Published in final edited form as:
IEEE Trans Med Imaging. 2021 November ; 40(11): 3249–3260. doi:10.1109/TMI.2021.3077857.

On Hallucinations in Tomographic Image Reconstruction

Sayantan Bhadra,
Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO 63130 USA

Varun A. Kelkar,
Department of Electrical and Computer Engineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA

Frank J. Brooks,
Department of Bioengineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA

Mark A. Anastasio [Senior Member, IEEE]
Department of Bioengineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA

Abstract

Tomographic image reconstruction is generally an ill-posed linear inverse problem. Such ill-posed inverse problems are typically regularized using prior knowledge of the sought-after object property. Recently, deep neural networks have been actively investigated for regularizing image reconstruction problems by learning a prior for the object properties from training images. However, an analysis of the prior information learned by these deep networks and their ability to generalize to data that may lie outside the training distribution is still being explored. An inaccurate prior might lead to false structures being hallucinated in the reconstructed image and that is a cause for serious concern in medical imaging. In this work, we propose to illustrate the effect of the prior imposed by a reconstruction method by decomposing the image estimate into generalized measurement and null components. The concept of a hallucination map is introduced for the general purpose of understanding the effect of the prior in regularized reconstruction methods. Numerical studies are conducted corresponding to a stylized tomographic imaging modality. The behavior of different reconstruction methods under the proposed formalism is discussed with the help of the numerical studies.

Index Terms—

Tomographic image reconstruction; image quality assessment; deep learning; hallucinations

(Corresponding author: Mark A. Anastasio, maa@illinois.edu.).
(Sayantan Bhadra and Varun A. Kelkar contributed equally to this work.)

This article has supplementary downloadable material available at https://doi.org/10.1109/TMI.2021.3077857, provided by the authors.

I. Introduction

IN TOMOGRAPHIC imaging, a reconstruction method is employed to estimate the sought-after object from a collection of measurements obtained from an imaging system [1]. Since the sought-after object is usually described as a continuous function and the measurements are discrete, image reconstruction methods usually seek a finite-dimensional estimate of the object. Moreover, it is often desirable to reconstruct images from as few measurements as possible, without compromising on the diagnostic quality of the image. For example, data-acquisition times in magnetic resonance imaging (MRI) can be reduced by undersampling the k-space [2]. In such situations the acquired measurements are said to be *sparse*, i.e., they are generally insufficient to uniquely specify a finite-dimensional approximation of the sought-after object, even in the absence of measurement noise or errors related to modeling the imaging system. This naturally implies that the inverse problem is ill-posed and some form of regularization needs to be performed with priors imposed on the sought-after object. Various methods have been proposed for regularization that can effectively mitigate the impact of measurement-incompleteness on image reconstruction. Among these methods, regularization using sparsity-promoting penalties has been employed widely [3]–[6].

Recently, there has been considerable focus on developing regularization strategies that seek to learn the prior distribution that describes the object to-be-imaged from existing data, instead of using hand-crafted priors such as sparsity-promoting penalties. Nascent deep learning-based methods have inspired a new wave of reconstruction methods that implicitly or explicitly learn the prior distribution from a set of training images in order to regularize the reconstruction problem [7]–[9]. However, such learning-based methods have also raised concerns regarding their robustness [10]–[13] and their ability to generalize to measurements that may lie outside the distribution of the training data [12], [14], [15]. This is particularly relevant in the field of medical imaging where novel abnormalities can be present in the observed measurement data that may not be encountered even with a large training dataset. Moreover, simulation studies have shown that deep learning-based reconstruction methods are inherently unstable, i.e. small perturbations in the measurement may produce large differences in the reconstructed image [11], [12].

The potential lack of generalization of deep learning-based reconstruction methods as well as their innate unstable nature may cause false structures to appear in the reconstructed image that are absent in the object being imaged. These false structures may arise due to the reconstruction method incorrectly estimating parts of the object that either did not contribute to the observed measurement data or cannot be recovered in a stable manner, a phenomenon that can be termed as *hallucination*. The presence of such false structures in reconstructed images can possibly lead to an incorrect medical diagnosis. Hence, there is an urgent need to investigate the nature and impact of false structures arising out of hallucinations from deep learning-based reconstruction methods for tomographic imaging.

The topic of image hallucinations has previously been studied within the context of image super-resolution [16]–[19]. In image super-resolution, the term hallucination generally refers to high-frequency features that are introduced into the high-resolution image but do not exist in the measured low-resolution image. Hallucinations can also be realized in more general inverse problems such as image reconstruction. In such cases, the structure of the imaging operator null space is generally more complicated and the hallucinations may not be confined to high-frequency structures [20]. However, a formal definition of hallucinations within the context of such inverse problems has not been reported.

This study proposes a way to mathematically formalize the concept of hallucinations for general linear imaging systems that is consistent with both the mathematical notion of a hallucination in image super-resolution and the intuitive notion of hallucinations as "artifacts or incorrect features that occur due to the prior that cannot be produced from the measurements". In addition, the notion of a *task-informed* or *specific* hallucination map is introduced. Through preliminary numerical studies, the behavior of different reconstruction methods under the proposed formalism is illustrated. It is shown that, in certain cases, traditional error maps are insufficient for visualizing and detecting specific hallucinations.

The remainder of this paper is organized as follows. In Section II, salient aspects of linear operator theory are reviewed, and the need for describing hallucinations based on the measurement and null space components is motivated. The concept of a hallucination map is introduced in Section III, along with a definition of specific hallucination maps. Sections IV and V describe the numerical studies performed to demonstrate the potential utility of proposed hallucination maps with a stylized tomographic imaging modality. Finally, a discussion and summary of the work is presented in Section VI.

II. Background

A. Imaging Models

A linear digital imaging system can be described as a continuous-to-discrete (CD) mapping [20], [21]:

$$\mathbf{g} = \mathscr{H} f(\mathbf{r}) + \mathbf{n}, \tag{1}$$

where $f(\mathbf{r}) \in \mathbb{L}_2\left(\mathbb{R}^d\right)$ is a function of continuous variables that represents the object being imaged, the vector $\mathbf{g} \in \mathbb{E}^M$ denotes the measured data samples and $\mathbf{n} \in \mathbb{E}^M$ is the measurement noise. The linear CD operator $\mathscr{H} : \mathbb{L}_2\left(\mathbb{R}^d\right) \to \mathbb{E}^M$ describes the action of the imaging system. In practice, discrete-to-discrete (DD) imaging models are often employed as approximations to the true CD imaging model. In a DD model, an N-dimensional approximation of $f(\mathbf{r})$ is utilized [20], [21]:

$$f_a(\mathbf{r}) = \sum_{n=1}^{N} [\theta]_n \psi_n(\mathbf{r}), \tag{2}$$

where the subscript a stands for approximate, $[\theta]_n$ is the n-th element of the coefficient vector $\theta \in \mathbb{E}^N$ and $\psi_n(\mathbf{r})$ is the n-th expansion function. On substitution from Eq. (2) in Eq. (1), the DD imaging system can be expressed as

$$\mathbf{g} \approx \mathscr{H} f_a(\mathbf{r}) + \mathbf{n} = \sum_{n=1}^{N} [\theta]_n \mathscr{H} \psi_n(\mathbf{r}) + \mathbf{n} \equiv \mathbf{H}\boldsymbol{\theta} + \mathbf{n}, \tag{3}$$

where $\mathbf{H}: \mathbb{E}^N \to \mathbb{E}^M$ is the system matrix constructed using $\mathscr{H}$ and $\{\psi_n(\mathbf{r})\}_{n=1}^N$. Image reconstruction methods based on Eq. (3) seek to estimate $\boldsymbol{\theta}$ from $\mathbf{g}$, after which the approximate object function $f_a(\mathbf{r})$ can be determined by use of Eq. (2). A well-known expansion function is the pixel expansion function. For two-dimensional objects $f(\mathbf{r})$ with $\mathbf{r} = (x, y)$, the pixel expansion function can be expressed as [21]:

$$\psi_n(\mathbf{r}) = \begin{cases} 1, \text{if} |x - x_n| \text{and} |y - y_n| \le \frac{\gamma}{2} \\ 0, \text{otherwise} \end{cases} \tag{4}$$

where $\mathbf{r}_n = (x_n, y_n)$ represents the coordinates of the n-th grid point of a uniform Cartesian lattice and γ denotes the spacing between the lattice points. When a pixel expansion function is employed, the corresponding coefficient vector $\boldsymbol{\theta}$ directly provides a digital image representation of the continuous object function $f_a(\mathbf{r})$.

B. Generalized Measurement and Null Components

For the DD imaging model described by Eq. (3), the properties of $\mathbf{H}$ affect the ability to estimate $\boldsymbol{\theta}$ uniquely and stably. In the absence of measurement noise, $\boldsymbol{\theta}$ can be determined uniquely from measurements $\mathbf{H}\boldsymbol{\theta}$ when $\mathbf{H}$ is injective or if $\boldsymbol{\theta}$ is known to lie in a subset S of $\mathbb{E}^N$ and the restriction $\mathbf{H}|_S$ is injective. The ability to stably reconstruct an estimate of $\boldsymbol{\theta}$ is also of fundamental importance. Stability is a way of quantifying how “close” two estimates $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2$ of $\boldsymbol{\theta}$ will be, if they are estimated from two “close” measurement vectors $\mathbf{g}_1$ and $\mathbf{g}_2$ respectively. For instance, $\mathbf{g}_1$ and $\mathbf{g}_2$ may correspond to the same object but differ due to them having two different measurement noise realizations. A popular notion of stability is based on how the ℓ_2-distance between $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ relates to that between $\mathbf{g}_1$ and $\mathbf{g}_2$ [22]:

$$\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2\|_2 \le \alpha \|\mathbf{g}_1 - \mathbf{g}_2\|_2, \tag{5}$$

where α is a constant that is additionally required to be smaller than a tolerance value ϵ.

