Unsupervised Learning From Incomplete Measurements for Inverse Problems

Anonymous Author(s)
Affiliation
Address
email

Abstract

In many real-world inverse problems, only incomplete measurement data are available for training which can pose a problem for learning a reconstruction function. Indeed, unsupervised learning using a fixed incomplete measurement process is impossible in general, as there is no information in the nullspace of the measurement operator. This limitation can be overcome by using measurements from multiple operators. While this idea has been successfully applied in various applications, a precise characterization of the conditions for learning is still lacking. In this paper, we fill this gap by presenting necessary and sufficient conditions for learning the underlying signal model needed for reconstruction which indicate the interplay between the number of distinct measurement operators, the number of measurements per operator, the dimension of the model and the dimension of the signals. Furthermore, we propose a novel and conceptually simple unsupervised learning loss which only requires access to incomplete measurement data and achieves a performance on par with supervised learning when the sufficient condition is verified. We validate our theoretical bounds and demonstrate the advantages of the proposed unsupervised loss compared to previous methods via a series of experiments on various imaging inverse problems, such as accelerated magnetic resonance imaging, compressed sensing and image inpainting.

1 Introduction

In multiple sensing applications, we observe measurements $y \in \mathbb{R}^m$ associated with a signal $x \in \mathcal{X} \subset \mathbb{R}^n$, through the forward process

$$y = Ax + \epsilon \tag{1}$$

where $A \in \mathbb{R}^{n \times n}$ is a linear measurement operator and ϵ denotes the noise affecting the measurements. This is the case of computed tomography [1], depth ranging [2] and non-line-of-sight imaging [3] to name a few. Estimating x from y is generally an ill-posed inverse problem due to the incomplete operator A with $m < n$ and the presence of noise. Knowledge of the signal model is required to make this problem well-posed.

In many cases, obtaining ground-truth reconstructions x to learn the reconstruction function $y \mapsto x$ might be very expensive or even impossible. For example, in medical imaging, it is not always possible to obtain fully sampled images of patients as they require long acquisition times. In astronomical imaging, it is impossible to obtain ground-truth references due to physical limitations. In electron-microscopy imaging [4], we can only measure 2D projections of a molecule. In these settings, we can only access measurements y for learning. Moreover, if the measurement process A is incomplete, it is fundamentally impossible to learn the model with only measurements y, as there is no information about the model in the nullspace of A. Thus, we end up with a chicken-and-egg

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

problem: in order to reconstruct x we need the reconstruction function, but to learn this function we require some reconstructed samples x.

This fundamental limitation can be overcome by using information from multiple incomplete sensing operators $A_1, \ldots, A_G$, the general principle being that each operator can provide additional information about the signal model if it has a different nullspace. For example, in the image inpainting problem, Studer and Baraniuk [5] used the fact that the set of missing pixels may vary between observed images to learn a sparse dictionary model and reconstruct the images. Yang et al. [6] used multiple operators to learn a Gaussian mixture model in the context of hyperspectral imaging and high-speed video. Bora et al. [7] exploited this idea for learning a generative model in various imaging problems such as deblurring and compressed sensing. Matrix completion methods [8] exploit a similar principle, as the missing entries of each column (i.e., signal) are generally different. Ideally we would like to learn the reconstruction function and signal model from only a small number of different measurement operators. We are thus motivated to determine typically how many such operators are required.

The problem can be formalized as follows. We first focus on the noiseless case to study the intrinsic identifiability problems associated to having only incomplete measurement data. The effect of noise will be discussed in Section 4. We assume that we observe a set of N training samples y_i, where the ith signal is observed via $A_{g_i} \in \mathbb{R}^{m \times n}$, one of G linear operators , i.e.,

$$y_i = A_{g_i} x_i \tag{2}$$

where $g_i \in \{1, \ldots, G\}$ and $i = 1, \ldots, N$. While we assume that the measurement operator A_{g_i} is known for all observed signals, it is important to note that we do not know a priori if two observations (y_i, A_{g_i}) and $(y_{i'}, A_{g_{i'}})$ are related to the same signal[1] x_i. There are two natural questions regarding this learning problem:

Q1. Signal Recovery Is there a unique signal $x \in \mathcal{X}$ which verifies the measurements $y = A_g x$? In other words, is the reconstruction function $f : (y, A_g) \mapsto x$ one-to-one?

Q2. Model Identification Can we uniquely identify the signal model from only measurements obtained via these incomplete operators?

In general, there can be a unique solution for neither problem, just one or both. There might be a unique solution for signal recovery if the model is known, but it might be impossible to learn the model in the first place (e.g., blind compressed sensing [9]). The converse is also possible, that is, uniquely identifying a model without having enough measurements per sample to uniquely identify the associated signal (e.g., subspace learning from rank-1 projections [10]).

The answer to **Q1** is well-known from generalized compressed sensing theory, see for example [11]. Unique signal recovery is possible if the signal model is low-dimensional, i.e., if A_g has $m > 2k$ measurements, where k is the model dimension. On the other hand, **Q2** has been mostly studied in the context of matrix completion, where the set of signals is assumed to lie in a low-dimensional subspace of $\mathbb{R}^n$. Bora et al. [7] presented some results in the general setting, but only for the case where $G = \infty$ which is quite restrictive. In this paper, we provide sharp necessary and sufficient conditions which hold for any low-dimensional distribution (beyond linear subspaces) and only require a finite number of operators G.

If the conditions for signal recovery and model identification are satisfied, we can expect to learn the reconstruction function from measurement data alone. We introduce a new unsupervised learning objective which can be used to learn the reconstruction function $f : (y, A_g) \mapsto x$, and provides performances on par with supervised learning when the sufficient conditions are met. The main contributions of this paper are as follows:

- We show that unsupervised learning from a finite number of incomplete measurement operators is only possible if the model is low-dimensional. More precisely, we show that $m \geq n/G$ measurements per operator are necessary for learning, and that for almost every set of G operators, $m > k + n/G$ measurements per operator are sufficient.
- We propose a new unsupervised loss for learning the reconstruction function that only requires incomplete measurement data, which empirically obtains a performance on par with fully supervised methods when the sufficient condition $m > k + n/G$ is met.

[1] If the signal distribution is not atomic, observing such a pair has probability 0.

- A series of experiments demonstrate that our bounds accurately characterize the performance of unsupervised approaches on synthetic and real datasets, and that the proposed unsupervised approach outperforms previous methods in various inverse problems.

1.1 Related Work

Blind Compressed Sensing The fundamental limitation of failing to learn a signal model from incomplete (compressed) measurements data goes back to blind compressed sensing [9] for the specific case of models exploiting sparsity on an orthogonal dictionary. In order to learn the dictionary from incomplete data, [9] imposed additional constraints on the dictionary, while some subsequent papers [12, 13] removed these assumptions by proposing to use multiple operators A_g as studied here. This paper can be seen as a generalization of such results to more general signal models.

