MELBA Journal Machine Learning for Biomedical Imaging

Benchmarking Self-Supervised Learning Methods for Accelerated MRI Reconstruction

Andrew Wang [1], Steven McDonagh [1], Mike Davies [1]

1 Institute for Imaging, Data and Communications, School of Engineering, University of Edinburgh

Abstract

Reconstructing MRI from highly undersampled measurements is crucial for accelerating medical imaging, but is challenging due to the ill-posedness of the inverse problem. While supervised deep learning (DL) approaches have shown remarkable success, they traditionally rely on fully-sampled ground truth (GT) images, which are expensive or impossible to obtain in real scenarios. This problem has created a recent surge in interest in self-supervised learning methods that do not require GT. Although recent methods are now fast approaching "oracle" supervised performance, the lack of systematic comparison and standard experimental setups are hindering targeted methodological research and precluding widespread trustworthy industry adoption. We present **SSIBench**, a modular and flexible comparison framework to unify and thoroughly benchmark **S**elf-**S**upervised **I**maging methods (SSI) without GT. We focus on end-to-end trained DL methods, which do not require long inference time, large datasets, or per-image training. We evaluate 21 such recent methods across seven realistic MRI scenarios on real data, showing a wide performance landscape whose method ranking differs across scenarios and metrics, exposing the need for further SSI research. To accelerate reproducible research and lower the barrier to entry, we provide the extensible benchmark and open-source reimplementations of all methods at https://github.com/Andrewwango/ssibench, allowing researchers to rapidly and fairly contribute and evaluate new methods on the standardised setup for potential leaderboard ranking, or benchmark existing methods on custom datasets, forward operators, or models, unlocking the application of SSI to other valuable nascent GT-free scientific imaging modalities.

Keywords

self-supervised learning, image reconstruction, MRI, inverse problems, benchmarking

Article informations

https://doi.org/10.59275/j.melba.2026-b6a3

Volume 2026, Received: 2026-02-22, Published 2026-08-29

Corresponding author: andrew.wang@ed.ac.uk

1. Introduction

Accelerating MRI reconstruction is crucial for reducing lengthy scan times, modelled as (Lustig et al., 2008):

$$\mathbf{y}_{i,c} = \mathbf{A}_{i,c}\mathbf{x}_i + \epsilon,\ \mathbf{A}_{i,c} = \mathbf{M}_i\mathbf{F}\mathbf{S}_c,\ \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \quad (1)$$

where $\mathbf{x}_i \in \mathcal{X} \in \mathbb{R}^n$ is the i-th underlying image, $\mathbf{y}_i = \{\mathbf{y}_{i,c}\}_{c=1}^C \in \mathcal{Y} \in \mathbb{R}^m$ are the *k-space* measurements, $\mathbf{A}_i = \{\mathbf{A}_{i,c}\}_{c=1}^C$ is the *forward operator* (or *acquisition method*), consisting of the undersampling mask $\mathbf{M}_i$, the c-th coil sensitivity map $\mathbf{S}_c$ and the Fourier operator $\mathbf{F}$, and $m = n/\alpha$ where α is the acceleration rate. Note we may drop the index i for clarity where it is unambiguous. The goal is to recover high quality images $\hat{\mathbf{x}}_i = f_\theta(\mathbf{y}_i, \mathbf{A}_i)$ via a *model* f_θ from partial, noisy measurements. However, this is challenging as the inverse problem is ill-posed, either due to undersampling $C = 1, m < n$, leading to a non-trivial null-space $\mathcal{N}(\mathbf{A}) \in \mathbb{R}^{n-m}$, or poor conditioning $\kappa(\mathbf{A}) \gg 1$.

Classical compressed sensing solves the problem using regularised optimisation (Lustig et al., 2008). However, this requires hand-crafted, image-specific priors, inference is lengthy and expensive, and acceleration rates may be limited (Heckel et al., 2024). Deep learning has been used to learn f_θ directly from data, and achieves impressive results compared to classical methods (Heckel et al., 2024; Klug et al., 2023; Tachella et al., 2022). Prevailing methods assume that underlying ground-truth (GT) images $\mathbf{x}$ are available in order to train with supervised learning $\mathcal{L}_{\text{sup}} = \ell(\hat{\mathbf{x}}, \mathbf{x})$ where $\ell(\cdot, \cdot)$ is any metric such as the mean squared

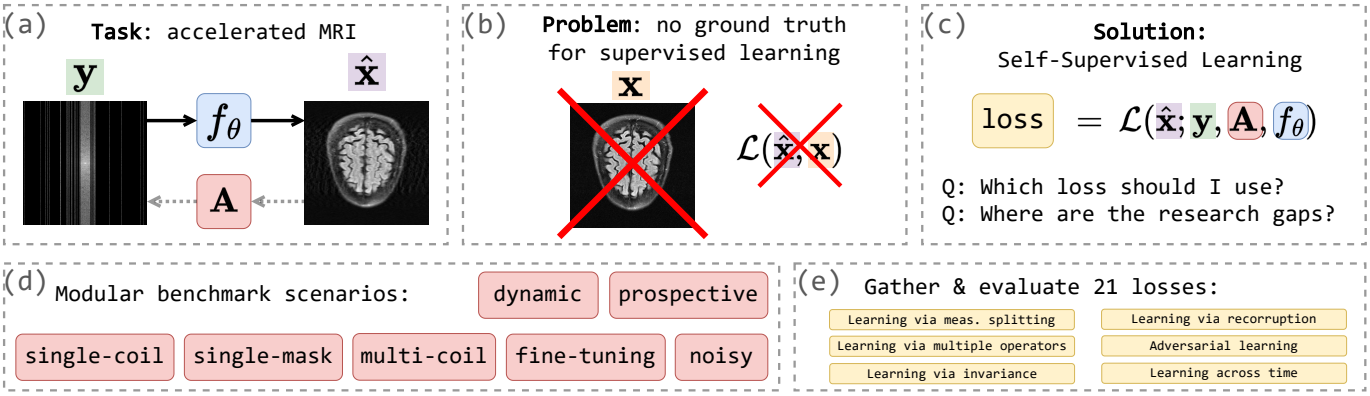

Figure 1: *(a)* Accelerated MRI reconstruction from undersampled images $\mathbf{y}$. *(b)* Ground truth (GT) images $\mathbf{x}$ are impossible or expensive to obtain. *(c)* There has been a recent surge in self-supervised imaging (SSI) methods that can learn without GT. *(d)* We introduce **SSIBench**, a modular benchmark over seven diverse MRI acquisition settings and anatomies, that gathers and evaluates 21 state-of-the-art end-to-end training loss functions *(e)*.

error (MSE).

However, fully-sampled GT measurements are impossible or expensive to obtain in real scenarios, *e.g.* when imaging moving organs (Wang and Davies, 2024), adaptation to new anatomies (Darestani et al., 2022), 4D MRI (Ong et al., 2020), low-field MRI (Janjušević et al., 2025), or where it leads to blurring in patient motion. Therefore, **self-supervised imaging (SSI) *i.e.* methods that learn without access to GT** $\mathbf{x}$ are needed. Note this differs from "self-supervision" in representation learning: here, images are output $\hat{\mathbf{x}} = f_\theta(\mathbf{y}, \mathbf{A})$ rather than input, and GT $\mathbf{x}$ refers to images rather than labels.

A vast array of SSI methods have been proposed in recent years for MRI reconstruction. However, they often fail to compare to existing methods (Cole et al., 2021; Chen et al., 2022a; Millard and Chiew, 2024), exacerbated by disjoint communities of researchers working on the same problem, *e.g.* from ML research, imaging, or various medical or other application domains. It is hence difficult to draw direct comparisons due to different or proprietary setups, datasets, a vast variety of models f_θ, forward operators $\mathbf{A}$, evaluation protocols or lack of transparent implementation of compared methods. Typically, each proposed method claims to perform best, preventing research from focussing on where true gaps lie, and impeding trustworthy translation of SSI into real-world settings.

Our benchmark unifies previously fractured efforts by providing a standardised, objective evaluation of methods, leading to trustworthy application. It also provides the opportunity for ML researchers to find unsolved gaps exposed in the common framework and clear performance landscape, and collectively contribute methodological advances to solve real-world and future imaging problems. By making the benchmark modular and accessible to the public, researchers can *a)* rapidly implement new methods or combine existing ones and fairly test these on the standard setup, *b)* evaluate them on custom datasets, models f_θ, or forward operators $\mathbf{A}$, or *c)* use as a blueprint for other challenging nascent scientific imaging problems within and beyond MRI, where lack-of-GT currently hurts progress.

Our benchmark investigates methods that satisfy critical inclusion criteria for practical adoption, such as speed, ease of training and generalisability (Heckel et al., 2024). Explicitly, we include:

End-to-end trained methods, since they provide fast inference by requiring only one neural function evaluation per image (NFE) $N = 1$, and can be easily trained/fine-tuned on an modest size dataset ($\sim$500 images) (Heckel et al., 2024) via feedforward inference on a discriminative model. We exclude generative methods with lengthy inference-time iterative procedures (Shafique et al., 2024) as these are infeasibly expensive in real-world clinical workflows, and that require large-scale datasets ($\sim 10^4$), since these are challenging to gather for many anatomies/acquisitions (such as 3D) and impede fine-tuning to new acquisitions. These include diffusion models (Daras et al., 2023; Kawar et al., 2024; Aali et al., 2024b) ($N \sim 10^{3 \text{ to } 4}$), unconditional GANs (Bora et al., 2018) ($N \sim 10^{2 \text{ to } 3}$), or flow matching (Luo et al., 2025) ($N \sim 10^{1 \text{ to } 2}$). We do include pilot results on SotA diffusion models pretrained on much larger quantities of measurements in section B.5, since these cannot be successfully trained on any of the scenarios' modest-sized datasets we consider here. We also exclude single-image methods that require retraining for *every* image, such as Deep Image Prior methods (Darestani and Heckel, 2021; Korkmaz et al., 2022; Zou et al., 2021) ($N \sim 10^{3 \text{ to } 4}$), implicit neural representations (Shen et al., 2024) ($N = 10^3$) or test-time training (Qin et al., 2023; Darestani et al., 2022; Joo et al., 2025) ($N \sim 10^2$).

Ground-truth-free *i.e.* self-supervised methods that learn from k-space measurements only $\{\mathbf{y}_i\}$, as GT images $\{\mathbf{x}_i\}$ are expensive or impossible to obtain during

training/pretraining of models, priors or features (Daras et al., 2024; Chung and Ye, 2022; Ericsson et al., 2022; He et al., 2020). Note that GT $\{\mathbf{x}_i\}$ is also required for learning from unpaired data $\{\mathbf{x}_i\}, \{\mathbf{y}_j\}$ (Oh et al., 2020; Li et al., 2024; Lei et al., 2021; Mukherjee et al., 2021) or weakly/semi-supervised learning $\{\mathbf{x}_i, \mathbf{y}_i\}, \{\mathbf{y}_j\}$ (Desai et al., 2022a).

Architecture-agnostic methods that do not depend on a specific model architecture, such that the learning is primarily guided by the loss function rather than the inductive bias of strong, specific architectural priors (Darestani and Heckel, 2021; Korkmaz et al., 2022; Tamir et al., 2020), and the loss function is the conceptual advance (Chen et al., 2021; Millard and Chiew, 2023). This way, our findings generalise to diverse future architectures, rather than limiting the applicability of our benchmark to current modality-specific architectures.

We provide extensive experimental results benchmarking these methods on accelerated MRI reconstruction without GT across seven acquisition scenarios & anatomies that are sufficiently general to be applied to future problems, and provide the loss reimplementations, living benchmark and training code at https://github.com/Andrewwango/ssibench.

1.1 Contributions

1. A comprehensive review unifying SotA self-supervised end-to-end trained deep learning methods for imaging;
2. Reproducible, well-documented and `pytested` reimplementations of 21 methods, and a modular benchmark site facilitating contribution and application to new problems;
3. Benchmarking experiments for accelerated MRI without GT across seven insightful scenarios on a standardised setup, where different methods have different strengths, and a decision framework for practitioners.

2. SSIBench - a benchmark for self-supervised imaging

2.1 Benchmark methodology

We provide a systematic review of existing SotA self-supervised end-to-end deep learning methods from the literature that fall under our scope mentioned above, and unify these by constructing **SSIBench** as follows. It is the design of the loss function $\mathcal{L}(\hat{\mathbf{x}}; \mathbf{y}, \mathbf{A}, f_\theta)$ that lets a method learn to recover information from the null-space (Tachella and Davies, 2026), and is the key conceptual advance that differs between methods (Millard and Chiew, 2023; Tachella et al., 2022). To provide a fair, substantive comparison and maximise compatibility of our benchmark to different setups, we evaluate the losses while fixing all other elements of the experimental setup, including the forward operator $\mathbf{A}$, the type, architecture and size of the model f_θ (such that all methods have the same inductive bias), data preprocessing stages, and the metrics. This decision is crucial for controlled comparisons of loss functions, so that the modular benchmark can easily be adapted with any model f_θ, forward operator $\mathbf{A}$ or dataset of the future.

In this framework, we gather 21 distinct ground-truth-free loss functions $\mathcal{L}(\hat{\mathbf{x}}; \mathbf{y}, \mathbf{A}, f_\theta)$ from across the ML and imaging literatures, summarised in table 1. We reimplement these since the compared methods' original codebases often adhere to differing software engineering principles and are often implemented for their specific experimental setups with poor generalisability to other setups. We provide details of their generalised implementations and sketches of the code mapped to their equations in section A.3. We contribute the losses into the DeepInverse library (Tachella et al., 2025b) following standard modern coding best practice of unit-testing, code-review and thorough documentation.

2.2 Benchmark scenarios

We consider seven common accelerated MRI experimental scenarios. This provides insight on the performance of the methods on (a) individual controlled tasks, which permit disentangling true reconstruction ability from other MRI sub-tasks *e.g.* coil sensitivity map estimation (Millard and Chiew, 2023; Yaman et al., 2020), and (b) diverse GT-free acquisition scenarios in real-world practice (Yu et al., 2022; Wang et al., 2023a; Desai et al., 2022c).

Scen. 1 (Single-coil) A basic but very important setup where no additional information is provided by multiple coils $C = 1$, so we can demonstrate performance on recovering data from the clear null-space of the rank-deficient forward operator $\mathcal{N}(A) \in \mathbb{R}^{n-m}$. No added noise $\sigma = 0$ and retrospectively undersampled to $6\times$ acceleration, with a random mask per sample $\mathbf{M}_i \sim \mathcal{M}$.

Scen. 2 (Noisy) Like scenario 1 (same $\mathbf{A}$, data), but simulates more realistic acquisition under thermal device noise $\sigma = 0.1$ (Millard and Chiew, 2024), where methods must perform joint reconstruction and denoising.

Scen. 3 (Single-mask) Like scenario 1, but with one fixed sampling mask $\mathbf{M}_i = \mathbf{M} \forall i$, common in clinical systems where a predetermined vendor-optimized sampling pattern is used to simplify hardware constraints and ensure consistency across patients.

Scen. 4 (Multi-coil) Like scenario 1, but $C = 4$ as in modern parallel imaging (Sriram et al., 2020). Since now m is a factor of $C\times$ larger, the effective null-space is smaller; see section B.2 for an analysis.

Table 1: Self-supervised losses gathered, reimplemented and evaluated in our benchmark framework. Above: losses for reconstruction. Below: losses for joint reconstruction and denoising. NFE: number of neural function f_θ evaluations during loss forward pass / inference.

Method	Ref	Loss	Code: `deepinv.loss.`_____	NFEs (train/test)
MC	(Senouf et al., 2019)	eq. (2)	`MCLoss()`	$1/1$
SSDU	(Yaman et al., 2022a)	eq. (3)	`SplittingLoss()`	$1/1$
Noise2Inverse	(Hendriksen et al., 2020)	eq. (4)	`SplittingLoss()`	$1/J$
Weighted-SSDU	(Millard and Chiew, 2023)	eq. (5)	`WeightedSplittingLoss()`	$1/1$
SSDU-Consistency	(Hu et al., 2021)	eq. (6)	`SplittingConsistencyLoss()`	$2/1$
MOC-SSDU	(Zhang et al., 2024)	eq. (8)+eq. (3)	`MOConsistencyLoss() + SplittingLoss()`	$2/1$
Adversarial	(Cole et al., 2021)	eq. (13)	`UnsupAdversarialGeneratorLoss()`	$1/1$
UAIR	(Pajot et al., 2018)	eq. (14)	`UAIRGeneratorLoss()`	$2/1$
VORTEX	(Desai et al., 2022a)	eq. (12)	`AugmentConsistencyLoss()`	$2/1$
ES	(Sechaud et al., 2025)	eq. (11)	`ESLoss()`	$1/1$
E2I	(Schut et al., 2025)	eq. (3)+eq. (9)	`SplittingLoss() + EILoss()`	$2/1$
EI	(Chen et al., 2021)	eq. (9)	`EILoss()`	$2/1$
MOI	(Tachella et al., 2022)	eq. (7)	`MOILoss()`	$2/1$
MO-EI	This work	eq. (10)	`MOEILoss()`	$2/1$
ENSURE	(Aggarwal et al., 2023)	eq. (16)	`ENSURELoss()`	$2/1$
DDSSL	(Quan et al., 2022)	eq. (18)+eq. (9)	`R2RLoss() + EILoss()`	$2/J$
Robust-SSDU	(Millard and Chiew, 2024)	eq. (17)	`RobustSplittingLoss()`	$1/1$
Noise2Recon-SSDU	(Desai et al., 2022b)	eq. (5)+eq. (12)	`WeightedSplittingLoss()` + `AugmentConsistencyLoss()`	$2/1$
ES-R2R	(Sechaud et al., 2025)	eq. (18) & eq. (11)	`R2RLoss()` & `ESLoss()`	$1/1$
Robust-EI	(Chen et al., 2022a)	eq. (9)+eq. (15)	`EILoss() + SUREGaussianLoss()`	$3/1$
Robust-MO-EI	This work	eq. (10)+eq. (15)	`MOEILoss() + SUREGaussianLoss()`	$3/1$

We assume known sensitivity maps to disentangle the estimation problem.

