Published in final edited form as:

IEEE Signal Process Mag. 2022 March ; 39(2): 28–44. doi:10.1109/msp.2021.3119273.

Unsupervised Deep Learning Methods for Biological Image Reconstruction and Enhancement: An overview from a signal processing perspective

Mehmet Akçakaya [Senior Member, IEEE],
Department of Electrical and Computer Engineering, and Center for Magnetic Resonance Research, University of Minnesota, USA.

Burhaneddin Yaman [Student Member, IEEE],
Department of Electrical and Computer Engineering, and Center for Magnetic Resonance Research, University of Minnesota, USA.

Hyungjin Chung [Fellow, IEEE],
Department of Bio and Brain Engineering, Korea Advanced Inst. of Science and Technology (KAIST), Korea.

Jong Chul Ye [Fellow, IEEE]
Department of Bio and Brain Engineering, Korea Advanced Inst. of Science and Technology (KAIST), Korea.

Abstract

Recently, deep learning approaches have become the main research frontier for biological image reconstruction and enhancement problems thanks to their high performance, along with their ultra-fast inference times. However, due to the difficulty of obtaining matched reference data for supervised learning, there has been increasing interest in unsupervised learning approaches that do not need paired reference data. In particular, self-supervised learning and generative models have been successfully used for various biological imaging applications. In this paper, we overview these approaches from a coherent perspective in the context of classical inverse problems, and discuss their applications to biological imaging, including electron, fluorescence and deconvolution microscopy, optical diffraction tomography and functional neuroimaging.

Keywords

Deep learning; unsupervised learning; biological imaging; image reconstruction

I. Introduction

Biological imaging techniques, such as optical microscopy, electron microscopy, x-ray crystallography have become indispensable tools for modern biological discoveries. Here, an image sensor measurement $\boldsymbol{y} \in \mathcal{Y}$ from an underlying unknown image $\boldsymbol{x} \in \mathcal{X}$ is usually described by

$$\boldsymbol{y} = H(\boldsymbol{x}) + \boldsymbol{w}, \tag{1}$$

where $\boldsymbol{w}$ is the measurement noise and $H:\mathcal{X}\mapsto\mathcal{Y}$ is a potentially nonlinear forward mapping arising from the corresponding imaging physics. In practice, the resulting inverse problem to obtain $\boldsymbol{x}$ from the sensor measurement $\boldsymbol{y}$ is ill-posed. Over the past several decades, many tools have been developed to address such ill-posed inverse problems, among which a popular one is the regularized least squares (RLS) that employs regularization (or penalty) terms to stabilize the inverse solution:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} c(\boldsymbol{x},\boldsymbol{y}) + R(\boldsymbol{x}) \quad \text{where } c(\boldsymbol{x},\boldsymbol{y}) \triangleq \left\|\boldsymbol{y} - H(\boldsymbol{x})\right\|_2^2. \tag{2}$$

In this objective function, the regularization term $R(\cdot)$ is usually designed in a top-down manner using mathematical and engineering principles, such as sparsity, total variation, or entropy-based methods, among others.

Recently, deep learning (DL) approaches have become mainstream for inverse problems in biological imaging, owing to their excellent performance and ultra-fast inference time compared to RLS. Most DL approaches are trained in a supervised manner, with paired input and ground-truth data, which often leads to a straightforward training procedure. Unfortunately, matched label data are not available in many applications. This is particularly problematic with biological imaging problems, as the unknown image itself is intended for scientific investigation that was not possible by other means.

To address this problem, two types of approaches have gained interest: self-supervised learning and generative model-based approaches. Self-supervised learning aims to generate supervisory labels automatically from the data itself to solve some tasks, and has found applications in many machine learning applications [1]. For regression tasks, such as image reconstruction and denoising, this is typically achieved by a form of hold-out masking, where parts of the raw or image data are hidden from the network and used in defining the training labels. For image denoising, it was shown that this idea can be used to train a deep learning approach from single noisy images [2]. Furthermore, with an appropriate choice of the holdout mask, the self-supervised training loss was shown to be within an additive constant of the supervised training loss [3], providing a theoretical grounding for their success for denoising applications. For image reconstruction, the use of self-supervised learning was proposed in [4] for physics-guided neural networks that solve the RLS problem, showing comparable quality to supervised deep learning. In this case, the masking is performed in a data fidelity step, decoupling it from the regularization problem, and also facilitating the use of different loss functions in the sensor domain. Self-supervised learning techniques have been applied in numerous biological imaging applications, such as fluorescence microscopy [3], electron microscopy [2], [5], and functional neuroimaging [6].