Matrix Completion Matrix completion consists of inferring missing entries of a data matrix $Y = [y_1, \dots, y_N]$, whose columns are generally inpainted samples from a low-dimensional distribution, i.e., $y_i = A_{g_i} x_i$ where the operators A_{g_i} randomly select a subset of m entries of the signal x_i. This problem can be viewed as the combination of model identification, i.e., identifying the low-rank subspace that the columns of $X = [x_1, \dots, x_N]$ belong to, and signal recovery, i.e., reconstructing the individual columns. Assuming that the samples belong to a k-dimensional subspace can be imposed by recovering a rank-k signal matrix X from Y. If the columns are sampled via G sufficiently different patterns A_{g_i} with the same number of entries m, a sufficient condition [14] for uniquely recovering almost every subspace model is[2] $m \geq (1 - 1/G)k + n/G$.

A similar condition was shown in [15] for the case of *high-rank* matrix completion [16], which arises when the samples x_i belong to a union of k-dimensional subspaces. We show that model identification is possible for almost every set of G operators with $m > k + n/G$ measurements, however the theory presented here goes beyond linear subspaces, being also valid for general low-dimensional models.

Deep Nets for Inverse Problems Despite providing very competitive results, most deep learning based solvers require measurements and signal pairs (x_i, y_i) (or at least clean signals x_i) in order to learn the reconstruction function $y \mapsto x$ from incomplete measurements. A first step to overcome this limitation is due to Noise2Noise [17], where the authors show that it is possible to learn from only noisy samples. However, their ideas only apply to denoising settings where there is a trivial nullspace, as the operator A is the identity matrix. This approach was extended in [18] to the case where two measurements are observed per signal, each associated with a different operator. Yaman et al. [19] and Artifact2Artifact [20] empirically showed that it is possible to exploit different measurement operators to learn the reconstruction function in the context of magnetic resonance imaging (MRI). AmbientGAN [7] proposed to learn a signal distribution from only incomplete measurements using multiple forward operators, however they only provide reconstruction guarantees for the case where an infinite number of operators A_g is available[3], a condition that is not met in practice.

Another line of work focuses on learning using measurements from a single incomplete operator. The works in [21, 22] use the large system limit properties of random compressed sensing operators to learn from measurements alone. The equivariant imaging approach [23, 24] leverages invariance of the signal set to a group of transformations to learn from general incomplete operators.

2 Signal Recovery Preliminaries

We denote the nullspace of A as $\mathcal{N}_A$. Its complement, the range space of the pseudo-inverse $A^\dagger$, is denoted as $\mathcal{R}_A$, where $\mathcal{R}_A \oplus \mathcal{N}_A = \mathbb{R}^n$ and $\oplus$ denotes the direct sum. Throughout the paper, we assume that the signals are sampled from a measure μ supported on the signal set $\mathcal{X} \subset \mathbb{R}^n$. Signal recovery has a unique solution if and only if the forward operator $x \mapsto y$ is one-to-one, i.e., if for every pair of signals $x_1, x_2 \in \mathcal{X}$ where $x_1 \neq x_2$ we have that

$$Ax_1 \neq Ax_2 \tag{3}$$

$$A(x_1 - x_2) \neq 0 \tag{4}$$

[2] A larger number of measurements $m = \mathcal{O}(k \log n)$ is required to guarantee a stable recovery when the number of patterns G is large [8].

[3] Their result relies on the Cramér-Wold theorem, which is discussed in Section 3.

In other words, there is no vector $x_1 - x_2 \neq 0$ in the nullspace of A. It is well-known that this is only possible if the signal set $\mathcal{X}$ is low-dimensional. There are multiple ways to define the notion of dimensionality of a set in $\mathbb{R}^n$. In this paper, we focus on the upper box-counting dimension which is defined for a compact subset $S \subset \mathbb{R}^n$ as

$$\text{boxdim}(S) = \limsup_{\epsilon \to 0} \frac{\log N(S, \epsilon)}{-\log \epsilon} \tag{5}$$

where $N(S, \epsilon)$ is the minimum number of closed balls of radius ϵ with respect to the norm $\|\cdot\|$ that are required to cover S. This definition of dimension covers both well-behaved models such as compact manifolds and more general low-dimensional sets. The mapping $x \mapsto y$ is one-to-one for almost every forward operator $A \in \mathbb{R}^{m \times n}$ if [25]

$$m > \text{boxdim}(\Delta\mathcal{X}) \tag{6}$$

where $\Delta\mathcal{X}$ denotes the normalized secant set which is defined as

$$\Delta\mathcal{X} = \{\Delta x \in \mathbb{R}^n |\ \Delta x = \frac{x_2 - x_1}{\|x_2 - x_1\|}, x_1, x_2 \in \mathcal{X},\ x_2 \neq x_1\}. \tag{7}$$

The term *almost every* means that the complement has Lebesgue measure 0 in the space of linear measurement operators $\mathbb{R}^{m \times n}$. The normalized secant set of models of dimension k generally has dimension $2k$, requiring $m > 2k$ measurements to ensure signal recovery. For example, the union of k-dimensional subspaces requires at least $2k$ measurements[4] to guarantee one-to-oneness [26]. This includes well-known models such as k-sparse models (e.g., convolutional sparse coding [27]) and co-sparse models (e.g., total variation [28]). In the regime $k < m \leq 2k$, the subset of signals where one-to-oneness fails is at most $(2k - m)$-dimensional [25].

3 Uniqueness of Any Model?

A natural first question when considering uniqueness of the model is: can we recover any measure μ observed via forward operators $A_1, \ldots, A_G$, even in the case where $\mathcal{X}$ is the full $\mathbb{R}^n$? We show that, in general, the answer is no.

Uniqueness can be analysed from the point of view of the characteristic function of μ, defined as $\varphi(w) = \mathbb{E}\{e^{\mathrm{i}w^\top x}\}$ where the expectation is taken with respect to μ and $\mathrm{i} = \sqrt{-1}$ is the imaginary unit. If two distributions have the same characteristic function, then they are necessarily the same almost everywhere. Each forward operator provides information about a subspace of the characteristic function as

$$\mathbb{E}\{e^{\mathrm{i}w^\top A_g^\dagger y}\} = \mathbb{E}\{e^{\mathrm{i}w^\top A_g^\dagger A_g x}\} \tag{8}$$

$$= \mathbb{E}\{e^{\mathrm{i}(A_g^\dagger A_g w)^\top x}\} \tag{9}$$

$$= \varphi(A_g^\dagger A_g w) \tag{10}$$

where $A_g^\dagger A_g$ is a linear projection onto the subspace $\mathcal{R}_{A_g}$. Given that $m < n$, the characteristic function is only observed in the subspaces $\mathcal{R}_{A_g}$ for all $g \in \{1, \ldots, G\}$. For any finite number of operators, the union of these subspaces does not cover the whole $\mathbb{R}^n$, and hence there is loss of information, i.e., the signal model cannot be uniquely identified.

In the case of an infinite number of operators $G = \infty$, the Cramér-Wold theorem guarantees uniqueness of the signal distribution if all possible one dimensional projections ($m = 1$) are available [29, 7]. However, in most practical settings we can only access a finite number of operators and many distributions will be non-identifiable.

