Equivariant imaging: Learning beyond the range space

Equivariant Imaging: Learning Beyond the Range Space

Dongdong Chen
School of Engineering
University of Edinburgh

Julián Tachella
School of Engineering
University of Edinburgh

Mike E. Davies
School of Engineering
University of Edinburgh

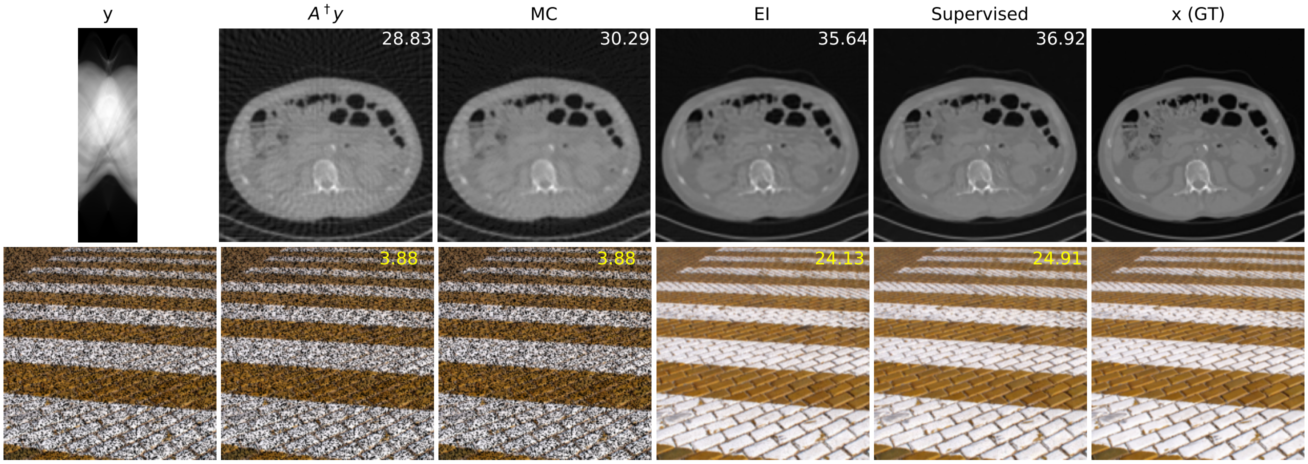

Figure 1: **Learning to image from only measurements**. Training an imaging network through just measurement consistency (MC) does not significantly improve the reconstruction over the simple pseudo-inverse ($A^\dagger y$). However, by enforcing invariance in the reconstructed image set, *equivariant imaging* (EI) performs almost as well as a fully supervised network. **Top**: sparse view CT reconstruction, **Bottom**: pixel inpainting. PSNR is shown in top right corner of the images.

Abstract

In various imaging problems, we only have access to compressed measurements of the underlying signals, hindering most learning-based strategies which usually require pairs of signals and associated measurements for training. Learning only from compressed measurements is impossible in general, as the compressed observations do not contain information outside the range of the forward sensing operator. We propose a new end-to-end self-supervised framework that overcomes this limitation by exploiting the equivariances present in natural signals. Our proposed learning strategy performs as well as fully supervised methods. Experiments demonstrate the potential of this framework on inverse problems including sparse-view X-ray computed tomography on real clinical data and image inpainting on natural images. Code has been made available at: `https://github.com/edongdongchen/EI`.

1. Introduction

Linear inverse problems are ubiquitous in computer vision and signal processing, appearing in multiple important applications such as super-resolution, image inpainting and computed tomography (CT). The goal in these problems consists of recovering a signal x from measurements y, that is inverting the forward process

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

which is generally a challenging task due to the ill-conditioned operator A and noise ϵ. In order to obtain a stable inversion, traditional approaches have used linear reconstruction, *i.e.* $A^\dagger y$, where the estimate is restricted to the range space of $A^\top$, or model-based approaches that reduce the set of plausible reconstructions x using prior information (*e.g.* sparsity). Leveraging the powerful representation properties of deep neural networks, a different approach is taken by recent end-to-end learning solutions which learn the inverse mapping directly from samples (x, y). However, all of these approaches require ground truth signals x for learning the reconstruction function $x = f(y)$, which hinders their applicability in many real-world scenarios where ground truth signals are either impossible or expensive to obtain. This limitation raises the natural question: *can we learn the reconstruction function without imposing strong priors and without knowing the ground-truth signals*?

Here, we show that typical properties of physical models such as rotation or shift invariance, constitute mild prior information that can be exploited to learn beyond the range space of $A^\top$. We present an end-to-end *equivariant imaging* framework which can learn the reconstruction function from compressed measurements y alone for a single forward operator. As shown in Figure 1, the equivariant imaging approach performs almost as well as having a dataset of ground truth signals x and significantly outperforms simply enforcing the measurement consistency $Af(y) = y$ in the training process. We show both theoretically and empirically that the proposed framework is able to identify the signal set and learn the reconstruction function from few training samples of compressed observations without accessing ground truth signals x. Experimental results demonstrate the potential of our framework through qualitative and quantitative evaluation on inverse problems. Specifically our contributions are as follows:

1. We present a conceptually simple *equivariant imaging* paradigm for solving inverse problems without ground truth. We show how invariances enable learning beyond the range space of the adjoint of the forward operator, providing necessary conditions to learn the reconstruction function.

2. We show that this framework can be easily incorporated into deep learning pipelines using an additional loss term enforcing the system equivariance.

3. We validate our approach on sparse-view CT reconstruction and image inpainting tasks, and show that our approach obtains reconstructions comparable to fully supervised networks trained with ground truth signals.

1.1. Related work

Model-based approaches The classical model-based approach for solving inverse problems [1, 2], constrains the space of plausible solutions using a fixed model based on prior knowledge (*e.g.* sparsity [3]). Although the model-based paradigm has typically nice theoretical properties, it presents two disadvantages: constructing a good prior model that captures the low-dimensionality of natural signals is generally a challenging task. Moreover, the reconstruction can be computationally expensive, since it requires running an optimization procedure at test time.

Deep learning approaches Departing from model-based strategies, the deep learning strategies aim to learn the reconstruction mapping from samples (x, y). This idea has been successfully applied to a wide variety of inverse problems, such as image denoising and inpainting [4, 5, 6], super-resolution [7, 8, 9], MRI reconstruction [10, 11] and CT image reconstruction [12, 13]. However, all of these approaches require access to training pairs (x, y) which might not be available in multiple real-world scenarios.

Learning with compressed observations In general, given a fixed forward model A with a non trivial nullspace, it is fundamentally impossible to learn the signal model beyond the range space of $A^\top$ using only compressed samples y. This idea traces back to *blind compressive sensing* [14], where it was shown that is impossible to learn a dictionary from compressed data without imposing strong assumptions on the set of plausible dictionaries.

Self-supervised learning More recently, there is a growing body of work on *self-supervised* learning exploring what can be learnt without ground truth. For example, there is a collection of studies in the mould of Noise2X [15, 16, 17, 18] where image denoising is performed without access to the ground truth. However, the denoising task does not have a non trivial nullspace since A is the identity. Although some follow-up works including [19] and [20] have tried to solve a more general situation, the former does not consider a nontrivial null space while the latter requires the exploitation of the diversity of multiple forward operators to learn a denoiser and eventually solves the inverse problem in an iterative model-based optimization [21]. Finally, some unconditional [22] and conditional [23] generative models, were proposed to learn to reconstruct from compressed samples, but again requiring multiple different forward operators. In contrast, we are able to learn this for a single forward operator.