The ability to estimate $\boldsymbol{\theta}$ stably can be analyzed through the lens of the singular value decomposition (SVD) of $\mathbf{H}$ [20]:

$$\mathbf{H} = \sum_{n=1}^{R} \sqrt{\mu_n} \mathbf{v}_n \mathbf{u}_n^{\dagger}. \tag{6}$$

Here, $\mathbf{u}_n$ and $\mathbf{v}_n$ are the singular vectors of $\mathbf{H}$ and $(\mu_n)^{1/2}$ are the singular values. The vector $\mathbf{u}_n^{\dagger}$ is the adjoint of $\mathbf{u}_n$ and the integer $R > 0$ denotes the rank of $\mathbf{H}$, where $\mathbf{H}$ is not necessarily full-rank. The singular values $(\mu_n)^{1/2}$ are ordered such that $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_R > 0$.

A pseudoinverse-based estimate of $\boldsymbol{\theta}$ can be computed as $\hat{\theta}_{pinv} \equiv \mathbf{H}^{+}\mathbf{g}$, where the linear operator $\mathbf{H}^{+}$ denotes the Moore-Penrose pseudoinverse of $\mathbf{H}$ that can be expressed as

$$\mathbf{H}^{+} = \sum_{n=1}^{R} \frac{1}{\sqrt{\mu_n}} \mathbf{u}_n \mathbf{v}_n^{\dagger}. \tag{7}$$

From Eq. (3), due to the linearity of $\mathbf{H}^{+}$, $\hat{\theta}_{pinv}$ can be represented as

$$\hat{\theta}_{pinv} = \mathbf{H}^{+}\mathbf{g} \approx \mathbf{H}^{+}(\mathbf{H}\theta + \mathbf{n}) = \mathbf{H}^{+}\mathbf{H}\theta + \mathbf{H}^{+}\mathbf{n}. \tag{8}$$

Due to the presence of the term $\mathbf{H}^{+}\mathbf{n}$ in Eq. (8), when the trailing singular values of $\mathbf{H}$ are small, α in Eq. (5) is large, leading to unstable estimates of $\boldsymbol{\theta}$. In this scenario, a truncated pseudoinverse can be defined as

$$\mathbf{H}_P^{+} = \sum_{n=1}^{P} \frac{1}{\sqrt{\mu_n}} \mathbf{u}_n \mathbf{v}_n^{\dagger}, \tag{9}$$

where the integer $P \leq R$ is chosen such that, for a given tolerance ϵ, $\mathbf{H}_P^{+}\mathbf{g}$ is a stable, linear estimate of $\boldsymbol{\theta}$ according to Eq. (5) with $\mu_P > 1/\epsilon^2 \geq \mu_{P+1}$. The truncated pseudoinverse can be used to form projection operators that project $\theta \in \mathbb{E}^N$ onto orthogonal subspaces – the 'generalized' null and measurement spaces [23]. The generalized null space of $\mathbf{H}$, denoted by $\mathcal{N}_P(\mathbf{H})$, is spanned by the singular vectors $\{\mathbf{u}_n\}_{n=P+1}^{N}$ that correspond to singular values satisfying $(\mu_n)^{1/2} \leq 1/\epsilon$. The orthogonal complement of the generalized null space is the generalized measurement space $\mathcal{N}_P^{\perp}(\mathbf{H})$.

Definition 1 (Generalized Measurement and Null Components:)—Let $\mathbf{H}$ and $\mathbf{H}_P^{+}$ denote the forward and truncated pseudoinverse operators, described in Equations (3) and (9) respectively. Let $\mathbf{H}_P$ denote the truncated forward operator, defined as

$$\mathbf{H}_P = \sum_{n=1}^{P} \sqrt{\mu_n} \mathbf{v}_n \mathbf{u}_n^{\dagger}. \tag{10}$$

Note that $\mathbf{H}_P^{+} = (\mathbf{H}_P)^{+}$. The coefficient vector $\boldsymbol{\theta}$ can be uniquely decomposed as $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{meas}} + \boldsymbol{\theta}_{\text{null}}$, where $\theta_{\text{meas}} \in \mathcal{N}_P^{\perp}(\mathbf{H})$ and $\theta_{\text{null}} \in \mathcal{N}_P(\mathbf{H})$ are specified as

$$\theta_{\text{meas}} = \mathscr{P}_{\text{meas}}\theta = \mathbf{H}_P^{+}\mathbf{H}\theta = \mathbf{H}_P^{+}\mathbf{H}_P\theta, \tag{11}$$

and

$$\theta_{\text{null}} = \mathscr{P}_{\text{null}}\theta = \left[\mathbf{I}_N - \mathbf{H}_P^{+}\mathbf{H}\right]\theta = \left[\mathbf{I}_N - \mathbf{H}_P^{+}\mathbf{H}_P\right]\theta. \tag{12}$$

Here, the projection operators $\mathscr{P}_{\text{meas}}$ and $\mathscr{P}_{\text{null}}$ project $\boldsymbol{\theta}$ to $\mathscr{N}_P^{\perp}(\mathbf{H})$ and $\mathscr{N}_P(\mathbf{H})$ [20], and $\mathbf{I}_N$ is the identity operator in $\mathbb{E}^N$.

In special cases where the singular values $(\mu_n)^{1/2}$ and the tolerance ϵ are such that $P = R$, the generalized null space is spanned by the singular vectors $\{\mathbf{u}_n\}_{n=R+1}^{N}$ with singular values $(\mu n)^{1/2} = 0$. In such cases, the generalized null space reduces to the true null space

$$\mathscr{N}_P(\mathbf{H}) = \mathscr{N}(\mathbf{H}) \equiv \left\{\theta \in \mathbb{E}^N \mid \mathbf{H}\theta = \mathbf{0}\right\}, \tag{13}$$

where $\mathbf{0}$ is the zero vector in $\mathbb{E}^M$. Correspondingly, the true measurement space is the orthogonal complement of the true null space. By definition, the true null space contains those object vectors that are mapped to the zero measurement data vector and hence are ‘invisible’ to the imaging system.

Having obtained the generalized measurement and null components of $\boldsymbol{\theta}$, the approximate object function $f_a(\mathbf{r})$ can also be decomposed into generalized measurement and null components:

$$\begin{aligned} f_a(\mathbf{r}) &= \sum_{n=1}^{N} [\theta]_n \psi_n(\mathbf{r}) \\ &= \sum_{n=1}^{N} [\theta_{\text{meas}}]_n \psi_n(\mathbf{r}) + \sum_{n=1}^{N} [\theta_{\text{null}}]_n \psi_n(\mathbf{r}) \\ &= f_{a,\text{meas}}(\mathbf{r}) + f_{a,\text{null}}(\mathbf{r}) \,. \end{aligned} \tag{14}$$

Note that for all $\mathbf{g}_1, \mathbf{g}_2 \in \mathbb{E}^M$, $\|\mathbf{H}_P^{+}\mathbf{g}_1 - \mathbf{H}_P^{+}\mathbf{g}_2\| \leq \left(1/(\mu_P)^{1/2}\right)\|\mathbf{g}_1 - \mathbf{g}_2\|$, whereas for all $\sigma \in \mathscr{N}_P(\mathbf{H})$, $\|\sigma\| \geq \|\mathbf{H}\sigma\|/(\mu_{P+1})^{1/2}$. Hence, for a given $\theta \in \mathbb{E}^N$, $\boldsymbol{\theta}_{\text{meas}}$ is the component of $\boldsymbol{\theta}$ that can be stably estimated via the truncated pseudoinverse from the measurement data. Contrarily, $\boldsymbol{\theta}_{\text{null}}$ cannot be stably estimated from the measurement data alone; additional information provided through priors and regularization is needed to estimate this component. These observations will be essential to the definitions of hallucinations that are provided later.

C. Regularization in Tomographic Image Reconstruction

As discussed above, in order to obtain a stable estimate of $\boldsymbol{\theta}$ from incomplete and/or noisy measurements, imposition of prior knowledge about the object is generally needed. A flexible method of incorporating priors in the estimation of $\boldsymbol{\theta}$ is through the Bayesian formalism, where $\boldsymbol{\theta}$, $\mathbf{g}$ and $\mathbf{n}$ are treated as instances of random variables with distributions p_{θ}, p_{g} and p_{n} respectively [22]. It is assumed that p_{θ}, i.e. the distribution over all objects is known, and is called the *prior*. By Bayes’ theorem, the posterior distribution $p_{\boldsymbol{\theta}|\text{g}}$, given by

$$p_{\theta\,|\,\text{g}}(\boldsymbol{\theta} \mid \mathbf{g}) = \frac{p_{\text{g}\,|\,\theta}(\mathbf{g} \mid \boldsymbol{\theta})\, p_{\theta}(\boldsymbol{\theta})}{p_{\text{g}}(\mathbf{g})}, \tag{15}$$

characterizes the probability density over all possible values of the object given the prior and the noise model. Estimates such as the maximum a posteriori (MAP) estimate argmax$_{\boldsymbol{\theta}}$ $p_{\boldsymbol{\theta}|\mathrm{g}}(\boldsymbol{\theta}|\mathbf{g})$ can then be obtained from the posterior.

Regularization via penalization is an alternative formalism to incorporate prior knowledge. Here, the image reconstruction task is formulated as an optimization problem such as [21]

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, C_d(\mathbf{g}, \mathbf{H}\boldsymbol{\theta}) + \lambda C_p(\boldsymbol{\theta}), \tag{16}$$

where the data fidelity term $C_d(\mathbf{g}, \mathbf{H}\boldsymbol{\theta})$ enforces the estimate $\hat{\boldsymbol{\theta}}$ when acted upon by $\mathbf{H}$ to agree with the observed measurement data $\mathbf{g}$ and the penalty term $C_p(\boldsymbol{\theta})$ encourages the solution to be consistent with the assumed prior. The hyper-parameter λ controls the trade-off between data fidelity and regularization. Often, the penalty term $C_p(\boldsymbol{\theta})$ is hand-crafted to encode priors such as the smoothness of natural images or sparsity of natural images in some transform domain [9]. The solution obtained through this formalism can be interpreted as the MAP estimate obtained from the Bayesian formalism in Eq. (15), with $p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \exp(-\lambda C_p(\boldsymbol{\theta}))$ and $p_{\mathrm{g}|\boldsymbol{\theta}}(\mathbf{g}|\boldsymbol{\theta}) = \exp(-C_d(\mathbf{g}, \mathbf{H}\boldsymbol{\theta}))$.