4 Uniqueness of Low-Dimensional Models

Most models appearing in signal processing and machine learning are assumed to be approximately low-dimensional, with a dimension k which is much lower than the ambient dimension n. As discussed in Section 2, the low-dimensional property is the key to obtain stable reconstructions, e.g., in compressed sensing. In the rest of the paper, we impose the following assumptions on the model:

[4] While the bound in (6) guarantees *unique* signal recovery, more measurements (e.g., an additional factor of $\mathcal{O}(\log n)$ measurements) are typically necessary in order to have a *stable* inverse $f : y \mapsto x$, i.e., possessing a certain Lipschitz constant. A detailed discussion can be found for example in [11].

A1 The signal set $\mathcal{X}$ is either

(a) A bounded set with box-counting dimension k.

(b) An unbounded conic set whose intersection with the unit sphere has box-counting dimension $k-1$.

This assumption has been widely adopted in the inverse problems literature, as it is a necessary assumption to guarantee signal recovery. Our definition of dimension covers most models used in practice, such as union of subspaces (simple subspace models, convolutional sparse coding models, k-sparse models), low-rank matrices and compact manifolds. It is worth noting that dimension is a property of the dataset and thus independent of the specific algorithm used for learning.

In the rest of the paper, we focus on conditions for the identification of the support $\mathcal{X}$ instead of the signal distribution μ, due to the following observation: if there is a one-to-one reconstruction function (which happens for almost every A with $m > 2k$ as explained in Section 2), uniqueness of the support implies uniqueness of μ. If $\mathcal{X}$ is known and there is a measurable one-to-one mapping from each observed measurement y to $\mathcal{X}$, then it is possible to obtain μ as the push-forward of the measurement distribution.

Before delving into the main theorem, we present a simple example which provides intuition of how a low-dimensional model can be learned via multiple projections A_g:

Learning a one-dimensional subspace Consider a toy signal model with support $\mathcal{X} \subset \mathbb{R}^3$ which consists of a one-dimensional linear subspace spanned by $\phi = [1, 1, 1]^\top$, and $G = 3$ measurement operators $A_1, A_2, A_3 \in \mathbb{R}^{2\times 3}$ which project the signals into the $x(3) = 0$, $x(2) = 0$ and $x(1) = 0$ planes respectively, where $x(i)$ denotes the ith entry of the vector x. The example is illustrated in Figure 1. The first operator A_1 imposes a constraint on $\mathcal{X}$, that is, every $x \in \mathcal{X}$ should verify $x(1) - x(2) = 0$. Without more operators providing additional information about $\mathcal{X}$, this constraint yields a plane containing $\mathcal{X}$, and there are infinitely many one-dimensional models that would fit the training data perfectly. However, the additional operator A_2 adds the constraint $x(2) - x(3) = 0$, which is sufficient to uniquely identify $\mathcal{X}$ as

$$\hat{\mathcal{X}} = \mathcal{X} = \{v \in \mathbb{R}^3 |\ v(1) - v(2) = v(2) - v(3) = 0\}$$

is the desired 1-dimensional subspace. Finally, note that in this case the operator A_3 does not restrict the signal set further, as the constraint $x(1) - x(3) = 0$ is verified by the other two constraints.

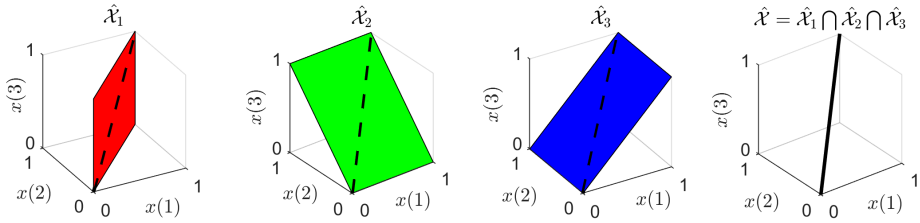


Figure 1: Toy example of a 1-dimensional subspace embedded in $\mathbb{R}^3$. If we only observe the projection of the signal set into the plane $x(3) = 0$, then there are infinite possible lines that are consistent with the measurements (red plane). Adding the projection into the $x(1) = 0$ plane, allows us to uniquely identify the signal model.

The ideas from the one-dimensional subspace example can be generalized and formalized as follows: for each projection A_g, we can constrain the model support $\mathcal{X}$ by considering the set

$$\hat{\mathcal{X}}_g = \{v \in \mathbb{R}^n |\ v = \hat{x}_g + u,\ \hat{x}_g \in \mathcal{X},\ u \in \mathcal{N}_{A_g}\} \tag{11}$$

which has dimension at most $n - (m - k)$. Note that the true signal model is a subset of $\hat{\mathcal{X}}_g$. The inferred signal set belongs to the intersection of these sets

$$\hat{\mathcal{X}} = \bigcap_{g \in \mathcal{G}} \hat{\mathcal{X}}_g \tag{12}$$

which can be expressed concisely as

$$\hat{\mathcal{X}} = \{v \in \mathbb{R}^n |\ A_g(x_g - v) = 0,\ g = 1, \ldots, G, x_1, \ldots, x_G \in \mathcal{X}\} \tag{13}$$

Even though we have derived the set $\hat{\mathcal{X}}$ from a purely geometrical argument, the constraints in (13) also offer a simple algebraic intuition: the inferred signal set consists of the points $v \in \mathbb{R}^n$ which verify the following system of equations

$$\begin{bmatrix} A_1 \\ \vdots \\ A_G \end{bmatrix} v = \begin{bmatrix} A_1 x_1 \\ \vdots \\ A_G x_G \end{bmatrix}. \tag{14}$$

for all possible choices of G points $x_1, \dots, x_G$ in $\mathcal{X}$. In other words, given a dataset of N incomplete measurements $\{A_{g_i} x_i\}_{i=1}^N$, it is possible to build $\hat{\mathcal{X}}$ by trying all the possible combinations of G samples[5] and keeping only the points v which are the solutions of (14).

It is trivial to see that $\mathcal{X} \subseteq \hat{\mathcal{X}}$, but when can we guarantee $\mathcal{X} = \hat{\mathcal{X}}$? As in the previous toy example, if there are not enough constraints, e.g., if we have a single A and no additional measurement operators, the inferred set will have a dimension larger than k, containing undesired aliases. In particular, we have the following lower bound on the minimum number of measurements:

Proposition 4.1 (Theorem 1 in [23])**.** *A necessary condition for model uniqueness is that*

$$\textit{rank}\left(\begin{bmatrix} A_1 \\ \vdots \\ A_G \end{bmatrix}\right) = n \tag{15}$$

and thus $m \geq n/G$.