2. Method

2.1. Problem Overview

We consider a linear imaging physics model $A : \mathbb{R}^n \to \mathbb{R}^m$, and the challenging setting in which only a set of N compressed observations $\{y_i\}_{i=1,\dots,N}$ are available for training. The learning task consists of learning a reconstruction function $f_\theta : \mathbb{R}^m \to \mathbb{R}^n$ such that $f_\theta(y) = x$. As the number of measurements is lower than the dimension of the signal space $m < n$, and the operator A has a non trivial nullspace.

Measurement consistency Given that we only have access to compressed data y, we can enforce that the inverse mapping f is consistent in the measurement domain. That is

$$Af(y) = y. \tag{2}$$

However, this constraint is not enough to learn the inverse mapping, as it cannot capture information about $\mathcal{X}$ outside the range of the operator $A^\top$. As shown in Section 3, there are multiple functions f_θ that can verify (2), even if we have infinitely many samples y_i.

Invariant set consistency In order to learn beyond the range space of $A^\top$, we can exploit some mild prior information about the set of plausible signals $\mathcal{X}$. We assume that the set presents some symmetries, *i.e.* that it is invariant to certain groups of transformations, such as shifts, rotations, reflections, etc. This assumption has been widely adopted both in multiple signal processing and computer vision applications. For example, it is commonly assumed that natural images are shift invariant. Another example is computed tomography data, where the same organ can be imaged at different angles making the problem invariant to rotation.

Under this assumption, the set of signals $\mathcal{X}$ is invariant to a certain group of transformations $\mathcal{G} = \{g_1, \dots, g_{|\mathcal{G}|}\}$ which are unitary matrices $T_g \in \mathbb{R}^{n\times n}$ such that for all $x \in \mathcal{X}$, we have

$$T_g x \in \mathcal{X} \tag{3}$$

for $\forall g \in \mathcal{G}$, and the sets $T_g\mathcal{X}$ and $\mathcal{X}$ are the same.

According to the invariance assumption in (3), the composition $f \circ A$ should then be equivariant to the transformation T_g, *i.e.*

$$f(AT_g x) = T_g f(Ax) \tag{4}$$

for all $x \in \mathcal{X}$, and all $g \in \mathcal{G}$. It is important to note that (4) does not require f to be invariant or equivariant, but rather the composition $f \circ A$ to be equivariant. As discussed in Section 3, as long as the range of $A^\top$ itself is not invariant to all T_g, this additional constraint on the inverse mapping f allows us to learn beyond the range space.

Invariant distribution consistency In most cases, not only is the signal set $\mathcal{X}$ invariant but also the distribution $p(x)$ defined on this set is invariant, i.e.

$$p(T_g x) = p(x) \tag{5}$$

for all $g \in \mathcal{G}$. Hence, we can also enforce this distributional constraint when learning the inverse mapping f.

2.2. Equivariant Imaging

We propose to use a trainable deep neural network $G_\theta : \mathbb{R}^n \to \mathbb{R}^n$ and an approximate linear inverse (for example a pseudo-inverse) $A^\dagger \in \mathbb{R}^{n\times m}$ to define the inverse mapping as $f_\theta = G_\theta \circ A^\dagger$. We emphasize that while in principle the form of f_θ is flexible, here we use the linear $A^\dagger$ to first project y into $\mathbb{R}^n$ to simplify the learning complexity. In practice $A^\dagger$ can be chosen to be any approximate inverse that is cheap to compute.

We propose a training strategy that enforces both the measurement consistency in (2) and the equivariance condition in (4) using only a dataset of compressed samples $\{y_i\}_{i=1,\dots,N}$. In the forward pass, we first compute $x^{(1)} = f_\theta(y)$ as an estimate of the actual ground truth x which is not available for learning. Note the data consistency between $Af_\theta(y)$ and y only ensures that $Ax^{(1)}$ stays close to the input measurement y but fails to learn beyond the range space of $A^\top$. According to the equivariance property in (4), we subsequently transform $x^{(2)} = T_g x^{(1)}$, for some $g \sim \mathcal{G}$, and pass it through $f \circ A$ to obtain $x^{(3)}$. The computations of $x^{(1)}$, $x^{(2)}$ and $x^{(3)}$ are illustrated in Figure 2.

Algorithm 1 Pseudocode of EI in a PyTorch-like style.

```
# A.forw, A.pinv: forward and pseudo inverse operators
# G: neural network
# T: transformations group
# a: alpha

for y in loader: # load a minibatch y with N samples
   # randomly select a transformation from T
   t = select(T)

   x1 = G(A.pinv(y)) # reconstruct x from y
   x2 = t(x1) # transform x1
   x3 = G(A.pinv(A.forw(x2))) # reconstruct x2

   # training loss, Eqn.(6)
   loss = MSELoss(A.forw(x1), y) # data consistency
        + alpha*MSELoss(x3, x2) # equivariance

   # update G network
   loss.backward()
   update(G.params)
```

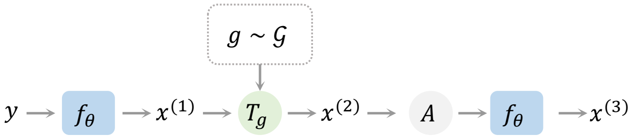

Figure 2: **Equivariant learning strategy.** $x^{(1)}$ represents the estimated image, while $x^{(2)}$ and $x^{(3)}$ represent $T_g x^{(1)}$ and the estimate of $x^{(2)}$ from the measurements $\tilde{y} = Ax^{(2)}$ respectively.

The network weights are updated according to the error between y and $Ax^{(1)}$, and the error between $x^{(2)}$ and $x^{(3)}$, by minimizing the following training loss

$$\arg\min_\theta \mathbb{E}_{y,g}\{\mathcal{L}(Ax^{(1)}, y) + \alpha\mathcal{L}(x^{(2)}, x^{(3)})\}, \tag{6}$$

where the first term is for data consistency and the second term is to impose equivariance, α is the trade-off parameter to control the strength of equivariance, and $\mathcal{L}$ is an error function.

After training, the learned reconstruction function $f_\theta = G_\theta \circ A^\dagger$ can be directly deployed either on the training samples of observations or on new previously unseen observations to recover their respective ground-truth signals. Algorithm 1 provides the pseudo-code of the *Equivariant Imaging (EI)* where $\mathcal{L}$ is the mean squared error (MSE).

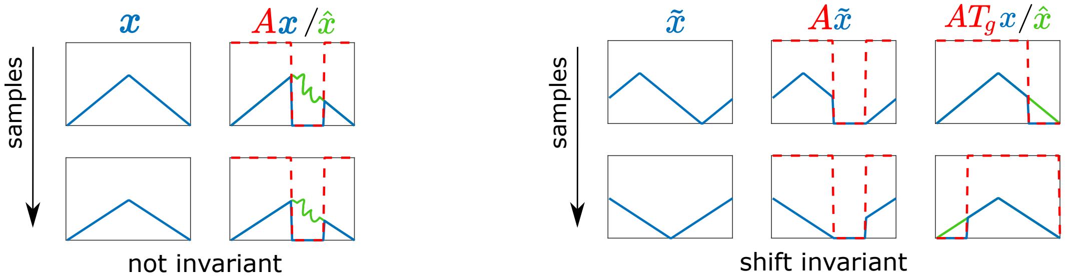

Figure 3: Learning with and without equivariance in a toy 1D signal inpainting task. The signal set consists of different scaling of a triangular signal. On the left, the dataset does not enjoy any invariance, and hence it is not possible to learn the data distribution in the nullspace of A. In this case, the network can inpaint the signal in an arbitrary way (in green), while achieving zero data consistency loss. On the right, the dataset is shift invariant. The range of A is shifted via the transformations T_g, and the network inpaints the signal correctly.