Scen. 5 (Fine-tuning) We test the widespread real-world practice of fine-tuning a foundation model (Terris et al., 2025) on out-of-domain data without GT on raw multicoil SKM-TEA (Desai et al., 2022c) knees at extreme 32× acc. disk sampling with estimated coil maps.

Scen. 6 (Dynamic) Learning to reconstruct 2D+t dynamic cine $10\times$ acquisition without GT on the real CMRxRecon (Wang et al., 2023a) cardiac dataset, as true GT is impossible to capture in dynamic MRI (Wang and Davies, 2024).

Scen. 7 (Prospective) Learning to reconstruct a single knee volume from 7× prospectively undersampled raw data (Yu et al., 2022), where fully-sampled GT truly does not exist.

2.3 Benchmarked methods

We formulate the 21 loss functions gathered from across the literature; reimplementation details are in section A.3.

Measurement consistency (MC) The most basic self-supervised loss (Senouf et al., 2019) simply computes:

$$\mathcal{L}_{\text{MC}} = \ell(\mathbf{A}\hat{\mathbf{x}}, \mathbf{y}), \tag{2}$$

The MC loss cannot recover information from the operator's null-space $\mathcal{N}(\mathbf{A})$ as any solutions $\mathbf{A}^\dagger\mathbf{y} + \mathbf{v}$ trivially satisfy $\mathcal{L}_{\text{MC}} = 0$ where $\mathbf{v} \in \mathcal{N}(\mathbf{A})$. Instead, MC has been used in the presence of strong inductive bias where regularisation is provided by very specific architectures (Darestani and Heckel, 2021; Korkmaz et al., 2022; Tamir et al., 2020) or initialisations (Darestani et al., 2022). However, to disentangle this we focus on comparing the performance of different loss functions while using identical networks: this is a common setting (Yaman et al., 2022b) where MC (Senouf et al., 2019) simply recovers $\mathbf{A}^\dagger\mathbf{y}$ (Pruessmann et al., 1999). Traditional regularising terms (*e.g.* sparsity) (Wang et al., 2020; Alçalar et al., 2024) can always be appended independently to any of the losses used in this paper such as MC, so we do not consider their usage here.

Learning from measurement splitting Methods such as SSDU (Yaman et al., 2020) and similar works (Yaman et al., 2022a; Huang et al., 2024; Klug et al., 2023), randomly split $\mathbf{y}$ into two sets at each instance during training:

$$\mathcal{L}_{\text{SSDU}} = \ell(\mathbf{M}^{(2)}\mathbf{A}f_\theta(\mathbf{M}^{(1)}\mathbf{y}, \mathbf{M}^{(1)}\mathbf{A}), \mathbf{M}^{(2)}\mathbf{y}),\ \hat{\mathbf{x}} = f_\theta(\mathbf{y}, \mathbf{A}) \tag{3}$$

where $\mathbf{M}^{(1)}$ is a randomly generated mask (spanning the whole measurement domain in expectation), $\mathbf{M}^{(1)} + \mathbf{M}^{(2)} = \mathbb{I}_m$, and ℓ is either a metric in the measurement domain and image domain. Methods requiring pairs of measurements of the same subject (Xia and Chakrabarti, 2019; Hu et al., 2024; Gan et al., 2021; Huang et al., 2019; Gan et al., 2022;

Liu et al., 2020) are equivalent to SSDU but with $2\times$ less acceleration. The same concept can be applied to a specific acquisition (Wang et al., 2024). However, for the estimator $\hat{\mathbf{x}}$ to correspond to the supervised MMSE, an inference-time adaptation must be made to average over $J > 1$ passes, which was used in both of Hendriksen et al. (2020); Gruber et al. (2024):

$$\hat{\mathbf{x}} = \frac{1}{J}\Sigma_{j=1}^{J} f_\theta(\mathbf{M}^{(j)}\mathbf{y}, \mathbf{M}^{(j)}\mathbf{A}). \tag{4}$$

Millard and Chiew (2023) bypass this averaging by weighting the SSDU loss metric $\ell(\cdot,\cdot)$ to directly recover the MMSE estimator, and by splitting $\mathbf{M}^{(j)} \sim \mathcal{M}$, where $\mathbf{P} = \mathbb{E}[\mathbf{M}], \tilde{\mathbf{P}} = \mathbb{E}[\mathbf{M}^{(j)}]$ *i.e.* the expectations of the imaging mask and the splitting mask, respectively:

$$\begin{aligned}\mathcal{L}_{\text{Weighted-SSDU}} = \mathcal{L}_{\text{SSDU}}, \quad \ell(a,b) := \|(1-\mathbf{K})^{-\frac{1}{2}}(a-b)\|_2^2,\\ \mathbf{K} = (\mathbb{I}_m - \tilde{\mathbf{P}}\mathbf{P})^{-1}(\mathbb{I}_m - \mathbf{P}).\end{aligned} \tag{5}$$

(Hu et al., 2021; Zhou et al., 2022; Wang et al., 2023b; Xu et al., 2025) split $\mathbf{y}$ with both $\mathbf{M}^{(1)}, \mathbf{M}^{(2)}$, compute a SSDU-style loss for each of these subsets, and also compute a consistency loss between the two reconstructions $\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(2)}$, where $\bar{\mathbf{A}} = (\mathbb{I}_m - \mathbf{M})\mathbf{F}\mathbf{S}$:

$$\begin{aligned}\mathcal{L}_{\text{SSDU-Consistency}} &= \ell(\mathbf{A}\hat{\mathbf{x}}^{(1)}, \mathbf{y}) + \ell(\mathbf{A}\hat{\mathbf{x}}^{(2)}, \mathbf{y})\\ &\quad + \ell(\bar{\mathbf{A}}\hat{\mathbf{x}}^{(1)}, \bar{\mathbf{A}}\hat{\mathbf{x}}^{(2)}),\\ \hat{\mathbf{x}}^{(k)} &= f_\theta(\mathbf{M}^{(k)}\mathbf{y}, \mathbf{M}^{(k)}\mathbf{A}).\end{aligned} \tag{6}$$

Learning from multiple operators These approaches leverage the fact that the image training set is seen through multiple $\mathbf{A}_i \sim \mathcal{A}$, where $\mathcal{A}$ is the operator set of same order as the mask set $|\mathcal{M}| = \binom{n}{m}$. Multi-Operator Imaging (Tachella et al., 2022) leverages this by drawing a random operator $\tilde{\mathbf{A}} \sim \mathcal{A}$ at each forward pass and regularises the MC loss with another loss that encourages consistency of f_θ over $\mathbf{A}_i$, shown to theoretically enable the function to learn in the null-space:

$$\mathcal{L}_{\text{MOI}} = \mathcal{L}_{\text{MC}} + \ell(\hat{\mathbf{x}}, f_\theta(\tilde{\mathbf{A}}\hat{\mathbf{x}}, \tilde{\mathbf{A}})), \ \hat{\mathbf{x}} = f_\theta(\mathbf{y}, \mathbf{A}). \tag{7}$$

(Zhang et al., 2024) constructs a very similar loss to MOI, enforcing multi-operator consistency on the measurement $\mathbf{y}$:

$$\mathcal{L}_{\text{MOC-SSDU}} = \mathcal{L}_{\text{SSDU}} + \ell(\mathbf{y}, \mathbf{A}f_\theta(\tilde{\mathbf{A}}\hat{\mathbf{x}}, \tilde{\mathbf{A}})). \tag{8}$$

Learning from invariance Equivariant Imaging (EI) (Chen et al., 2021, 2022a; Wang and Davies, 2025, 2024) leverages the natural assumption that the image set $\mathcal{X}$ is invariant to a group G of transformations $g \circ \mathbf{x} \in \mathcal{X}, \forall \mathbf{x} \in \mathcal{X}, g \in G$. The image set can then be interpreted as being observed through a set of multiple transformed operators $\mathbf{A} \circ g(\cdot)$ of order $|G|$, allowing the solver to "see" into the null-space. The assumption is constrained using:

$$\mathcal{L}_{\text{EI}} = \mathcal{L}_{\text{MC}} + \ell(\mathbf{T}_g\hat{\mathbf{x}}, f_\theta(\mathbf{A}\mathbf{T}_g\hat{\mathbf{x}}, \mathbf{A})), \ \hat{\mathbf{x}} = f_\theta(\mathbf{y}, \mathbf{A}), \tag{9}$$

where $\mathbf{T}_g$ is an action of G, where, for MRI images, we use 2D rotations for G following Chen et al. (2022a).

Combining two methods Our benchmark makes a new loss function apparent, which we can rapidly implement and test. Since adding more, well-targeted regularisation typically improves the stability and generalisation of the model, and that the operators leveraged by MOI (multiple physical operators) and EI (multiple virtual operators) are complementary, we propose a hybrid Multi-Operator Equivariant Imaging loss that unifies these strategies. Assume that the image set is imaged by the *augmented operator set* $\mathcal{A}_G = \{\mathbf{A}_i\mathbf{T}_g \ \forall \mathbf{A}_i \in \mathcal{A}, g \in G\}$, the orbit space of $\mathcal{A}$ under G. Since $|\mathcal{A}_G| \approx |\mathcal{A}||G|$ is much larger than in MOI or EI, we expect to provide more regularisation and thereby improve the performance. We leverage the augmented operator set using the loss function:

$$\mathcal{L}_{\text{MO-EI}} = \mathcal{L}_{\text{MC}} + \ell(\mathbf{T}_g\hat{\mathbf{x}}, f_\theta(\tilde{\mathbf{A}}_g\hat{\mathbf{x}}, \tilde{\mathbf{A}})), \ \tilde{\mathbf{A}}_g \sim \mathcal{A}_G. \tag{10}$$

For G, since we have soft deformable tissue (Gan et al., 2022), we expect $\mathcal{X}$ be invariant to the group of perturbative C^1-diffeomorphisms. See section A.3 for details.

Another possible combination, was proposed in Equivariant Splitting (Sechaud et al., 2025) that circumvents the problem of measurement splitting methods when the splitting masks do not span the full domain in expectation, by constraining equivariance of f_θ, enlargening the set of operators virtually:

$$\mathcal{L}_{\text{ES}} = \mathbb{E}_g\mathcal{L}_{\text{SSDU}} \quad \text{where} \quad \mathbf{A} \to \mathbf{A}\mathbf{T}_g. \tag{11}$$

A similar combination of the above was proposed in Equivariance2Inverse (Schut et al., 2025), which simply takes a sum of the measurement splitting (eq. (3)) and EI (eq. (9)) losses.

Data augmentation EI is related to data augmentation (Chen et al., 2023), which, when GT $\mathbf{x}$ is available, can be used to constrain invariance or equivariance on $\mathbf{x}$ (Fabian et al., 2021). Data augmentation consistency (DAC) methods (Xie et al., 2020) such as VORTEX (Desai et al., 2022a) build on this in the semi-supervised context by performing data augmentation on $\mathbf{y}$ when GT is not available, chiefly to provide robustness to OOD data:

$$\mathcal{L}_{\text{VORTEX}} = \begin{cases} \mathcal{L}_{\text{sup}}(f_\theta(\mathbf{y}, \mathbf{A}), \mathbf{x}) & \text{if GT exists} \\ \ell(\mathbf{T}_2 f_\theta(y, \mathbf{A}), f_\theta(\mathbf{T}_1\mathbf{y}, \mathbf{A}\mathbf{T}_2^{-1})) & \text{otherwise} \end{cases} \tag{12}$$

where $\mathbf{T}_1, \mathbf{T}_2$ are random transforms in k-space (*e.g.* add noise, phase shift) and image space respectively. We test the ability of DAC methods in the fully GT free case, replacing $\mathcal{L}_{\text{sup}}$ with $\mathcal{L}_{\text{MC}}$ to enforce MC.

Adversarial losses Adversarial losses have been proposed to reconstruct MRI without GT via the dual training of f_θ and a discriminator D. While unconditional, generative adversarial networks (Bora et al., 2018) are not feedforward at inference time, conditional methods using an adversarial loss (Cole et al., 2021):

$$\mathcal{L}_{\text{adversarial}} = D(\tilde{\mathbf{A}} f_\theta(\mathbf{y}, \mathbf{A}), \tilde{\mathbf{y}}), \tag{13}$$

where $\tilde{\mathbf{y}} \sim \mathcal{Y}$ *i.e.* another randomly sampled "real" measurement from the training dataset $\mathcal{Y}$. The aim of D is to distinguish between reconstructed measurements and training measurements, such that the aim of f_θ is to reconstruct high quality measurements that are indistinguishable from "real" measurements. Note that it is unclear how this loss can learn the unknown signal model and recover information from the null-space. UAIR (Pajot et al., 2018) is a related method originally proposed for inpainting, using a multi-step feedforward loss consisting of adversarial and MC terms:

$$\hat{\mathbf{y}} = \tilde{\mathbf{A}} f_\theta(\mathbf{y}, \mathbf{A}),\ \mathcal{L}_{\text{UAIR}} = D(\hat{\mathbf{y}}, \mathbf{y}) + \ell(\tilde{\mathbf{A}} f_\theta(\hat{\mathbf{y}}, \tilde{\mathbf{A}}), \hat{\mathbf{y}}). \tag{14}$$

Joint reconstruction and denoising For Scen. 2 (noisy), we consider composite losses that jointly learn to reconstruct and denoise from partial noisy measurements $\sigma > 0$ alone (note that denoising can always independently be done as a separate preprocessing step (Aali et al., 2024a)). Several self-supervised denoising losses exist in the literature (Tachella et al., 2025a); most combinations of these with reconstruction losses are yet to be explored. Stein's Unbiased Risk Estimator (SURE) (Stein, 1981; Ramani et al., 2008) can be used in place of $\mathcal{L}_{\text{MC}}$, shown to be an unbiased estimator of $\mathcal{L}_{\text{MC}}$ (Chen et al., 2022a). It can be seen as further "recorrupting" data:

$$\mathcal{L}_{\text{SURE}} = \mathcal{L}_{\text{MC}} + \frac{2\sigma^2}{\tau}\mathbf{b}^\top(\mathbf{A} f_\theta(\mathbf{y} + \tau\mathbf{b}, \mathbf{A}) - A f_\theta(\mathbf{y}, \mathbf{A})) - \sigma^2, \tag{15}$$

where $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, τ is a hyperparameter and $\mathbb{E}_{\mathbf{y}}[\mathcal{L}_{\text{SURE}}] = \mathbb{E}_{\mathbf{x},\mathbf{y}}[\ell(\mathbf{A} f_\theta(\mathbf{y}), \mathbf{A}\mathbf{x})]$. Robust-EI (Chen et al., 2022a) then simply use $\mathcal{L}_{\text{Robust-EI}} = \mathcal{L}_{\text{SURE}} + \mathcal{L}_{\text{EI}}$ to train. SURE has been generalised to the case of rank-deficient $\mathbf{A}$ in GSURE (Eldar, 2009) and LDAMP-SURE (Metzler et al., 2020; Zhussip et al., 2019). However, ENSURE (Aggarwal et al., 2023) is shown to improve on these by attempting to leverage multi-operator information of $\mathcal{A}$:

$$\mathcal{L}_{\text{ENSURE}} = \|\mathbf{R}(\mathbf{A}^\dagger\mathbf{y} - f_\theta(\mathbf{y}, \mathbf{A}))\|_2^2 + \frac{2\sigma^2}{\tau}\mathbf{b}^\top\Big(f_\theta(\mathbf{A}^\top\mathbf{y} + \tau\mathbf{b}, \mathbf{A}) - f_\theta(\mathbf{A}^\top\mathbf{y})\Big) \tag{16}$$

where $\mathbf{R} = \mathbb{E}\left[\mathbf{P}\right]^{-\frac{1}{2}}\mathbf{P}, \mathbf{P} = \mathbf{A}^\dagger\mathbf{A}$. Note that when $\sigma \to 0$, ENSURE simply reduces to $\mathcal{L}_{\text{MC}}$.