Another class of unsupervised learning approaches are based on generative models [7], such as generative adversarial nets (GAN) that have attracted significant attention in the machine learning community by providing a way to generate a target data distribution from a random distribution. In the paper on *f*-GAN [8], the authors show that a general class of so-called *f*-GAN can be derived by minimizing the statistical distance in terms of *f*-divergence, and the original GAN is a special case of *f*-GAN, when the Jensen-Shannon divergence is used

as the statistical distance measure. Similarly, the so-called Wasserstein GAN (W-GAN) can be regarded as another statistical distance minimization approach, where the statistical distance is measured by Wasserstein-1 distance [7]. Inspired by these observations, cycle-consistent GAN (cycleGAN) [9], which imposes one-to-one correspondence to address the mode-collapsing behavior, was shown to be similarly obtained when the statistical distances in both measurement space and the image space can be simultaneously minimized [10]. The cycleGAN formulation has been applied for various biological imaging problems, such as deconvolution microscopy [11] and super-resolution microscopy [10], where the forward model is known or partially known.

Given the success of these unsupervised learning approaches, one of the fundamental questions is how these seemingly different approaches relate to each other and even to the classic inverse problem approaches. The main aim of this paper is therefore to offer a coherent perspective to understand this exciting area of research.

This paper is composed as follows. In Section II, classical approaches of biological image reconstruction problems and modern supervised learning approaches are introduced, and the need for unsupervised learning approaches in biological imaging applications is explained. Section III then overviews the self-supervised learning techniques, which is followed by generative model-based unsupervised learning approaches in Section IV. Section V discusses open problems in unsupervised learning methods, which is followed by conclusion in Section VI.

II. Background on Biological Image Reconstruction and Enhancement

A. Conventional solutions to the regularized least squares problem

The objective function of the RLS problem in Eq. (2) forms the basis of most conventional algorithms for inverse problems in biological imaging. As this objective function does not often have a closed form solution, especially when using compressibility-based regularizers, iterative algorithms are typically used.

For the generic form of the problem, where $H(\cdot)$ can be non-linear, gradient descent is a commonly used algorithm for solution:

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} - \eta_k \nabla_x c\left(\boldsymbol{x}^{(k-1)}, \boldsymbol{y}\right) - \eta_k \nabla_x R\left(\boldsymbol{x}^{(k-1)}\right), \tag{3}$$

where $\boldsymbol{x}^{(k)}$ is the solution at the k^{th} iteration, and η_k is the gradient step. While gradient descent remains popular, it requires taking the derivative of the regularization term, which may not be straightforward in a number of scenarios. Thus, alternative methods have been proposed for the types of objective function in Eq. (2), relying on the use of the so-called proximal operator associated with $R(\cdot)$. These methods encompass proximal gradient descent and its variants, and variable splitting methods, such as alternating direction method of multipliers and variable splitting with quadratic penalty. Among these, variable splitting approaches are popular due to their fast convergence rates and performance in a number of applications even with non-convex objective functions. In particular, variable splitting

approaches decouple the $c(\boldsymbol{x}, \boldsymbol{y})$ and $R(\boldsymbol{x})$ terms by introducing an auxiliary variable $\boldsymbol{z}$ constrained to be equal to $\boldsymbol{x}$, as:

$$\arg \min_{x,\, z} c(\boldsymbol{x}, \boldsymbol{y}) + R(z) \;\; \text{s.t.} \;\; \boldsymbol{x} = z \tag{4}$$

This constrained optimization problem can be solved in different ways, with the simplest being the introduction of a quadratic penalty that leads to the following alternating minimization:

$$z^{(k-1)} = \arg \min_{z} \mu \left\| \boldsymbol{x}^{(k-1)} - z \right\|^2 + R(z) \tag{5a}$$

$$\boldsymbol{x}^{(k)} = \arg \min_{\boldsymbol{x}} \left\| \boldsymbol{y} - H(\boldsymbol{x}) \right\|^2 + \mu \left\| \boldsymbol{x} - z^{(k-1)} \right\|^2 \tag{5b}$$