Regularization can also be interpreted as restricting the possible solutions to a subset $S_\mu \subset \mathbb{E}^N$, with S_μ being a member of a family of subsets parameterized by μ. The reconstruction procedure can then be represented by a possibly nonlinear mapping $\mathscr{R}_\mu: \mathbb{E}^M \to S_\mu$, with the image estimate given by $\hat{\boldsymbol{\theta}} = \mathscr{R}_\mu(\mathbf{g})$. Ideally, it is desirable that $\mathscr{R}_\mu$ satisfies the stability criterion described in Eq. (5).

Recently, methods that implicitly learn a regularizer from existing data have been proposed. Methods based on dictionary learning and learning sparsifying transforms were some of the earliest applications of such data-driven regularization [24]–[27]. However, the most actively investigated data-driven regularization methods involve learning from training data by use of deep neural networks, popularly known as deep learning [7], [28]. Deep learning has been employed in different ways to explicitly or implicitly impose priors in image reconstruction problems. For example, within the context of an end-to-end learned reconstruction mapping, a prior is imposed that is implicitly specified by the distribution of training data and network topology. A comprehensive survey of the current state of deep learning-based methods in tomographic image reconstruction can be found in recent reviews, [9], [29], [30].

However, there have been growing concerns regarding the ability of data-driven and learning based reconstruction methods to generalize to measurements that lie outside the training distribution [10]–[12], [31]. Moreover, deep learning-based reconstruction methods have been shown to not be uniformly stable, in the sense that certain imperceptible perturbations in the measurements may lead to large fluctuations in the reconstructed estimate [11], [12]. Such phenomena may lead to false structures appearing in the reconstructed image that do not exist in the object being imaged, and cannot be recovered stably from the original measurement data.

III. Definition of Hallucination Maps

When comparing or evaluating image reconstruction methods, it may be useful to visualize and quantify false structures that cannot be stably reconstructed from the measurements. Such structures have been colloquially referred to as being ‘hallucinated’ and are attributable to use of an imperfect reconstruction prior. Error maps that display the difference between the reconstructed image estimate and the true object are commonly employed to assess reconstruction errors. Artifacts revealed by error maps encompass a broad range of deviations that can appear in a reconstructed image with respect to its depiction of the object function being imaged. For example, incorrect modeling of the system matrix **H** or measurement noise can lead to artifacts. Consequently, as demonstrated in Fig. 1, it may not be possible to isolate and label the artifacts attributable to the reconstruction prior from the error map alone. A possible way to circumvent this is to compute separate error maps for the null and measurement components of the reconstructed image estimate. However, precise definitions for hallucinations in these sub-spaces have been lacking.

In order to visualize and quantify hallucinations in tomographic images, measurement and null space hallucination maps are formally defined below. The proposed definitions are general and can be applied to analyze hallucinations produced by any reconstruction method that seeks to invert a linear imaging model. The defined hallucination maps will permit isolation of image artifacts that cannot be stably reconstructed from the measurement data and are attributable to the implicit or explicit reconstruction prior.

A. Hallucination Map in the Generalized Measurement Space

Let $\widehat{\theta}$ denote the estimate of the coefficient vector $\boldsymbol{\theta}$ obtained from **g** by use of an image reconstruction method. It is desirable that the projection of $\widehat{\theta}$ onto the generalized measurement space $\mathcal{N}_P^{\perp}(\mathbf{H})$, i.e. $\widehat{\theta}_{\text{meas}}$, should be near the truncated pseudoinverse solution $\widehat{\theta}_{\text{tp}} \equiv \mathbf{H}_P^{+}\mathbf{g}$. This would ensure that $\widehat{\theta}_{\text{meas}}$ is consistent with the estimate of $\boldsymbol{\theta}$ that can be stably recovered from **g**. However, due to the imposed regularization in a reconstruction method, there may be discrepancies in $\widehat{\theta}_{\text{meas}}$ with respect to the stable estimate $\widehat{\theta}_{\text{tp}}$ in the generalized measurement space $\mathcal{N}_P^{\perp}(\mathbf{H})$. In order to quantify such differences, a hallucination map in the generalized measurement space is defined as follows.

Definition 2 (Generalized Measurement Space Hallucination Map): As previously defined, let $\widehat{\theta}$ be an image estimate obtained by use of a reconstruction method and let $\widehat{\theta}_{\text{tp}}$ be the truncated pseudoinverse solution. The hallucination map in the measurement space is defined as,

$$\widehat{\theta}_{\text{meas}}^{\text{HM}} \equiv \widehat{\theta}_{\text{meas}} - \widehat{\theta}_{\text{tp}}\,. \tag{17}$$

It should be noted that the computation of the hallucination map in the generalized measurement space requires no knowledge of the true object and simply reveals errors in the measurement component of $\widehat{\theta}$ with respect to the stably computed estimate $\widehat{\theta}_{\text{tp}}$.

For use in cases where pixel expansion functions are not employed, it is useful to translate the definition of hallucination maps to the subspace of the object space $\mathbb{L}_2\left(\mathbb{R}^d\right)$ spanned by a generic basis $\{\psi_n(\mathbf{r})\}_{i=1}^{N}$. By use of Eq. (2), the estimate of $f_a(\mathbf{r})$ can be represented as

$$\hat{f}_a(\mathbf{r}) = \sum_{n=1}^{N} [\hat{\theta}]_n \psi_n(\mathbf{r}) . \tag{18}$$

The hallucination map $\hat{f}^{\mathrm{HM}}_{a,\,\mathrm{meas}}(\mathbf{r})$ can be defined in the space $\mathbb{L}_2\left(\mathbb{R}^d\right)$ as

$$\hat{f}^{\mathrm{HM}}_{a,\,\mathrm{meas}}(\mathbf{r}) \equiv \sum_{n=1}^{N} \left[\hat{\theta}^{\mathrm{HM}}_{\mathrm{meas}}\right]_n \psi_n(\mathbf{r}) . \tag{19}$$

B. Hallucination Map in the Generalized Null Space

As reviewed in Section II-B, to estimate the generalized null vector $\boldsymbol{\theta}_{\mathrm{null}}$ from $\mathbf{g}$, reconstruction methods that impose appropriate priors are required. Hence, to accurately capture the effect of the prior on the reconstructed image, a definition of hallucinations must satisfy the following two desiderata:

1. The definition must involve the assessment of how accurate the estimate $\hat{\theta}_{\mathrm{null}} = \mathcal{P}_{\mathrm{null}}\hat{\theta}$ is as compared to the true generalized null vector $\boldsymbol{\theta}_{\mathrm{null}}$.
2. Since no prior is used in obtaining $\hat{\theta}_{\mathrm{tp}}$, the definition must ensure that $\hat{\theta}_{\mathrm{tp}}$ does not have any null space hallucinations.

With these in mind, a hallucination map $\hat{\theta}^{\mathrm{HM}}_{\mathrm{null}}$ in the generalized null space $\mathcal{N}_P(\mathbf{H})$ is defined as follows.

Definition 3 (Generalized Null Space Hallucination Map): Consider a pixelwise indicator function $\mathbb{1}\colon \mathbb{R}^N \to \mathbb{R}^N$ such that for any $\vartheta \in \mathbb{R}^N$

$$[\mathbb{1}(\vartheta)]_n = \begin{cases} 1, & \text{if}[\vartheta]_n \neq 0 \\ 0, & \text{if}[\vartheta]_n = 0. \end{cases} \tag{20}$$

Then, the hallucination map $\boldsymbol{\theta}^{\mathrm{HM}}_{\mathrm{null}} \in \mathbb{E}^N$ can be defined as

$$\hat{\boldsymbol{\theta}}^{\mathrm{HM}}_{\mathrm{null}} \equiv \mathbb{1}\left(\hat{\boldsymbol{\theta}}_{\mathrm{null}}\right) \odot \left(\hat{\boldsymbol{\theta}}_{\mathrm{null}} - \boldsymbol{\theta}_{\mathrm{null}}\right), \tag{21}$$

where $\odot$ denotes the Hadamard product or element-wise multiplication. Note that the indicator function in the definition ensures that $\hat{\theta}_{\mathrm{tp}}$ does not possess any null space hallucinations, since no prior was imposed.

It is important to highlight that, for the computation of the hallucination map in the generalized null space, one must have full knowledge of the generalized null component of the true object. This is in contrast to the hallucinations in the generalized measurement space, where the knowledge of the generalized measurement component of the true object is not required. This simply reflects that, according to the provided definitions, the generalized null space hallucination maps depict errors in the reconstructed null component of the object, while the generalized measurement space hallucination maps depict errors in the component of the object that can be stably reconstructed via a truncated pseudoinverse operator from the observed measurement data.

This difference in the two definitions is associated with the fact that $\hat{\boldsymbol{\theta}}_{\mathrm{tp}}$ is close to $\mathbf{H}_P^+\mathbf{H}\boldsymbol{\theta}$ if the measurement noise is small in the sense of Eq. (5), and/or the model error is negligible. Hence, the proposed definition of $\hat{\boldsymbol{\theta}}_{\mathrm{meas}}^{\mathrm{HM}}$ is able to reveal the effect of the prior on the reconstructed generalized measurement space component, without requiring the true object. In this sense, there is no analog of a stably reconstructed component like $\hat{\boldsymbol{\theta}}_{\mathrm{tp}}$ in the null space; hence invoking the true null component is necessary for defining $\hat{\boldsymbol{\theta}}_{\mathrm{null}}^{\mathrm{HM}}$. Note that due to our definition, $\hat{\boldsymbol{\theta}}_{\mathrm{meas}}^{\mathrm{HM}}$ may also be influenced by the different noise propagation characteristics of the methods employed to form $\hat{\boldsymbol{\theta}}_{\mathrm{tp}}$ and $\hat{\boldsymbol{\theta}}$ and therefore may not solely quantify errors associated with the prior.