Adversarial extension We can add an additional penalty to enforce the invariant distribution consistency (5). Inspired by generative adversarial networks [24], we use a discriminator network D and adopt an adversarial training strategy to further enforce that $x^{(1)}$ and $x^{(2)}$ are identically distributed. The resulting adversarial equivariance learning strategy consists in solving the following optimization:

$$\min_G \max_D \mathbb{E}_{y,g}\{\mathcal{L}(Ax^{(1)},y)+\alpha\mathcal{L}(x^{(2)},x^{(3)}) \\ +\beta\mathcal{L}_{\text{adv}}(x^{(1)},x^{(2)})\}, \tag{7}$$

and $\mathcal{L}_{\text{adv}}(x^{(1)},x^{(2)}) = \mathbb{E}_{x^{(1)}}\{D(x^{(1)})\} + \mathbb{E}_{x^{(2)}}\{1 - D(x^{(2)})\}$ *i.e.* a least square adversarial loss [25] is adopted and β is to control the strength of invariant distribution constraint. Our experimental findings (see in Supplemental material) suggest that the adversarial invariance learning only provides a very slight improvement against the equivariant learning in (6). Thus in the next sections we mainly focus on equivariant learning.

3. Theoretical analysis

We start with some basic definitions. A measurement operator $A \in \mathbb{R}^{m\times n}$ with $m < n$ has a non-trivial linear nullspace $\mathcal{N}_A \subseteq \mathbb{R}^n$ of dimension at least $n-m$, such that $\forall v \in \mathcal{N}_A$ we have $Av = 0$. The complement of $\mathcal{N}_A$ is the range space $\mathcal{R}_A = \text{range}(A^\top)$, such that $\mathcal{R}_A \oplus \mathcal{N}_A = \mathbb{R}^n$, which verifies that $\forall v \in \mathcal{R}_A$ we have $Av \neq 0$.

Learning without invariance The problem of learning the signal set $\mathcal{X}$ only using compressed samples was first explored in the context of blind compressive sensing [14], for the special case where $\mathcal{X}$ is modelled with a sparse dictionary. The authors in [14] showed that learning is impossible in general, becoming only possible when strong assumptions on the set of plausible dictionaries are imposed. A similar result can be stated in a more general setting, showing that there are multiple possible reconstruction functions f that satisfy measurement consistency:

Proposition 1 Any reconstruction function $f(y): \mathbb{R}^m \mapsto \mathbb{R}^n$ of the form

$$f(y) = A^\dagger y + v(y) \tag{8}$$

where $v(y): \mathbb{R}^m \mapsto \mathcal{N}$ is any function whose image belongs to the nullspace of A verifies the measurement consistency requirement.

Proof: For f any form (8) the measurement consistency can be expressed as $Af(y) = AA^\dagger y + Av(y)$ where the first term is simply y as $AA^\dagger$ is the identity matrix, and $Av(y) = 0$ for any $v(y)$ in the nullspace of A. □

For example, the function $v(x)$ can be as simply as $v = 0$ and the resulting f will be measurement consistent. Interestingly, some previous supervised approaches [26, 27] separate the learning of the range and the nullspace components. Proposition 1 shows that without ground truth signals, there is no information to learn the nullspace component.

Learning with invariance In the proposed equivariant imaging paradigm, each observation can equally be thought of as a new observation with a new measurement operator $A_g = AT_g$, as

$$y = Ax = AT_gT_g^\top x = A_g\tilde{x} \tag{9}$$

where $\tilde{x} = T_g^\top x$ is also a signal in $\mathcal{X}$. Hence, the invariance property allows us to see in the range of the operators A_g, or equivalently, *rotate* the range space $\mathcal{R}_A$ through the action of the group $\mathcal{G}$, *i.e.*

$$\mathcal{R}_{A_g} = \text{range}(T_g^\top A^\top) = T_g^\top \mathcal{R}_A. \tag{10}$$

This idea is illustrated in Figure 3 for inpainting a simple 1D signal model. A necessary condition to recover a unique model $\mathcal{X}$ is that the concatenation of operators A_g spans the full space $\mathbb{R}^n$:

Theorem 1: A necessary condition for recovering the signal model $\mathcal{X}$ from compressed observations is that the matrix

$$M = \begin{bmatrix} AT_1 \\ \vdots \\ AT_{|\mathcal{G}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{G}|m \times n} \tag{11}$$

is of rank n.

Proof: Assume the best case scenario where we have an oracle access to the measurements associated with the different transformations of the same signal[1] x, that is $y_g = AT_g x$ for all g. Stacking all the measurements together into $\tilde{y} \in \mathbb{R}^{|\mathcal{G}|m}$, we observe $\tilde{y} = Mx$ and hence M needs to be of rank n in order to recover x. □

This necessary condition provides a lower bound on how big the group $\mathcal{G}$ has to be, *i.e.* at least satisfy $m|\mathcal{G}| \geq n$. For example, if the model is invariant to single reflections ($|\mathcal{G}| = 2$), we need at least $m \geq n/2$. Moreover, this condition also tells us that the range space $\mathcal{R}_A$ cannot be invariant to $\mathcal{G}$.

Corollary 1: A necessary condition for recovering the signal model $\mathcal{X}$ from compressed observations is that the range $\mathcal{R}_A$ with $m < n$ is not invariant to the action of $\mathcal{G}$, i.e., there is $g \in \mathcal{G}$ such that

$$\mathcal{R}_A \neq \mathcal{R}_{AT_g} \tag{12}$$

Proof: If $\mathcal{R}_A = \mathcal{R}_{AT_g}$ for all g then $\mathcal{N}_A = \mathcal{N}_{AT_g}$ for all g. From (11) we have that M shares the same null space and is therefore rank $m < n$. □

Corollary 1 tells us that not any combination of A and $\mathcal{G}$ is useful for learning beyond the range space. For example, shift invariance cannot be used to learn from Fourier based measurement operators (which is the case in deblurring, super-resolution and magnetic resonance imaging), as $A^\top$ would be invariant to the shifts. It is worth noting that the necessary condition in Theorem 1 will in general not be sufficient. For example, for shift invariant models, a forward matrix composed of a single localized measurement $A = [1, 0, \ldots, 0]^\top$ verifies the necessary condition but might not enough to learn a complex model $\mathcal{X}$.

4. Experiments

We show experimentally the performance of the proposed method for diverse image reconstruction problems. Due to space limitations, we present a few examples here and include more in the Supplementary Material (SM).

4.1. Setup and Implementation

We evaluate the proposed approach on two inverse imaging problems: *sparse-view CT image reconstruction* and *image inpainting*, where the measurement operator A in both tasks are fixed and have non-trivial nullspaces, illustrating the models' ability to learn beyond the range space. We designed our experiments to address the following questions: (i) how well does the equivariant imaging paradigm compare to fully supervised learning? (ii) how does it compare to measurement consistent only learning (*i.e.* with the equivariance loss term removed)?

Throughout the experiments, we use a U-Net [28] to build G_θ with a residual connection at the output, *i.e.* $G_\theta = I + G_\theta^{\text{res}}$ and $f_\theta(y) = A^\dagger y + G_\theta^{\text{res}}(A^\dagger y)$, to explicitly let the learning target of G_θ^{res} recover the nullspace component of x. We compare our method (EI) with four different learning strategies: measurement-consistency only (MC) with the equivariance term in (6) removed; the adversarial extension of EI (EI$_{\text{adv}}$) in (7) using the discriminator network from [8]; supervised learning (Sup) [12] that minimizes $\mathbb{E}_y\{\mathcal{L}(f_\theta(y), x)\}$ using ground truth signal-measurement pairs; and EI regularized supervised learning (EI$_{\text{sup}}$) with the data consistency term replaced by $\mathcal{L}(f_\theta(y), x)$ in (6). For a fair comparison with EI, no data augmentation of ground truth signals are conducted for both supervised learning methods, Sup and EI$_{\text{sup}}$. We use the residual U-Net architecture for all the counterpart learning methods to ensure all methods have the same inductive bias from the neural network architecture. Note that while there are many options to determine the optimal network architecture such as exploring different convolutions [29, 5, 30, 6] or different depths [31], these aspects are somewhat orthogonal to the *learning beyond the range space* question.