where $\boldsymbol{x}^{(0)} = -\eta \nabla_{\boldsymbol{x}} c(\boldsymbol{0}, \boldsymbol{y})$ can be initialized with a single gradient descent step on the data consistency term and $z^{(k)}$ is an intermediate optimization variable. The sub-problems in Eq. (5a) and (5b) correspond to a proximal operation and a data consistency step, respectively. While for generic $H(\cdot)$ and $R(\cdot)$, convergence cannot be guaranteed, under certain conditions, which are more relaxed for gradient descent, convergence can be established. Nonetheless, both gradient descent, and algorithms that utilize the alternating data consistency and proximal operation iteratively have found extensive use in inverse problems in biological imaging. Moreover, plug-and-play (PnP) and regularization by denoising (RED) approaches show that powerful denoisers can be used as a prior for achieving state-of-the-art performance for solving inverse problems, even if they do not necessarily have closed form expressions. Unfortunately, the main drawbacks of these methods include lengthy computation times due to their iterative nature, and sensitivity to hyper-parameter choices, which often limit their routine use in practice.

B. Deep learning based reconstruction and enhancement with supervised training

Deep learning (DL) methods have recently gained popularity as an alternative for estimating $\boldsymbol{x}$ from the measurement model in Eq. (1). In the broadest terms, these techniques learn a parametrized nonlinear function that maps the measurements to an image estimate. Early methods that utilized DL for reconstruction focused on directly outputting an image estimate from (a function of) the measurement data, $\boldsymbol{y}$, using a neural network. These DL methods, classified under image enhancement strategies, learn a function $F_{\theta_e}(\boldsymbol{y})$. In particular, the input to the neural network is $\boldsymbol{y}$ if the measurements are in image domain or a function of $\boldsymbol{y}$, such as the adjoint of $H(\cdot)$ applied to $\boldsymbol{y}$ for linear measurement systems, if the measurements are in a different sensor domain. The main distinctive feature of these enhancement-type methods is that $H(\cdot)$ is not explicitly used by the neural network, except potentially for generating the input to the neural network. As such, the neural network has to learn the whole inverse problem solution without the forward operator. While this leads to very fast runtime, these methods may face issues with generalizability especially when $H(\cdot)$ varies from one sample to another [12].

An alternative line of DL methods fall under the category of physics-guided or physics-driven methods. These methods aim to solve the objective function in Eq. (2) explicitly using $H(\cdot)$, and implicitly learning an improved regularization term $R(\cdot)$ through the use of neural networks. These methods rely on the concept of algorithm unrolling [12], where a conventional iterative algorithm for solving Eq. (2) is unrolled for a fixed number of iterations, K. For instance, for the variable splitting algorithm described in Eq. (5a)–(5b), the unrolled algorithm consists of an alternating cascade of K pairs of proximal and data consistency operations. In unrolled networks, the proximal operation in Eq. (5a) is implicitly implemented by a neural network, while the data consistency operation in Eq. (5b) is implemented by conventional methods that explicitly use $H(\cdot)$, such as gradient descent with the only learnable parameter being the gradient step size. These physics-guided methods have recently become the state-of-the-art in a number of image reconstruction problems, including large-scale medical imaging reconstruction challenges, largely due to their more interpretable nature and ability for improved generalization when faced with changes in the forward operator $H(\cdot)$ across samples [12]. Thus, the final unrolled network can be described by a function $F_{\theta_r}(\boldsymbol{y}; H)$ that explicitly incorporates the forward operator and is parametrized by $\boldsymbol{\theta}_r$.

For both of these deep learning approaches, supervised training, which utilizes pairs of input and ground-truth data, remains a popular approach for inverse problems in biological imaging. For a unified notation among enhancement and reconstruction approaches, we use $F_{\boldsymbol{\theta}}(\boldsymbol{y})$ to denote the network output for measurements $\boldsymbol{y}$. In supervised learning, the goal is to minimize a loss of the form

$$\min_{\theta} \mathbb{E}_{\boldsymbol{x}, \boldsymbol{y}} \mathscr{L}(\boldsymbol{x}, F_{\theta}(\boldsymbol{y})), \tag{6}$$

where $\mathscr{L}(\cdot, \cdot)$ is a loss function that quantitatively characterizes how well the neural network $F_{\boldsymbol{\theta}}(\cdot)$ predicts the ground truth data for the given input.