It should also be noted that the errors introduced by the prior in the measurement space can be remedied by adopting a reconstruction method that penalizes measurement space hallucinations without any prior knowledge of the object, e.g., via a data consistency constraint [12] or null space shuttle procedure [23]. Accordingly, for such constrained image reconstruction methods, analyzing hallucinations in the null space is critical towards understanding the effect of the prior on the image estimate.

Similar to the hallucination map in the generalized measurement space, the hallucination map $\hat{f}_{a,\mathrm{null}}^{\mathrm{HM}}(\mathbf{r})$ can be defined as

$$\hat{f}_{a,\mathrm{null}}^{\mathrm{HM}}(\mathbf{r}) \equiv \sum_{n=1}^{N} \left[\hat{\theta}_{\mathrm{null}}^{\mathrm{HM}}\right]_n \psi_n(\mathbf{r})\,. \tag{22}$$

According to the proposed definitions, the truncated pseudoinverse solution $\hat{\boldsymbol{\theta}}_{\mathrm{tp}}$ has zero hallucination in both the generalized measurement space and null space. However, that does not necessarily imply that $\hat{\boldsymbol{\theta}}_{\mathrm{tp}}$ is without artifacts, since $\hat{\boldsymbol{\theta}}_{\mathrm{tp}}$ ignores $\boldsymbol{\theta}_{\mathrm{null}}$ completely. The computation of $\hat{\boldsymbol{\theta}}_{\mathrm{tp}}$ leads to the recovery of only $\boldsymbol{\theta}_{\mathrm{meas}}$ that can be estimated stably. When other regularized reconstruction methods attempt to reduce artifacts by imposing priors to estimate $\boldsymbol{\theta}_{\mathrm{null}}$, a trade-off is made between the estimation of $\boldsymbol{\theta}_{\mathrm{meas}}$ and $\boldsymbol{\theta}_{\mathrm{null}}$ that can potentially lead to hallucinations in the generalized measurement space and null space.

C. Specific Hallucination Maps

The use of objective, or task-based, measures of image quality for evaluating imaging systems has been widely advocated [20]. However, the hallucination maps as defined in

Section III do not incorporate any task-specific information. In particular, $\widehat{\theta}_{\text{null}}^{\text{HM}}$ may contain an abundance of structures or textures, some of which may not confound an observer on a specified diagnostic task. Hence, it may be useful to identify those structures or textures in the hallucination maps that are task-relevant. One possible way to accomplish this is to process the hallucination map via an image processing transformation T, such that potentially task-relevant features or textures are localized while others are suppressed [32], [33]. Formally, this can be described as:

$$\widehat{\theta}_{\text{null}}^{\text{SHM}} = T\widehat{\theta}_{\text{null}}^{\text{HM}}, \tag{23}$$

where the processed pixel map $\widehat{\theta}_{\text{null}}^{\text{SHM}}$ that preserves task-specific information is referred to as a *specific hallucination map*. Note that the design of the transformation T is application-dependent, as it should localize those structures or textures from the hallucination map that are relevant to a specified task. Moreover, the specification of the observer (which could be a human or computational procedure) who will perform the task should also influence the design of T, as the extent to which hallucinations impact observer performance will vary. While requiring significant effort to formulate, specific hallucination maps open up the possibility of comparing reconstruction methods based on their propensities for creating hallucinations that influence task-performance.

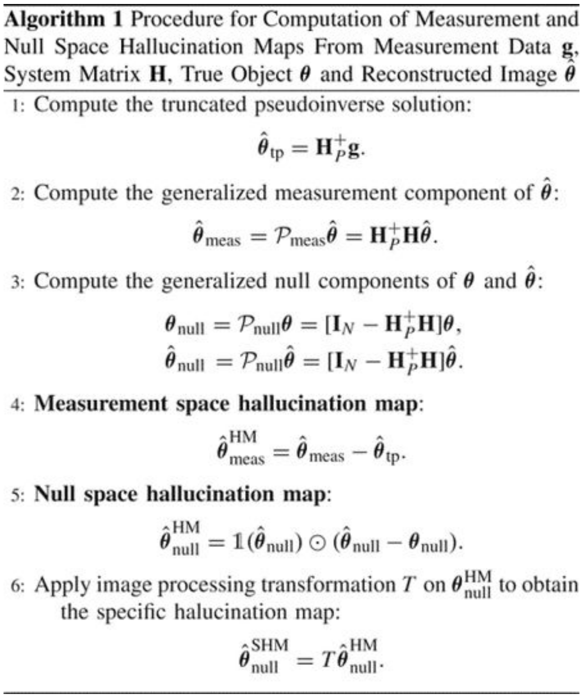
The complete procedure for computing measurement and null space hallucination maps, as well as the specific hallucination map, is presented in Algorithm 1.

IV. Numerical studies

Numerical studies were conducted to demonstrate the utility of the proposed hallucination maps. Although the focus of these preliminary studies is on null space hallucination maps, the presented analyses could readily be repeated by use of measurement space hallucination maps. Hallucination maps were employed to compare the behavior of data-driven and model-based image reconstruction methods under different conditions.

A. Stylized Imaging System

A stylized two-dimensional (2D) single-coil magnetic resonance (MR) imaging system was considered. It should be noted that the assumed imaging operator was not intended to accurately model a real-world MR imager. Instead, the purpose of the presented simulation studies is only to demonstrate the potential utility of hallucination maps. Hence physical factors such as coil sensitivity and bias field inhomogeneity were not considered. Fully-sampled k-space data were emulated by applying the 2D Fast Fourier Transform (FFT) on the digital objects described below. Independent and identically distributed (iid) Gaussian noise was added to the real and imaginary components of the complex-valued k-space data [34] in the training dataset for the U-Net as well as in the test data during evaluation with different reconstruction methods. Additionally, in the test dataset, zero-mean random uniform phase noise [35] was introduced into the k-space measurements to emulate modeling errors [20]. A uniform Cartesian undersampling mask with an undersampling factor of 3 was applied on the fully-sampled k-space data to obtain undersampled measurements, as shown in Fig. S.1 in the Supplementary file. The k-space lines that were not sampled were subsequently zero-filled. The Moore-Penrose pseudoinverse $\mathbf{H}^{+}$ was applied by performing the inverse 2D Fast Fourier Transform (IFFT) on the zero-filled k-space data. Since the true pseudoinverse was considered without any truncation of singular values, the hallucination map in the generalized null space in our studies corresponds to the hallucination map in the true null space.



Algorithm 1 Procedure for Computation of Measurement and Null Space Hallucination Maps From Measurement Data $\mathbf{g}$, System Matrix $\mathbf{H}$, True Object $\boldsymbol{\theta}$ and Reconstructed Image $\hat{\boldsymbol{\theta}}$

```
Compute the truncated pseudoinverse solution:
    $\hat{\boldsymbol{\theta}}_{\mathrm{tp}} = \mathbf{H}_{P}^{+}\mathbf{g}.$
Compute the generalized measurement component of $\hat{\boldsymbol{\theta}}$:
    $\hat{\boldsymbol{\theta}}_{\mathrm{meas}} = \mathcal{P}_{\mathrm{meas}}\hat{\boldsymbol{\theta}} = \mathbf{H}_{P}^{+}\mathbf{H}\hat{\boldsymbol{\theta}}.$
Compute the generalized null components of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$:
    $\boldsymbol{\theta}_{\mathrm{null}} = \mathcal{P}_{\mathrm{null}}\boldsymbol{\theta} = [\mathbf{I}_{N} - \mathbf{H}_{P}^{+}\mathbf{H}]\boldsymbol{\theta},$
    $\hat{\boldsymbol{\theta}}_{\mathrm{null}} = \mathcal{P}_{\mathrm{null}}\hat{\boldsymbol{\theta}} = [\mathbf{I}_{N} - \mathbf{H}_{P}^{+}\mathbf{H}]\hat{\boldsymbol{\theta}}.$
**Measurement space hallucination map**:
    $\hat{\boldsymbol{\theta}}_{\mathrm{meas}}^{\mathrm{HM}} = \hat{\boldsymbol{\theta}}_{\mathrm{meas}} - \hat{\boldsymbol{\theta}}_{\mathrm{tp}}.$
**Null space hallucination map**:
    $\hat{\boldsymbol{\theta}}_{\mathrm{null}}^{\mathrm{HM}} = \mathbb{1}(\hat{\boldsymbol{\theta}}_{\mathrm{null}}) \odot (\hat{\boldsymbol{\theta}}_{\mathrm{null}} - \boldsymbol{\theta}_{\mathrm{null}}).$
Apply image processing transformation $T$ on $\boldsymbol{\theta}_{\mathrm{null}}^{\mathrm{HM}}$ to obtain
  the specific halucination map:
    $\hat{\boldsymbol{\theta}}_{\mathrm{null}}^{\mathrm{SHM}} = T\hat{\boldsymbol{\theta}}_{\mathrm{null}}^{\mathrm{HM}}.$
```

B. Reconstruction Methods

Both data-driven and non-data-driven image reconstruction methods were investigated. The data-driven method considered was a U-Net based method [36]–[38], which learns a mapping from an initial image estimate that contains artifacts due to undersampling to an accurate estimate of the true object. In our studies, the initial image estimate that was input to the U-Net was obtained by applying the pseudoinverse on the k-space data. Two different non-data-driven reconstruction methods were considered. The first method, which is known as penalized least-squares with total variation (PLS-TV) [6], involves solving a least-squares optimization problem with a total variation penalty [6]. The second method is known as deep image prior (DIP) [39], [40], where the reconstructed estimate is constrained to lie in the range of an untrained deep network [28] such that the estimate agrees with the observed measurements in a least-squares sense. These reconstruction methods are described in detail in Section S.II of the Supplementary file.

C. Training, Validation and Test Data

For the U-Net based reconstruction method, training was performed on 2D axial adult brain MRI images from the NYU fastMRI Initiative database [41]. These will be referred to as the in-distribution (IND) images. The training and validation datasets contained 2500 and

500 images, respectively. For testing, both IND and out-of-distribution (OOD) images were considered. The OOD images were obtained from a pediatric epilepsy resection MRI dataset [42]. Both the IND and OOD testing datasets contained 69 images. It should be noted that the OOD images differed from the IND images in several aspects, such as the nature of the objects (adult for IND and pediatric for OOD) as well as the use of different MR systems employed to obtain the true object images in each case. All images were of dimension 320 × 320.