We demonstrate that the equivariant imaging approach is straightforward and can be easily extended to existing deep models without modifying the architectures. All methods are implemented in PyTorch and optimized by Adam [32]. We tuned the α for specific inverse problems (see SM for training details).

4.2. Sparse-view CT

The imaging physics model of X-ray computed tomography (CT) is the discrete `radon` transform. The physics model A is the `radon` transformation where 50 views (angles) are uniformly subsampled to generate the sparse-view sinograms (observations) y. The Filtered back projection (FBP) function, *i.e.* `iradon`, is used to approximate $A^\dagger$. In this task, we exploit the invariance of the CT images to rotations[2], and $\mathcal{G}$ is the group of rotations by 1 degree ($|\mathcal{G}|$=360). We use the CT100 dataset [33], a public real CT clinic dataset which comprises 100 real in-vivo CT images collected from the cancer imaging archive[3] which consist of the middle slice of CT images taken from 69 different

[1] This is also the case for the simplest signal model where $\mathcal{X}$ is composed of a single atom.

[2] It is worth noting that shift invariance is not useful for the CT case, as the forward operator is shift invariant itself (see Corollary 1).

[3] `https://wiki.cancerimagingarchive.net/display/Public/TCGA-LUAD`

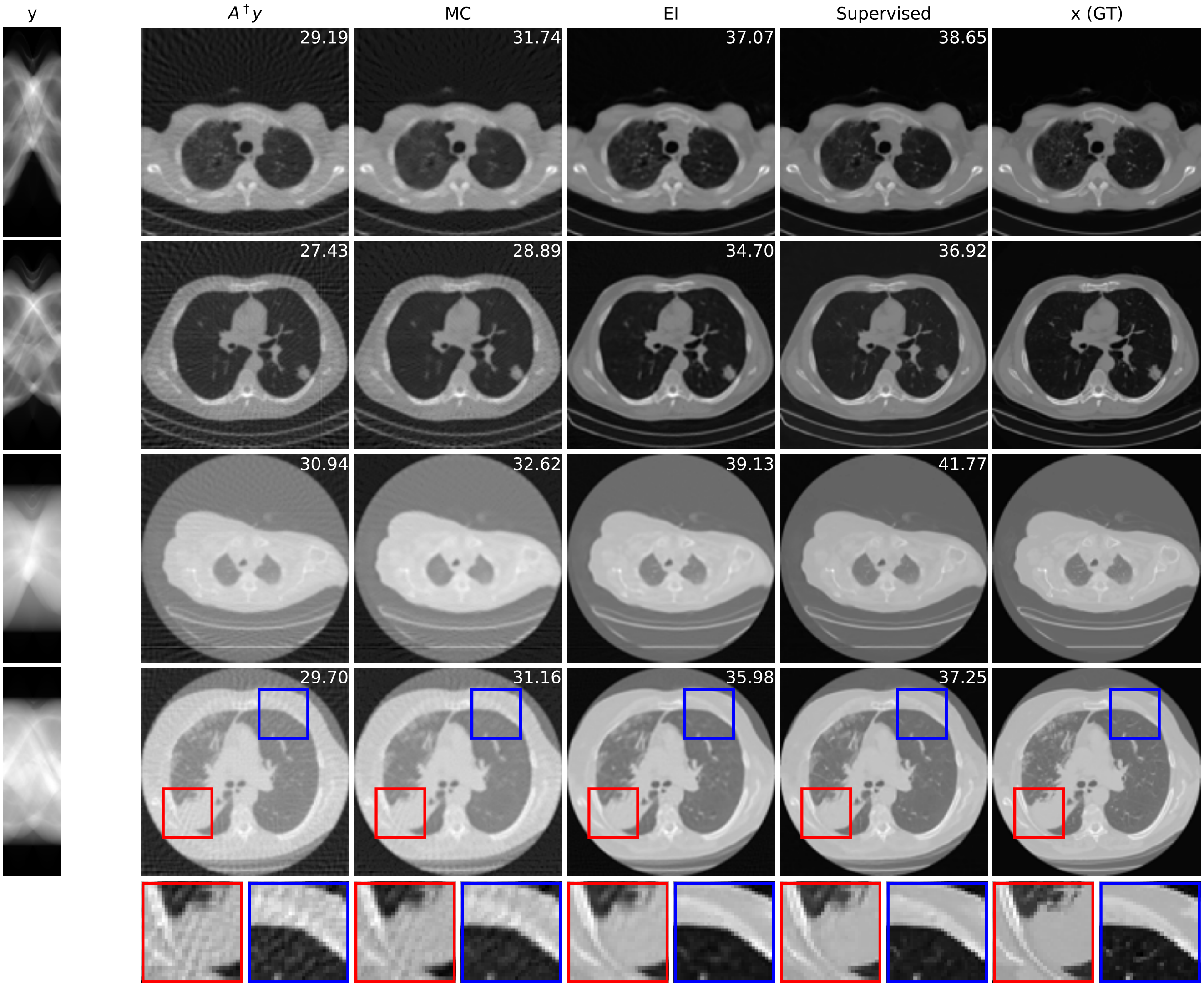

Figure 4: Examples of sparse-view CT image reconstruction on the unseen test observations. We train the supervised model (FBPConvNet [12]) with observation-groundtruth pairs while train our equivariance learned model with observations alone. We adopt the *random rotation* as the transformation T for our equivariance learning. We obtained results comparable to supervised learning in artifacts-removal. Corresponding PSNR are shown in images.

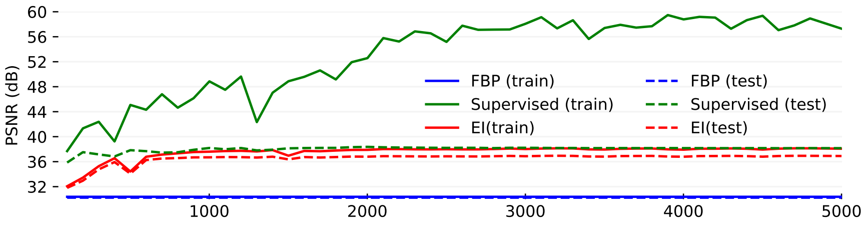

Figure 5: Reconstruction performance (PSNR) as a function of training epoch for supervised trained FBPConvNet and our method (learn without groundtruth) on sparse-view CT observations for training and testing.

patients. The CT images are resized to 128×128 pixels and we then apply the `radon` function on them to generate the 50-views sinograms. We used the first 90 sinograms for training while the remaining 10 sinograms for testing. Note in this task, the supervised trained residual U-Net is just the FBPConvNet proposed in [12] which has been demonstrated to be very effective in supervised learning for sparse-view CT image reconstruction. We train our model with equivariance strength $\alpha = 10^2$ (see SM for more results and the equivariance strength effect). using the sinograms