In practice, the mapping function in Eq. (6) is approximated by minimizing the empirical loss on a large database. Consider a database of N pairs of input and reference data, $\{\boldsymbol{y}^n, \boldsymbol{x}^n_{\text{ref}}\}_{n=1}^{N}$. Supervised learning approaches aim to learn the parameters $\boldsymbol{\theta}$ of the function $F_{\boldsymbol{\theta}}(\cdot)$. In particular, during training, $\boldsymbol{\theta}$ are adjusted to minimize the difference between the network output and the ground-truth reference. More formally, training is performed by minimizing

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} \mathscr{L}(\boldsymbol{x}^n_{\text{ref}}, F_{\theta}(\boldsymbol{y}^n)). \tag{7}$$

Note that the loss function does not need to be related to the negative log-likelihood, $c(\boldsymbol{x}, \boldsymbol{y})$ of the RLS problem given in Eq. (2). While the mean squared error (MSE) loss, $\frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{x}^n_{\text{ref}} - F_{\theta}(\boldsymbol{y}^n)\|^2$, remains popular, a variety of other loss functions such as $\mathscr{L}_1$, adversarial and perceptual losses are used for supervised deep learning approaches.

C. Motivation for unsupervised deep learning approaches

While supervised deep learning approaches outperform classical methods and provide state-of-the-art results in many settings, acquisition of reference ground-truth images are either challenging or infeasible in many biological applications.

For example, in transmission electron microscopy (TEM), acquired projections are inherently low-contrast. A common approach for high-contrast images is to acquire defocused images which in turn reduces the resolution. Moreover, in TEM, acquisition of the clean reference images are not feasible due to limited electron dose used during acquisition to avoid sample destruction. Similarly, in scanning electron microscopy (SEM), the lengthy acquisition times for imaging large volumes remains a main limitation. While it is desirable to speed up the acquisitions, such acceleration degrades the acquired image quality [5]. Fluorescence microscopy is commonly used for live-cell imaging, but the intense illumination and long exposure during imaging can lead to photobleaching and phototoxicity. Hence, safer live-cell imaging requires lower intensity and exposure. However, this causes noise amplification in the resulting images, rendering it impractical for analysis. These challenges are not unique to listed microscopy applications. In many other biological applications, such as optical diffraction tomography, functional magnetic resonance imaging or super resolution microscopy, such challenges exist in similar forms. Hence, unsupervised deep learning approaches are essential for addressing the training of deep learning reconstruction methods in biological imaging applications.

III. Self-supervised learning methods

A. Overview

Self-supervised learning encompasses a number of approaches, including colorization, geometric transformations, content encoding, hold-out masking and momentum contrast [1]. Among these methods, hold-out masking is the most commonly used strategy for regression-type problems, including image denoising and reconstruction. In these methods, parts of the image or raw measurement/sensor data are hidden from the neural network during training, and instead are used to automatically define supervisory training labels from the data itself. An overview of this strategy for denoising is shown in Fig. 1. While the masking idea is similar, there is a subtle difference between the denoising and reconstruction problems. In denoising, $H(\cdot)$ is the identity operator, thus all the pixels in the image are accessible, albeit in a noise-degraded state. This allows for a theoretical characterization of self-supervised learning loss with respect to the supervised learning loss, verifying the practicality of self-supervision. This has also led to attention for self-supervised denoising from the broader computer vision community. On the other hand, theoretical results have not been established for image reconstruction due to the incomplete nature of available data, yet reported empirical results from variety of DL algorithms, especially physics-guided ones incorporating the forward operator, show that it can achieve similar reconstruction quality as supervised learning algorithms. In order to capture these inherent differences between the two problems, we will next separately discuss self-supervised deep learning for denoising and reconstruction methods.

B. Self-supervised deep learning for denoising

1) Background on denoising using deep learning: Image denoising concerns a special case of the acquisition model in Eq. (1), where $H(\cdot)$ is the identity operator. In this case, the objective function for the inverse problem in Eq. (2) becomes $\arg\min_x \|\boldsymbol{y}-\boldsymbol{x}\|_2^2 + R(\boldsymbol{x})$. In deep learning methods for denoising, this proximal operation is replaced by a neural network, which estimates a denoised image $\hat{\boldsymbol{x}}_{\text{denoised}} = F_{\theta_d}(\boldsymbol{y})$ through a $\boldsymbol{\theta}_d$-parametrized function. While supervised deep learning methods provide state-of-the-art results for denoising applications, absence of clean target images render the supervised approaches inoperative for a number of biological imaging problems as discussed earlier.