SSDU-style methods have also been adapted to the noisy case using Noisier2Noise-style recorruption losses (Moran et al., 2019). Robust-SSDU (Millard and Chiew, 2024) constructs a loss attempting to remove an additional recorruption $\tilde{\mathbf{y}} \sim \mathcal{N}(\mathbf{y}, \alpha^2\sigma^2\mathbf{I})$:

$$\mathcal{L}_{\text{Robust-SSDU}} = \mathcal{L}_{\text{Weighted-SSDU}}(\tilde{\mathbf{y}}; \mathbf{y}) + \|(1 + \tfrac{1}{\alpha^2})\mathbf{M}^{(1)}\mathbf{M}(\mathbf{A} f_\theta(\tilde{\mathbf{y}}, \mathbf{A}) - \mathbf{y})\|_2^2 \tag{17}$$

where α is a hyperparameter. Similarly, Noise2Recon-SSDU (Desai et al., 2022b) appends to SSDU a recorruption constructed as a noising special-case of VORTEX (Desai et al., 2022a) $\mathcal{L}_{\text{Noise2Recon-SSDU}} = \mathcal{L}_{\text{SSDU}} + \mathcal{L}_{\text{VORTEX}}$ where $\mathbf{T}_1 = \mathcal{N}, \mathbf{T}_2 = \mathbf{I}$. DDSSL (Quan et al., 2022) also adds a random noise special-case of $\mathcal{L}_{\text{EI},\mathbf{T}=\mathcal{N}}$ to the Recorrupted2Recorrupted loss (Pang et al., 2021):

$$\mathcal{L}_{\text{R2R}} = \|\mathbf{A} f_\theta(\mathbf{y} + \alpha\omega, \mathbf{A}) - \left(\mathbf{y} - \frac{\omega}{\alpha}\right)\|_2^2. \tag{18}$$

Dynamic losses For Scen. 6 (dynamic MRI), we adapt the losses to learn from temporal patterns. For SSDU losses we let $\mathbf{M}_1$ vary randomly in time (Acar et al., 2021), and DDEI (Wang and Davies, 2024) adds temporal symmetries to the EI group $\mathcal{G}$.

2.4 Related work

Surveys for imaging inverse problems Recent surveys exist covering inverse problems with deep learning (Ongie et al., 2020), for MRI reconstruction (Heckel et al., 2024; Chen et al., 2022b) and without GT (Zeng et al., 2021; Akçakaya et al., 2022). While (Chen et al., 2022b; Heckel et al., 2024; Ongie et al., 2020; Akçakaya et al., 2022; Xu et al., 2026) provide valuable general overviews, they discuss only a subset of existing self-supervised methods (*e.g.* zero to five), or methods from only one category mentioned in section 2.3, such as measurement splitting. While GT-free MRI surveys (Zeng et al., 2021; Akçakaya et al.,

2022) highlight the merits of self-supervised imaging, they lack a principled comparison of the core theoretical differences between methods, comparing instead entire pipelines, conflating therefore mostly orthogonal problems including model architecture f_θ or forward operator $\mathbf{A}$. Furthermore, no experimental results are provided by (Heckel et al., 2024; Zeng et al., 2021; Akçakaya et al., 2022). Previous individual methodological works either do not report comparison with other methods *e.g.* (Cole et al., 2021; Chen et al., 2022a), or compare to very few (Millard and Chiew, 2024; Desai et al., 2022a; Aggarwal et al., 2023), limiting their applicability; a comprehensive benchmark comparison is lacking. Papers commonly do not provide public implementation details of competitor methods, potentially reducing the trustworthiness of the comparisons, especially when competitors' codebases are directly used without disentangling the method from the experimental setup $(f_\theta, \mathbf{A})$. On the contrary, we highlight the key conceptual difference, the loss function, and benchmark this while keeping constant all other elements of the imaging pipeline, providing benchmark results across multiple fixed setup scenarios $(f_\theta, \mathbf{A})$. Furthermore, we reimplement the compared methods clearly in a common codebase, bypassing the practical barriers of code uninteroperability, to encourage reproducible research.

Benchmarks for imaging inverse problems Numerous imaging benchmarks on real data have been proposed for tasks such as MRI or CT (Zbontar et al., 2018; Desai et al., 2022c; Kiss et al., 2025; Zheng et al., 2025). However, to the best of our knowledge, in any domain, no benchmarks on image reconstruction from partial measurements without GT exist. Such a benchmark is crucial for enabling fair, reproducible comparison of self-supervised reconstruction methods and driving progress toward reliable GT-free imaging.

3. Experimental setup

We evaluate all table 1 methods using our reimplementations across the various scenarios. Each scenario uses a fixed setup across all methods, and this setup varies across scenarios, including varied models, anatomy, dataset, scanner vendor, acceleration and acquisition noise.

Data For scenarios 1-4 we retrospectively simulate $\mathbf{y}$ from slices of the popular FastMRI brain dataset (Zbontar et al., 2018), in order to fully control the forward operator $\mathbf{A}$ (its distribution is needed for (Yaman et al., 2022a; Millard and Chiew, 2023; Hendriksen et al., 2020; Tachella et al., 2022)) and have GT $\mathbf{x}$ for evaluation. The specific forward operator $\mathbf{A}$ varies for each benchmarked scenario. For the multi-coil case we assume known coil maps to focus on the reconstruction task and not introduce artifacts associated with map estimation. For external data Scenarios 5, 6 & 7, we use respectively: raw multicoil sagittal knee k-space slices from the SKM-TEA (Desai et al., 2022c) test set with estimated coil maps via JSENSE (Ying and Sheng, 2007), raw prospectively-undersampled multicoil axial knee k-space slices from Yu et al. (2022), and cardiac 2D+t sequences from the CMRxRecon (Wang et al., 2023a) test set.

Model f_θ From the vast architectural literature (Heckel et al., 2024) for scenarios 1-4 we choose an unrolled network architecture, which balances being specific for efficient image reconstruction yet sufficiently general to not favour specific losses. We choose the popular MoDL (Aggarwal et al., 2019), unrolling finite (three) steps of half-quadratic splitting (Zhang et al., 2022), with a U-Net backbone (Ronneberger et al., 2015). Identical hyperparameters are used for all loss functions to provide valid comparisons. For Scenarios 5 & 7 we instead fine-tune a physics-informed foundation model architecture (Terris et al., 2025) and for Scenario 6 we train a convolutional recurrent network (Qin et al., 2019).

Evaluation We report PSNR and SSIM as used in FastMRI (Zbontar et al., 2018), as well as LPIPS for perceptual quality (Zhang et al., 2018). We compare to the baseline $\mathbf{A}^\dagger\mathbf{y}$, corresponding to the zero-filled (ZF) in single-coil MRI and SENSE (Pruessmann et al., 1999) in multi-coil MRI (we avoid comparisons to iterative methods (Lustig et al., 2008)). We also train a gold-standard model with the supervised oracle loss $\ell(\hat{\mathbf{x}}, \mathbf{x})$. Further setup details and loss and training code are in section A. Benchmark code and contribution instructions are at github.com/Andrewwango/ssibench.

Extending We highlight that our modular benchmark framework facilitates replacing any element of the setup used here: other SotA architectures, popular image quality metrics, or forward operator scenarios. To evidence ease of benchmark extensibility, further experiments investigate varying each benchmark module: pilot results on SotA GT-free pretrained diffusion models, testing performance scaling as unrolling depth increases or on different architectures (VarNet (Hammernik et al., 2018), a flat CNN (Zhang et al., 2017), and transformer (Zamir et al., 2022)), component ablations on MO-EI and SSDU, and ease of transferability to a sparse-view CT task and an environmental imaging task.

3.1 Results

Scenario 1 (single-coil) Results are shown in table 2 and fig. 2. As expected, MC cannot improve the ZF performance, as eq. (2) cannot learn information from the null-space; results presented in (Senouf et al., 2019) may instead heavily rely on a specific network's inductive bias. EI and MOI both show good results but suffer from strong ar-

Table 2: Benchmarking test set results ($\mu \pm \sigma$) for scenarios 1, 3, 4 and 5. **Best unsupervised method in bold**.

Scenario	**(1)** Brain MRI 6× acc.			**(3)** Single-mask		**(4)** Multicoil $C = 4$		**(5)** SKM-TEA knee fine-tune 32×		
Loss	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	LPIPS↓
Zero-shot	–	–	–	–	–	–	–	$26.20_{\pm.54}$	$.5081_{\pm.0226}$	$.5273_{\pm.0379}$
Zero-filled	$27.67_{\pm2.40}$	$.7862_{\pm.05}$	$.3270_{\pm.0426}$	$28.02_{\pm2.26}$	$.7900_{\pm.05}$	$27.82_{\pm2.41}$	$.7988_{\pm.05}$	$25.29_{\pm.50}$	$.5896_{\pm.02}$	$.4913_{\pm.0257}$
MC	$27.66_{\pm2.40}$	$.7858_{\pm.0550}$	$.3273_{\pm.0426}$	$28.02_{\pm2.26}$	$.7900_{\pm.0496}$	$28.96_{\pm2.59}$	$.8271_{\pm.0495}$	$27.75_{\pm.65}$	$\mathbf{.5965}_{\pm.0277}$	$\mathbf{.2713}_{\pm.0235}$
SSDU	$27.98_{\pm1.43}$	$.7485_{\pm.0667}$	$.1965_{\pm.0417}$	$21.89_{\pm1.07}$	$.6288_{\pm.0510}$	$31.47_{\pm2.67}$	$.8705_{\pm.0534}$	$28.17_{\pm.62}$	$.5800_{\pm.0174}$	$.4265_{\pm.0371}$
Noise2Inverse	$28.42_{\pm2.02}$	$.7853_{\pm.0735}$	$.2347_{\pm.0416}$	$24.63_{\pm2.07}$	$.6559_{\pm.0676}$	$30.93_{\pm2.65}$	$.8589_{\pm.0570}$	$27.45_{\pm.64}$	$.5465_{\pm.0175}$	$.4509_{\pm.0367}$
Weighted-SSDU	$29.93_{\pm1.66}$	$.8355_{\pm.0626}$	$\mathbf{.1304}_{\pm.0434}$	$30.14_{\pm1.64}$	$.8454_{\pm.0615}$	$\mathbf{33.03}_{\pm2.36}$	$.8991_{\pm.0479}$	$\mathbf{28.83}_{\pm.60}$	$.5901_{\pm.0180}$	$.3964_{\pm.0394}$
SSDU-Consistency	$30.81_{\pm2.58}$	$.8495_{\pm.0581}$	$.1845_{\pm.0370}$	$31.05_{\pm2.50}$	$.8614_{\pm.0567}$	$32.30_{\pm2.58}$	$.8949_{\pm.0425}$	$28.18_{\pm.61}$	$.5900_{\pm.0167}$	$.4088_{\pm.0304}$
MOC-SSDU	$30.43_{\pm2.71}$	$.8198_{\pm.0854}$	$.1578_{\pm.0568}$	$27.85_{\pm1.86}$	$.7717_{\pm.0833}$	$31.80_{\pm2.69}$	$.8761_{\pm.0529}$	$28.28_{\pm.61}$	$.5823_{\pm.0170}$	$.4238_{\pm.0375}$
Adversarial	$18.52^{*}_{\pm.31}$	$.4732_{\pm.0388}$	$.3927_{\pm.0413}$	$26.53_{\pm1.62}$	$.7013_{\pm.0370}$	$17.47^{*}_{\pm1.93}$	$.6464_{\pm.0590}$	$15.60^{*}_{\pm1.06}$	$.3355_{\pm.0268}$	$.6608_{\pm.0410}$
UAIR	$14.00^{*}_{\pm1.88}$	$.3715_{\pm.0824}$	$.4826_{\pm.0584}$	$18.44^{*}_{\pm1.61}$	$.5388_{\pm.0542}$	$15.26^{*}_{\pm2.16}$	$.3453_{\pm.0540}$	$16.07^{*}_{\pm.40}$	$.3902_{\pm.0231}$	$.7461_{\pm.0778}$
VORTEX	$27.75_{\pm2.35}$	$.7898_{\pm.0543}$	$.3201_{\pm.0405}$	$28.07_{\pm2.26}$	$.7916_{\pm.0507}$	$23.59_{\pm1.17}$	$.5846_{\pm.0469}$	$25.01_{\pm1.07}$	$.5766_{\pm.0291}$	$.5517_{\pm.0508}$
ES	$29.80_{\pm2.09}$	$.8105_{\pm.0591}$	$.3309_{\pm.0508}$	$30.26_{\pm1.97}$	$.8242_{\pm.0644}$	$30.82_{\pm2.48}$	$.8370_{\pm.0653}$	$25.96_{\pm.90}$	$.5150_{\pm.0179}$	$.4354_{\pm.0308}$
E2I	$29.92_{\pm1.89}$	$.8122_{\pm.0587}$	$.2972_{\pm.0516}$	$25.45_{\pm1.10}$	$.6760_{\pm.0435}$	$32.12_{\pm2.67}$	$\mathbf{.9116}_{\pm.0590}$	$27.68_{\pm.61}$	$.5500_{\pm.0176}$	$.4696_{\pm.0329}$
EI	$30.26_{\pm2.61}$	$.8523_{\pm.0542}$	$.2429_{\pm.0423}$	$\mathbf{31.99}_{\pm2.80}$	$\mathbf{.8806}_{\pm.0486}$	$31.66_{\pm2.74}$	$.8769_{\pm.0494}$	$25.18_{\pm.56}$	$.4676_{\pm.0194}$	$.5952_{\pm.0342}$
MOI	$30.29_{\pm2.88}$	$.8651_{\pm.0528}$	$.2341_{\pm.0472}$	$31.60_{\pm2.74}$	$.8789_{\pm.0477}$	$31.37_{\pm2.86}$	$.8810_{\pm.0502}$	$25.58_{\pm.57}$	$.4801_{\pm.0199}$	$.5742_{\pm.0354}$
MO-EI	$\mathbf{32.14}_{\pm\mathbf{2.73}}$	$\mathbf{.8846}_{\pm.0498}$	$.1803_{\pm.0406}$	$31.11_{\pm2.69}$	$.8713_{\pm.0496}$	$31.56_{\pm2.85}$	$.8836_{\pm.0477}$	$25.41_{\pm.57}$	$.4756_{\pm.0200}$	$.5841_{\pm.0374}$
(Supervised)	$33.15_{\pm2.76}$	$.9032_{\pm.0435}$	$.1185_{\pm.0335}$	$34.03_{\pm2.49}$	$.9040_{\pm.0435}$	$33.89_{\pm2.78}$	$.9147_{\pm.0365}$	$29.43_{\pm.59}$	$.6203_{\pm.0142}$	$.3259_{\pm.0382}$

Table 3: **Scenario 2** (joint denoising & recon) test results.

Loss	PSNR ↑	SSIM ↑	LPIPS↓
Zero-filled	$24.34_{\pm1.01}$	$.4428_{\pm.03}$	$.4623_{\pm.0530}$
ENSURE	$26.29_{\pm1.51}$	$.5856_{\pm.0360}$	$.3117_{\pm.0474}$
Robust-SSDU	$27.42_{\pm1.13}$	$.6159_{\pm.0425}$	$\mathbf{.3040}_{\pm.0589}$
Noise2Recon-SSDU	$27.84_{\pm1.78}$	$.7661_{\pm.0727}$	$.3267_{\pm.0601}$
DDSSL	$28.25_{\pm2.30}$	$.7836_{\pm.0548}$	$.3786_{\pm.0652}$
ES-R2R	$28.14_{\pm1.51}$	$.6377_{\pm.0381}$	$.3782_{\pm.0462}$
E2I	$25.80_{\pm.80}$	$.4668_{\pm.0471}$	$.4681_{\pm.0540}$
Robust-EI	$29.07_{\pm2.37}$	$.8227_{\pm.0578}$	$.3405_{\pm.0621}$
Robust-MO-EI	$\mathbf{29.72}_{\pm2.44}$	$\mathbf{.8409}_{\pm.0575}$	$.3311_{\pm.0614}$
Non-robust MO-EI	$26.12_{\pm1.92}$	$.6002_{\pm.0602}$	$.3865_{\pm.0518}$
Non-robust Weighted-SSDU	$25.91_{\pm.94}$	$.5477_{\pm.0511}$	$.3174_{\pm.0734}$
(Supervised)	$30.19_{\pm2.10}$	$.8411_{\pm.0552}$	$.2532_{\pm.0562}$

Table 4: **Scenario 6** ($10\times$ acc. dynamic cardiac MRI).

Loss	PSNR ↑	SSIM ↑	LPIPS↓
Zero-filled	$30.41_{\pm1.07}$	$.8041_{\pm.03}$	$.2137_{\pm.0248}$
MC	$30.41_{\pm1.07}$	$.8041_{\pm.0299}$	$.2137_{\pm.0248}$
t-SSDU	$30.18_{\pm1.38}$	$.8212_{\pm.0272}$	$.1992_{\pm.0272}$
Weighted-t-SSDU	$29.65_{\pm1.56}$	$.7836_{\pm.0290}$	$.1901_{\pm.0274}$
t-SSDU-Consistency	$31.82_{\pm1.15}$	$.8477_{\pm.0236}$	$.1714_{\pm.0225}$
MOC-t-SSDU	$\mathbf{31.88}_{\pm1.19}$	$.8485_{\pm.0235}$	$\mathbf{.1572}_{\pm.0192}$
Diffeo-EI	$31.81_{\pm1.39}$	$\mathbf{.8548}_{\pm.0244}$	$.1691_{\pm.0247}$
EI	$31.22_{\pm1.20}$	$.8431_{\pm.0247}$	$.1931_{\pm.0246}$
MOI	$31.47_{\pm1.25}$	$.8471_{\pm.0239}$	$.1938_{\pm.0259}$
DDEI	$31.36_{\pm1.17}$	$.8450_{\pm.0248}$	$.1853_{\pm.0245}$
(Supervised)	$33.95_{\pm1.75}$	$.8750_{\pm.0244}$	$.1038_{\pm.0227}$

tifacts in the brain images, but MOI slightly outperforms EI because its set of virtual operators in eq. (7) is likely larger than that of eq. (9), so that the per-sample null-space is better covered per training step. MO-EI improves significantly on both these previous SotA methods (see section B.1 for statistical test), approaching the oracle supervised learning performance, as eq. (10) is able to recover information from the null-space via a larger set of virtual operators than that of EI or MOI. SSDU suppresses artifacts and recovers sharper edges quite well, but its quantitative performance is relatively poor, because eq. (3) constructs an estimator conditioned on a small subset of the available measurements. Even though the loss is equal to the supervised loss associated with this masked subset in expectation, it is inefficient and thus may perform worse in finite data settings, and the estimator $f_\theta(\mathbf{y}, \mathbf{A})$ over-estimates the MMSE estimator at inference-time. Other methods based on splitting improve on this, as they recover the MMSE at inference-time. We attribute the very poor performance of the adversarial losses (marked with *) to the well-known instability and sensitivity of training f_θ and D (Salimans et al., 2016): after extensive experimentation we were unable to reproduce the results from (Cole et al., 2021) (which use a proprietary dataset), which may require heavy hyperparameter tuning and specific architectures. Since UAIR (Pajot et al., 2018) does not show results on accelerated MRI, we assume their results do not transfer, and further work is needed to find better theoretically-justified adversarial implementation choices to reach stronger conclusions for adversarial methods. VORTEX learns only marginal information compared to MC, showing that measurement-domain data augmentations alone cannot recover information from the null-space.