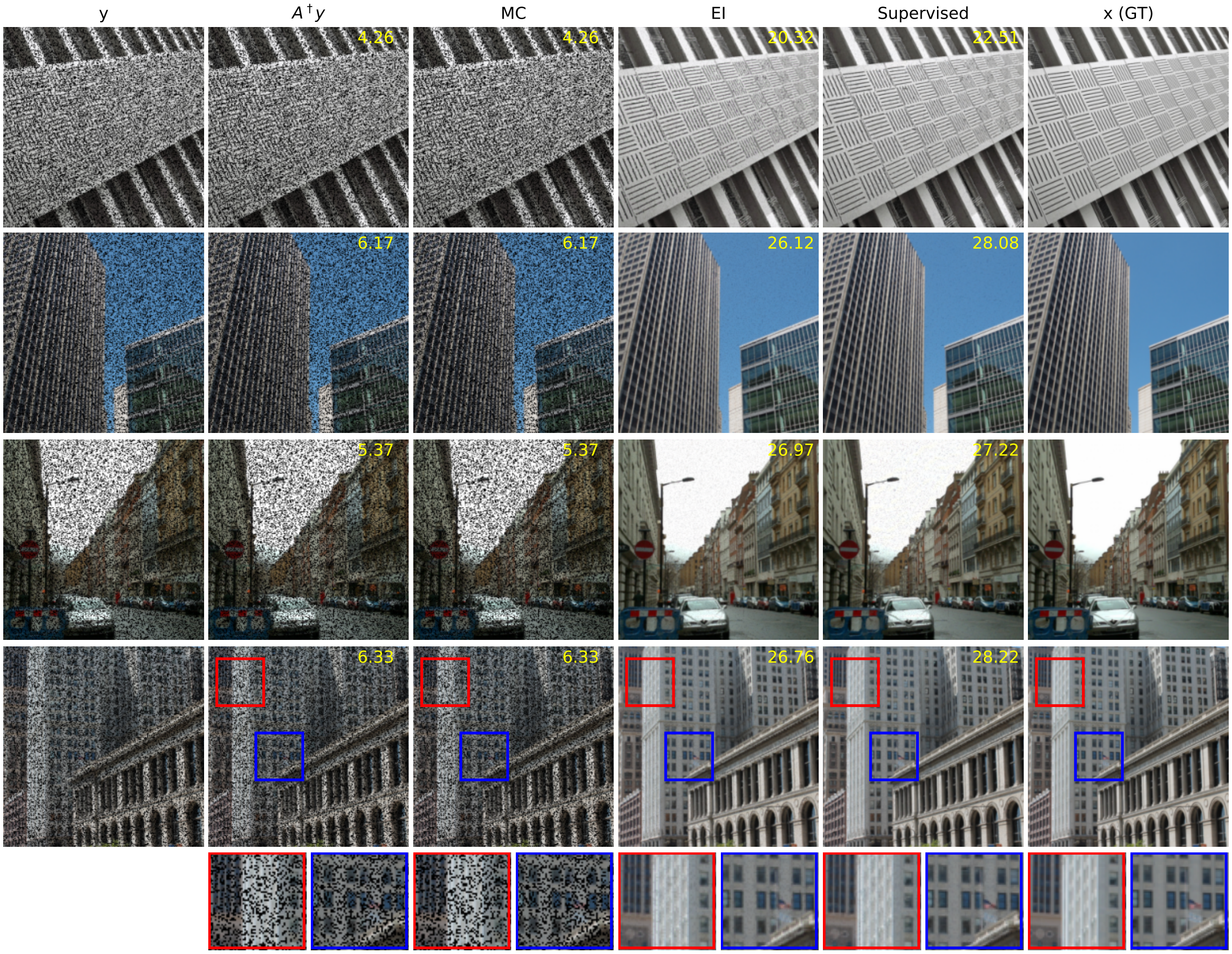

Figure 6: Examples of pixelwise image inpainting on the unseen test observations. We train the supervised model [12] with observation-groundtruth pairs while train our equivariance learned model with observations alone. We adopt the *random shift* as the transformation T for our equivariance learning. We obtained results comparable to supervised learning in recovering missing pixels. Corresponding PSNR are shown in images.

y alone while the FBPConvNet is trained with the ground truth pairs (x, y).

A qualitative comparison is presented in Figure 4. The sparse-view FBP contains the line artifacts. Both the FBPConvNet and our methods significantly reduce these artifacts, giving visually indistinguishable results. Figure 5 shows the value of PSNR of reconstruction on the training measurements and test measurements and we have the following observations: (i) We would naturally expect the network trained with ground truth data to perform the best. However, we note that the equivariant test error is almost as good despite having no access to ground truth images and only learning on the sparse sinogram data. Furthermore the EI solution is about 7 dB better than the FBP, clearly demonstrating the correct learning of the null space component of the image. (ii) We note that there is a significant gap between training and test error for the FBPConvNet, suggesting that the network may be overfitting. We do not observe this in the EI learning. This can be explained by the fact that the EI constrains the network to a much small class of functions (those that are equivariant on the data) and thus can be expected to have better generalization properties.

We also compared the EI with its adversarial extension in (7) and the supervised learning regularized by equivariance objective. The quantitative results are given in table 1 below. First, MC learning obtains a small improvement in performance over FBP which may be attributable to the fact that FBP is only an approximation to $A^\dagger$. Alternatively it may be due to the inductive bias of the neural network architecture [34]. Second, the adversarial extension provides a slight improvement to EI and similarly the EI regularization helps the vanilla supervised learning obtain a further

	FBP	MC	EI	EI_{adv}	Sup	EI_{sup}
50-views CT	30.24	31.01	36.94	36.96	38.17	38.79
	$A^\dagger y$	MC	EI	EI_{adv}	Sup	EI_{sup}
Inpainting	5.84	5.84	25.14	23.26	26.51	26.75

Table 1: Reconstruction performance (PSNR) of 50-views CT reconstruction and image inpainting for different methods on the CT100 and Urban100 test measurements, respectively.

0.6 dB improvement. These results suggest that it is indeed possible to learn to reconstruct challenging inverse problems with only access to measurement data.

4.3. Image inpainting

As a proof-of-concept of the generality of the method, we also applied our method on an image inpainting task with a fixed set of deleted pixels. This is relevant for example to the problem of reconstructing images from cameras with hot or dead pixels.

In the image inpainting task, the corrupted measurement is given by $y = b \odot x$ where b is a binary mask, $\odot$ is the Hadamard product, and the associated operator $A = \text{diag}(b)$ and $A = A^\dagger$. Here, we consider *pixelwise inpainting* where we *randomly* drop 30% of pixel measurements. We train our model by applying *random shift* transformations. We evaluate the reconstruction performance of our approach and other learning methods using the Urban100 [35] natural image dataset. For each image, we cropped a 512x512 pixel area at the center and then resized it to 256x256 for the ground truth image. The first 90 measurements are for training while the last 10 measurements are for testing.

The reconstruction comparisons are presented in Figure 6 and Table 1. We have the following observations: First, the MC reconstruction is exactly equal to $A^\dagger y$ as the exact pseudo inverse is used, the reconstruction quality of MC is very poor as it completely failed to learn the nullspace at all. Second, the EI reconstruction is about 20 dB better than $A^\dagger y$ and MC reconstruction, the missing pixels are recovered well, again demonstrating the correct learning of the null space component of the image (the adversarial EI was not competitive in this application). Finally, there is only a 1.37 dB gap between the reconstruction of EI and the fully supervised model. As with the CT imaging, we again find the generalization error of EI is also much smaller than for the supervised model (see SM).

5. Discussion

The equivariant imaging framework presented here is conceptually different from recent ideas on invariant networks [36, 37, 38] where the goal is to train an invariant neural network for classification problems, which generally performs better than a non-invariant one [39]. In contrast, the equivariant imaging goal is to make the composition $f_\theta \circ A$ equivariant but not necessarily f_θ, promoting invariance across the complete imaging system. Moreover, our framework also differs from standard data augmentation techniques, as no augmentation can be done directly on the compressed samples y. The proposed method overcomes the fundamental limitation of only having range space information, effectively solving challenging inverse problems without the need of ground truth training signals. As shown in the experiments, our equivariant constraint can also be applied in the fully supervised setting to improve the performance of the networks.