Proof. In order to have model uniqueness, the system in (14) should only admit a solution if $v = x_1 = \cdots = x_G$. If the rank condition in (15) is not satisfied, there is more than one solution for any choice of $x_1, \dots, x_G \in \mathcal{X}$. □

Note that this necessary condition does not take into account the dimension of the model. As discussed in Section 3, a sufficient condition for model uniqueness must depend on the dimension of the signal set k. Our main theorem shows that k additional measurements per operator are sufficient for model identification:

Theorem 4.2. *For almost every set of G mappings $A_1, \dots, A_G \in \mathbb{R}^{m \times n}$, under assumption* **A1** *the signal model $\mathcal{X}$ can be uniquely recovered if the number of measurements verify $m > k + n/G$.*

The proof is included in Appendix A. If we have a large number of independent operators $G \geq n$, Theorem 4.2 states that only $m > k + 1$ measurements are sufficient for model identification, which is slightly smaller (if the model is not trivial, i.e., $k > 1$) than the number of measurements typically needed for signal recovery $m > 2k$. In this case, it is possible to uniquely identify the model, without necessarily having a unique reconstruction of each observed signal. However, as discussed in Section 2, for $k < m \leq 2k$, the subset of signals which cannot be uniquely recovered is at most $(2k - m)$-dimensional.

Operators with Different Number of Measurements The results of the previous subsections can be easily extended to the setting where each measurement operator has a different number of measurements, i.e. $A_1 \in \mathbb{R}^{m_1 \times n}, \dots, A_G \in \mathbb{R}^{m_G \times n}$. In this case, the necessary condition in Proposition 4.1 is $\sum_{g=1}^G m_g \geq n$, and the sufficient condition in Theorem 4.2 is $\frac{1}{G}\sum_{g=1}^G m_g > k + n/G$. As the proofs mirror the ones of Proposition 4.1 and Theorem 4.2, we leave the details to the reader.

Noisy measurement data Surprisingly, the results of this section are also theoretically valid if the measurements are corrupted by independent additive noise ϵ, i.e., $y = A_g x + \epsilon$, as long as the noise distribution is *known* and has a nowhere zero characteristic function (e.g., Gaussian noise):

Proposition 4.3. *For a fixed noise distribution, if its characteristic function is nowhere zero, then there is a one-to-one mapping between the space of clean measurement distributions and noise measurement distributions.*

[5] Despite providing a good intuition, this procedure for estimating $\mathcal{X}$ is far from being practical as it would require an infinite number of observed samples if the dimension of the signal set is not trivial $k > 0$.

The proof is included in Appendix A. If the clean measurement distribution can be uniquely identified from the noisy distribution, the results in Theorem 4.2 and Proposition 4.1 also carry over to the noisy setting. Note that this only guarantees model identifiability and makes no claims on the sample complexity of any learning process.

5 Algorithms

Unsupervised algorithms mainly come in two flavours: we can first learn a model $\hat{\mathcal{X}}$ to then reconstruct by projecting measurements into this set, or we can attempt to directly learn the reconstruction function parameterized by a deep network.

5.1 Learn the Model and Reconstruct

Dictionary and Subspace Learning If $\mathcal{X}$ is (approximately) a union of subspaces [9] or a single subspace [8], we can learn a model by

$$\arg\min_{z,D} \mathbb{E}_{(y,g)} \|y - A_g D z\|^2 + \rho_1(D) + \rho_2(z) \tag{16}$$

where $\rho_1(D)$ and $\rho_2(z)$ are regularisation terms that promote low-dimensional solutions, e.g., sparse codes z if $\mathcal{X}$ is a union of subspaces. At test time, the dictionary is fixed and the optimization is performed over the codes only.

AmbientGAN Complex datasets are often better modelled by a generative network $f : \mathbb{R}^k \mapsto \mathbb{R}^n$ whose input is a low-dimensional latent code $z \in \mathbb{R}^k$. The generative model can be learned using an adversarial strategy, i.e.,

$$\arg\min_{f} \max_{d} \mathbb{E}_{(y,g)} q\{d(A_g^\dagger y)\} + \mathbb{E}_z \mathbb{E}_g q\left\{1 - d\left(A_g^\dagger A_g f(z)\right)\right\} \tag{17}$$

where $d : \mathbb{R}^n \mapsto \mathbb{R}^n$ is the discriminator network which compares measurements in the image domain, z is usually sampled from a Gaussian distribution, and $q(t) = \log(t)$ for standard GANs and $q(t) = t$ for Wasserstein GANs. At test time, the reconstruction can be obtained by finding the latent code that best fits the measurements $\hat{x} = f(\arg\min_z \|y - A_g f(z)\|^2)$, as in [30].

5.2 Learn to Reconstruct

Another approach consists in learning directly the reconstruction mapping $f : \mathbb{R}^m \times \mathbb{R}^{m\times n} \mapsto \mathbb{R}^n$ whose inputs are the noisy measurement y and the associated operator A_g, and the output is the reconstructed signal x. The reconstruction can have either a denoiser form, $f(y, A_g) = \tilde{f}(A_g^\dagger y)$ where $\tilde{f}$ is independent of A_g [1], or a more complex unrolled structure with many denoising and gradient steps [31].

Measurement Splitting Inspired by the Noise2Noise approaches [17, 18], some self-supervised methods [19, 20] split each measurement into two parts, $y^\top = [y_1^\top, y_2^\top]$, such that the input is y_1 and the target is y_2. These methods can be summarised as minimising the following loss

$$\arg\min_{f} \mathbb{E}_{(y,g)} \|y_2 - A_{g,2} f(y_1, A_{g,1})\|^2 \tag{18}$$

where $y_1 = A_{g,1}x + \epsilon_1$ and $y_2 = A_{g,2}x + \epsilon_2$. This approach suffers from the fact that f does not use all the available information in a given measurement, as it attempts to solve a harder reconstruction problem associated with $y_1 = A_{g,1}x + \epsilon_1$. As reconstruction networks often fail to generalise to operators with more measurements [32], this method can suffer from suboptimal reconstructions at test time.

Proposed Method The analysis in Section 4 shows that model identifiability necessarily requires that reconstructed signals are consistent with all operators $A_1, \ldots, A_G$. Thus, we propose an unsupervised loss that ensures consistency across all projections A_g, that is

$$\arg\min_{f} \mathbb{E}_{(y,g)} \left\{\|y - A_g f(y, A_g)\|^2 + \mathbb{E}_s \|\hat{x} - f(A_s \hat{x}, A_s)\|^2\right\} \tag{19}$$

where $\hat{x} = f(y, A_g)$. The first term ensures measurement consistency $y = A_g f(y, A_g)$, whereas the second term enforces consistency across operators, i.e., $f(y, A_g) = f(A_s f(y, A_g), A_s)$ for all $g \neq s$. Crucially, the second term prevents the network from learning the trivial pseudo-inverse $f(y, A_g) = A_g^\dagger y$. In practice, we choose an operator A_s uniformly at random per minibatch. Thus, compared the supervised case, we only require an additional evaluation of f and A_s per minibatch. We coin this approach multi-operator imaging (MOI).

This loss overcomes the main disadvantages of previous approaches: it doesn't require training a discriminator network, and it learns to reconstruct using all the available information in each measurement y.