Noise2Noise (N2N) was among the first works that tackled this challenge, where a neural network was trained on pairs of noisy images and yielded results on par with their supervised counterparts. Given pairs of noisy images arising from the same clean target image each with its own i.i.d. zero-mean random noise components ($\boldsymbol{y} = \boldsymbol{x} + w, \hat{\boldsymbol{y}} = \boldsymbol{x} + \hat{w}$), N2N aims to minimize an MSE loss of the form

$$\min_{\theta_d} \mathbb{E}_{\hat{\boldsymbol{y}},\boldsymbol{y}} \left\| F_{\theta_d}(\boldsymbol{y}) - \hat{\boldsymbol{y}} \right\|^2 = \min_{\theta_d} \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\| F_{\theta_d}(\boldsymbol{y}) - \boldsymbol{x} \right\|^2 + \mathbb{E}_{\hat{w}} \left\| \hat{w} \right\|^2 - 2\mathbb{E}\left\langle \hat{w}, F_{\theta_d}(\boldsymbol{y}) - \boldsymbol{x} \right\rangle \tag{8}$$

$$= \min_{\theta_d} \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\| F_{\theta_d}(\boldsymbol{y}) - \boldsymbol{x} \right\|^2 + \mathbb{E}_{\hat{w}} \left\| \hat{w} \right\|^2, \tag{9}$$

where the last term in Eq. (8) becomes zero since $\mathbb{E}\hat{w} = 0$. Note that the last term in Eq. (9) does not depend on $\boldsymbol{\theta}_d$. Hence, the $\theta_d^\star$ that minimize the N2N loss, $\mathbb{E}_{\boldsymbol{x},\boldsymbol{y},\hat{w}} \left\| F_{\theta_d}(\boldsymbol{y}) - (\boldsymbol{x} + \hat{w}) \right\|^2$, is also a minimizer of the supervised loss $\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\| F_{\theta_d}(\boldsymbol{y}) - \boldsymbol{x} \right\|^2$. We note that different loss functions such as L_1 loss can also be used with N2N [13].

In practice, training is performed by minimizing empirical loss on a database with N pairs of noisy images $\left\{ \boldsymbol{y}^n = \boldsymbol{x}^n + w^n, \hat{\boldsymbol{y}}^n = \boldsymbol{x}^n + \hat{w}^n \right\}_{n=1}^N$. N2N trains a neural network for denoising by minimizing

$$\min_{\theta_d} \sum_{n=1}^{N} \left\| F_{\theta_d}(\boldsymbol{y}^n) - \hat{\boldsymbol{y}}^n \right\|^2. \tag{10}$$

The key assumption of N2N is that the expected value of the noisy image pairs are equivalent to the clean target image. While N2N eliminates the need for acquiring noisy/clean pairs used for supervised training, which is either challenging or impossible in most applications, the N2N requirement for pairs of noisy measurements may nonetheless be infeasible in some biological applications.

2) Self-supervised training for deep learning-based denoising: Self-supervised learning methods for image denoising build on the intuitions from the N2N strategy, while

enabling training from single noisy measurements in the absence of clean or paired noisy images. Following the N2N strategy, the self-supervised loss can be generally stated as

$$\min_{\theta_d} \mathbb{E}_{\boldsymbol{y}} \left\| F_{\theta_d}(\boldsymbol{y}) - \boldsymbol{y} \right\|^2 . \tag{11}$$

However, the naive application of Eq. (11) leads to the denoising function F_{θ_d} to be identity.

Noise2Void (N2V) was the first work to propose the use of masking to train such a neural network. Concurrently, Noise2Self (N2S) proposed the idea of $\mathcal{J}$-invariance to theoretically characterize how the function F_{θ_d} can be learned without collapsing to the identity function. To this end, consider an image with m pixels, and define a partition (or index set) of an image as $J \subseteq \{1, \ldots, m\}$. Further, let $\boldsymbol{x}_J$ denote the pixel values of the image on the partition defined by J. With this notation, $\mathcal{J}$-invariance was defined as follows [3]: For a given set of partitions of an image $\mathcal{J} = \{J_1, \ldots, J_N\}$, where $\sum_{i=1}^{N} |J_i| = m$, a function $F_{\theta_d}: \mathbb{R}^m \to \mathbb{R}^m$ is $\mathcal{J}$-invariant if the value of $F_{\theta_d}(\boldsymbol{y})J$ does not depend on the value of y_J for all $J \in \mathcal{J}$. In essence, the pixels of an image are split into two disjoint sets J and J^c with $|J|+|J|^c = m$, and $\mathcal{J}$-invariant denoising function $F_{\theta_d}(\boldsymbol{y})J$ uses pixels in $\boldsymbol{y}_{J^c}$ to predict a denoised version of $\boldsymbol{y}_J$. The objective self-supervised loss function over J-invariant functions can be written as [3]