Scenario 2 (noisy) Results are shown in table 3 and figure in fig. 9. ENSURE denoises a moderate amount, but strong undersampling artifacts remain. Noise2Recon-SSDU and Robust-SSDU perform fairly well, resulting in clearer reconstructions with fairly sharp edges. The SURE Robust-EI/MO-EI methods are almost artifact-free with high metrics, because of the combination of null-space loss and SURE eq. (15) which approximates the clean supervised loss in expectation, but edges are less sharp due to the averaging. All loss functions without the denoising terms perform poorly as expected.

Scenario 3 (single-mask) Results are shown in table 2 and figure in fig. 10. As expected, the best-performing method is EI, which recovers information from the null-space

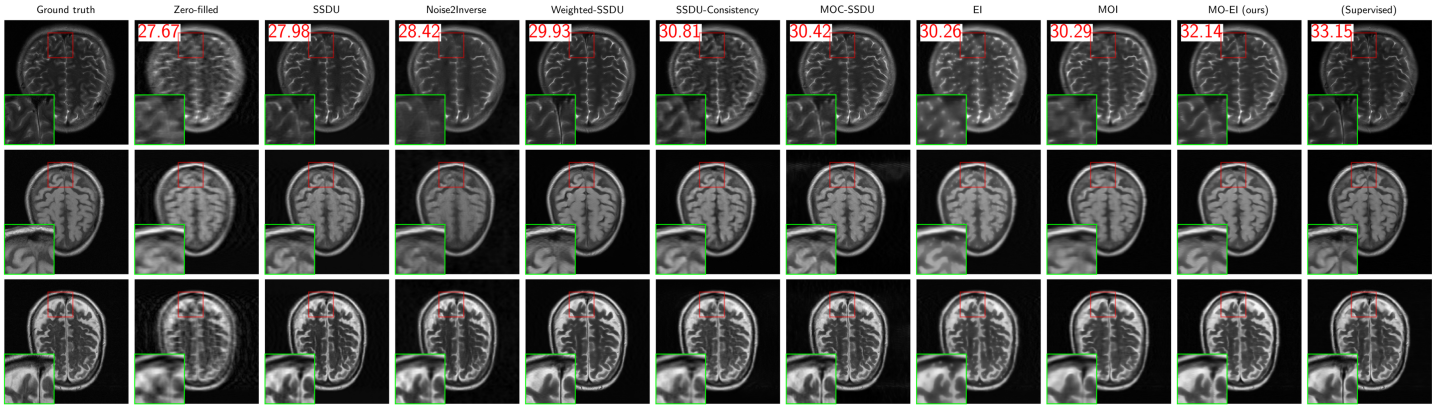

Figure 2: $6\times$ acc. (scenario 1) recons with test-set PSNR, showing selected highest-performing methods. See table 2 for more.

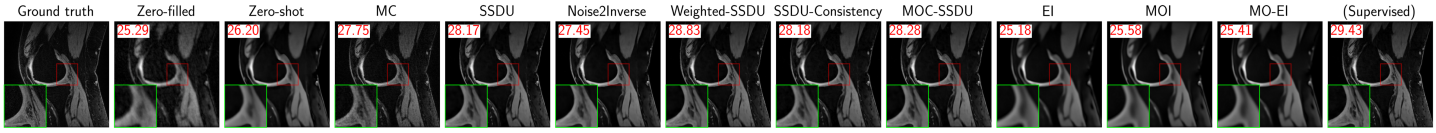

Figure 3: $32\times$ acc. SKM-TEA (Desai et al., 2022c) fine-tuning (scenario 5) recons, showing selected highest-performing methods. See table 2 for more.

but is the only loss *without* the assumption that the image set is seen through multiple masks. Other methods which do require this assumption still learn some information over the ZF solution, demonstrating the challenge of disentangling actual learning from network inductive bias.

Scenario 4 (multi-coil) Results are shown in table 2 and figure in fig. 11. As expected, MC learns the SENSE $\mathbf{A}^\dagger\mathbf{y}$, as it constructs an estimator that uses only the data fidelity term that is used in the pseudo-inverse. Interestingly, splitting-based methods consistently outpeform EI, MOI and MO-EI, with weighted-SSDU approaching supervised learning. While in previous scenarios, splitting methods struggled to recover information from larger null-spaces, here $\mathbf{A}$ has a higher effective rank (analysis in section B.2), which highlights instead their strength on sharp edges.

Scenario 5 (fine-tuning) Results are shown in table 2 and fig. 3. The zero-shot base foundation model (Terris et al., 2025), trained with supervised learning, produces out-of-domain, cartoonish artifacts in the knees (fig. 3, 3$^{\text{rd}}$ image). Weighted-SSDU recovers much of the original sharp detail, as its loss eq. (5) is equal in expectation to the supervised loss. However, EI/MOI null-space losses now *do not* improve the performance, since they cannot recover any more information from the null-space than that recovered by the well-trained MMSE foundation model. Interestingly, the foundation model's inductive bias allows pure MC to improve performance, although reconstructions remain noisy.

Scenario 6 (dynamic) Results are shown in table 4 and figure in fig. 12. As before, EI/MOI-style losses recover information from the null-space and remove $10\times$ undersampling artifacts. The addition of the diffeo transform for EI improves the results since the cardiac sequence displays strong deformable frame self-similarity.

Scenario 7 (prospective) Axial reconstructions are shown in fig. 4; since no GT is present, we do not report quantitative results. We observe that the best methods recover sharp details in the knees, such as in the patellar cartilage or the muscle fascia, and remove the residual noise compared to the zero-filled reconstructions; since the noise level is relatively high, the SENSE reconstruction performs poorly. Similar to Scenario 5, only Weighted-SSDU recovers the fine cartilage detail, whereas the EI/MOI methods catastrophically fail in fine-tuning.

Further results Further experiments and ablations are reported in section B. **Ablation**: best performing implementation choices correspond to the theoretically optimal choices, while performance, and thus our conclusions, are sensitive to suboptimal choices. **Diffusion** (table 8): posterior sampling methods achieve good perceptual but poorer distortion performance compared to MMSE estimators as expected (these models were *pretrained on much larger datasets* so results should not be directly compared). **Unrolling**: performances scale uniformly with unrolling depth, suggesting that our results would scale to setups with deeper models. **Model architecture**: with either a $50\times$ smaller architecture and flat CNN, or shallower transformer-based architecture, Scenario 1 performance decreases but ranking

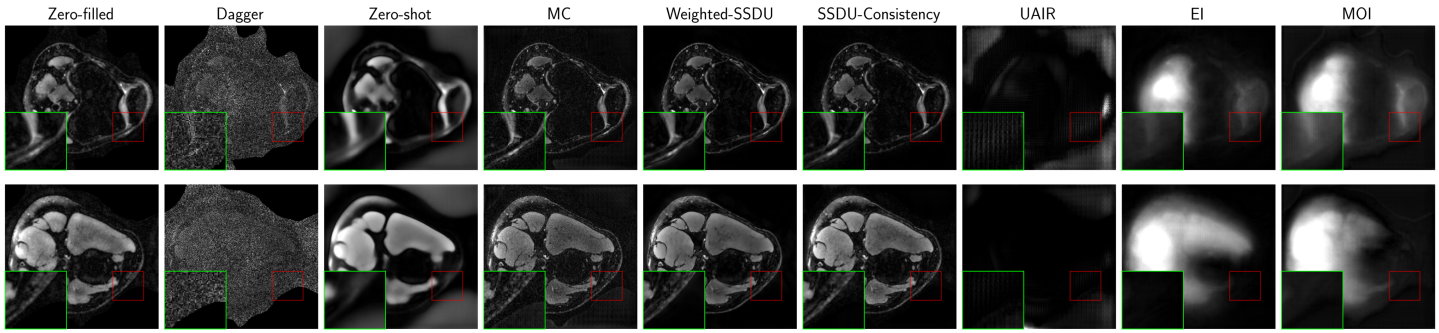

Figure 4: 7× acc. prospectively undersampled knees (Yu et al., 2022) (scenario 7) : fine-tuning recons, showing selected highest-performing methods.

is broadly preserved, suggesting that loss choice is orthogonal to architecture design for this scenario.

4. Discussion and Conclusions

Design We close an important gap in self-supervised imaging (SSI) for MRI reconstruction without GT by designing a modular and comprehensive benchmark, **SSIBench**, that unifies 21 distinct end-to-end training losses.This systematic comparison is crucial for targeted future research and industrial adoption. We formulate seven realistic MRI scenarios that each test different capabilities of methods in diverse acquisition settings, architectures and anatomies.

Insights Our extensive experiments on standardised setups show a wide range in performance, with different methods leading (*e.g.* by 2 dB) on different scenarios and metrics. A possible decision framework for practitioners, based on the results of the tested scenarios, is:

- Highly ill-posed ($\alpha \gg C$): use MO-EI, MOI or EI;
- Multicoil ($\alpha \lesssim C$): use weighted-SSDU;
- Fine-tuning pretrained models: use weighted-SSDU;
- Sampling pattern is fixed for all patients: use EI;
- High acquisition noise: use above losses with appended denoising terms such as SURE.

These observations motivate further ML research by exposing the ongoing need for a reliable method that can lead on all scenarios. The benchmark facilitates the systematic combination and prototyping of loss components to discover more effective configurations such as MO-EI, paving the way for future exploration of even more promising ones.

Although one might suspect that fixing the model hides architectural bias, further experiments (section B) suggests that Scenario 1 performance trends do generalise across model architectures and sizes and their choice is thus orthogonal to the loss function for this scenario.

Limitations In future we aim to include a wider range of datasets *e.g.* raw noisy low-field data (Lyu et al., 2023), other forward operators $\mathbf{A}$ *e.g.* non-Cartesian sampling and more complex models f_θ (Tanabene et al., 2024), a larger set of (Desai et al., 2022c; Breger et al., 2025), or other realistic MRI subtasks *e.g.* joint coil map estimation (Hu et al., 2024). Toward further expansion on these fronts, our modular setup facilitates swapping out any of these components to accommodate diverse research configurations. We demonstrate this capability by extending the benchmark framework beyond MRI and provide a proof-of-concept example benchmarking SSI for a task in environmental imaging and CT in section B.7.

Furthermore, while end-to-end trained methods do not require large datasets, long inference times nor per-image training, and thus favour clinical adoption, SSIBench is limited by not including other promising methods that today are not yet clinically feasible, but may soon become so. Future work thus would add these nascent methods, such as few-step diffusion models or generalisable implicit neural representations.

We note that research in self-supervised imaging is an active and fast-moving field. Therefore, the paper inevitably presents a snapshot of methods taken at the time of writing. Crucially, the creation of the modular open-source benchmark allows future methods to be added and extended as new methods become practical.

Benchmark resource and outlook Our benchmark site allows researchers to easily use reimplementations of all losses and contribute. This lowers the barrier to entry and enables ML researchers and practitioners to a) contribute and compare new ML methods on a standard evaluation framework, and b) rapidly and fairly test these methods on their own setups. We encourage the community to generalise our insights and use our benchmark resource as a blueprint to unlock self-supervised learning for emerging, exciting domains within MRI and beyond, while remaining mindful of potential misuse of high quality images.

Acknowledgments

This work was supported by the School of Engineering at the University of Edinburgh.

Ethical Standards

This study was conducted retrospectively using human subject data made available in open access by Zbontar et al. (2018); Desai et al. (2022c); Wang et al. (2023a); Yu et al. (2022). Ethical approval was not required as confirmed by the licenses attached with the open access data.

Conflicts of Interest

The authors have no relevant financial or non-financial interests to disclose.

Data availability

The data supporting the findings of this study are openly available from Zbontar et al. (2018); Desai et al. (2022c); Wang et al. (2023a); Yu et al. (2022), subject to their respective licenses. The code is available at `https://github.com/Andrewwango/ssibench` and in DeepInverse (Tachella et al., 2025b) at `http://deepinv.org`.

References

Asad Aali, Marius Arvinte, Sidharth Kumar, Yamin I. Arefeen, and Jonathan I. Tamir. GSURE Denoising enables training of higher quality generative priors for accelerated Multi-Coil MRI Reconstruction. *ISMRM 2024*, April 2024a.

Asad Aali, Giannis Daras, Brett Levac, Sidharth Kumar, Alex Dimakis, and Jon Tamir. Ambient Diffusion Posterior Sampling: Solving Inverse Problems with Diffusion Models Trained on Corrupted Data. In *International Conference on Learning Representations*, October 2024b.

Mert Acar, Tolga Çukur, and Ilkay Öksüz. Self-supervised Dynamic MRI Reconstruction. In *Machine Learning for Medical Image Reconstruction*, pages 35–44. Springer International Publishing, 2021. .

Hemant K. Aggarwal, Merry P. Mani, and Mathews Jacob. MoDL: Model-Based Deep Learning Architecture for Inverse Problems. *IEEE Transactions on Medical Imaging*, 38(2):394–405, February 2019. .

Hemant Kumar Aggarwal, Aniket Pramanik, Maneesh John, and Mathews Jacob. ENSURE: A General Approach for Unsupervised Training of Deep Image Reconstruction Algorithms. *IEEE Transactions on Medical Imaging*, 42 (4):1133–1144, April 2023. .

Mehmet Akçakaya, Burhaneddin Yaman, Hyungjin Chung, and Jong Chul Ye. Unsupervised Deep Learning Methods for Biological Image Reconstruction and Enhancement: An overview from a signal processing perspective. *IEEE Signal Processing Magazine*, 39(2):28–44, March 2022. .

Yaşar Utku Alçalar, Merve Gülle, and Mehmet Akçakaya. A Convex Compressibility-Inspired Unsupervised Loss Function for Physics-Driven Deep Learning Reconstruction. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, May 2024. .

Yochai Blau and Tomer Michaeli. The Perception-Distortion Tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018.

Theo Bodrito, Alexandre Zouaoui, Jocelyn Chanussot, and Julien Mairal. A Trainable Spectral-Spatial Sparse Coding Model for Hyperspectral Image Restoration. In *Advances in Neural Information Processing Systems*, volume 34, pages 5430–5442, 2021.

Ashish Bora, Eric Price, and Alexandros G. Dimakis. AmbientGAN: Generative models from lossy measurements. In *International Conference on Learning Representations*, February 2018.

Anna Breger, Ander Biguri, Malena Sabaté Landman, Ian Selby, Nicole Amberg, Elisabeth Brunner, Janek Gröhl, Sepideh Hatamikia, Clemens Karner, Lipeng Ning, Sören Dittmer, Michael Roberts, AIX-COVNET Collaboration, and Carola-Bibiane Schönlieb. A study of why we need to reassess full reference image quality assessment with medical images. *Journal of Imaging Informatics in Medicine*, March 2025. .

Akshay S. Chaudhari, Christopher M. Sandino, Elizabeth K. Cole, David B. Larson, Garry E. Gold, Shreyas S. Vasanawala, Matthew P. Lungren, Brian A. Hargreaves, and Curtis P. Langlotz. Prospective Deployment of Deep Learning in MRI: A Framework for Important Considerations, Challenges, and Recommendations for Best Practices. *Journal of Magnetic Resonance Imaging*, 54 (2):357–371, 2021. .

Dongdong Chen, Julián Tachella, and Mike E. Davies. Equivariant Imaging: Learning Beyond the Range Space. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. .

Dongdong Chen, Julián Tachella, and Mike E. Davies. Robust Equivariant Imaging: a fully unsupervised framework

for learning to image from noisy and partial measurements. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022a. .

Dongdong Chen, Mike Davies, Matthias J. Ehrhardt, Carola-Bibiane Schönlieb, Ferdia Sherry, and Julián Tachella. Imaging With Equivariant Deep Learning: From unrolled network design to fully unsupervised learning. *IEEE Signal Processing Magazine*, 40(1):134–147, January 2023. .