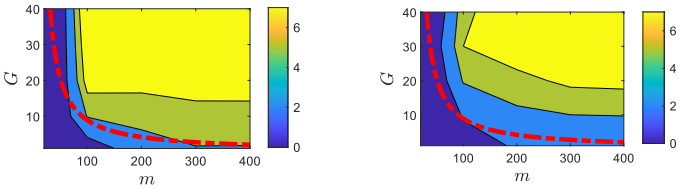
6 Experiments

In this section, we present a series of experiments where the goal is to learn the reconstruction function with deep networks using real datasets. Experiments on low-dimensional subspace learning using synthetic datasets are presented in Appendix C.1. All our experiments were performed using an internal cluster of 4 NVIDIA RTX 3090 GPUs with a total compute time of approximately 48 hs.

Compressed Sensing and Inpainting with MNIST We evaluate the theoretical bounds on the MNIST dataset, based on the well known approximation of its box-counting dimension $k \approx 12$ [33]. The dataset contains $N = 60000$ training samples, and these are partitioned such that N/G different samples are observed via each operator. The forward operators are compressed sensing (CS) matrices with entries sampled from a Gaussian distribution with zero mean and variance m^{-1}. The test set consists of 10000 samples, which are also randomly divided into G parts, one per operator. To evaluate the theoretical bounds, we attempt to minimize the impact of the inductive bias of the networks' architecture [34, 35] by using a network with 5 fully connected layers and relu non-linearities.

Figure 2a shows the average test peak-signal-to-noise ratio (PSNR) achieved by the model trained using the proposed MOI loss for $G = 1, 10, 20, 30, 40$ and $m = 50, 100, 200, 300, 400$. The results follow closely the bound presented in Section 4 which is indicated by the red dashed line, as the network is only able to learn the reconstruction mapping when the sufficient condition $m > k + n/G$ is verified. In sampling regimes below this condition, the performance is similar to simply applying the pseudo-inverse $A_g^\dagger$.

We also evaluate the reconstruction for a different number G of random inpainting masks and different rates m. The inpainting operators have a diagonal structure which has zero measure in $\mathbb{R}^{m \times n}$, however our sufficient condition still provides a reasonable lower bound on predicting the performance, as shown in Figure 2b. It is likely that due to the coherence between measurement operators and images (both operators and MNIST images are sparse), more measurements are required to obtain good reconstructions than in the CS case.



(a) Compressed Sensing (b) Inpainting

Figure 2: Average test PSNR improvement in dB over the pseudo-inverse for the MNIST dataset using the proposed training loss, for different number of CS operators or inpainting masks and measurements per operator. The curve in red shows the necessary condition of Theorem 4.2, $m > k + n/G$.

Inpainting with CelebA We evaluate the unsupervised methods in Section 5 on the CelebA dataset [36], which is split into 32556 images for training and 32556 images for testing. We use the same U-Net architecture (see Appendix B for details) for MOI and supervised learning, and the DCGAN architecture [37] for AmbientGAN as in [7]. We observed that using the U-Net for AmbientGAN's generator achieves worse results than the DCGAN architecture. Reconstructed test images are shown in Figure 3 and average test PSNR is presented in Table 1. The proposed method

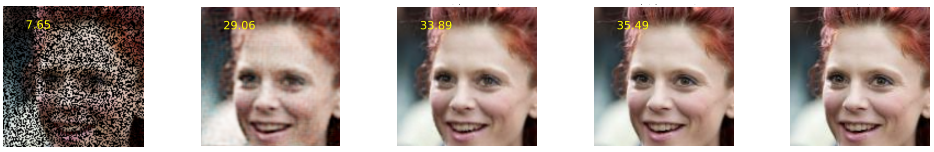


Figure 3: Examples of reconstructed test images for the inpainting task using the CelebA dataset. From left to right: pseudo-inverse $A^\dagger y$, AmbientGAN, MOI, supervised and ground-truth. PSNR values are reported in yellow.

	$A^\dagger y$	AmbientGAN	MOI (ours)	Supervised
Inpainting/CelebA	9.05±1.65	29.57±1.24	**34.05±3.77**	36.21±3.76
	$A^\dagger y$	Meas. Splitting	MOI (ours)	Supervised
Acc. MRI/FastMRI	25.77±2.71	28.72±1.64	**29.51±1.85**	31.45±1.98

Table 1: Comparison of supervised and unsupervised learning methods for inpainting and accelerated MRI. Reported values correspond to average PSNR in dB on the testing set.

obtains an improvement of more than 4 dB with respect to AmbientGAN and falls only 2.1 dB behind fully-supervised learning.

Accelerated MRI with FastMRI Finally, we consider the FastMRI dataset [38], where the set of forward operators A_g consist of different sets of single-coil k-space measurements, with $4\times$ acceleration, i.e., $m/n = 0.25$. We used 900 images for training and 74 for testing, which we split across $G = 40$ operators. We compare measurement splitting, MOI and supervised learning, all using the same U-Net architecture. For the measurement splitting method, we follow the strategy in [19], and choose to assign a random subset representing 60% of the measurements in A_g to $A_{g,1}$ and the remaining to $A_{g,2}$. We observed that a model trained with measurement splitting obtained less test error using reduced measurements associated with $A_{g,1}$ instead of the full measurements A_g (see Appendix B), so we report the best results using $A_{g,1}$. As observed in [32], the network fails to generalize to operators with more measurements. Average test PSNR is presented in Table 1. Reconstructed test images are shown in Appendix C. The proposed method obtains .8 dB improvement over measurement splitting and is approximately 2 dB worse than the fully supervised setting.

7 Limitations

Theorem 4.2 does not cover cases where the operators A_g present some problem specific constraints (e.g., they are inpainting matrices) as well as cases where the signal model is only approximately low dimensional. Note however that Proposition 4.1 applies to constrained operators. We leave the study of sufficient conditions for these particular cases for future work. The proposed loss might not be effective in problems where learning the reconstruction function is impossible, e.g., due to very high noise affecting the measurements [4]. In this particular case, it might still be possible to learn a generative model as in AmbientGAN.

8 Conclusions

We have presented sampling theorems for the unsupervised learning of signal models from incomplete measurements using multiple measurement operators. Our bounds characterize the interplay between the fundamental properties of the problem: the ambient dimension, the data dimension and the number of measurement operators. Moreover, the bounds are agnostic of the learning algorithms and provide useful necessary and sufficient conditions for designing principled sensing strategies. Furthermore, we presented a new practical unsupervised learning loss which learns to reconstruct incomplete measurement data from multiple operators, outperforming previously proposed unsupervised methods. The proposed strategy avoids the adversarial training in AmbientGAN, which can suffer from mode collapse [39], and, contrary to measurement splitting, is trained using full operators A_g. Our results shed light into the setting where access to ground truth data cannot be guaranteed which is of extreme importance in various applications.

References

[1] 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, 2017.

[2] Joshua Rapp, Julian Tachella, Yoann Altmann, Stephen McLaughlin, and Vivek K Goyal. Advances in single-photon lidar for autonomous vehicles: Working principles, challenges, and recent advances. *IEEE Signal Processing Magazine*, 37(4):62–71, 2020.

[3] Matthew O'Toole, David B Lindell, and Gordon Wetzstein. Confocal non-line-of-sight imaging based on the light-cone transform. *Nature*, 555(7696):338–341, 2018.