$$\begin{aligned} \mathbb{E}_{\boldsymbol{y}} \left\| F_{\theta_d}(\mathbf{y}) - \boldsymbol{y} \right\|^2 &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\| F_{\theta_d}(\mathbf{y}) - \boldsymbol{x} \right\|^2 + \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\| \boldsymbol{y} - \boldsymbol{x} \right\|^2 \\ &\quad - 2\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\langle F_{\theta_d}(\mathbf{y}) - \boldsymbol{y}, \boldsymbol{y} - \boldsymbol{x} \right\rangle \end{aligned} \tag{12}$$

$$\begin{aligned} &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\| F_{\theta_d}(\mathbf{y}) - \boldsymbol{x} \right\|^2 + \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\| \boldsymbol{y} - \boldsymbol{x} \right\|^2 \\ &\quad - 2\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{y}\,|\,\boldsymbol{x}} \left\langle F_{\boldsymbol{\theta}_d}(\mathbf{y}) - \boldsymbol{y}, \boldsymbol{y} - \boldsymbol{x} \right\rangle \end{aligned} \tag{13}$$

$$= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\| F_{\theta_d}(\mathbf{y}) - \boldsymbol{x} \right\|^2 + \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\| \boldsymbol{y} - \boldsymbol{x} \right\|^2 . \tag{14}$$

Note that for each pixel j in Eq. (13), the random variables $F_{\theta_d}(\mathbf{y})_j \mid \boldsymbol{x}$ and $\boldsymbol{y}_j|\boldsymbol{x}$ are independent if F_{θ_d} is $\mathcal{J}$-invariant, while the noise is zero-mean by assumption. Hence, the third term in Eq. (13) vanishes. Eq. (14) shows that minimizing a self-supervised loss function over $\mathcal{J}$-invariant functions is equivalent to minimizing a supervised loss up to a constant term (variance of the noise). Thus, self-supervised denoising approaches learns a $\mathcal{J}$-invariant denoising function F_{θ_d} over a database of single noisy images by minimizing the self-supervised loss

$$\arg\min_{\theta_d} \sum_{n=1}^{N} \sum_{J \in \mathcal{J}} \left\| F_{\theta_d}\left(\mathbf{y}_{J^c}^n\right) - \boldsymbol{y}_J^n \right\|^2 . \tag{15}$$

Implementation-wise, it is not straightforward to just set the pixels specified by J to zero, since this will affect the way convolutions will be computed. Thus, during training of self-supervised techniques such as N2V or N2S, the network takes $\boldsymbol{y}_J{}^c = \mathbf{1}_J{}^c \boldsymbol{y} + \mathbf{1}_J \kappa(\boldsymbol{y})$ as input [3], where $\kappa(\cdot)$ is a function assigning new values to masked pixel locations, J. The new pixel values in J indices of the network input are either a result of a local averaging filter that excludes the center, or random values drawn from a uniform random distribution [3]. In the former case, $\mathcal{J}$-invariance can be achieved by using a uniform grid structure for the masks J, where the spacing is determined by the kernel size of the averaging filter, while for the latter case, a uniform random selection of J may suffice [3].

At inference time, two approaches can be adapted: 1) inputting the full noisy image on the trained network, 2) inputting a partition $\mathcal{J}$ containing $|\mathcal{J}|$ sets and averaging them.

C. Self-supervised learning for image reconstruction

Self-supervised learning for image reconstruction neural networks provides a method for training without paired measurement and reference data. One important line of work entails a method called self-supervised learning via data undersampling (SSDU) [4], which generalizes the hold-out masking of Section III–B2 for physics-guided image reconstruction.