Yutong Chen, Carola-Bibiane Schönlieb, Pietro Liò, Tim Leiner, Pier Luigi Dragotti, Ge Wang, Daniel Rueckert, David Firmin, and Guang Yang. AI-Based Reconstruction for Fast MRI—A Systematic Review and Meta-Analysis. *Proceedings of the IEEE*, 110(2):224–245, February 2022b. .

Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated MRI. *Medical Image Analysis*, 80: 102479, August 2022. .

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *The Eleventh International Conference on Learning Representations*, September 2022.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26(6):1045–1057, December 2013. .

Elizabeth K. Cole, Frank Ong, Shreyas S. Vasanawala, and John M. Pauly. Fast Unsupervised MRI Reconstruction Without Fully-Sampled Ground Truth Data Using Generative Adversarial Networks. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3971–3980, October 2021. .

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alexandros G. Dimakis, and Adam Klivans. Ambient Diffusion: Learning Clean Distributions from Corrupted Data, May 2023. arXiv:2305.19256 [cs, math].

Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G. Dimakis, and Mauricio Delbracio. A Survey on Diffusion Models for Inverse Problems, September 2024. arXiv:2410.00083 [cs].

Mohammad Zalbagi Darestani and Reinhard Heckel. Accelerated MRI with Un-trained Neural Networks, April 2021. arXiv:2007.02471 [eess].

Mohammad Zalbagi Darestani, Jiayu Liu, and Reinhard Heckel. Test-Time Training Can Close the Natural Distribution Shift Performance Gap in Deep Learning Based Compressed Sensing. In *Proceedings of the 39th International Conference on Machine Learning*, pages 4754–4776. PMLR, June 2022.

Arjun D. Desai, Beliz Gunel, Batu M. Ozturkler, Harris Beg, Shreyas Vasanawala, Brian A. Hargreaves, Christopher Ré, John M. Pauly, and Akshay S. Chaudhari. VORTEX: Physics-Driven Data Augmentations Using Consistency Training for Robust Accelerated MRI Reconstruction, June 2022a. arXiv:2111.02549 [eess].

Arjun D. Desai, Batu M. Ozturkler, Christopher M. Sandino, Robert Boutin, Marc Willis, Shreyas Vasanawala, Brian A. Hargreaves, Christopher M. Ré, John M. Pauly, and Akshay S. Chaudhari. Noise2Recon: Enabling Joint MRI Reconstruction and Denoising with Semi-Supervised and Self-Supervised Learning, October 2022b. arXiv:2110.00075 [eess].

Arjun D. Desai, Andrew M. Schmidt, Elka B. Rubin, Christopher M. Sandino, Marianne S. Black, Valentina Mazzoli, Kathryn J. Stevens, Robert Boutin, Christopher Ré, Garry E. Gold, Brian A. Hargreaves, and Akshay S. Chaudhari. SKM-TEA: A Dataset for Accelerated MRI Reconstruction with Dense Image Labels for Quantitative Clinical Evaluation, March 2022c. arXiv:2203.06823 [eess].

Yonina C. Eldar. Generalized SURE for Exponential Families: Applications to Regularization. *IEEE Transactions on Signal Processing*, 57(2):471–481, February 2009. .

Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. Self-Supervised Representation Learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, May 2022. .

Zalan Fabian, Reinhard Heckel, and Mahdi Soltanolkotabi. Data augmentation for deep learning based accelerated MRI reconstruction with limited data. In *Proceedings of the 38th International Conference on Machine Learning*, pages 3057–3067, July 2021.

Oren Freifeld, Søren Hauberg, Kayhan Batmanghelich, and Jonn W. Fisher. Transformations Based on Continuous Piecewise-Affine Velocity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2496–2509, December 2017. .

Weijie Gan, Yu Sun, Cihat Eldeniz, Jiaming Liu, Hongyu An, and Ulugbek S. Kamilov. Deep Image Reconstruction Using Unregistered Measurements Without Groundtruth. In

2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pages 1531–1534, April 2021. .

Weijie Gan, Yu Sun, Cihat Eldeniz, Jiaming Liu, Hongyu An, and Ulugbek S. Kamilov. Deformation-Compensated Learning for Image Reconstruction Without Ground Truth. *IEEE Transactions on Medical Imaging*, 41(9):2371–2384, September 2022. ISSN 1558-254X. .

Nadja Gruber, Johannes Schwab, Elke Gizewski, and Markus Haltmeier. Sparse2Inverse: Self-supervised inversion of sparse-view CT data, February 2024. URL http://arxiv.org/abs/2402.16921. arXiv:2402.16921 [eess].

Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P. Recht, Daniel K. Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated MRI data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018. .

Zhuonan He, Cong Quan, Siyuan Wang, Yuanzheng Zhu, Minghui Zhang, Yanjie Zhu, Qiegen Liu, Zhuonan He, Cong Quan, Siyuan Wang, Yuanzheng Zhu, Minghui Zhang, Yanjie Zhu, and Qiegen Liu. A Comparative Study of Unsupervised Deep Learning Methods for MRI Reconstruction. *Investigative Magnetic Resonance Imaging*, pages 179–195, 2020.

Reinhard Heckel, Mathews Jacob, Akshay Chaudhari, Or Perlman, and Efrat Shimron. Deep learning for accelerated and robust MRI reconstruction. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 37(3): 335–368, July 2024. .

Allard Adriaan Hendriksen, Daniël Maria Pelt, and K. Joost Batenburg. Noise2Inverse: Self-Supervised Deep Convolutional Denoising for Tomography. *IEEE Transactions on Computational Imaging*, 6:1320–1335, 2020. .

Chen Hu, Cheng Li, Haifeng Wang, Qiegen Liu, Hairong Zheng, and Shanshan Wang. Self-supervised Learning for MRI Reconstruction with a Parallel Network Training Framework. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 382–391, 2021. .

Yuyang Hu, Weijie Gan, Chunwei Ying, Tongyao Wang, Cihat Eldeniz, Jiaming Liu, Yasheng Chen, Hongyu An, and Ulugbek S. Kamilov. SPICER: Self-Supervised Learning for MRI with Automatic Coil Sensitivity Estimation and Reconstruction, June 2024. arXiv:2210.02584 [eess].

Peizhou Huang, Chaoyi Zhang, Hongyu Li, Sunil Kumar Gaire, Ruiying Liu, Xiaoliang Zhang, Xiaojuan Li, and Leslie Ying. Deep MRI Reconstruction without Ground Truth for Training. In *Proc. Intl. Soc. Mag. Reson. Med. 27*, 2019.

Peizhou Huang, Chaoyi Zhang, Xiaoliang Zhang, Xiaojuan Li, Liang Dong, and Leslie Ying. Self-Supervised Deep Unrolled Reconstruction Using Regularization by Denoising. *IEEE Transactions on Medical Imaging*, 43(3):1203–1213, March 2024. .

Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon Tamir. Robust Compressed Sensing MRI with Deep Generative Priors. In *Advances in Neural Information Processing Systems*, 2021.

Nikola Janjušević, Jingjia Chen, Luke Ginocchio, Mary Bruno, Yuhui Huang, Yao Wang, Hersh Chandarana, and Li Feng. Self-Supervised Noise Adaptive MRI Denoising via Repetition to Repetition (Rep2Rep) Learning, April 2025. arXiv:2504.17698 [eess].

Kyong Hwan Jin, Michael T. McCann, Emmanuel Froustey, and Michael Unser. Deep Convolutional Neural Network for Inverse Problems in Imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, September 2017. .

Jinho Joo, Hyeseong Kim, Hyeyeon Won, Deukhee Lee, Taejoon Eo, and Dosik Hwang. AeSPa : Attention-guided Self-supervised Parallel Imaging for MRI Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the Design Space of Diffusion-Based Generative Models. *Advances in Neural Information Processing Systems*, 35:26565–26577, December 2022.

Bahjat Kawar, Noam Elata, Tomer Michaeli, and Michael Elad. GSURE-Based Diffusion Model Training with Corrupted Data, June 2024. arXiv:2305.13128 [eess].

Maximilian B. Kiss, Ander Biguri, Zakhar Shumaylov, Ferdia Sherry, K. Joost Batenburg, Carola-Bibiane Schönlieb, and Felix Lucka. Benchmarking learned algorithms for computed tomography image reconstruction tasks. *Applied Mathematics for Modern Challenges*, 3(0):1–43, February 2025. .

Tobit Klug, Dogukan Atik, and Reinhard Heckel. Analyzing the Sample Complexity of Self-Supervised Image Reconstruction Methods. *Advances in Neural Information Processing Systems*, 36:65869–65893, December 2023.

Yilmaz Korkmaz, Salman U. H. Dar, Mahmut Yurt, Muzaffer Özbey, and Tolga Çukur. Unsupervised MRI Reconstruction via Zero-Shot Learned Adversarial Transformers.

IEEE Transactions on Medical Imaging, 41(7):1747–1763, July 2022. .

Ke Lei, Morteza Mardani, John M. Pauly, and Shreyas S. Vasanawala. Wasserstein GANs for MR Imaging: From Paired to Unpaired Training. *IEEE Transactions on Medical Imaging*, 40(1):105–115, January 2021. .

Bowen Li, Zhiwen Wang, Ziyuan Yang, Wenjun Xia, and Yi Zhang. Progressive dual-domain-transfer cycleGAN for unsupervised MRI reconstruction. *Neurocomputing*, 563:126934, January 2024. .

Jiaming Liu, Yu Sun, Cihat Eldeniz, Weijie Gan, Hongyu An, and Ulugbek S. Kamilov. RARE: Image Reconstruction Using Deep Priors Learned Without Groundtruth. *IEEE Journal of Selected Topics in Signal Processing*, 14(6): 1088–1099, October 2020. .

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs Created Equal? A Large-Scale Study. In *Advances in Neural Information Processing Systems*, 2018.

Xinzhe Luo, Yingzhen Li, and Chen Qin. Unsupervised Accelerated MRI Reconstruction via Ground-Truth-Free Flow Matching, February 2025. arXiv:2502.17174 [eess].

Michael Lustig, David L. Donoho, Juan M. Santos, and John M. Pauly. Compressed Sensing MRI. *IEEE Signal Processing Magazine*, 25(2):72–82, March 2008. .

Mengye Lyu, Lifeng Mei, Shoujin Huang, Sixing Liu, Yi Li, Kexin Yang, Yilong Liu, Yu Dong, Linzheng Dong, and Ed X. Wu. M4Raw: A multi-contrast, multi-repetition, multi-channel MRI k-space dataset for low-field MRI research. *Scientific Data*, 10(1):264, May 2023. .

Xiangchao Meng, Yiming Xiong, Feng Shao, Huanfeng Shen, Weiwei Sun, Gang Yang, Qiangqiang Yuan, Randi Fu, and Hongyan Zhang. A Large-Scale Benchmark Data Set for Evaluating Pansharpening Performance: Overview and Implementation. *IEEE Geoscience and Remote Sensing Magazine*, March 2021. .

Christopher A. Metzler, Ali Mousavi, Reinhard Heckel, and Richard G. Baraniuk. Unsupervised Learning with Stein's Unbiased Risk Estimator, July 2020. arXiv:1805.10531 [stat].

Charles Millard and Mark Chiew. A Theoretical Framework for Self-Supervised MR Image Reconstruction Using Sub-Sampling via Variable Density Noisier2Noise. *IEEE Transactions on Computational Imaging*, 9:707–720, 2023. .

Charles Millard and Mark Chiew. Clean self-supervised MRI reconstruction from noisy, sub-sampled training data with Robust SSDU, June 2024. arXiv:2210.01696 [eess].

Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2Noise: Learning to Denoise from Unpaired Noisy Data, October 2019. arXiv:1910.11908 [cs, eess].

Subhadip Mukherjee, Marcello Carioni, Ozan Öktem, and Carola-Bibiane Schönlieb. End-to-end reconstruction meets data-driven regularization for inverse problems. In *Advances in Neural Information Processing Systems*, volume 34, pages 21413–21425, 2021.

Gyutaek Oh, Byeongsu Sim, HyungJin Chung, Leonard Sunwoo, and Jong Chul Ye. Unpaired Deep Learning for Accelerated MRI Using Optimal Transport Driven CycleGAN. *IEEE Transactions on Computational Imaging*, 6:1285–1296, 2020. .

Frank Ong and Michael Lustig. SigPy: a python package for high performance iterative reconstruction. In *Proceedings of the ISMRM 27th Annual Meeting, Montreal, Quebec, Canada*, volume 4819, 2019.

Frank Ong, Xucheng Zhu, Joseph Y. Cheng, Kevin M. Johnson, Peder E. Z. Larson, Shreyas S. Vasanawala, and Michael Lustig. Extreme MRI: Large-scale volumetric dynamic imaging from continuous non-gated acquisitions. *Magnetic Resonance in Medicine*, 84(4):1763–1780, 2020. .

Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, and Rebecca Willett. Deep Learning Techniques for Inverse Problems in Imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, May 2020. .

Arthur Pajot, Emmanuel de Bezenac, and Patrick Gallinari. Unsupervised Adversarial Image Reconstruction. In *International Conference on Learning Representations*, September 2018.

Li Pang, Xiangyu Rui, Long Cui, Hongzhong Wang, Deyu Meng, and Xiangyong Cao. HIR-Diff: Unsupervised Hyperspectral Image Restoration Via Improved Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3005–3014, 2024.

Tongyao Pang, Huan Zheng, Yuhui Quan, and Hui Ji. Recorrupted-to-Recorrupted: Unsupervised Deep Learning for Image Denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2043–2052, 2021.

Klaas P. Pruessmann, Markus Weiger, Markus B. Scheidegger, and Peter Boesiger. SENSE: Sensitivity encoding for fast MRI. *Magnetic Resonance in Medicine*, 42(5): 952–962, 1999. .

Chen Qin, Jo Schlemper, Jose Caballero, Anthony N. Price, Joseph V. Hajnal, and Daniel Rueckert. Convolutional Recurrent Neural Networks for Dynamic MR Image Reconstruction. *IEEE Transactions on Medical Imaging*, 38 (1):280–290, January 2019. .

Xinran Qin, Yuhui Quan, Tongyao Pang, and Hui Ji. Ground-Truth Free Meta-Learning for Deep Compressive Sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9947–9956, 2023.

Yuhui Quan, Xinran Qin, Tongyao Pang, and Hui Ji. Dual-Domain Self-supervised Learning and Model Adaption for Deep Compressive Imaging. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 409–426, Cham, 2022. Springer Nature Switzerland. .

Sathish Ramani, Thierry Blu, and Michael Unser. Monte-Carlo Sure: A Black-Box Optimization of Regularization Parameters for General Denoising Algorithms. *IEEE Transactions on Image Processing*, 17(9):1540–1554, September 2008. .

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, 2015. .

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Dirk Elias Schut, Adriaan Graas, Robert van Liere, and Tristan van Leeuwen. Equivariance2Inverse: A Practical Self-Supervised CT Reconstruction Method Benchmarked on Real, Limited-Angle, and Blurred Data, October 2025.

Victor Sechaud, Jérémy Scanvic, Quentin Barthélemy, Patrice Abry, and Julián Tachella. Equivariant Splitting: Self-supervised learning from incomplete data. In *Proceedings of The Fourteenth International Conference on Learning Representations*, October 2025.

Ortal Senouf, Sanketh Vedula, Tomer Weiss, Alex Bronstein, Oleg Michailovich, and Michael Zibulevsky. Self-supervised Learning of Inverse Problem Solvers in Medical Imaging. In Qian Wang, Fausto Milletari, Hien V. Nguyen, Shadi Albarqouni, M. Jorge Cardoso, Nicola Rieke, Ziyue Xu, Konstantinos Kamnitsas, Vishal Patel, Badri Roysam, Steve Jiang, Kevin Zhou, Khoa Luu, and Ngan Le, editors, *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 111–119, Cham, 2019. Springer International Publishing. .

Muhammad Shafique, Sizhuo Liu, Philip Schniter, and Rizwan Ahmad. MRI Recovery with Self-Calibrated Denoisers without Fully-Sampled Data. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 38(1):53–66, October 2024. .

Liyue Shen, John Pauly, and Lei Xing. NeRP: Implicit Neural Representation Learning With Prior Embedding for Sparsely Sampled Image Reconstruction. *IEEE Transactions on Neural Networks and Learning Systems*, 35 (1):770–782, January 2024. .

Efrat Shimron, Jonathan I. Tamir, Ke Wang, and Michael Lustig. Implicit data crimes: Machine learning bias arising from misuse of public data. *Proceedings of the National Academy of Sciences*, 119(13), March 2022. .

Emil Y. Sidky and Xiaochuan Pan. Report on the AAPM deep-learning sparse-view CT grand challenge. *Medical Physics*, 49(8):4935–4943, 2022. .

Oleksii Sidorov and Jon Yngve Hardeberg. Deep Hyperspectral Prior: Single-Image Denoising, Inpainting, Super-Resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3844–3851, October 2019. .

Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C. Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-End Variational Networks for Accelerated MRI Reconstruction. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 64–73, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59713-9. .

Charles M. Stein. Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9(6): 1135–1151, November 1981. .

Julián Tachella and Mike Davies. Self-supervised learning from noisy and incomplete data. *Foundations and Trends in Signal Processing*, 20(2):85–184, April 2026. ISSN 1932-8346. . URL https://doi.org/10.1108/FTSIG-10-2025-0133.

Julián Tachella, Dongdong Chen, and Mike Davies. Unsupervised Learning From Incomplete Measurements for Inverse Problems. *Advances in Neural Information Processing Systems*, 35:4983–4995, December 2022.