The equivariant imaging framework admits many straightforward extensions. For example, while we have shown how to use shift-invariance and rotation-invariance to solve the inpainting task and CT reconstruction. We believe that there are many other imaging tasks that could benefit from equivariant imaging. It would also be very interesting to investigate whether the benefits seen here can be extended to nonlinear imaging problems.

Mixed types of group transformations can also be applied at the training time and may help improve convergence time and performance. However, as we have shown, the strength of different transformations will depend on the nature of the signal model and the physics operator.

We have also found that equivariant imaging can be used to improve the performance for *single image reconstruction* and have reported some preliminary results in the Supplementary material. However, as single image reconstruction itself relies heavily on the strong inductive bias of the network [34] the role is EI in this scenario is less clear.

6. Conclusions

We have introduced a novel self-supervised learning strategy that can learn to solve an ill-posed inverse problem from only the observed measurements, without having any knowledge of the underlying signal distribution, other than assuming that it is invariant to the action of a group of transformations. This relates to an important question on the use of deep learning in scientific imaging [40]: can networks learn to image structures and patterns for which no ground truth images yet exist? We believe that the EI framework suggests that with the addition of the basic physical principle of invariance, such data-driven discovery is indeed possible.

Acknowledgments

This work is supported by the ERC C-SENSE project (ERCADG-2015-694888).

References

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[2] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 57, no. 11, pp. 1413–1457, 2004.

[3] M. K. Ng, P. Weiss, and X. Yuan, "Solving constrained total-variation image restoration and reconstruction problems via alternating direction methods," *SIAM journal on Scientific Computing*, vol. 32, no. 5, pp. 2710–2736, 2010.

[4] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in neural information processing systems*, 2016, pp. 2802–2810.

[5] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4471–4480.

[6] ——, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 5505–5514.

[7] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[8] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[9] A. Lugmayr, M. Danelljan, L. Van Gool, and R. Timofte, "Srflow: Learning the super-resolution space with normalizing flow," in *Proceedings of the European Conference on Computer Vision (ECCV).* Springer, 2020, pp. 715–732.

[10] M. Mardani, Q. Sun, D. Donoho, V. Papyan, H. Monajemi, S. Vasanawala, and J. Pauly, "Neural proximal gradient descent for compressive imaging," in *Advances in Neural Information Processing Systems*, 2018, pp. 9573–9583.

[11] D. Chen, M. E. Davies, and M. Golbabaee, "Compressive mr fingerprinting reconstruction with neural proximal gradient iterations," in *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, 2020.

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

[13] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, "Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1348–1357, 2018.

[14] S. Gleichman and Y. C. Eldar, "Blind compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6958–6975, 2011.

[15] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2noise: Learning image restoration without clean data," in *International Conference on Machine Learning*, 2018, pp. 2965–2974.

[16] J. Batson and L. Royer, "Noise2self: Blind denoising by self-supervision," *arXiv preprint arXiv:1901.11365*, 2019.

[17] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2void-learning denoising from single noisy images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2129–2137.

[18] N. Moran, D. Schmidt, Y. Zhong, and P. Coady, "Noisier2noise: Learning to denoise from unpaired noisy data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 064–12 072.

[19] A. A. Hendriksen, D. M. Pelt, and K. J. Batenburg, "Noise2inverse: Self-supervised deep convolutional denoising for linear inverse problems in imaging," *arXiv preprint arXiv:2001.11801*, 2020.

[20] J. Liu, Y. Sun, C. Eldeniz, W. Gan, H. An, and U. S. Kamilov, "Rare: Image reconstruction using deep priors learned without ground truth," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[21] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (RED)," *SIAM Journal on Imaging Sciences*, vol. 10, no. 4, pp. 1804–1844, 2017.

[22] A. Bora, E. Price, and A. G. Dimakis, "Ambientgan: Generative models from lossy measurements," in *International Conference on Learning Representations*, 2018.

[23] A. Pajot, E. de Bezenac, and P. Gallinari, "Unsupervised adversarial image reconstruction," in *International Conference on Learning Representations*, 2019.

[24] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[25] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017, pp. 2794–2802.

[26] J. Schwab, S. Antholzer, and M. Haltmeier, "Deep null space learning for inverse problems: Convergence analysis and rates," *Inverse Problems*, 2019.

[27] D. Chen and M. E. Davies, "Deep decomposition learning for inverse imaging problems," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[29] J. S. Ren, L. Xu, Q. Yan, and W. Sun, "Shepard convolutional neural networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, 2015, pp. 901–909.

[30] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 85–100.

[31] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9446–9454.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[34] J. Tachella, J. Tang, and M. 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)*, 2021, pp. 8618–8627.

[35] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5197–5206.

[36] K. Lenc and A. Vedaldi, "Understanding image representations by measuring their equivariance and equivalence," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015, pp. 991–999.

[37] U. Schmidt and S. Roth, "Learning rotation-aware features: From invariant priors to equivariant descriptors," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2050–2057.

[38] A. Foster, R. Pukdee, and T. Rainforth, "Improving transformation invariance in contrastive representation learning," *arXiv preprint arXiv:2010.09515*, 2020.

[39] J. Sokolic, R. Giryes, G. Sapiro, and M. Rodrigues, "Generalization error of invariant classifiers," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1094–1103.

[40] C. Belthangady and L. Royer, "Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction," *Nature Methods*, vol. 16, 07 2019.

A. Training Details

We first provide the details of the network architectures and hyperparameters of Figs. 4-6 and Table 1 of the main paper. We implemented the algorithms and operators (*e.g.* `radon` and `iradon`) in Python with PyTorch 1.6 and trained the models on NVIDIA 1080ti and 2080ti GPUs. Figure 7 illustrates the architecture of the residual U-Net used [28] in our paper.

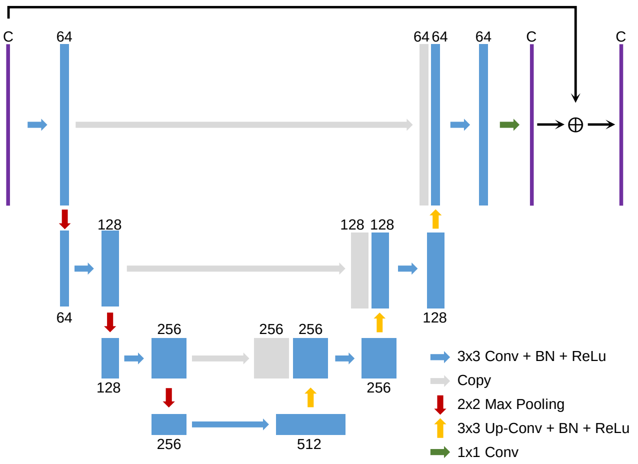

Figure 7: The residual U-Net [28] used in the paper. The number of input and output channels is denoted as C, with $C = 1$ in the CT task and $C = 3$ in the inpainting task.

For the sparse-view CT task, we used the Adam optimizer with a batch size of 2 and an initial learning rate of 0.0005. The weight decay is 10^{-8}. The distribution strength β is 10^{-8} for EI$_{adv}$. We trained the networks for 5000 epochs, keeping the learning rate constant for the first 2000 epochs and then shrinking it by a factor of 0.1 every 1000 epochs. More reconstruction examples are presented in Figure 9.