[4] Harshit Gupta, Thong H Phan, Jaejun Yoo, and Michael Unser. Multi-CryoGAN: Reconstruction of continuous conformations in cryo-em using generative adversarial networks. In *European Conference on Computer Vision*, pages 429–444. Springer, 2020.

[5] Christoph Studer and Richard G Baraniuk. Dictionary learning from sparsely corrupted or compressed signals. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3341–3344. IEEE, 2012.

[6] Jianbo Yang, Xuejun Liao, Xin Yuan, Patrick Llull, David J. Brady, Guillermo Sapiro, and Lawrence Carin. Compressive sensing by learning a gaussian mixture model from measurements. *IEEE Transactions on Image Processing*, 24(1):106–119, 2015.

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

[8] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[9] Sivan Gleichman and Yonina C Eldar. Blind compressed sensing. *IEEE Transactions on Information Theory*, 57(10):6958–6975, 2011.

[10] Y. Chen, Y. Chi, and A. J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

[11] Anthony Bourrier, Mike E Davies, Tomer Peleg, Patrick Pérez, and Rémi Gribonval. Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems. *IEEE Transactions on Information Theory*, 60(12):7928–7946, 2014.

[12] Jorge Silva, Minhua Chen, Yonina C Eldar, Guillermo Sapiro, and Lawrence Carin. Blind compressed sensing over a structured union of subspaces. *arXiv preprint arXiv:1103.2469*, 2011.

[13] Mohammad Aghagolzadeh and Hayder Radha. New guarantees for blind compressed sensing. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1227–1234. IEEE, 2015.

[14] Daniel L Pimentel-Alarcón, Nigel Boston, and Robert D Nowak. A characterization of deterministic sampling patterns for low-rank matrix completion. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):623–636, 2016.

[15] Daniel Pimentel-Alarcon and Robert Nowak. The information-theoretic requirements of subspace clustering with missing data. In *International Conference on Machine Learning*, pages 802–810. PMLR, 2016.

[16] Brian Eriksson, Laura Balzano, and Robert Nowak. High-rank matrix completion. In *Artificial Intelligence and Statistics*, pages 373–381. PMLR, 2012.

[17] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, Timo Aila, et al. Noise2noise. In *International Conference on Machine Learning*. PMLR, 2018.

[18] Zhihao Xia and Ayan Chakrabarti. Training image estimators without image ground truth. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[19] 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.

[20] 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, 2020.

[21] Christopher A Metzler, Ali Mousavi, Reinhard Heckel, and Richard G Baraniuk. Unsupervised learning with stein's unbiased risk estimator. *arXiv preprint arXiv:1805.10531*, 2018.

[22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10255–10264, 2019.

[23] Dongdong Chen, Julián Tachella, and Mike E Davies. Equivariant imaging: Learning beyond the range space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4379–4388, October 2021.

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5647–5656, 2022.

[25] Tim Sauer, James A Yorke, and Martin Casdagli. Embedology. *Journal of statistical Physics*, 65(3):579–616, 1991.

[26] T. Blumensath and M. E. Davies. Sampling theorems for signals from the union of finite-dimensional linear subspaces. *IEEE Transactions on Information Theory*, 55(4):1872–1882, 2009.

[27] Hilton Bristow, Anders Eriksson, and Simon Lucey. Fast convolutional sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 391–398, 2013.

[28] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[29] Harald Cramér and Herman Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, 1(4):290–294, 1936.

[30] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning*, pages 537–546. PMLR, 2017.

[31] Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

[32] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of ai. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020.

[33] Matthias Hein and Jean-Yves Audibert. Intrinsic dimensionality estimation of submanifolds in rd. In *Proceedings of the 22nd international conference on Machine learning*, pages 289–296, 2005.

[34] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

[35] Julian Tachella, Junqi Tang, and Mike Davies. The neural tangent link between CNN denoisers and non-local filters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8618–8627, June 2021.

[36] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[37] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[38] Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J. Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzalv, Adriana Romero, Michael Rabbat, Pascal Vincent, James Pinkerton, Duo Wang, Nafissa Yakubova, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated mr image reconstruction using machine learning. *Radiology: Artificial Intelligence*, 2(1):e190007, 2020. PMID: 32076662.

[39] Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do GANs learn the distribution? some theory and empirics. In *International Conference on Learning Representations*, 2018.

[40] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] We used existing models and datasets.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] All datasets considered in the paper are publicly available.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not use any crowdsourcing or conducted research with human subjects.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs

The proof of Theorem 4.2 utilises the following technical lemma:

Lemma A.1 (Lemmas 4.5 and 4.6 in [25])**.** *Let S be a bounded subset of $\mathbb{R}^n$, and let $G_0, G_1, \ldots, G_t$ be Lipschitz maps from S to $\mathbb{R}^m$. For each integer $r \geq 0$, let S_r be the subset of $z \in S$ such that the rank of the $m \times t$ matrix*

$$\Phi_z = [G_1(z), \ldots, G_t(z)] \tag{20}$$

is r, and let boxdim$(S_r) = k_r$. For each $\alpha \in \mathbb{R}^t$ define $G_\alpha(z) = G_0 + \Phi_z \alpha$. If for all integers $r \geq 0$ we have that $r > k_r$, then $G_\alpha^{-1}(0)$ is empty for almost every $\alpha \in \mathbb{R}^t$.

Proof. The proof of Lemma A.1 follows standard covering arguments and may be sketched as follows. From the dimensionality assumption, the set S_r can be essentially covered by $\mathcal{O}(\epsilon^{-k_r})$ ϵ-balls. Furthermore, for any $z \in S_r$, the probability (measured with respect to $\alpha \in \mathbb{R}^t$) that $G_\alpha(z)$ maps to the neighborhood of 0 scales as ϵ^r. Hence the probability of this happening for any of the points in the cover scales as ϵ^{r-k_r}. If we take $r > k_r$ then the probability of such an event tends to zero as we shrink ϵ. Full details can be found in the proofs in [25]. □

We can now present the proof of Theorem 4.2:

Proof. In order to have model uniqueness, we require that the inferred signal set $\hat{\mathcal{X}}$ defined as

$$\hat{\mathcal{X}} = \{v \in \mathbb{R}^n |\ A_g(x_g - v) = 0,\ g = 1, \ldots, G, x_1, \ldots, x_G \in \mathcal{X}\} \tag{21}$$

equals the true set $\mathcal{X}$, or equivalently that their difference

$$\hat{\mathcal{X}} \setminus \mathcal{X} = \{v \in \mathbb{R}^n \setminus \mathcal{X} |\ A_1(x_1 - v) = \cdots = A_G(x_G - v) = 0,\ x_1, \ldots, x_G \in \mathcal{X}\} \tag{22}$$

is empty, where $\setminus$ denotes set difference. Let $S \subset \mathbb{R}^{n(G+1)}$ be the set of all vectors $z = [v, x_1, \ldots, x_G]^\top$ with $v \in \mathbb{R}^n \setminus \mathcal{X}$ and $x_1, \ldots, x_G \in \mathcal{X}$. The difference set defined in (22) is empty if and only if for any $z \in S$ we have