After creating the training, validation and test datasets, neural network training was performed with the IND training and validation datasets for the U-Net method. At test time, images were reconstructed from both IND and OOD test datasets using the U-Net, PLS-TV and DIP methods. Details regarding the implementation of these methods are presented in Section S.II of the Supplementary file.

D. Computation of Hallucination Maps

After images were reconstructed from the testing data, null space hallucination maps $\hat{\boldsymbol{\theta}}_{\mathrm{null}}^{\mathrm{HM}}$ were computed. The quantities $\hat{\boldsymbol{\theta}}_{\mathrm{null}}$ and $\boldsymbol{\theta}_{\mathrm{null}}$, as required by Eq. (21), were computed according to Eq. (12). Subsequently, specific null space hallucination maps $\hat{\boldsymbol{\theta}}_{\mathrm{null}}^{\mathrm{SHM}}$ were also computed. In this preliminary study, these maps were designed for the purpose of localizing regions where coherent structures, as opposed to random errors, were present in $\hat{\boldsymbol{\theta}}_{\mathrm{null}}^{\mathrm{HM}}$. Such structured hallucinations could be relevant to certain signal detection tasks. To accomplish this, the transformation T in Eq. (23) was implemented as follows. First, the region of support of each object was identified using Otsu's method [43] and binary support masks were formed for each object. The support masks were applied on the $\hat{\boldsymbol{\theta}}_{\mathrm{null}}^{\mathrm{HM}}$ such that errors in the reconstructed image that lie outside the region of support could be ignored. Subsequently, histogram equalization was performed. A 2D Gaussian filter with kernel width of 7 was applied on the histogram-equalized map in order to obtain a smooth distribution of intensities across the hallucination map. The width of the Gaussian filter was chosen heuristically in this study. Finally, a binary threshold was applied where the cut-off value was set to the 95-th percentile of intensity values in the processed map, such that intensities below the threshold were set to zero and intensities above the threshold were set to 1. From the thresholded maps, connected components that had a size of less than 100 pixels (≈ 0.1% of total number of pixels in each image) were eliminated to remove localized regions with negligible dimensions, resulting in the specific hallucination maps $\hat{\boldsymbol{\theta}}_{\mathrm{null}}^{\mathrm{SHM}}$ for our studies. This procedure for computing the action of T was repeated for all $\hat{\boldsymbol{\theta}}_{\mathrm{null}}^{\mathrm{HM}}$ computed from both the IND and OOD test datasets for each reconstruction method. It should be noted that this procedure serves only as a simplistic example of the computation of a specific hallucination map, and there is no suggestion that it is optimal in any sense.

Finally, conventional error maps were computed as the difference between the reconstructed estimate $\hat{\boldsymbol{\theta}}$ and the true object $\boldsymbol{\theta}$. In order to demonstrate the potential utility of the specific hallucination maps over processed versions of conventional error maps, *specific error maps* were formed by acting T on the error maps. The codes employed in our numerical studies are available at https://github.com/compimaging-sci/hallucinations-tomo-recon.

V. Results

The numerical results are organized as follows. First, an illustration of hallucination maps is provided for different reconstruction methods, in order to demonstrate their utility in highlighting false structures that may be introduced due to the imposed prior. Differences in the null space hallucination maps corresponding to the data-driven U-Net method when applied to IND and OOD data are examined. This is followed by a demonstration of the difference in the quantitative performance of the U-Net method on IND and OOD data. The performance of the U-Net is compared with the non-data driven methods in our studies – PLS-TV and DIP – in terms of metrics derived by use of null space hallucination maps.

A. Differences Between Error and Hallucination Maps

Reconstructed images and corresponding error maps and null space hallucination maps from an IND measurement are shown in Fig. 2. It can be observed that, for all the reconstruction methods, the error map and the null space hallucination map have different characteristics in some regions of the image. This is because the error map contains false structures due to hallucinations as well as all other factors, whereas the null space hallucination map only contains errors due to the imposed prior. These differences can also be observed from the computed specific error maps and specific null space hallucination maps. As expected, the U-Net method performs well, leading to mostly low intensity regions in the null space hallucination map. In one of the regions that is featured in the specific hallucination map for all the reconstruction methods, it can be seen that the U-Net has lower hallucinations since it is able to faithfully recover fine structures in the region. Such fine structures were oversmoothed in the reconstructed images that were obtained by use of the PLS-TV and DIP methods, leading to higher hallucinations. On the other hand, all the reconstructed images also contain a distinct false structure that is revealed in the specific error map but not the specific hallucination map. This is an example of a false structure that can exist in reconstructed images, but may not necessarily be classified as a hallucination.

To further demonstrate the different characteristics of error maps and null space hallucination maps for this IND study, scatter plots of the centroids of the detected regions in each type of map corresponding to the ensemble of IND reconstructed images from all three reconstruction methods are shown in Fig. 3 (top row). From these scatter plots, it can be observed that there is a high amount of variance in the locations of the detected regions in the specific error maps as compared to the detected regions in the specific hallucination maps. The latter typically appear in similar regions across the ensemble of reconstructed images for all the methods. Furthermore, the concentrations of centroids for the detected regions in both types of maps have some degree of non-overlap. These observations reflect the fact that, due to additional sources of error such as measurement noise and model error that are also typically random in nature, the regions in the reconstructed images that are revealed by the error map can sometimes be different from those revealed by the null space hallucination map that considers error only due to an inaccurate prior.

As the distribution shifts to OOD, as shown in Fig. 4, the null space hallucination map for the U-Net method appears comparable to the hallucination maps obtained by use of PLS-TV and DIP. False structures that can be identified as hallucinations appear in the image reconstructed by the U-Net method. The higher error for the U-Net method is a result of the change of distribution and the method's inability to generalize well to data that are significantly out of distribution with respect to the training data. The change of distribution results in significant inaccuracies in the null component of the reconstructed estimate produced by the U-Net. Under such circumstances, it can be useful to identify and localize hallucinations due to inaccuracies in the imposed data-driven regularization through the null space hallucinations.

As shown in Fig. 4 and consistent with the IND results discussed above, the localized regions detected in the specific error map and specific hallucination map for the OOD cases are generally different. Scatter plots of the centroids of the detected regions in the specific error maps and specific hallucination maps confirm this and are displayed in Fig. 3 (bottom row). For all the reconstruction methods, the error map centroids again have a higher variance and are located away from clusters of hallucination map centroids in some regions. In other words, under such circumstances, one cannot rely on only the error maps without considering the corresponding hallucination maps in order to estimate where hallucinations due to the imposed prior are likely to be localized in a reconstructed image.

Although hallucination maps can reveal false structures, the impact of the false structures on specific applications requires further analysis. For example, a false structure may be classified as a *false positive structure* or a *false negative structure* [44], [45]. A false positive structure is one which is absent in the true object but present in the reconstructed image, whereas a false negative structure denotes the opposite. While an important topic, the classification of hallucinations is beyond the scope of this paper.

B. Investigation of Structured Hallucinations

Additional studies were conducted to validate that the specific hallucination maps actually revealed regions in the image that contain significant errors. To accomplish this, two empirical probability distribution functions (PDFs) were estimated that describe the average SSIM values computed over two non-overlapping regions in the reconstructed images for the OOD case. One region corresponded to the support of the specific hallucination maps described above and the second region was spanned by all other pixels in the image. The two empirical PDFs are shown in Fig. 5(a) and reveal that the mode of the distribution corresponding to the SSIM averaged over the structured hallucination regions is demonstrably lower than that describing the average SSIM values over the background regions.

The empirical PDFs that described the SSIM value averaged over the structured hallucination regions were also compared for each of the three reconstruction methods. As shown in Fig. 5(b), for the IND case, the images reconstructed by use of the U-Net had significantly higher SSIM values, on average, in the structured hallucination regions as compared to both the PLS-TV and DIP methods. This can be attributed to network training with a sufficiently large amount of IND data. However, for the OOD case in Fig.

5(c), because null space hallucinations increased for the U-Net method, the corresponding reconstructed images had lower SSIM values on average as compared with DIP in the support of the null space hallucination maps. The medians of ensemble SSIM values in these support regions for all the reconstruction methods with IND and OOD data are shown in Table I. It should be noted that, for both the IND and OOD cases, the DIP method was implemented with the same network architecture as the U-Net based method. Thus, when there is a shift in the testing data distribution, some data-driven methods such as the U-Net method may not provide any significant improvement in the estimate of the null component compared to model-based methods that do not employ training data. However, the data-driven methods involve the additional risk of hallucinating false structures. These observations gained through hallucination maps provide insight into the impact of the data-driven nature of the prior imposed by pre-trained neural networks.

C. Bias Maps and Hallucinations

A *bias map*, defined as

$$\mathbf{b} \equiv \mathbb{E}\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}, \tag{24}$$

determines the expected deviation of an image estimate from the true object, and as such, may include contributions from an incorrect prior, as well as those from incorrect measurement and noise models. Hence, the bias map may be correlated with the hallucination maps, but may display significant differences from it based on the average behavior of the inaccuracies in the measurement and noise models. For example, Fig. 6 shows the bias map computed using a dataset of images estimated from simulated undersampled MRI measurements with fixed phase noise and iid Gaussian additive noise, along with the error map and the null space hallucination map for an IND and an OOD image. The corresponding true objects are shown in Figs. 2 and 1(b) respectively. Fig. 6 shows that the bias map retains clusters of artifacts from the error map that are due to the phase noise. Hence, although the bias maps are correlated with both the hallucination maps and the error map, each provides a different kind of information.

VI. Summary and Conclusion

While regularization via sparsity-promoting penalties in an optimization-based reconstruction framework is commonly employed, emerging learning-based methods that employ deep neural networks have shown the potential to improve reconstructed image quality further by learning priors from existing data. However, an analysis of the prior information learned by deep networks and their ability to generalize to data that may lie outside the training distribution is still being explored. Additionally, there are open questions and concerns about the stability of such networks when applied for image reconstruction. While it has been understood that use of an inaccurate prior might lead to false structures, or hallucinations, being introduced in the reconstructed image, formal definitions for hallucinations within the context of tomographic image reconstruction have not been reported.