For the inpainting task, we also used Adam but with a batch size of 1 and an initial learning rate of 0.001. The weight decay is 10^{-8}. The distribution strength β is 10^{-8} for EI$_{adv}$. We trained the networks for 2000 epochs, shrinking the learning rate by a factor of 0.1 every 500 epochs. Figure 8 shows the peak signal-to-noise ratio (PSNR) of the reconstructions on the training and test measurements. Again, the generalization error of EI is smaller than for the supervised model. More reconstruction examples are presented in Figure 10.

B. More results

Effect of the equivariance hyperparameter α Table 2 shows EI reconstruction performance (PSNR) with different equivariance strength values (α in Eqn. (6) of the main paper). It performs reasonably well when $\alpha = 100$ for the CT task and $\alpha = 1$ for the inpainting task. When α is too small, the performance drops considerably; at the extreme of no equivariance ($\alpha = 0$), the model fails to learn. These results support our motivation of equivariant imaging.

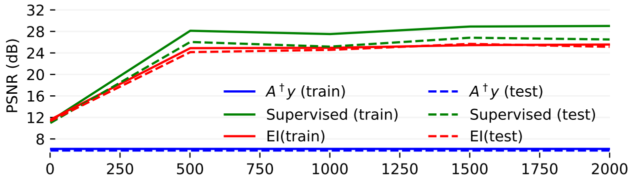

Figure 8: Reconstruction performance (PSNR) as a function of training epoch for the supervised model [12] and our method (no ground truth) on inpainting task measurements for training and testing.

α	0	1	10	100	1000
50-views CT	31.01	36.78	36.88	36.94	33.31

α	0	0.1	1	10
Inpainting	5.84	23.42	25.14	22.96

Table 2: Effect of the equivariance hyperparameter α on the reconstruction performance (PSNR) in the 50-views CT reconstruction (CT100 dataset) and image inpainting (Urban100 dataset) tasks.

Effect of the networks' inductive bias In the deep image prior (DIP) paper, the authors showed that some specific convolutional networks can be trained to fit a single image by only enforcing measurement consistency [31]. The DIP approach relies heavily on the choice of the network architecture (generally an autoencoder), and does not work with various popular architectures (*e.g.* those with skip-connections). Moreover, this approach is constrained to a single image and cannot incorporate additional training data.

In contrast, we show that our method can learn beyond the range space without heavily relying on the inductive bias of an specific autoencoder architecture. Moreover, we show that EI outperforms the best DIP architecture as it leverages the full compressed training dataset. We compare our method with the DIP on the 50-views CT image reconstruction task. For our method, we use the same residual U-Net as in the other experiments. We build the DIP using two architectures: the same residual U-Net used in EI (which we denote DIP-1) and the best autoencoder network suggested in [31] (which we denote DIP-2). Following [31], we input iid Gaussian noise to both DIP-1 (1 channel) and DIP-2

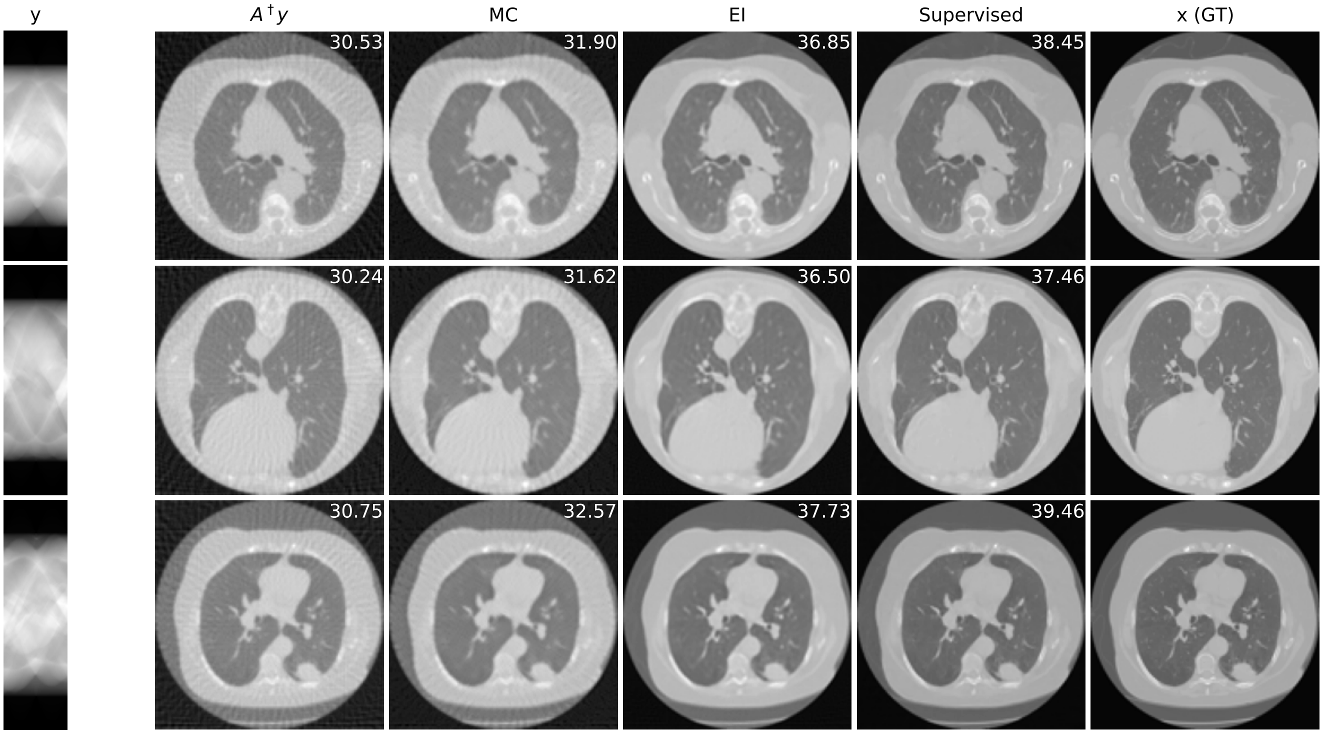

Figure 9: More examples of sparse-view CT image reconstruction on the unseen test measurements. We train the supervised model (FBPConvNet [12]) with measurement/ground truth pairs while we train the equivariance learned model with measurements alone. We adopt *random rotations* as the transformation T for our equivariance learning. We obtained results comparable to supervised learning in artifacts-removal. Corresponding PSNR are shown in images.

(32 channels). Our model is trained using the hyperparameters for sparse-view CT (see Section A). We train DIP-1 and DIP-2 using 5000 training iterations and a learning rate of 0.001. As shown in Figure 11, our method outperforms the DIP methods. DIP-2 performs significantly better than DIP-1 due to the inductive bias of that autoencoder architecture. In contrast, our method works very well even with the residual U-Net. Moreover, our model also outperforms DIP-2 by 5 dB.

Equivariant imaging using a single training image We are interested in whether the proposed method works for single image reconstruction, *i.e.* reconstructing a single compressed measurement. Here we provide some preliminary results. As an example, we compared our method with the DIP on the inpainting task for single image reconstruction. We trained all 3 models (EI, DIP-1, DIP-2) using 5000 training iterations and a learning rate of 0.001 on a single measurement input. The results are presented in Figure 12. We observe that our method works very well for this single image reconstruction task and outperforms both DIP-1 and DIP-2. In addition, DIP-1 performs worse than DIP-2 due to the residual architecture with skip-connections. Again, our model is not so dependent on the inductive bias of network and works well when using the residual connections. We note that although our method is able to learn with a single measurement, the role of equivariance in this scenario needs to be explored more, and we leave this for future work.