$$\underbrace{\begin{bmatrix} -A_1 & A_1 & & \\ \vdots & & \ddots & \\ -A_G & & & A_G \end{bmatrix}}_{G_\alpha \in \mathbb{R}^{mG \times n(G+1)}} \underbrace{\begin{bmatrix} v \\ x_1 \\ \vdots \\ x_G \end{bmatrix}}_{z \in S} \neq 0 \tag{23}$$

$$G_\alpha(z) \neq 0 \tag{24}$$

where G_α maps $z \in S$ to $\mathbb{R}^{mG}$. Let $\alpha = [\text{vec}(A_1)^\top, \ldots, \text{vec}(A_G)^\top]^\top \in \mathbb{R}^{mnG}$, then as a function of α we can also write (23) as

$$\begin{bmatrix} (x_1 - v)^\top \otimes I_m & & \\ & \ddots & \\ & & (x_G - v)^\top \otimes I_m \end{bmatrix} \alpha \neq 0 \tag{25}$$

where $\otimes$ is the Kronecker product and we used the fact that $A(x_g - v) = (x_g - v)^\top \otimes I_m \text{vec}(A)$. As v does not belong to the signal set, the matrix on the left hand side of (25) has rank mG for all $z \in S$. We treat the cases of bounded and conic signal sets separately, showing in both cases that, for almost every $\alpha \in \mathbb{R}^{mnG}$, the condition in (25) holds for all $z \in S$ if $m > k + n/G$:

Bounded signal set Let S_ρ be a subset of S defined as

$$S_\rho = \{z \in \mathbb{R}^{n(G+1)} \ |z = [v^\top, x_1^\top, \ldots, x_G^\top]^\top, x_1, \ldots, x_G \in \mathcal{X},\ \|v\|_2 \leq \rho\}. \tag{26}$$

As S_ρ is bounded, we have boxdim$(S_\rho) \leq kG + n$. Thus, if $mG > kG + n$, Lemma A.1 states that for almost every α, (25) holds for all $z \in S_\rho$. As S can be decomposed as a countable union of S_ρ of increasing radius, i.e., $S = \bigcup_{\rho \in \mathbb{N}} S_\rho$, and a countable union of events of measure zero has measure zero, then for almost every α all $z \in S$ verifies (25) if $m > k + n/G$.

Conic signal set If the signal set is conic, then S is also conic. Hence, due to the linearity of (23) with respect to z, there exists $z \in S$ which does not verify (23) if and only if for any bounded set B containing an open neighbourhood of 0, there exists a $z \in S \cap B$ which does not verify (23). As $\text{boxdim}(S \cap B) \leq Gk + n$, Lemma A.1 states that for almost every α, all $z \in S$ verifies (25) as long as $m > k + nG$.

□

We end this section with the proof of Proposition 4.1:

Proof. Consider the noisy measurements associated to the gth operator A_g, as $z = y + \epsilon$, where z are the observed noisy measurements, y are the clean measurements and ϵ is additive noise (independent of y). The characteristic function of the sum of two independent random variables is given by the multiplication of their characteristic functions, i.e.,

$$\varphi_z(w) = \varphi_y(w)\varphi_\epsilon(w) \tag{27}$$

where φ_z, φ_y and φ_ϵ are the characteristic functions the noisy measurement, clean measurements and noise distributions, respectively. If the characteristic function of the noise distribution is nowhere zero, we can uniquely identify the characteristic function of the clean measurement distribution as

$$\varphi_y(w) = \varphi_z(w)/\varphi_\epsilon(w) \tag{28}$$

The clean measurement distribution is fully characterized by its characteristic function $\varphi_y(w)$. We end the proof by noting that the same reasoning applies to the measurements of every operator A_g with $g \in \{1, \ldots, G\}$. □

B Training Details

Algorithm 1 provides the pseudo-code of the proposed multi-operator imaging (MOI) method. The training details for each task are as follows:

Algorithm 1: Pseudocode of MOI in a PyTorch-like style.

```
# 𝒢: forward operators 𝒢={1,...,G}
# f: reconstruction function (e.g., neural network)
for y, A_g in loader: # load a minibatch y with N samples and its corresponding operator A_g
   # randomly select a operator from 𝒢/g
   s = select(𝒢/g)
   x1 = f(y, A_g) # reconstruct x from y
   x2 = f(A_s(x1), A_s) # reconstruct x1
   # MOI training loss, Eqn.(18)
   loss = MSELoss(A_g(x1), y) # measurement consistency
         + MSELoss(x2, x1) # cross-operator consistency
   # update f network
   loss.backward()
   update(f.params)
```

Compressed Sensing and Inpainting with MNIST. In both cases, we use the Adam optimizer with a batch size of 128 and weight decay of 10^{-8}. We use a fully connected network with 5 layers, where the number neurons in each layer are $784, 1000, 32, 1000, 784$ respectively. The nonlinearity is relu and the network has a residual connection between the input and output. For the CS task, we use an initial learning rate of 10^{-4} and train the networks for 1000 epochs, keeping the learning rate constant for the first 800 epochs and then shrinking it by a factor of 0.1. For the inpainting task, we use an initial learning rate of 5×10^{-4} and train the networks for 500 epochs, keeping the learning rate constant for the first 300 epochs and then shrinking it by a factor of 0.1.

Inpainting with CelebA. The CelebA dataset contains more than 200K celebrity images, each with 40 binary attributes. We pick the attribute *smile* to evaluate the proposed method. The center part of the aligned images in the CelebA dataset are cropped to 128×128. We divide the selected images into two subsets for training and testing. There are 32557 images in each subset, which we split across $G = 40$ operators. For the inpainting task, we use the U-Net architecture (see Figure 4) to implement

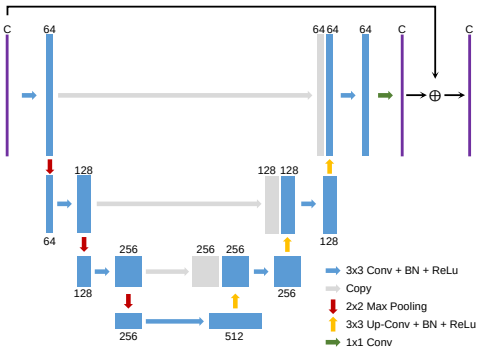


Figure 4: The residual U-Net used in the paper. The number of input and output channels is denoted as C, such that $C = 2$ for MRI and $C = 3$ for inpainting.

the reconstruction function $f(y, A)$, and the DCGAN architecture for AmbientGAN. Using the U-Net architecture for AmbientGAN's generator obtains an average test PSNR of 27.5 ± 1.3 dB, which is 2.1 dB below the performance obtained by the DCGAN generator reported in Section 6. We use Adam with a batch size of 20, an initial learning rate of 5×10^{-4} and a weight decay of 10^{-8}. We train the networks for 300 epochs, shrinking the learning rate by a factor of 0.1 after the first 200 epochs.