In this work, by use of concepts from linear operator theory, formal definitions for hallucination maps in linear tomographic imaging problems are introduced. These provide the opportunity to isolate and visualize image hallucinations that are contained within the measurement or null spaces of a linear imaging operator. The measurement space hallucination map permits the analysis of errors in the measurement space component of a reconstructed object estimate with respect to the component of the object that can be stably computed from a given set of measurement data. Alternatively, the null space hallucination map permits analysis of errors in the null space component of a reconstructed object estimate with respect to the true object null space component. These errors are caused solely by the reconstruction prior. Both maps can be employed to systematically investigate the impact of different priors utilized in image reconstruction methods. Finally, the notion of a specific hallucination map was also introduced, which can be formulated to reveal hallucinations that are relevant to a specified image-based inference.

Numerical studies were performed with simulated undersampled measurements from a stylized single-coil MRI system. Both data-driven and non-data-driven methods were investigated to demonstrate the utility of the proposed hallucination maps. It was observed that null space hallucination maps can be particularly useful as compared to traditional error maps when assessing the effect of data-driven regularization strategies with out-of-distribution data. Furthermore, it was shown that structured hallucinations with data-driven methods that are caused due to a shift in the data distribution may ultimately lead to significant artifacts in the reconstructed image.

The computation of the projection operations as described in Eq. (11) and Eq. (12) via the SVD may be infeasible for large-scale problems. Wilson and Barrett [46] proposed an iterative method to compute θ_{meas} and θ_{null} without explicit computation of the SVD of $\mathbf{H}$. Alternatively, randomized SVD [47] is a relatively computationally efficient algorithm that can be employed to estimate these quantities. Kuo *et al.* [48] recently proposed a method to learn null space projection operations that can significantly reduce the computational burden. It may also be expected that the importance of analyzing hallucinations in image reconstruction can further stimulate the development of efficient methods for implementing projection operators. The development of such computationally efficient methods for large-scale problems remains an active area of research.

It should be noted that the proposed definition of hallucination maps is general and can be applied to any linear imaging system and reconstruction method, provided that the computation of the projection operators $\mathscr{P}_{\text{meas}}$ and $\mathscr{P}_{\text{null}}$ is feasible. Depending on the sampling pattern involved in the data acquisition process, different system matrices $\mathbf{H}$ will have different null space characteristics. This, in turn, may lead to different properties in the corresponding hallucination maps that would allow a comparison of reconstruction methods under a variety of data acquisition strategies.

The proposed framework is most useful in situations where the generalized null component of the true object is significant and hence strong priors need to be incorporated in the reconstruction method via regularization. If the generalized null component is relatively small compared to the generalized measurement component, the need for strong regularization during reconstruction is diminished. This, in turn, would imply that hallucinations are likely to be minimal or non-existent due to the imposed weak regularization and hence computing hallucination maps may not be necessary. In such situations, computing only the error map may be sufficient to assess the reconstruction method.

There remain important topics for future investigation. Beyond the framework presented, it will be important to derive objective figures-of-merit (FOMs) from ensembles of hallucination maps. Furthermore, the probability of occurrence of hallucinations can be potentially quantified from ensembles of hallucination maps. While understanding the interplay between hallucinations and image reconstruction priors is important in preliminary studies, ultimately, image reconstruction methods should be objectively evaluated with consideration of all physical and statistical factors.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by the National Institutes of Health (NIH) Awards under Grant EB020604, Grant EB023045, Grant NS102213, and Grant EB028652; and in part by the National Science Foundation (NSF) Award under Grant DMS1614305.

References

[1]. Kak AC, Slaney M, and Wang G, “Principles of computerized tomographic imaging,” Med. Phys, vol. 29, no. 1, p. 107, 2002.

[2]. Yang AC-Y, Kretzler M, Sudarski S, Gulani V, and Seiberlich N, “Sparse reconstruction techniques in MRI: Methods, applications, and challenges to clinical adoption,” Investigative Radiol., vol. 51, no. 6, p. 349, 2016.

[3]. Donoho DL, “For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution,” Commun. Pure Appl. Math., J. Issued Courant Inst. Math. Sci, vol. 59, no. 6, pp. 797–829, 2006.

[4]. Donoho DL and Elad M, “Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization,” Proc. Nat. Acad. Sci. USA, vol. 100, no. 5, pp. 2197–2202, 2003. [PubMed: 16576749]

[5]. Candès EJ, Romberg JK, and Tao T, “Stable signal recovery from incomplete and inaccurate measurements,” Commun. Pure Appl. Math., J. Issued Courant Inst. Math. Sci, vol. 59, no. 8, pp. 1207–1223, 2006.

[6]. Sidky EY and Pan X, “Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization,” Phys. Med. Biol, vol. 53, no. 17, p. 4777, 9. 2008. [PubMed: 18701771]

[7]. Wang G, “A perspective on deep imaging,” IEEE Access, vol. 4, pp. 8914–8924, 2016.

[8]. McCann MT, Jin KH, and Unser M, “Convolutional neural networks for inverse problems in imaging: A review,” IEEE Signal Process. Mag, vol. 34, no. 6, pp. 85–95, 11. 2017.

[9]. Ravishankar S, Ye JC, and Fessler JA, “Image reconstruction: From sparsity to data-adaptive methods and machine learning,” Proc. IEEE, vol. 108, no. 1, pp. 86–109, 1. 2020.

[10]. Huang Y, Würfl T, Breininger K, Liu L, Lauritsch G, and Maier A, “Some investigations on robustness of deep learning in limited angle tomography,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent Cham, Switzerland: Springer, 2018, pp. 145–153.

[11]. Gottschling NM, Antun V, Adcock B, and Hansen AC, “The troublesome kernel: Why deep learning for inverse problems is typically unstable,” 2020, arXiv:2001.01258. [Online]. Available: http://arxiv.org/abs/2001.01258

[12]. Antun V, Renna F, Poon C, Adcock B, and Hansen AC, “On instabilities of deep learning in image reconstruction and the potential costs of AI,” Proc. Nat. Acad. Sci. USA, vol. 117, no. 48, pp. 30088–30095, 12. 2020. [PubMed: 32393633]

[13]. Laves M-H, Tölle M, and Ortmaier T, “Uncertainty estimation in medical image denoising with Bayesian deep image prior,” in Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis. Cham, Switzerland: Springer, 2020, pp. 81–96.

[14]. Asim M, Daniels M, Leong O, Ahmed A, and Hand P, “Invertible generative models for inverse problems: Mitigating representation error and dataset bias,” in Proc. Int. Conf. Mach. Learn, 2020, pp. 399–409.

[15]. Kelkar VA, Bhadra S, and Anastasio MA, “Compressible latent-space invertible networks for generative model-constrained image reconstruction,” IEEE Trans. Comput. Imag, vol. 7, pp. 209–223, 2021.

[16]. Baker S and Kanade T, “Limits on super-resolution and how to break them,” IEEE Trans. Pattern Anal. Mach. Intell, vol. 24, no. 9, pp. 1167–1183, 9. 2002.

[17]. Liu W, Lin D, and Tang X, “Hallucinating faces: TensorPatch super-resolution and coupled residue compensation,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 2, 6. 2005, pp. 478–484.

[18]. Wang N, Tao D, Gao X, Li X, and Li J, “A comprehensive survey to face hallucination,” Int. J. Comput. Vis, vol. 106, no. 1, pp. 9–30, 1. 2014.

[19]. Fawzi A, Samulowitz H, Turaga D, and Frossard P, “Image inpainting through neural networks hallucinations,” in Proc. IEEE 12th Image, Video, Multidimensional Signal Process. Workshop (IVMSP), 7. 2016, pp. 1–5.

[20]. Barrett HH and Myers KJ, Foundations of Image Science. Hoboken, NJ, USA: Wiley, 2013.

[21]. Anastasio MA and Schoonover RW, “Basic principles of inverse problems for optical scientists,” in Digital Encyclopedia of Applied Physics. Hoboken, NJ, USA: Wiley, 2003, pp. 1–24.

[22]. Bal G. (2012). Introduction to Inverse Problems. [Online]. Available: https://statistics.uchicago.edu/~guillaumebal/PAPERS/IntroductionInverseProblems.pdf

[23]. Deal MM and Nolet G, “Nullspace shuttles,” Geophys. J. Int, vol. 124, no. 2, pp. 372–380, 2. 1996.

[24]. Ravishankar S and Bresler Y, “MR image reconstruction from highly undersampled k-space data by dictionary learning,” IEEE Trans. Med. Imag, vol. 30, no. 5, pp. 1028–1041, 5 2011.

[25]. Tiwari S, Kaur K, and Arya K, “A study on dictionary learning based image reconstruction techniques for big medical data,” in Handbook of Multimedia Information Security: Techniques and Applications. Cham, Switzerland: Springer, 2019, pp. 377–393.

[26]. Ravishankar S and Bresler Y, “Learning sparsifying transforms,” IEEE Trans. Signal Process, vol. 61, no. 5, pp. 1072–1086, 3. 2013.

[27]. Ravishankar S and Bresler Y, “Data-driven learning of a union of sparsifying transforms model for blind compressed sensing,” IEEE Trans. Comput. Imag, vol. 2, no. 3, pp. 294–309, 9. 2016.

[28]. Goodfellow I, Bengio Y, Courville A, and Bengio Y, Deep Learning, vol. 1. Cambridge, MA, USA: MIT Press, 2016.

[29]. McCann MT and Unser M, “Biomedical image reconstruction: From the foundations to deep neural networks,” 2019, arXiv:1901.03565. [Online]. Available: http://arxiv.org/abs/1901.03565

[30]. Hammernik K and Knoll F, “Machine learning for image reconstruction,” in Handbook of Medical Image Computing and Computer Assisted Intervention. Amsterdam, The Netherlands: Elsevier, 2020, pp. 25–64.

[31]. Stayman JW, Prince JL, and Siewerdsen JH, “Information propagation in prior-image-based reconstruction,” in Proc. Int. Conf. Image Formation X-Ray Comput. Tomogr., Int. Conf. Image Formation X-Ray Comput. Tomogr, 2012, p. 334.