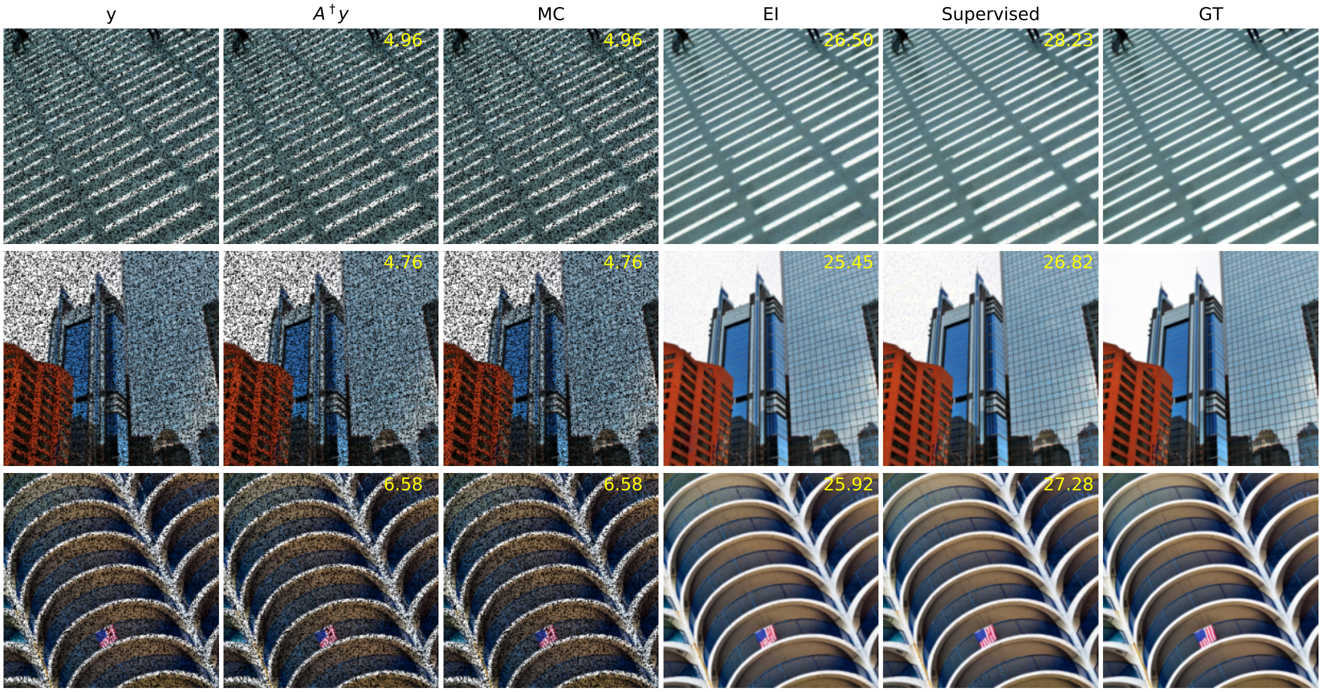

Figure 10: More examples of image inpainting reconstruction on the unseen test measurements. We train the supervised model [12] with measurement/ground truth pairs while we train the equivariance learned model with measurements alone. We adopt *random shifts* as the transformation T for our equivariance learning. We obtained results comparable to supervised learning in recovering missing pixels. Corresponding PSNR are shown in images.

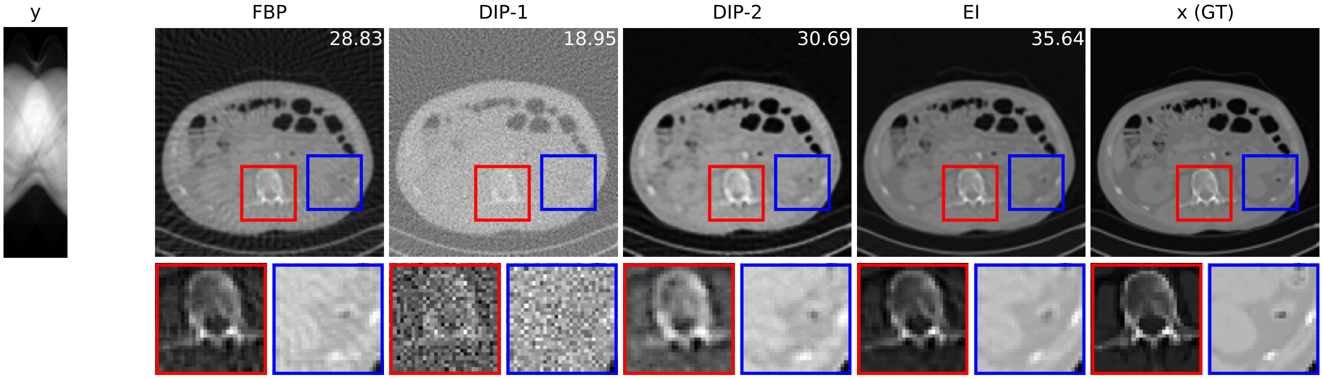

Figure 11: Comparison between EI and DIP on 50-views CT reconstruction. We denote DIP-1 and DIP-2 as the DIP learned models trained with residual U-Net (same as EI) and Encoder-Decoder (the best architecture for DIP as suggested in [31]), respectively. We trained EI on a measurement set and direct apply the trained model on the given new measurement here. Both DIP methods are trained using the given measurement here. Corresponding PSNR are shown in images.

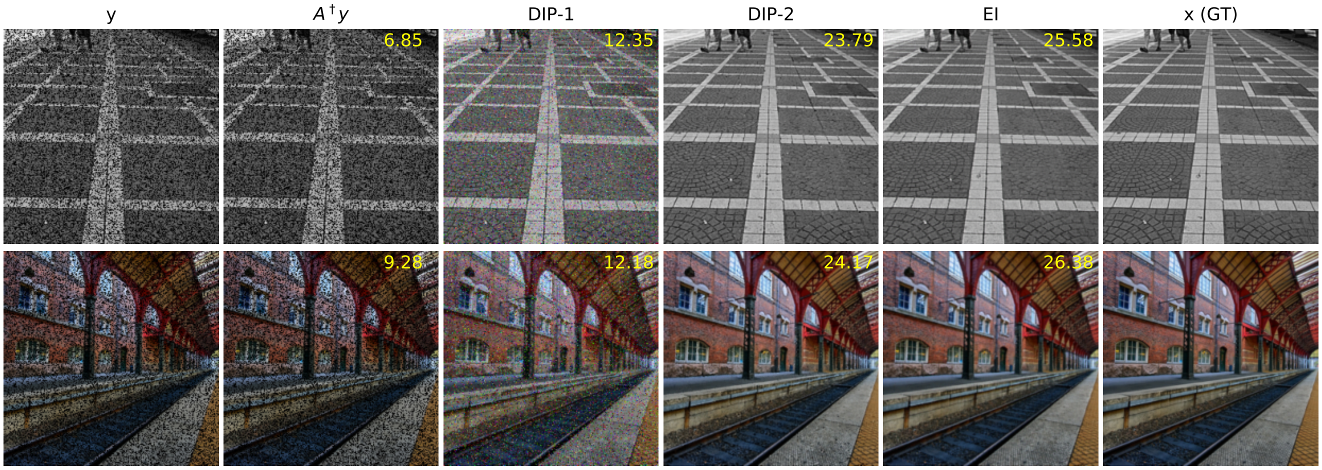

Figure 12: Comparison between EI and DIP for single image reconstruction on the inpainting task. We denote DIP-1 and DIP-2 as the DIP learned models trained with residual U-Net (same as EI) and Encoder-Decoder (the best architecture for DIP as suggested in [31]), respectively. All the models are trained with the given single compressed measurement data y. Corresponding PSNR are shown in images.