Accelerated MRI with fastMRI. For the $4\times$ accelerated MRI task, we use Adam with a batch size of 4 and an initial learning rate of 5×10^{-4}. We also use the U-Net in Figure 4 to implement the reconstruction function $f(y, A)$. We train the networks for 500 epochs, keeping the learning rate constant for the first 300 epochs and then shrinking it by a factor of 0.1. In all experiments, we use complex-valued data and treat real and imaginary parts of the images as separate channels. For the purpose of visualization, we display only the magnitude images. Reconstructed test images are shown in Figure 5.

The measurement splitting method performs worse at test time if we use the full measurements y associated with A_g instead of the reduced measurements y_1 associated with $A_{g,1}$. Using the full measurements obtains an average test PSNR of 27.3 dB, i.e., a decrease of 1.4 dB with respect to the performance using $A_{g,1}$ reported in the main paper.

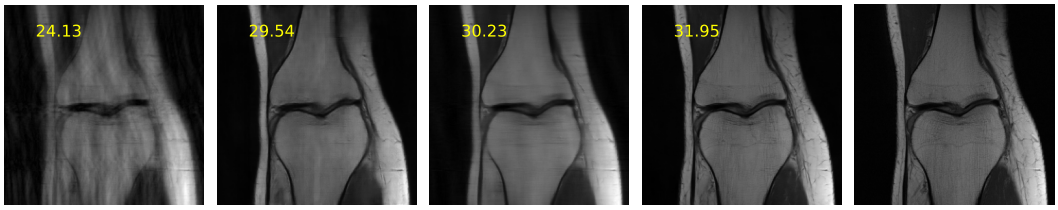


Figure 5: Examples of reconstructed test images for the accelerated MRI task using the fastMRI dataset using the denoiser architecture. From left to right: pseudo-inverse $A^\dagger y$, measurement splitting, MOI, supervised and ground-truth.

C Additional Experiments

C.1 Subspace Learning

We consider the problem of learning a k-dimensional subspace model from partial observations, where the signals x_i are generated from a standard Gaussian distribution on the low-dimensional subspace. The observations y_i are obtained by randomly choosing one out of G operators $A_1, \ldots, A_G \in \mathbb{R}^{m \times n}$, each composed of iid Gaussian entries of mean 0 and variance n^{-1}. In order to recover the signal

matrix $X = [x_1, \ldots, x_N]$, we solve the following low-rank matrix recovery problem

$$\begin{aligned} &\arg\min_{X} \|X\|_* \\ &\text{s.t. } A_{g_i} x_i = y_i \quad \forall i = 1, \ldots, N \end{aligned} \tag{29}$$

where $\|\cdot\|_*$ denotes the nuclear norm. A recovery is considered successful if $\frac{\sum_i \|\hat{x}_i - x_i\|^2}{\sum_i \|x_i\|^2} < 10^{-1}$, where $\hat{x}_i$ is the estimated signal for the ith sample. We use a standard matrix completion algorithm [40] to solve (29). The ambient dimension is fixed at $n = 50$, and the experiment is repeated for $k = 1, 10, 40$. For each experiment we set $N = 150k$ in order to have enough samples to estimate the subspaces [14]. Figure 6 shows the probability of recovery over 25 Monte Carlo trials for different numbers of measurements m and operators G. The reconstruction probability exhibits a sharp transition which follows the bound presented in Theorem 4.2, i.e., $m > k + n/G$.

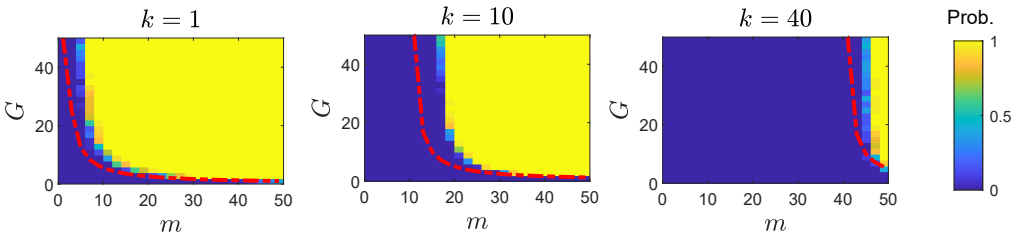


Figure 6: Reconstruction probability of a k-dimensional subspace using incomplete measurements arising from G independent operators for different k. The curve in red shows the bound of Theorem 4.2, $m > k + n/G$.

C.2 Learning with Unrolled Network Architecture

Section 6 of the main paper presents results comparing the proposed unsupervised loss and competing approaches using the denoising architecture, i.e. $f(y, A) = \tilde{f}(A^\dagger y)$, where $\tilde{f}$ is the U-Net presented in Figure 4. Here we are interested in investigating the performance of MOI and other competing approaches using a more complex unrolled architecture. We use an unrolled proximal gradient descent (PGD) network (see Algorithm (30)) with $T = 3$ iterations. The step size is initialized as $\eta^{(t)} = 0.4$ and is then learned during training. We employ 3 U-Net networks using the architecture presented in Figure 4 to implement $f^{(t)}$ for $t = 1, 2, 3$ (no weight sharing across PGD iterations). The average test PSNR is shown in Table 2 and reconstructed images are shown in Figure 7. All training approaches obtain a better performance than using the denoising network due to the architectural improvements. MOI obtains 1.9 dB improvement over measurement splitting and is 1 dB worse than supervised learning.

Unrolled Proximal Gradient Descent (PGD)
input: y, A
$x^{(0)} \leftarrow A^\dagger y$
for $t = 0, 1, \cdots, T-1$:
$x^{(t+1)} \leftarrow f^{(t+1)}(x^{(t)} - \eta^{(t)} A^\top (Ax^{(t)} - y))$
end for
return $f(y, A) := x^{(T)}$ (30)

Unrolling (PGD-3)	$A^\dagger y$	Meas. Splitting	MOI (ours)	Supervised
Acc. MRI/FastMRI	25.77±2.71	29.47±2.02	**31.39±2.17**	32.42±2.44

Table 2: Comparison of supervised and unsupervised learning methods using an unrolled architecture for the accelerated MRI. Reported values correspond to average PSNR in dB on the testing set.

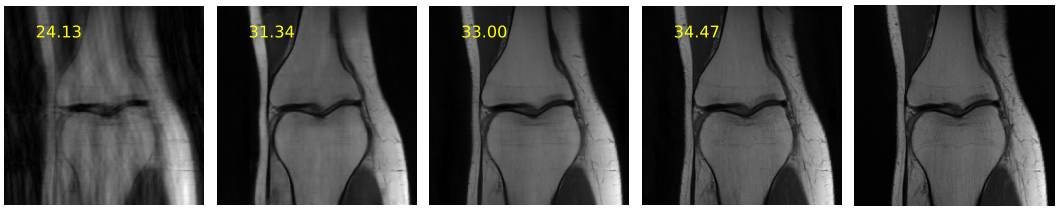


Figure 7: Examples of reconstructed test images for the accelerated MRI task using an unrolled network architecture (PGD-3). From left to right: pseudo-inverse $A^\dagger y$, measurement splitting, MOI, supervised and ground-truth. PSNR values are reported in yellow.