Julián Tachella, Mike Davies, and Laurent Jacques. UNSURE: self-supervised learning with Unknown Noise level and Stein's Unbiased Risk Estimate, February 2025a. arXiv:2409.01985 [stat].

Julián Tachella, Matthieu Terris, Samuel Hurault, Andrew Wang, Leo Davy, Jérémy Scanvic, Victor Sechaud, Romain Vo, Thomas Moreau, Thomas Davies, Dongdong Chen, Nils Laurent, Brayan Monroy, Jonathan Dong, Zhiyuan Hu, Minh-Hai Nguyen, Florian Sarron, Pierre Weiss, Paul Escande, Mathurin Massias, Thibaut Modrzyk, Brett Levac, Tobías I. Liaudat, Maxime Song, Johannes Hertrich, Sebastian Neumayer, and Georg Schramm. DeepInverse: A Python package for solving imaging inverse problems with deep learning. *Journal of Open Source Software*, 10(115):8923, November 2025b. ISSN 2475-9066. . URL `https://joss.theoj.org/papers/10.21105/joss.08923`.

Jonathan I. Tamir, Stella X. Yu, and Michael Lustig. Unsupervised Deep Basis Pursuit: Learning inverse problems without ground-truth data, February 2020. arXiv:1910.13110 [eess].

Asma Tanabene, Chaithya Giliyar Radhakrishna, Aurélien Massire, Mariappan S. Nadar, and Philippe Ciuciu. Benchmarking 3D multi-coil NC-PDNet MRI reconstruction, November 2024. arXiv:2411.05883 [eess].

Matthieu Terris, Samuel Hurault, Maxime Song, and Julián Tachella. Reconstruct Anything Model a lightweight general model for computational imaging. In *Proceedings of The Fourteenth International Conference on Learning Representations*, October 2025.

Shashanka Ubaru and Yousef Saad. Fast methods for estimating the Numerical rank of large matrices. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 468–477. PMLR, June 2016.

Martin Uecker, Peng Lai, Mark J. Murphy, Patrick Virtue, Michael Elad, John M. Pauly, Shreyas S. Vasanawala, and Michael Lustig. ESPIRiT—an eigenvalue approach to autocalibrating parallel MRI: Where SENSE meets GRAPPA. *Magnetic Resonance in Medicine*, 71(3):990–1001, 2014. .

Alan Q. Wang, Adrian V. Dalca, and Mert R. Sabuncu. Neural Network-Based Reconstruction in Compressed Sensing MRI Without Fully-Sampled Training Data. In Farah Deeba, Patricia Johnson, Tobias Würfl, and Jong Chul Ye, editors, *Machine Learning for Medical Image Reconstruction*, pages 27–37, 2020. ISBN 978-3-030-61598-7. .

Andrew Wang and Mike Davies. Fully Unsupervised Dynamic MRI Reconstruction via Diffeo-Temporal Equivariance, October 2024. arXiv:2410.08646 [eess].

Andrew Wang and Mike Davies. Perspective-Equivariance for Unsupervised Imaging with Camera Geometry. In Alessio Del Bue, Cristian Canton, Jordi Pont-Tuset, and Tatiana Tommasi, editors, *Computer Vision – ECCV 2024 Workshops*, pages 116–134, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-91585-7. .

Chengyan Wang, Jun Lyu, Shuo Wang, Chen Qin, Kunyuan Guo, Xinyu Zhang, Xiaotong Yu, Yan Li, Fanwen Wang, Jianhua Jin, Zhang Shi, Ziqiang Xu, Yapeng Tian, Sha Hua, Zhensen Chen, Meng Liu, Mengting Sun, Xutong Kuang, Kang Wang, Haoran Wang, Hao Li, Yinghua Chu, Guang Yang, Wenjia Bai, Xiahai Zhuang, He Wang, Jing Qin, and Xiaobo Qu. CMRxRecon: An open cardiac MRI dataset for the competition of accelerated image reconstruction, September 2023a.

Frederic Wang, Han Qi, Alfredo De Goyeneche, Reinhard Heckel, Michael Lustig, and Efrat Shimron. K-band: Self-supervised MRI Reconstruction via Stochastic Gradient Descent over K-space Subsets, May 2024. arXiv:2308.02958 [eess].

Puyang Wang, Pengfei Guo, Keyi Chai, Jinyuan Zhou, Daguang Xu, and Shanshan Jiang. SDUM: A Scalable Deep Unrolled Model for Universal MRI Reconstruction, December 2025.

Shanshan Wang, Ruoyou Wu, Cheng Li, Juan Zou, Ziyao Zhang, Qiegen Liu, Yan Xi, and Hairong Zheng. PARCEL: Physics-Based Unsupervised Contrastive Representation Learning for Multi-Coil MR Imaging. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(5):2659–2670, September 2023b. .

Zhihao Xia and Ayan Chakrabarti. Training Image Estimators without Image Ground Truth. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised Data Augmentation for Consistency Training, November 2020. arXiv:1904.12848 [cs].

Siying Xu, Marcel Früh, Kerstin Hammernik, Andreas Lingg, Jens Kübler, Patrick Krumm, Daniel Rueckert, Sergios Gatidis, and Thomas Küstner. Self-Supervised Feature Learning for Cardiac Cine MR Image Reconstruction. *IEEE Transactions on Medical Imaging*, 44(9), 2025.

Siying Xu, Kerstin Hammernik, Daniel Rueckert, Sergios Gatidis, and Thomas Küstner. Towards a Unified Theoretical Framework for Self-Supervised MRI Reconstruction, February 2026. URL http://arxiv.org/abs/2601.04775. arXiv:2601.04775 [eess].

Burhaneddin Yaman, Seyed Amir Hossein Hosseini, Steen Moeller, Jutta Ellermann, Kâmil Uğurbil, and Mehmet Akçakaya. Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data. *Magnetic Resonance in Medicine*, 84(6): 3172–3191, 2020. .

Burhaneddin Yaman, Hongyi Gu, Seyed Amir Hossein Hosseini, Omer Burak Demirel, Steen Moeller, Jutta Ellermann, Kâmil Uğurbil, and Mehmet Akçakaya. Multi-mask self-supervised learning for physics-guided neural networks in highly accelerated magnetic resonance imaging. *NMR in Biomedicine*, 35(12):e4798, 2022a. .

Burhaneddin Yaman, Seyed Amir Hossein Hosseini, and Mehmet Akcakaya. Zero-Shot Self-Supervised Learning for MRI Reconstruction. In *International Conference on Learning Representations*, January 2022b.

Leslie Ying and Jinhua Sheng. Joint image reconstruction and sensitivity estimation in SENSE (JSENSE). *Magnetic Resonance in Medicine*, 57(6):1196–1202, 2007.

Thomas Yu, Tom Hilbert, Gian Franco Piredda, Arun Joseph, Gabriele Bonanno, Salim Zenkhri, Patrick Omoumi, Meritxell Bach Cuadra, Erick Canales Rodriguez, Tobias Kober, and Jean-Philippe Thiran. Validation and Generalizability of Self-Supervised Image Reconstruction Methods for Undersampled MRI. *Machine Learning for Biomedical Imaging*, 1(September 2022 issue):1–31, September 2022.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming–Hsuan Yang. Restormer: Efficient Transformer for High-Resolution Image Restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5718–5729, June 2022.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An Open Dataset and Benchmarks for Accelerated MRI, November 2018.

Gushan Zeng, Yi Guo, Jiaying Zhan, Zi Wang, Zongying Lai, Xiaofeng Du, Xiaobo Qu, and Di Guo. A review on deep learning MRI reconstruction without fully sampled k-space. *BMC Medical Imaging*, 21(1):195, December 2021. .

Chi Zhang, Omer Burak Demirel, and Mehmet Akçakaya. Cycle-Consistent Self-Supervised Learning for Improved Highly-Accelerated MRI Reconstruction. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, May 2024. .

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, July 2017.

Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-Play Image Restoration With Deep Denoiser Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6360–6376, October 2022. .

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018. .

Hongkai Zheng, Wenda Chu, Bingliang Zhang, Zihui Wu, Austin Wang, Berthy T. Feng, Caifeng Zou, Yu Sun, Nikola Kovachki, Zachary E. Ross, Katherine L. Bouman, and Yisong Yue. InverseBench: Benchmarking Plug-and-Play Diffusion Priors for Inverse Problems in Physical Sciences, September 2025. URL http://arxiv.org/abs/2503.11043. arXiv:2503.11043 [cs].

Bo Zhou, Jo Schlemper, Neel Dey, Seyed Sadegh Mohseni Salehi, Kevin Sheth, Chi Liu, James S. Duncan, and Michal Sofka. Dual-domain self-supervised learning for accelerated non-Cartesian MRI reconstruction. *Medical Image Analysis*, 81:102538, October 2022. .

Magauiya Zhussip, Shakarim Soltanayev, and Se Young Chun. Training Deep Learning Based Image Denoisers From Undersampled Measurements Without Ground Truth and Without Image Prior. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10247–10256, June 2019. .

Qing Zou, Abdul Haseeb Ahmed, Prashant Nagpal, Stanley Kruger, and Mathews Jacob. Dynamic Imaging Using a Deep Generative SToRM (Gen-SToRM) Model. *IEEE Transactions on Medical Imaging*, 40(11):3102–3112, November 2021. .

Appendix A. Further experimental details

A.1 Data preprocessing details

We use a subset of the FastMRI (Zbontar et al., 2018) dataset for our experiments, and our dataset is reproducible by following the instructions given on the benchmark website. We take the middle axial slice of 455 brain volumes, use the provided 320×320 root-sum-square reconstructions as GT, normalise them to $[0, 1]$ and retrospectively simulate measurements using an undersampled Fourier transform using the DeepInverse library (Tachella et al., 2025b). We randomly create train-test datasets with 80-20 split, since the original FastMRI test set does not have GT for evaluation. For the subsampling masks $\mathbf{M}_i \sim \mathcal{M}$, we use random 1D Gaussian masking with $6\times$ acceleration and 8% autocalibration (ACS) lines kept in the centre. For the noisy scenario we add i.i.d. Gaussian measurement noise with $\sigma = 0.1$, and for the single-operator scenario we fix one $\mathbf{M} \sim \mathcal{M}$. For the multi-coil scenario we assume known sensitivity coil maps to decouple the parameter estimation problem from reconstruction, and assume a fixed number of coils C so that the number of measurements m is constant over the dataset. This is necessary as coil sensitivity estimation introduces its own artifacts, particularly in low image-density areas outside the main region of interest, which we do not wish to introduce into our setup. Therefore, we simulate coil maps using `sigpy` (Ong and Lustig, 2019) with $C = 4$ and simulate measurements using eq. (1).

The SKM-TEA (Desai et al., 2022c) dataset contains raw quantitative (qDESS) acquisitions from a GE MR750 3T scanner (GE, Waukesha, WI) and estimated coil maps via JSENSE (Ying and Sheng, 2007). We take raw multicoil data from 100 slices across 5 volumes and retrospectively simulate 2D Poisson disk undersampling at $32\times$ acc. as per the original paper. The CMRxRecon (Wang et al., 2023a) dataset contains cardiac cine sequences of length 12. We take raw data from 100 slice sequences from the training set, split by patient ID, and apply k-t-space Gaussian undersampling at $10\times$ acc. For dynamic MRI we swap the model f_θ for the CRNN (Qin et al., 2019). The knee volume in Yu et al. (2022) is prospectively acquired from a Siemens MAGNETOM Prisma Fit 3T scanner (Siemens Healthcare, Erlangen, Germany) with a T2 SPACE sequence using a $7\times$ 3D Poison disk mask with 18 coils; the coil maps are estimated with ESPIRiT (Uecker et al., 2014). We take slices in the axial dimension for the 2D experiment.

Note that while retrospective simulation is necessary for a controlled benchmarking experiment, future experiments would evaluate benchmarked methods on prospectively acquired data (Shimron et al., 2022; Chaudhari et al., 2021; Yu et al., 2022).

Note that all images are normalised to the range $[0, 1]$ before plotting.

A.2 Model and hyperparameter details

The unrolled network f_θ unrolls the iterative bi-level optimisation problem provided by half-quadratic splitting (Zhang et al., 2022) for u iterations, originally proposed for MRI in (Aggarwal et al., 2019), and we take $u = 3$. During inference, this model f_θ requires u forward passes of the denoiser. For the denoiser of our unrolled network, we use a residual U-Net (Ronneberger et al., 2015) with no batch norm of depth 4, with a total of 8.6M parameters. We train with the Adam optimiser at $1e-3$ learning rate, and we step this per method down to $1e-5$ to achieve convergence. All models are trained with batch size of 4 using 1 NVIDIA GeForce RTX 3090, taking around 15GB memory, with an average training time of 6 hours. Average training time per batch is 0.5s, and average inference time of the trained models f_θ is 0.1s per batch.

Experimental results (mean, sample standard deviation) are reported on held-out test sets for each dataset. Since end-to-end trained models are deterministic (*i.e.* their input is the measurement, rather than a random latent), inference runs are not repeated. Simulation of retrospective random undersampling masks, random noise, dataset random split and model initialisation used fixed seeds for all experiments to keep the setup fixed and repeatable across runs. Note that the reported results may be a function of the seeds used during training and are thus indicative. Furthermore, fixing certain hyperparameters in the architecture, loss functions and training recipe potentially affects the rankings of some methods across scenarios; future work would perform more thorough hyperparameter search across these.

A.3 Individual loss implementation details

We provide details of the benchmarked loss functions' reimplementations below. All components of our modular experiments are implemented using DeepInverse (Tachella et al., 2025b), and full code is at github.com/Andrewwango/ssibench.

General notes Some losses compute additional forward passes of f_θ per training iteration. For these, more Monte Carlo passes per iteration could improve performance but we uniformly limit this to one additional pass for all methods for efficiency. Secondly, the training metric ℓ in benchmarked methods' original papers varies between L1 and L2 (MSE); since this is not instrumental in recovering information, and many methods' theory rely on the MSE, we use L2 for standardised evaluation. Thirdly, where there is more than one term in the loss we weight the terms evenly.

Measurement consistency: We simply take ℓ as the MSE in kspace as per (Senouf et al., 2019) with $\beta > 0, \alpha = \gamma = 0$. In the multi-coil case, we take ℓ as the MSE in the backprojected $\mathbf{A}^\top$ space as we find this provides better results.

Measurement splitting: We use a 2D mask with split ratio $\rho = 0.6$ following Yaman et al. (2020), drawn randomly each instance during training following Yaman et al. (2022a). For weighted-SSDU we use an independent 1D Gaussian mask following Millard and Chiew (2023). At inference time, we do not perform additional correction for both Millard and Chiew (2023, 2024) since our network maps directly to images $f_\theta : \mathcal{Y} \to \mathcal{X}$. We ablate and justify the splitting mask choices in section B.6.2. For (Hu et al., 2021), the authors train two networks, but we follow their ablation and share these weights to keep the number of parameters constant for a fair setup, follow the partitioning used above, and remove the ISTA-Net-specific loss terms. Note that (Wang et al., 2023b) is equivalent to the above but takes the consistency loss on embeddings of $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2$ instead. Both (Wang et al., 2023b; Zhou et al., 2022) are equivalent to the above but with the addition of a superfluous $\mathcal{L}_{\text{MC}}$.

Proposition (Equivalence of unpaired Artifact2Artifact as sum of SSDU losses)**.** Consider the Artifact2Artifact (Liu et al., 2020) scenario where $\mathbf{y}_i^{(k)}$ are independent sets of measurements of the same subject i (*e.g.* arriving in a stream) such that $\mathbf{y}_i = \bigcup_k \mathbf{y}_i^{(k)} \in \mathbb{R}^m$ is the full undersampled measurement from one acquisition i (i.e. there are no more measurements of the same subject). $\mathbf{A}_i^{(k)\top}\mathbf{y}_i^{(k)}$ are then the independent "artifact"-corrupted images. The Artifact2Artifact loss draws a pair k, l randomly at each iteration and constructs the loss

$$\mathcal{L}_{\text{A2A}}^{(k,l)} = \ell(\mathbf{A}_i^{(l)} f_\theta(\mathbf{y}_i^{(k)}, \mathbf{A}_i^{(k)}), \mathbf{y}_i^{(l)}),$$

where $\mathbf{A}_i^{(k)}, \mathbf{A}_i^{(l)}$ are the operators associated with each measurement. We can then write this as the overlapping SSDU loss (Yaman et al., 2020, Fig. 5) by setting $\mathbf{A}_i^{(k)} = \mathbf{M}_k\mathbf{A}_i, \mathbf{A}_i^{(l)} = \mathbf{M}_l\mathbf{A}_i, \mathbf{y}_i^{(k)} = \mathbf{M}_k\mathbf{y}_i, \mathbf{y}_i^{(l)} = \mathbf{M}_l\mathbf{y}_i$ where $\mathbf{M}_k, \mathbf{M}_l$ are two overlapping mask subsampling sets. Then

$$\mathcal{L}_{\text{A2A}} = \sum_k \sum_{l \neq k} \mathcal{L}_{\text{SSDU}}^{(k,l)}.$$

Learning from invariance For losses involving equivariant imaging, we draw randomly a group transform at each iteration; for rotation we define the group as $G = \text{SO}(\mathbb{R}^2)$ following Chen et al. (2021).