[32]. Castellano G, Bonilha L, Li LM, and Cendes F, “Texture analysis of medical images,” Clin. Radiol, vol. 59, no. 12, pp. 1061–1069, 12. 2004. [PubMed: 15556588]

[33]. Chowdhary CL and Acharjya DP, “Segmentation and feature extraction in medical imaging: A systematic review,” Procedia Comput. Sci, vol. 167, pp. 26–36, 1. 2020.

[34]. Aja-Fernández S and Vegas-Sánchez-Ferrero G, Statistical Analysis of Noise in MRI. Cham, Switzerland: Springer, 2016.

[35]. Xiaoyu F, Qiusheng L, and Baoshun S, “Compressed sensing MRI with phase noise disturbance based on adaptive tight frame and total variation,” IEEE Access, vol. 5, pp. 19311–19321, 2017.

[36]. Jin KH, McCann MT, Froustey E, and Unser M, “Deep convolutional neural network for inverse problems in imaging,” IEEE Trans. Image Process, vol. 26, no. 9, pp. 4509–4522, 9. 2017. [PubMed: 28641250]

[37]. Han Y and Ye JC, “Framing U-Net via deep convolutional framelets: Application to sparse-view CT,” IEEE Trans. Med. Imag, vol. 37, no. 6, pp. 1418–1429, 6. 2018.

[38]. Hyun CM, Kim HP, Lee SM, Lee S, and Seo JK, “Deep learning for undersampled MRI reconstruction,” Phys. Med. Biol, vol. 63, no. 13, 6. 2018, Art. no. 135007.

[39]. Lempitsky V, Vedaldi A, and Ulyanov D, “Deep image prior,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, 6. 2018, pp. 9446–9454.

[40]. Van Veen D, Jalal A, Soltanolkotabi M, Price E, Vishwanath S, and Dimakis AG, “Compressed sensing with deep image prior and learned regularization,” 2018, arXiv:1806.06438. [Online]. Available: http://arxiv.org/abs/1806.06438

[41]. Zbontar J et al., “fastMRI: An open dataset and benchmarks for accelerated MRI,” 2018, arXiv:1811.08839. [Online]. Available: http://arxiv.org/abs/1811.08839

[42]. Maallo AMS, Freud E, Liu TT, Patterson C, and Behrmann M, “Effects of unilateral cortical resection of the visual cortex on bilateral human white matter,” NeuroImage, vol. 207, 2. 2020, Art. no. 116345.

[43]. Jain AK, Fundamentals of Digital Image Processing. Upper Saddle River, NJ, USA: Prentice-Hall, 1989.

[44]. Huang Y, Preuhs A, Manhart M, Lauritsch G, and Maier A, “Data consistent CT reconstruction from insufficient data with learned prior images,” 2020, arXiv:2005.10034. [Online]. Available: http://arxiv.org/abs/2005.10034

[45]. Cheng K, Calivá F, Shah R, Han M, Majumdar S, and Pedoia V, “Addressing the false negative problem of deep learning MRI reconstruction models by adversarial attacks and robust training,” in Proc. Med. Imag. Deep Learn, 2020, pp. 121–135.

[46]. Wilson DW and Barrett HH, “Decomposition of images and objects into measurement and null components,” Opt. Exp, vol. 2, no. 6, pp. 254–260, 1998.

[47]. Halko N, Martinsson PG, and Tropp JA, “Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions,” SIAM Rev., vol. 53, no. 2, pp. 217–288, 1. 2011.

[48]. Kuo J, Granstedt J, Villa U, and Anastasio MA, “Learning a projection operator onto the null space of a linear imaging operator,” Proc. SPIE, vol. 11595, 2. 2021, Art. no. 115953X.

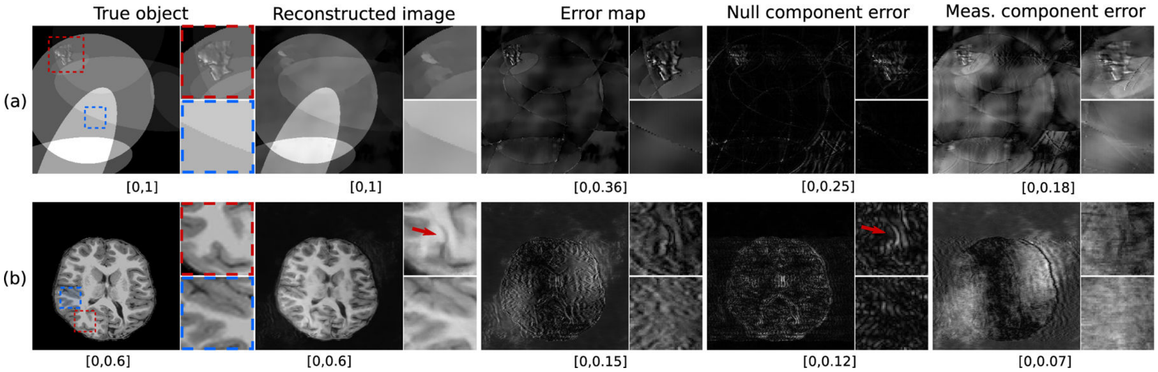


Fig. 1.

From left-to-right are examples of a true object, a reconstructed estimate of the object produced by use of a U-Net from tomographic measurements, the total error map, the error in the null component of the reconstructed object, and the error in the measurement component of the reconstructed object. The two rows correspond to different objects. In each case, the true object is outside the respective training data distribution of the U-Net and phase noise was added to the measurements prior to image reconstruction.

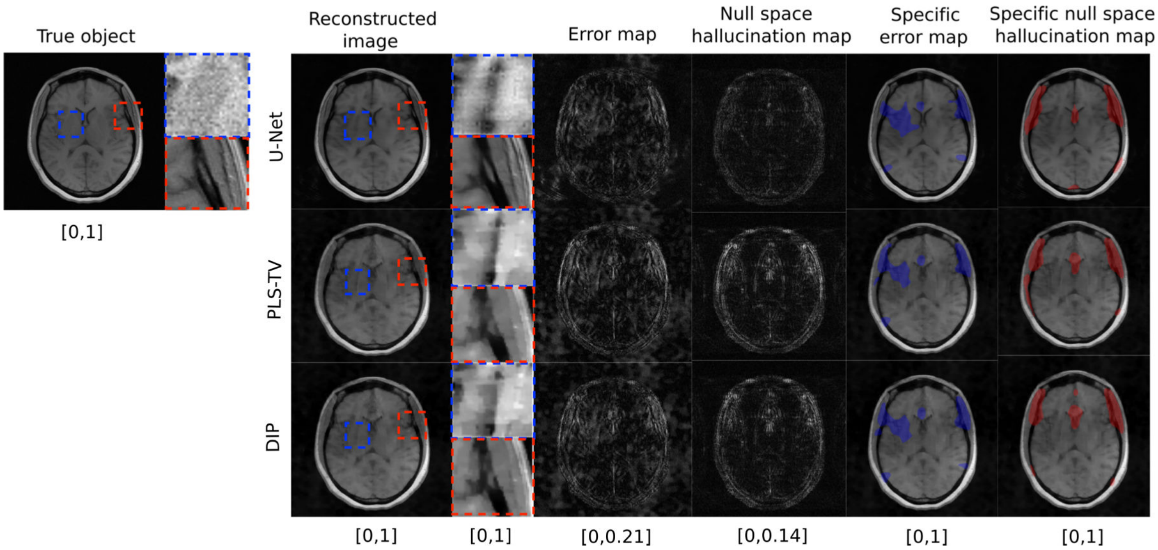


Fig. 2.

Example of a true object and reconstructed images along with error maps and hallucination maps (null space) for IND data with different reconstruction methods – U-Net (top), PLS-TV (middle) and DIP (bottom). Expanded regions are shown to the right of the reconstructed images. The specific error map (blue) and specific null space hallucinations map (red) are overlaid on the reconstructed images for each method. The image estimated by the U-Net method has visibly lower hallucinations in the null space compared to PLS-TV and DIP. The region within the red bounding box is one of the locations that contains hallucinations for all the reconstruction methods. In this region, the U-Net method shows mild hallucinations compared to PLS-TV and DIP. Fine structures in this region appear to be oversmoothed in the image estimates obtained by use of PLS-TV and DIP. A false structure is also shown (within the blue bounding box region) that appears for all the reconstruction methods due to the phase noise and not due to the imposed prior, and hence cannot be classified as a hallucination.

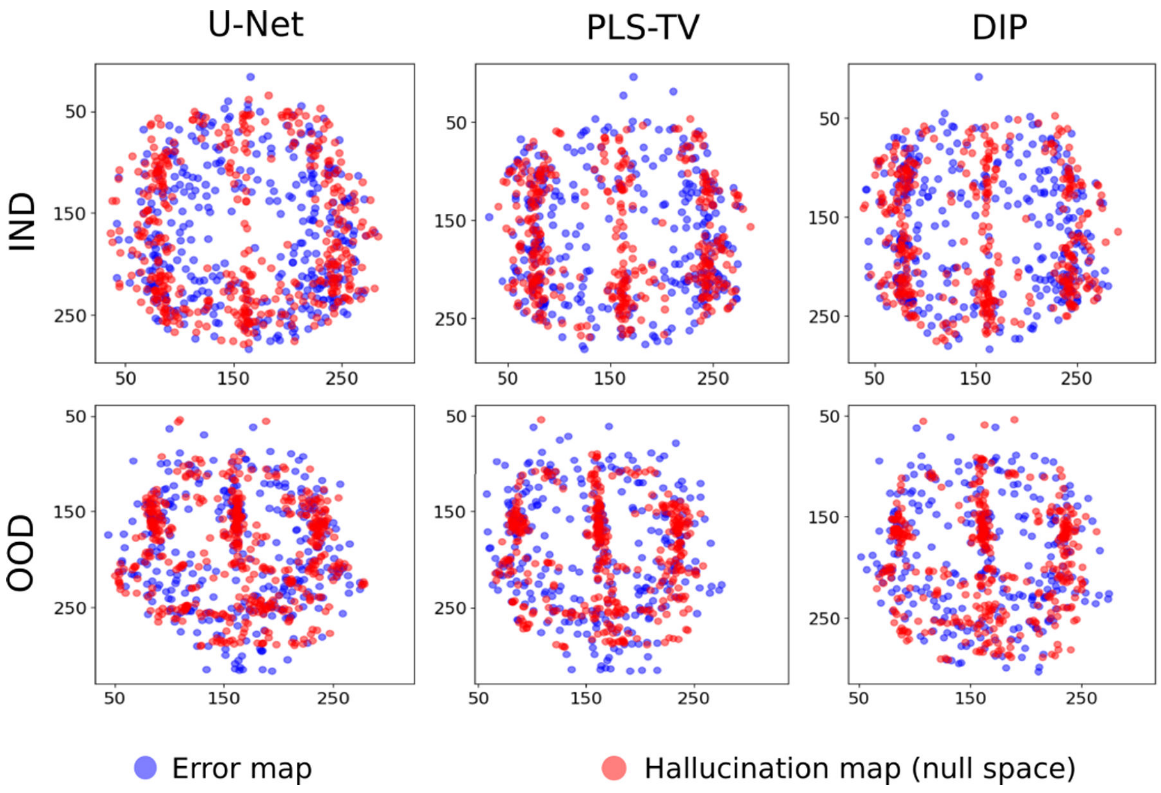


Fig. 3.