For m-dimensional $\boldsymbol{y}$, consider an index set $\Theta \subseteq \{1, \ldots, m\}$ of all the available measurement coordinates. In physics-guided DL reconstruction, the measurements interact with the neural network through the data consistency operations. To this end, let $H_\Theta(\cdot)$ be the operator that outputs the measurement coordinates corresponding to the index set Θ. In SSDU, hold-out masking is applied through these data consistency operations. Thus, while the index set Θ is used in the data consistency units of the unrolled network, the loss itself is calculated in the sensor domain on the indices specified by Θ^C [4]. Hence, SSDU minimizes the following self-supervised loss

$$\min_{\theta_r} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}\left(\boldsymbol{y}^n_{\Theta^C}, H^n_{\Theta^C}\left(F_{\theta_r}\left(\boldsymbol{y}^n_\Theta, H^n_\Theta\right)\right)\right), \tag{16}$$

where the output of the network is transformed back to the measurement domain by applying the forward operator $H^n_{\Theta^C}$ at corresponding unseen locations in the training, Θ^C. An overview of this strategy is given in Fig. 2.

Note that unlike in the denoising scenario, the measurements for reconstruction can be in different sensor domains, and thus the training algorithm does not have access to all the pixels of the image. Thus, the concept of $\mathcal{J}$-invariance is not applicable in this setting. Therefore, from a practical perspective, Θ is chosen randomly. In [4], which focused on a Fourier-based sensor domain, a variable density masking approach based on Gaussian probability densities was chosen. This inherently enabled a denser sampling of the low-frequency content in Fourier space, which contain most of the energy for images, for use in the data consistency units. However, a Gaussian density for masking requires a hyper-parameter controlling its variance. Thus, in later works, SSDU was extended to a multi-mask setting [14], where multiple index sets $\{\Theta_l\}_{l=1}^{L}$ were used to define the loss

$$\min_{\theta_r} \frac{1}{N} \sum_{n=1}^{N} \sum_{l=1}^{L} \mathscr{L}\left(\boldsymbol{y}^n_{\Theta_l^C}, H^n_{\Theta_l^C}\left(F_{\theta_r}\left(\boldsymbol{y}^n_{\Theta_l}; H^n_{\Theta_l}\right)\right)\right). \tag{17}$$

When utilizing multiple hold-out masks for the data consistency units, uniform random selection of the masks becomes a natural choice, also eliminating the need for an additional hyper-parameter. Furthermore, the use of multiple $\{\Theta_l\}_{l=1}^{L}$ also leads to an improved performance, especially as $H(\cdot)$ becomes increasingly ill-posed [14]. During inference time, SSDU-trained reconstruction uses all available m measurements in $\boldsymbol{y}$ in the data consistency units for maximal performance [4].

Note that because the masking happens in the data consistency term, the implementation is simplified to removing the relevant indices of the measurements for the data consistency components, and does not require a modification of the regularization neural network component or its input, unlike in the denoising scenario. This also enables a broader range of options for the loss $\mathscr{L}$. While the negative log-likelihood, $c(\boldsymbol{x}, \boldsymbol{y})$ of the RLS problem is an option, more advanced losses that better capture relevant features have been used [4].

Apart from the hold-out masking strategy discussed here, there is a line of work that performs self-supervision using a strategy akin to that described in Eq. (11), where all the measurements are used in the network and for defining the loss [15]. More formally, such approaches aim to minimize a loss function of the form

$$\min_{\theta_e} \frac{1}{N} \sum_{n=1}^{N} \mathscr{L}\left(\boldsymbol{y}^n, H^n\left(F_{\theta_e}\left(\boldsymbol{y}^n; H^n\right)\right)\right). \tag{18}$$

We note that $\boldsymbol{y}$ denotes all the acquired measurements and H transforms the network output $F_{\theta_e}(\cdot)$ to sensor domain. However, the performance of such naive application of self-supervised learning approaches suffers from noise amplification due to overfitting [4].

D. Biological Applications

1) Denoising: Even though N2N requires two independent noisy realizations of the target image for unsupervised training, which may be hard to meet in general, it has been applied to light and electron microscopy under Gaussian or Poisson noise scenarios. In cryo-TEM, the acquired datasets are inherently noisy, since the electron dose is restricted to avoid sample destruction [5]. Cryo-CARE [5] was the first work to show that the N2N can be applied to cryo-TEM data for denoising. Cryo-CARE was further applied on 3D cryo-electron tomogram (cryo-ET) data showing its ability to denoise whole tomographic volumes. Several other works have also extended N2N for denoising cryo-EM data.