Data augmentation For VORTEX (Desai et al., 2022a), noting that the results of their various variants are very similar, for $\mathbf{T}_1$ we use random noise and their random phase errors with $\sigma = \alpha = 1$ and for $\mathbf{T}_2$ we use random maximum $\pm 10\%$ shifts and random $\pm 15°$ rotations. We also observe reduced performance when adding more transformations including full $\pm 360°$ rotation, scaling, or shearing. We do not perform curriculum learning as the original paper does not suggest it significantly affects performance. We observe that when using the VORTEX consistency term on its own without any supervised or MC loss, it learns the trivial $f_\theta(\mathbf{y}, \mathbf{A}) = \mathbf{0}$, showing the superfluity of VORTEX on our experiments on in-domain data. Note that Noise2Recon (Desai et al., 2022b) is a special case VORTEX with $\mathbf{T}_2 = \mathbf{I}$ and $\mathbf{T}_1$ is random noise.

Adversarial losses For the adversarial losses, we use the same model f_θ as the generator, and the simple convolutional discriminator with skip connections used in Cole et al. (2021). We perform generator and discriminator steps with ratio 1. Noting that there is a wide range of different GAN flavours, we use the MSE in the adversarial loss calculations following LSGAN (Lucic et al., 2018).

MO-EI The MO-EI loss function is visualised diagrammatically in fig. 5. We let $G = \text{Diffeo}(\mathbb{R}^2)$. As these are very general, we relax the full assumption and instead enforce approximate equivariance to small perturbative distortions by taking the subset of continuous piecewise-affine-based diffeomorphic transforms from Freifeld et al. (2017); Wang and Davies (2024), visualised in fig. 6. For MO-EI, we ablate the choice of transform in section B.6.1.

Remark (MO-EI generalises MOI and EI)**.** *By letting G be the trivial group, we recover MOI (Tachella et al., 2022), and by letting $\mathcal{A} = \{\mathbf{A}\}$, we recover EI (Chen et al., 2021).*

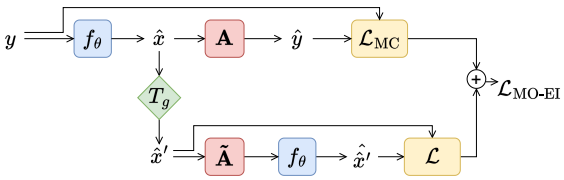

Figure 5: The multi-operator equivariant imaging (MO-EI) loss function.

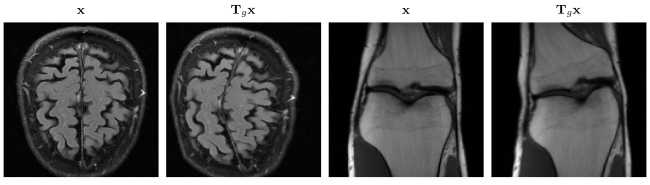

Figure 6: Diffeomorphic transforms.

Appendix B. Further experimental results

B.1 Statistical significance results

To compare the performance of MO-EI with two next-best methods in scenario 1, and likewise for other scenarios, we conduct two statistical tests: paired t-test (where normality of the differences is validated in fig. 7), and Wilcoxon signed-rank. Results are shown in table 5, suggesting, to a

0.5% significance level, that there is sufficient evidence to reject the null hypothesis and accept the alternative hypothesis that the top-performing method has higher average performance for scenarios 1-5. Note that scenario 7 lacks GT and thus no quantitative metrics were reported.

Table 5: Statistical significance test p-values comparing test-set PSNR of top method and next-best method(s) for scenarios 1-6.

Scen.	Comparison	Paired t	Wilcoxon
1	MO-EI ↔ MOI	3.7×10^{-48}	6.0×10^{-17}
1	MO-EI ↔ SSDU-Consistency	6.4×10^{-44}	6.0×10^{-17}
2	Robust-MO-EI ↔ Robust-EI	6.9×10^{-33}	6.0×10^{-17}
3	EI ↔ MOI	2.5×10^{-26}	8.9×10^{-17}
4	Weighted-SSDU ↔ SSDU-Consistency	1.9×10^{-11}	1.4×10^{-10}
5	Weighted-SSDU ↔ MOC-SSDU	3.1×10^{-9}	3.9×10^{-3}
6	MOC-t-SSDU ↔ t-SSDU-Consistency	5.3×10^{-2}	8.9×10^{-2}

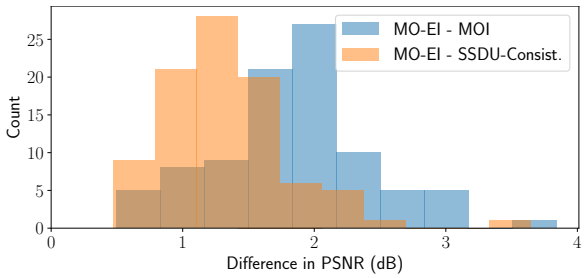

Figure 7: Histograms of paired diffs for scen. 1.

B.2 Rank of single-coil vs multi-coil operators

We further investigate the results of the multi-coil scenario by measuring the size of the null-space compared to the single-coil scenario. We consider a toy $4\times$ accelerated MRI example on 32×32 images, so $n = 32 \times 32 = 1024$, $m = nC/4 = 256$ for single-coil and $C = 4, m = 1024$ for multi-coil. We then compute the numerical SVD of the operator $\mathbf{A}$; results are in fig. 8. The single-coil operator (scenario 1) has exactly $m = \text{rank}(\mathbf{A})$ non-zero singluar values as expected, whereas the multi-coil operator (scenario 4), now incorporating both masks and sensitivity maps, has many more non-zero singular values, increasing its effective rank (Ubaru and Saad, 2016) and decreasing the size of the "effective null-space". For example, consider the case $\sigma = 0.1$; the effective $\text{rank}_{\sigma^2}(\mathbf{A}) = 686$, so the effective acceleration is $n/m = 1.49 < 4$. This reduces the difficulty of learning in the null-space; we leave for further work tuning the acceleration rate to achieve a similar effective rank as the single-coil case.

B.3 Scenarios 2, 3, 4 & 6

We provide sample reconstructions in figs. 9 to 12 to complement the quantitative results in tables 2 to 4.

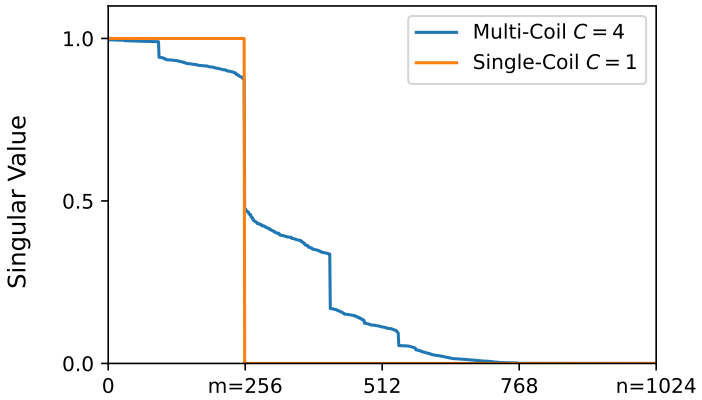

Figure 8: Numerical singular values of single-coil vs multi-coil operators, sorted in descending order.

B.4 Effect of model architecture

B.4.1 Unrolled depth

We investigate the effect of unrolling depth u (and hence time and memory cost) on performance scaling, starting at $u = 3$ (*i.e.* the model on which all results are reported in the main paper) up to $u = 9$. Results on scenario 1 are shown in fig. 13 on two self-supervised losses and the supervised loss, showing a largely monotonic increase in performance as u increases, suggesting that results presented in the main paper would scale with more expensive models, and that there is a trade-off between model cost and performance.

B.4.2 Effect of model architecture size

The benchmark results on scenarios 1–4 use an 8.6M parameter model. Although not large by modern standards, this setting may be too large for implementations in constrained hardware configurations. We thus evaluate our benchmark using a model with <1M parameters which would be more suitable for a constrained clinical environment, *e.g.* without extensive GPU compute, by providing a pilot study where we simply swap our benchmark 'model' for a 190K parameter unrolled VarNet (Hammernik et al., 2018) with a lightweight CNN denoiser (Zhang et al., 2017). Scenario 1 results in table 6 show that although the Scenario 1 performance is uniformly decreased, the ranking of the methods is broadly preserved, evidencing the orthogonality of loss choice and model architecture design for this scenario.

B.4.3 Alternative architectures

While convolutional nets remain popular for imaging tasks (Heckel et al., 2024), attention-based networks have recently demonstrated competing performance (Wang et al., 2025) at the cost of higher computational cost and parameter count due to the self-attention. We swap out the benchmark model for a shallow Restormer (Zamir et al., 2022) architecture and reduce the image size by half to fit

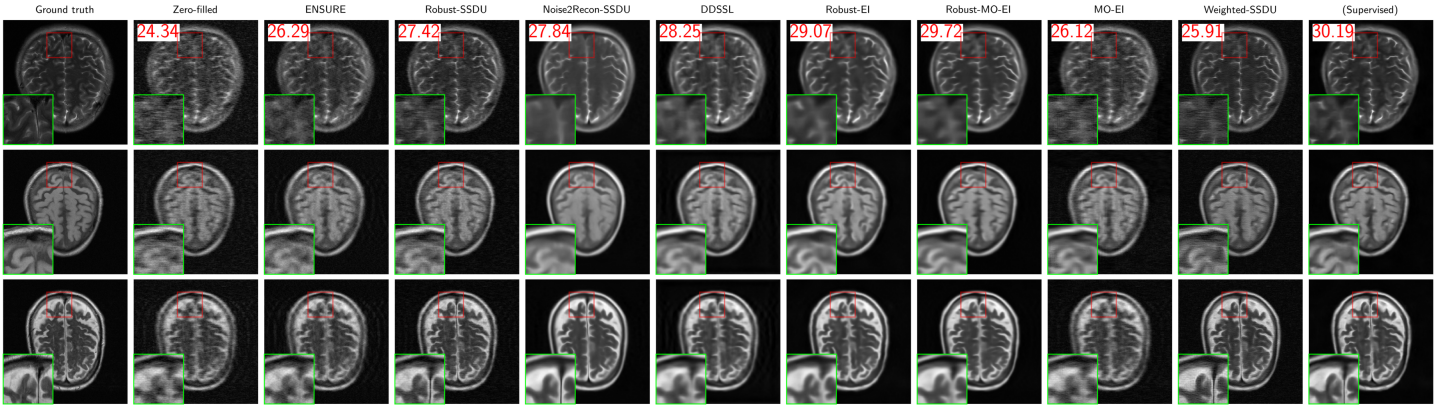

Figure 9: Joint denoise/recon (scenario 2) with test-set PSNR, showing selected highest-performing methods. See table 3 for more.

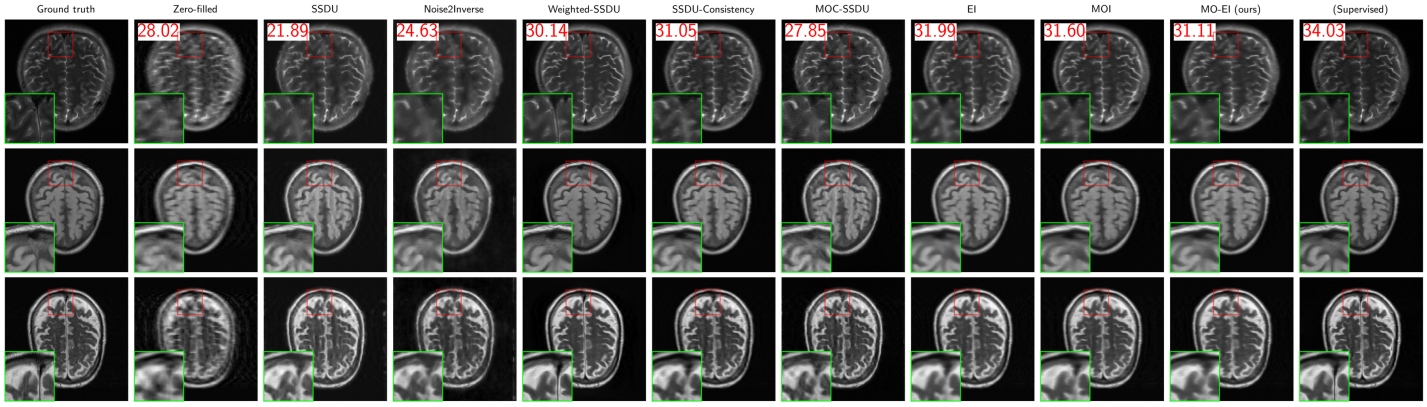

Figure 10: Single-mask (scenario 3) recons with test-set PSNR, showing selected highest-performing methods. See table 2 for more.

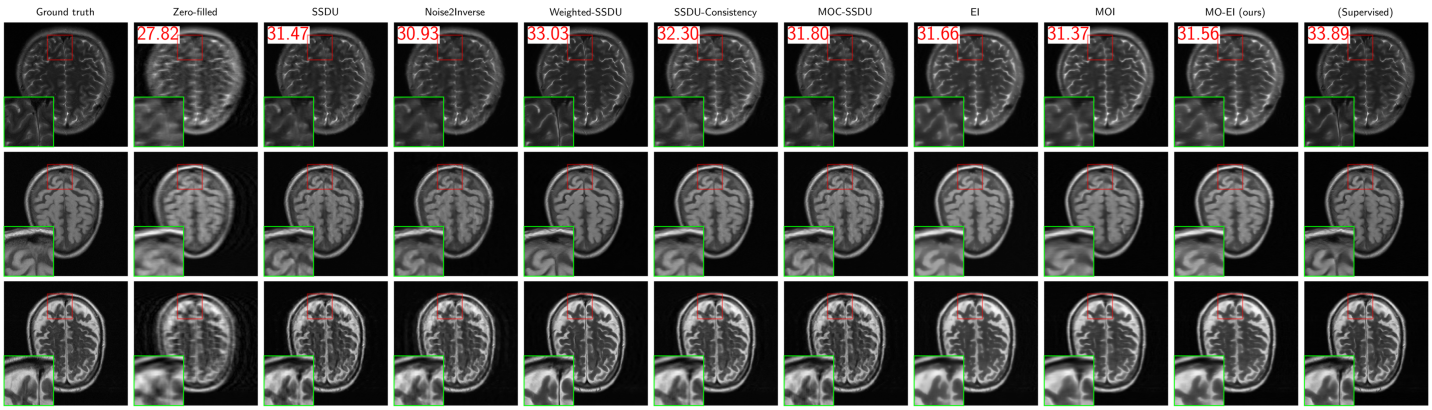

Figure 11: Multicoil brain recons (scenario 4) with test-set PSNR, showing selected highest-performing methods. See table 2 for more.

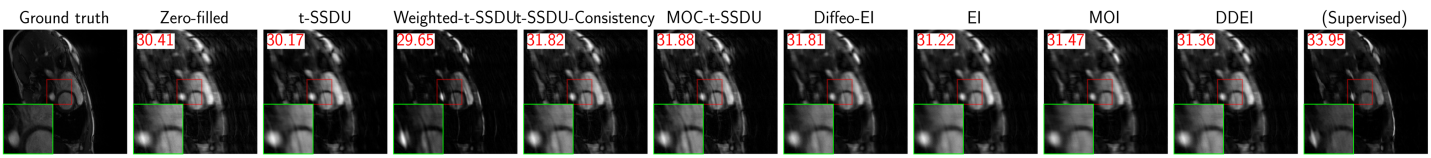

Figure 12: $10\times$ acc. CMRxRecon (Wang et al., 2023a) dynamic cardiac MRI (scenario 6) recons, showing 0th frame for selected highest-performing methods. See table 4 for more.

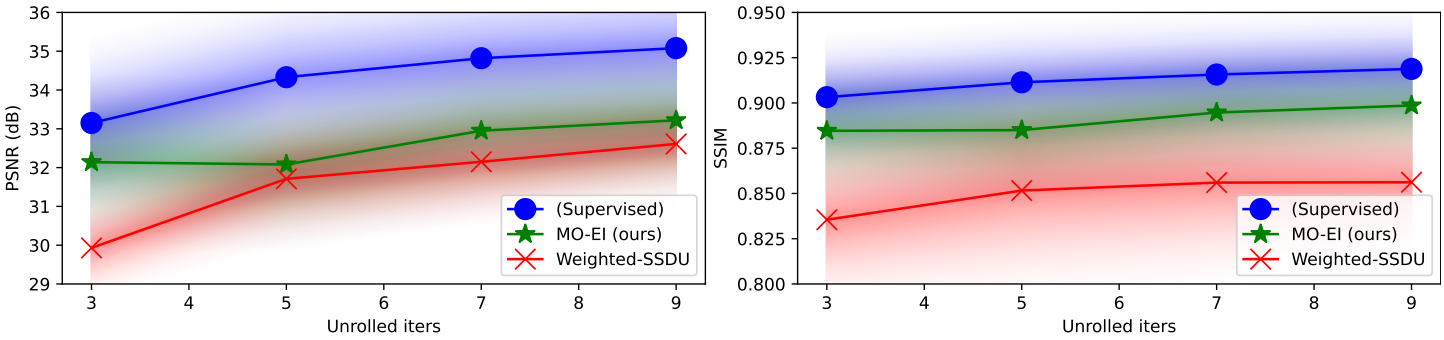

Figure 13: Quantitative results on scenario 1 for Weighted-SSDU, MO-EI, and oracle supervised losses for increasing number of unrolled iterations. Left: PSNR, right: SSIM. Shaded areas repesent one σ.