Scatter plots for centroids of localized regions in specific error maps and specific null space hallucination maps with different reconstruction methods for IND (top) and OOD (bottom) data. Note that for each type of data distribution and for all the reconstruction methods, the centroids of the regions detected from the error map have a higher variance compared to the hallucination map as well as some degree of non-overlap.

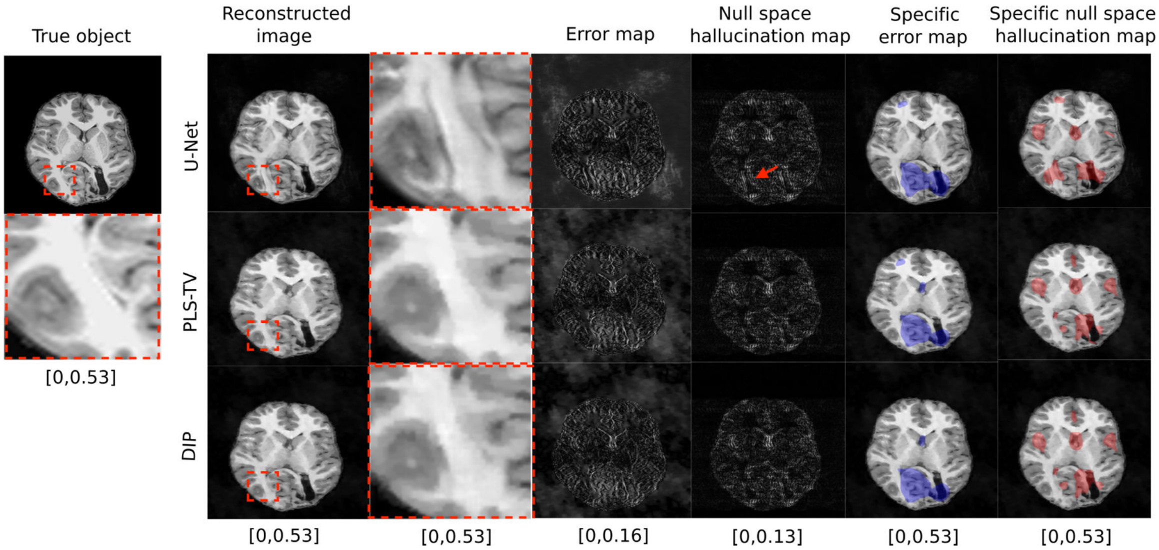


Fig. 4.

Example of true object and reconstructed images along with error map and hallucination maps (null space) for OOD data with different reconstruction methods – U-Net (top), PLS-TV (middle) and DIP (bottom). Expanded regions are shown to the right of the reconstructed images. The specific error map (blue) and specific null space hallucinations map (red) are overlaid on the reconstructed images for each method. The image estimated by the U-Net method has some distinct false structures (region within red bounding box) that do not exist in the reconstructed images obtained by using PLS-TV and DIP. This region is also highlighted in the specific null space hallucination map for the U-Net method which suggests that the false structure is a hallucination.

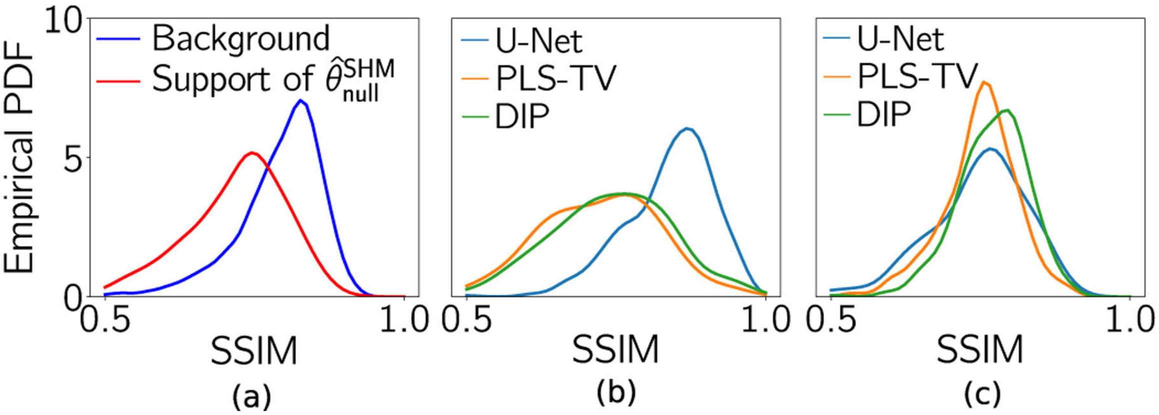


Fig. 5.

(a) Empirical PDF of SSIM values in the structured hallucination regions (support of $\hat{\theta}_{\text{null}}^{\text{SHM}}$) and the regions spanned by the remaining pixels in the support of the image (background), respectively, for the U-Net method with OOD data. (b) and (c) Empirical PDFs of SSIM values in the structured hallucination regions for all three reconstruction methods with IND and OOD data, respectively.

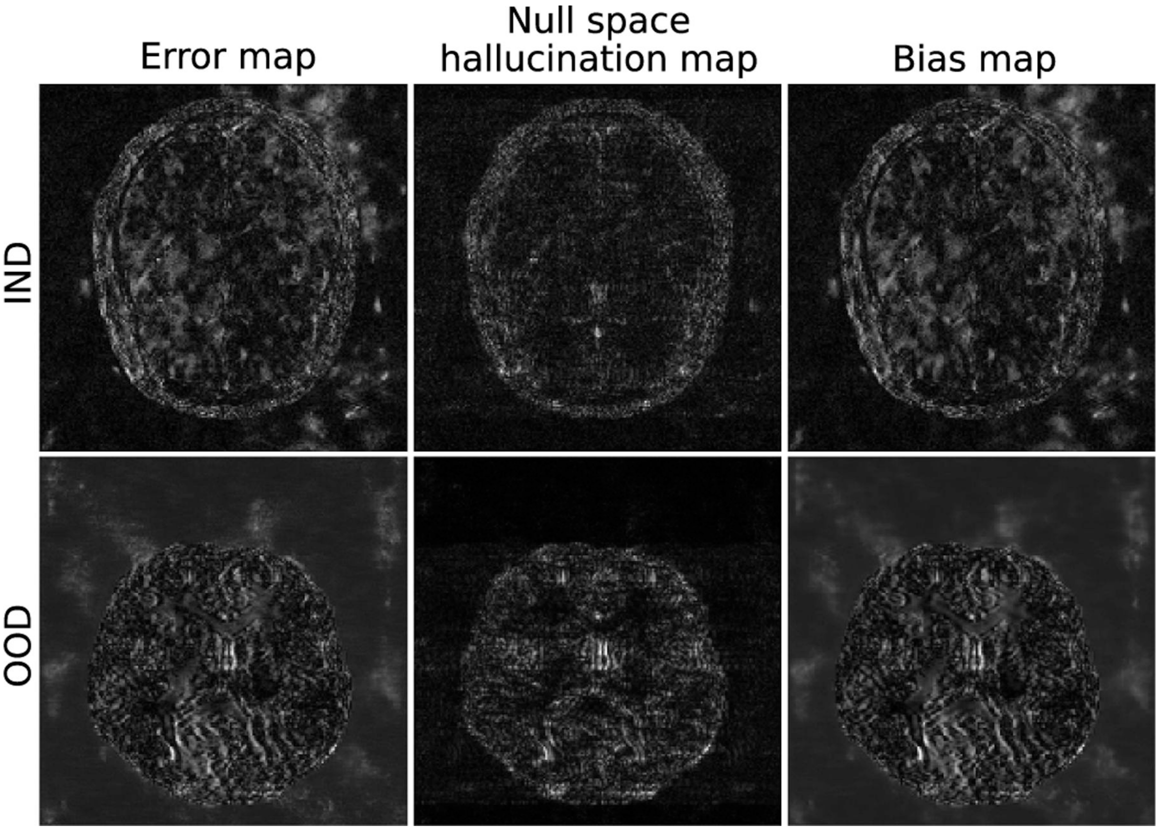


Fig. 6.

An error map, a null space hallucination map and a bias map for IND and OOD images estimated by use of the U-Net method. The corresponding true objects are shown in Figs. 2 and 1(b) respectively. The bias map was computed over a dataset of 100 images estimated from a single set of simulated measurements with fixed phase noise and different realizations of the iid Gaussian noise. The bias map contains contributions from both the model error, as well as inaccuracies in the prior.

TABLE I

Median of Ensemble SSIM Values in Support Region of Specific Null Space Hallucination Maps

Data distribution	U-Net	PLS-TV	DIP
IND	**0.84**	0.71	0.73
OOD	0.75	0.73	**0.76**