned primal-dual reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1322–1332, Jun. 2018.
[53] K. Hammernik, J. Schlemper, C. Qin, J. Duan, R. M. Summers, and D. Rueckert, "Σ-net: Systematic evaluation of iterative deep neural networks for fast parallel MR Image reconstruction," 2019, *arXiv:1912.09278*.
[54] I. Y. Chun, Z. Huang, H. Lim, and J. Fessler, "Momentum-Net: Fast and convergent iterative neural network for inverse problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 29, 2020, doi: 10.1109/TPAMI.2020.3012955.
[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[56] J. Sietsma and R. J. Dow, "Creating artificial neural networks that generalize," *Neural Netw.*, vol. 4, no. 1, pp. 67–79, 1991.
[57] L. Holmstrom and P. Koistinen, "Using additive noise in backpropagation training," *IEEE Trans. Neural Netw.*, vol. 3, no. 1, pp. 24–38, Jan. 1992.
[58] C. M. Bishop, "Training with noise is equivalent to Tikhonov regularization," *Neural Comput.*, vol. 7, no. 1, pp. 108–116, 1995.
[59] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, 2010.
[60] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenom.*, vol. 60, no. 1–4, pp. 259–268, 1992.
[61] A. Chambolle and P.-L. Lions, "Image recovery via total variation minimization and related problems," *Numerische Mathematik*, vol. 76, no. 2, pp. 167–188, 1997.
[62] M. Benning and M. Burger, "Modern regularization methods for inverse problems," *Acta Numerica*, vol. 27, pp. 1–111, 2018.
[63] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
[64] D. Needell and R. Ward, "Near-optimal compressed sensing guarantees for total variation minimization," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3941–3949, Oct. 2013.
[65] C. Poon, "On the role of total variation in compressed sensing," *SIAM J. Imag. Sci.*, vol. 8, no. 1, pp. 682–720, 2015.
[66] M. Genzel, M. März, and R. Seidel, "Compressed sensing with 1D total variation: Breaking sample complexity barriers via non-uniform recovery," *Inf. Inference*, 2021, *arXiv:2001.09952*.
[67] R. Glowinski and A. Marroco, "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires," *RAIRO Math. Model. Num.*, vol. 9, no. R2, pp. 41–76, 1975.
[68] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Comput. Math. Appl.*, vol. 2, no. 1, pp. 17–40, 1976.
[69] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 39–57.
[70] A. Agrawal, B. Amos, S. Barratt, S. Boyd, S. Diamond, and J. Z. Kolter, "Differentiable convex optimization layers," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 9562–9574.
[71] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[72] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Inf. Inference*, vol. 3, no. 3, pp. 224–294, 2014.
[73] J. Adler and O. Öktem, "Solving ill-posed inverse problems using iterative deep neural networks," *Inverse Problems*, vol. 33, no. 12, 2017, Art. no. 124007.
[74] E. Y. Sidky and X. Pan, "Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization," *Phys. Med. Biol.*, vol. 53, no. 17, pp. 4777–4807, 2008.
[75] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[76] P. Ernst, "Pytorch implementation of scikit-image's radon function, version 0.1.4," 2020. [Online]. Available: https://github.com/phernst/pytorch_radon
[77] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," 2016, *arXiv:1605.07277*.
[78] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[79] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*. New York, NY, USA: Springer, 2006.
[80] J. L. Mueller and S. Siltanen, *Linear and Nonlinear Inverse Problems With Practical Applications*. Philadelphia, PA, USA: SIAM, 2012.
[81] V. Antun, M. J. Colbrook, and A. C. Hansen, "Can stable and accurate neural networks be computed? On the barriers of deep learning and Smale's 18th problem," 2021, *arXiv:2101.08286*.
[82] A. S. Bandeira, D. G. Mixon, and B. Recht, "Compressive classification and the rare eclipse problem," in *Proc. 2nd Int. MATHEON Conf. Compressed Sens. Appl.*, 2017, pp. 197–220.
[83] C. Boyer, N. Chauffert, P. Ciuciu, J. Kahn, and P. Weiss, "On the generation of sampling schemes for magnetic resonance imaging," *SIAM J. Imag. Sci.*, vol. 9, no. 4, pp. 2039–2072, 2016.
[84] G. R. Machado, E. Silva, and R. R. Goldschmidt, "Adversarial machine learning in image classification: A survey towards the defender's perspective," 2020, *arXiv:2009.03728*.
[85] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9446–9454.
[86] H. Li, J. Schwab, S. Antholzer, and M. Haltmeier, "NETT: Solving inverse problems with deep neural networks," *Inverse Problems*, vol. 36, no. 6, 2020, Art. no. 065005.
[87] F. Knoll *et al.*, "Advancing machine learning for MR image reconstruction with an open competition: Overview of the 2019 fastMRI challenge," *Magn. Reson. Med.*, vol. 84, no. 6, pp. 3054–3070, 2020.
[88] M. J. Muckley *et al.*, "State-of-the-art machine learning MRI reconstruction in 2020: Results of the second fastMRI challenge," 2020, *arXiv:2012.06318*.
[89] K. Cheng, F. Calivá, R. Shah, M. Han, S. Majumdar, and V. Pedoia, "Addressing the false negative problem of deep learning MRI reconstruction models by adversarial attacks and robust training," in *Proc. 3rd Conf. Med. Imag. Deep Learn.*, 2020, pp. 121–135.
[90] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

Martin Genzel received the bachelor's, master's, and PhD degrees in mathematics from the Technical University Berlin, Berlin, Germany, in 2013, 2015, and 2019, respectively. He is currently a postdoctoral researcher with Utrecht University, The Netherlands. His research is focusing on topics at the interface of applied mathematics, signal processing, and machine learning, in particular, inverse problems, compressed sensing, high-dimensional statistics, and deep learning.

Jan Macdonald received the bachelor's and master's degrees in mathematics from the Technical University Berlin, Berlin, Germany, in 2014 and 2017, respectively. He is currently working toward the PhD degree at the Technical University Berlin, Berlin, Germany, working on topics in applied mathematics, inverse problems, and machine learning, in particular with applications in medical imaging.

Maximilian März received the bachelor's degree in mathematics from the University of Konstanz, Konstanz, Germany, in 2012, and the master's degree in mathematics from the Technical University Berlin, Berlin, Germany, in 2016. He is currently working toward the PhD degree with the Technical University Berlin, Berlin, Germany, where his work revolves around sparsity models for compressed sensing and the application of machine learning to inverse problems.