Table 6: Effect of model size: benchmark results on scenario 1 but with a 50× smaller model (VarNet with flat CNN).

Loss	PSNR ↑	SSIM ↑	LPIPS ↓
Zero-filled	$27.67_{\pm 2.40}$	$.7862_{\pm .05}$	$.3270_{\pm .0426}$
SSDU	$24.24_{\pm 1.50}$	$.6057_{\pm .0575}$	$.3640_{\pm .0527}$
Weighted-SSDU	$24.22_{\pm 1.28}$	$.6237_{\pm .0761}$	$.3697_{\pm .0427}$
SSDU-Consistency	$27.96_{\pm 2.30}$	$.7941_{\pm .0568}$	$.2801_{\pm .0500}$
MOC-SSDU	$26.16_{\pm 1.54}$	$.6670_{\pm .0627}$	$.3597_{\pm .0432}$
EI	$29.12_{\pm 2.56}$	$\mathbf{.8349}_{\pm .0546}$	$.2714_{\pm .0380}$
MOI	$28.68_{\pm 2.55}$	$.8275_{\pm .0552}$	$.2935_{\pm .0486}$
MO-EI	$\mathbf{29.22}_{\pm 2.68}$	$.8348_{\pm .0557}$	$\mathbf{.2629}_{\pm .0493}$
(Supervised)	$30.86_{\pm 2.50}$	$.8633_{\pm .0530}$	$.1598_{\pm .0399}$

the training in GPU memory. Scenario 1 results in table 7 again show broadly preserved ranking, albeit with a significant performance drop, due to the usage of a shallower network to fit into memory.

Table 7: Benchmark results with Restormer.

Loss	PSNR ↑	SSIM ↑	LPIPS ↓
Zero-filled	$25.17_{\pm 2.22}$	$.7381_{\pm .05}$	$.1956_{\pm .0398}$
MC	$25.16_{\pm 2.22}$	$.7373_{\pm .0490}$	$.1941_{\pm .0400}$
SSDU	$25.18_{\pm 1.15}$	$.7503_{\pm .0511}$	$.1715_{\pm .0345}$
Noise2Inverse	$25.22_{\pm 3.17}$	$.7520_{\pm .0823}$	$.1870_{\pm .0450}$
Weighted-SSDU	$25.48_{\pm 1.10}$	$.7789_{\pm .0450}$	$.1450_{\pm .0359}$
SSDU-Consistency	$24.92_{\pm 1.66}$	$.8070_{\pm .0368}$	$.1316_{\pm .0316}$
MOC-SSDU	$23.28_{\pm 1.74}$	$.7301_{\pm .0473}$	$.1181_{\pm .0301}$
UAIR	$0.90_{\pm 1.20}$	$.0921_{\pm .0214}$	$.2886_{\pm .0429}$
VORTEX	$25.06_{\pm 2.19}$	$.7282_{\pm .0472}$	$.1897_{\pm .0397}$
EI	$27.15_{\pm 2.73}$	$.8180_{\pm .0434}$	$.1213_{\pm .0366}$
MOI	$27.43_{\pm 2.63}$	$.8358_{\pm .0399}$	$.1438_{\pm .0434}$
MO-EI	$27.22_{\pm 2.80}$	$.8319_{\pm .0432}$	$.1307_{\pm .0392}$

B.5 Pretrained diffusion model results

We evidence the extensibility of our framework by including pilot results in table 8 on SotA diffusion methods (Chung et al., 2022) using recently proposed GT-free pretrained models (Aali et al., 2024b). Note that it is infeasible to train these diffusion models (Karras et al., 2022) on the modest-size dataset used in any of the scenarios considered in the main paper. Therefore, these results cannot be directly compared to other methods from the main paper, since the pretrained models from (Aali et al., 2024b) are trained on datasets of much larger size ($\sim 10^4$ images), perhaps covering our test images. We apply the diffusion methods to scenarios 1-4 in table 8. Since diffusion methods sample a point estimate from the posterior, it is expected that they achieve good perceptual performance, but poorer distortion metrics compared to MMSE estimators (Blau and Michaeli, 2018). In the in-distribution setting, the one-step MMSE reconstruction improves the distortion result at the expense of perceptual performance, mirroring the end-to-end loss performances. The noisy performance suggests that the model struggles to generalise to this setting. DPS and ALD methods require a large number of iterations, so they are sampled using 500 denoising iterations (NFEs) per image (Aali et al., 2024b).

B.6 Ablations

B.6.1 Ablation of MO-EI loss

The components of our proposed loss function are ablated by interpolating between existing methods MOI and EI and the proposed method MO-EI, and report results in table 9. We note that the base MOI and EI have similar results, and adding the two independent components of a) MO-EI and b) the diffeomorphic transforms helps, but combining these two provides the best performance in our proposed method, as per our theoretical framework.

B.6.2 Ablation of SSDU loss

We ablate for the components of SSDU (Yaman et al., 2022a) and weighted-SSDU (Millard and Chiew, 2023) and report results in table 10. While (Yaman et al., 2022a) construct variable-density 2D partitioning masks with an ACS block, we were only able to achieve good performance using uniform-density partitioning (*i.e.* Bernoulli) without an ACS block. This aligns with theory, since the lack of ACS in $\mathbf{M}_1$ means that both $\mathbf{M}_1$ and $\mathbf{M}_2$ contain low-frequency

Table 8: Pilot results on diffusion sampling methods with models pretrained without ground truth, applied to benchmark scenarios. Note that the results cannot be directly compared to tables from the main paper, as these models, while pretrained on measurement only data, required substantially larger datasets for training. DPS = diffusion posterior sampling (Chung et al., 2022; Aali et al., 2024b), ALD = annealed Langevin dynamics (Jalal et al., 2021; Aali et al., 2024b).

Scenario	**(1)** Brain MRI 6× acc.			**(2)** Noisy $\sigma = .1$			**(3)** Single-mask			**(4)** Multicoil $C = 4$		
Loss	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Zero-filled	$27.65_{\pm 2.40}$	$0.7857_{\pm 0.0533}$	$0.3270_{\pm 0.0471}$	$24.38_{\pm 1.03}$	$0.4438_{\pm 0.0348}$	$0.4636_{\pm 0.0517}$	$28.23_{\pm 2.29}$	$0.7951_{\pm 0.0495}$	$0.3494_{\pm 0.0491}$	$27.81_{\pm 2.40}$	$0.7984_{\pm 0.0511}$	$0.3175_{\pm 0.0475}$
Ambient-DPS	$29.49_{\pm 1.52}$	$0.7368_{\pm 0.0504}$	$0.1726_{\pm 0.0395}$	$\mathbf{26.10}_{\pm 0.98}$	$\mathbf{0.5421}_{\pm 0.0483}$	$0.3306_{\pm 0.0535}$	$29.27_{\pm 1.54}$	$0.7272_{\pm 0.0512}$	$0.1792_{\pm 0.0407}$	$29.98_{\pm 1.66}$	$0.7552_{\pm 0.0506}$	$\mathbf{0.1532}_{\pm 0.0361}$
Ambient-ALD	$29.52_{\pm 1.60}$	$0.7367_{\pm 0.0518}$	$\mathbf{0.1707}_{\pm 0.0376}$	$\mathbf{26.10}_{\pm 0.98}$	$0.5420_{\pm 0.0472}$	$\mathbf{0.3300}_{\pm 0.0548}$	$29.38_{\pm 1.54}$	$0.7319_{\pm 0.0513}$	$\mathbf{0.1753}_{\pm 0.0409}$	$29.97_{\pm 1.60}$	$0.7550_{\pm 0.0491}$	$0.1538_{\pm 0.0349}$
Ambient-One-Step	$\mathbf{31.34}_{\pm 2.29}$	$\mathbf{0.8673}_{\pm 0.0480}$	$0.1743_{\pm 0.0459}$	$24.93_{\pm 0.96}$	$0.4655_{\pm 0.0372}$	$0.4303_{\pm 0.0547}$	$\mathbf{31.51}_{\pm 2.17}$	$\mathbf{0.8686}_{\pm 0.0467}$	$0.2000_{\pm 0.0509}$	$\mathbf{31.53}_{\pm 2.21}$	$\mathbf{0.8741}_{\pm 0.0459}$	$0.1685_{\pm 0.0461}$

Table 9: Ablation of components of the proposed MO-EI loss function.

Loss	PSNR ↑	SSIM ↑
MOI	$30.29_{\pm 2.88}$	$.8651_{\pm .0528}$
EI (rotate)	$30.26_{\pm 2.61}$	$.8523_{\pm .0542}$
EI (diffeo)	$31.26_{\pm 2.88}$	$.8741_{\pm .0522}$
MO-EI (rotate)	$30.62_{\pm 2.70}$	$.8575_{\pm .0553}$
MO-EI (diffeo)	$\mathbf{32.14}_{\pm 2.73}$	$\mathbf{.8846}_{\pm .0498}$

Table 10: Ablation of components of SSDU (Yaman et al., 2022a) and weighted-SSDU (Millard and Chiew, 2023) loss functions. * = methods reported in main paper.

Loss	PSNR ↑	SSIM ↑
SSDU (uniform-density no ACS)*	$27.98_{\pm 1.43}$	$.7485_{\pm .0667}$
SSDU (variable-density with ACS)	$18.27_{\pm 1.01}$	$.4876_{\pm .0598}$
Weighted-SSDU (1D partition)*	$\mathbf{29.93}_{\pm 1.66}$	$\mathbf{.8355}_{\pm .0626}$
Weighted-SSDU (2D partition)	$28.62_{\pm 1.59}$	$.7489_{\pm .0692}$
Weighted-SSDU (1D with ACS)	$13.41_{\pm .63}$	$.3875_{\pm .0479}$
SSDU (1D partition)	$27.38_{\pm 1.87}$	$.8055_{\pm .0658}$

lines; otherwise, $\mathbf{y}_2$ would contain almost no low-frequency image content and thus diverge the loss. We reproduce the weighted-SSDU results reported in (Millard and Chiew, 2023) reflecting their theory: results improve when adding the weighting even when using the original 2D partitioning, and improves further when using 1D partitioning masks, since these $\mathbf{M}_1$ are then drawn from the same original distribution of $\mathbf{M}$, but only if these also do not have ACS lines.

B.7 Transfer of benchmark to different imaging modalities

B.7.1 Hyperspectral imaging

We demonstrate the transferability of our modular benchmark on a different scientific imaging modality where GT images do not exist. Hyperspectral imaging aims to obtain high quality multi-channel images of the Earth for remote sensing and environmental monitoring, but observations may be degraded due to impulse noise arising from missing lines and thermal noise (Pang et al., 2024; Bodrito et al., 2021; Sidorov and Hardeberg, 2019). Hyperspectral restoration is therefore an important preprocessing task, but GT is unavailable (Bodrito et al., 2021; Sidorov and Hardeberg, 2019) because of the non-stationary nature of satellites.

Our modular benchmark framework facilitates evaluating SSI methods on any imaging inverse problem such as the hyperspectral restoration problem, and we demonstrate a proof-of-concept showing the ease of adapting the benchmark to a different problem domain. Concretely, we simply swap out the following modules:

1. `physics = deepinv.physics.MRI()` → `deepinv.physics.Inpainting(noise_model=GaussianNoise())`, *i.e.* random inpainting with 1D masks (50% dropout) with additive noise ($\sigma = 0.1$);
2. `dataset = FastMRISliceDataset(...)` → `NBUDataset(...)`, where we use a subset of 150 $256 \times 256 \times 4$ WorldView-2 satellite patches provided by Meng et al. (2021).

Qualitative and quantitative results are shown in fig. 14 and table 11, benchmarking self-supervised losses for joint reconstruction and denoising, with the no-learning result showing the input since $\mathbf{A}^\top \mathbf{y} = \mathbf{y}$. The results show the superiority of Robust-EI, with ENSURE failing to recover information in the null-space, and Robust-SSDU and Noise2Recon-SSDU less able to produce spatially and spectrally faithful reconstructions. We conclude that Robust-MO-EI performs less well due to the inappropriate application of diffeomorphic invariance to urban satellite images, where straight lines should remain straight.

B.7.2 Sparse-view CT

We also perform the same transferability experiment as above but on a different medical task, sparse-view computed tomography reconstruction, where, similar to MRI, one acquires fewer projections to reduce the acquisition time, leading to a worse image quality (Sidky and Pan, 2022). We consider a 45 angle, 2D tomography problem analogous to the benchmark Scenario 1, where the forward operator is a subsampled Radon transform. We train and test the benchmarked losses on a FBPConvNet-style architecture

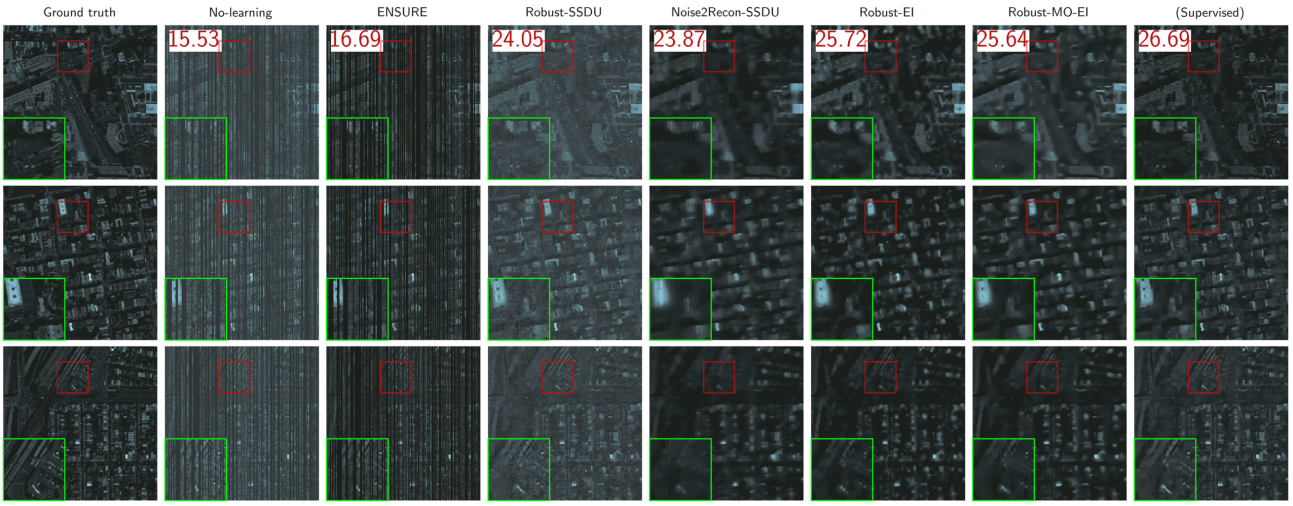

Figure 14: Sample reconstructions for hyperspectral image restoration.

Table 11: Test-set results for hyperspectral image restoration.

Loss	PSNR ↑	SSIM ↑
Zero-filled	$15.53_{\pm.59}$	$.1435_{\pm.01}$
ENSURE	$16.69_{\pm.75}$	$.2074_{\pm.0155}$
Robust-SSDU	$24.05_{\pm.31}$	$.4965_{\pm.0236}$
Noise2Recon-SSDU	$23.87_{\pm.75}$	$.5216_{\pm.0366}$
Robust-EI	$\mathbf{25.72}_{\pm.72}$	$\mathbf{.6601}_{\pm.0243}$
Robust-MO-EI	$25.64_{\pm.72}$	$.6535_{\pm.0242}$
(Supervised)	$26.69_{\pm.70}$	$.7282_{\pm.0201}$

Table 12: Test-set results for sparse-view CT.

Loss	PSNR ↑	SSIM ↑	LPIPS ↓
Backprojection	$17.46_{\pm2.19}$	$.4283_{\pm.11}$	$.2171_{\pm.0436}$
SSDU	$34.70_{\pm2.56}$	$.8890_{\pm.0344}$	$.1277_{\pm.0417}$
Noise2Inverse	$30.07_{\pm3.00}$	$.8215_{\pm.0556}$	$.2334_{\pm.0570}$
SSDU-Consistency	$34.85_{\pm2.69}$	$.8937_{\pm.0349}$	$.1292_{\pm.0421}$
MOC-SSDU	$34.87_{\pm2.78}$	$.8899_{\pm.0378}$	$.1320_{\pm.0436}$
Adversarial	$21.60_{\pm.17}$	$.3252_{\pm.2280}$	$.1322_{\pm.0437}$
UAIR	$22.57_{\pm2.47}$	$.8478_{\pm.0351}$	$.1448_{\pm.0484}$
VORTEX	$\text{nan}_{\pm.00}$	$\text{nan}_{\pm.0000}$	$\text{nan}_{\pm.0000}$
EI	$34.88_{\pm2.78}$	$.8897_{\pm.0378}$	$.1320_{\pm.0437}$
MOI	$34.88_{\pm2.78}$	$.8898_{\pm.0378}$	$.1321_{\pm.0436}$
MO-EI	$34.89_{\pm2.78}$	$.8902_{\pm.0376}$	$.1320_{\pm.0435}$

where the input to the U-Net is the filtered back-projection (Jin et al., 2017), and simulate measurements from the CT100 dataset, a dataset of 100 in-vivo CT images from the cancer imaging archive (Clark et al., 2013). The results in table 12 show a wide spread in performance, with similar rankings to that in accelerated MRI.