N2V was the first work showing the denoising can be performed from single noisy measurements. N2V has been extensively applied to EM datasets showing improved reconstruction quality compared to conventional blind denoising methods such as BM3D [2]. In follow-up works, Bayesian post-processing has been used to incorporate pixel-wise Gaussian or histogram-based noise models [16] for further improvements in the denoising

performance. However, their application is limited as it requires the knowledge of the noise model, which might be challenging to know as a prior in number of applications. Moreover, the noise could be a mixture of noise type hence further hindering their applications. A follow-up work on [16] show that the prior noise model knowledge requirement in probabilistic N2V models can be tackled by learning the noise model directly from the noisy image itself via bootstrapping [17]. Another extension of this method, called structured N2V, was also proposed to mask a larger area rather than a single pixel for removing structured noise in microscopy applications. Similarly, Noise2Self and its variants have also been applied to various microscopy datasets [3].

Fig. 3 shows denoising results using a conventional denoising algorithm BM3D, and self-supervised learning algorithm Noise2Self on two different microscopy datasets. These datasets contain only single noisy images, hence supervised deep learning and N2N can not be applied. Results show that self-supervised learning approaches visually improve the denoising performance compared to conventional denoising algorithms.

2) Reconstruction: DL-based ground-truth free reconstruction strategies has been applied in variety of medical imaging applications. SSDU was one of the first self-supervised methods to be applied for physics-guided medical imaging reconstruction in MRI [4]. Concurrently, there were approaches inspired by N2N that was used in non-Cartesian MRI [18], where pairs of undersampled measurements were used for training. Similar to the denoising scenario, a main limitation of these methods is the requirement of pairs of measurements, which may be challenging in some imaging applications. Furthermore, the naive self-supervised learning strategy of Eq. (18) was also used for MRI reconstruction, by using all acquired measurements for both input to the network and defining the loss [15]. However, this approach suffered from noise amplification, as expected.

While such self-supervised methods have found use in medical imaging, their utility in biological imaging are just being explored. Recent work has started using such self-supervised deep learning methods to functional MRI, which remains a critical biological imaging tool for neuroscientific discoveries that expand our understanding of human perception and cognition. In a recent work [6], multi-mask SSDU was applied to a Human Connectome Project style fMRI acquisition that was prospectively accelerated by 5-fold simultaneous multi-slice imaging and 2-fold in-plane undersampling. Note that ground-truth data for such high spatiotemporal resolution acquisitions cannot be acquired in practice, thus prohibiting the use of supervised learning. The results shown in Fig. 4 indicate that the self-supervised deep learning method based on multi-mask SSDU significantly outperforms the conventional reconstruction approaches, both qualitatively in terms of visual quality, and quantitatively in terms of temporal signal-to-noise ratio.

IV. Generative model-based methods

A. Overview

Generative models cover a large spectrum of research activities, which include variational autoencoder (VAE), generative adversarial network (GAN), normalizing flow, optimal transport (OT), among others [7]. Due to their popularity, there are so many variations, so

one of the main goals of this section is to provide a coherent geometric picture of generative models.

Specifically, our unified geometric view starts from Fig. 5. Here, the ambient image space is $\mathcal{X}$, where we can take samples with the real data distribution μ. If the latent space is $\mathcal{Z}$, the generator G can be treated as a mapping from the latent space to the ambient space, $G: \mathcal{Z} \mapsto \mathcal{X}$, often realized by a deep network with parameter $\boldsymbol{\theta}$, i.e. $G \triangleq G_\theta$. Let ζ be a fixed distribution on the latent space, such as uniform or Gaussian distribution. The generator $G_{\boldsymbol{\theta}}$ pushes forward ζ to a distribution $\mu_{\boldsymbol{\theta}} = G_{\boldsymbol{\theta}\#}\zeta$ in the ambient space $\mathcal{X}$. Then, the goal of the generative model training is to make $\mu_{\boldsymbol{\theta}}$ as close as possible to the real data distribution μ. Additionally, for the case of auto-encoding type generative models (e.g. VAE), the generator works as a decoder $G_\theta: \mathcal{Z} \mapsto \mathcal{X}$, while another neural network-encoder $F_\phi: \mathcal{X} \mapsto \mathcal{Z}$ maps from sample space to the latent space. Accordingly, the additional constraint is again to minimize the distance $d(\zeta_\varphi, \zeta)$.

Using this unified geometric model, we can show that various types of generative models only differ in their choices of distances between $\mu_{\boldsymbol{\theta}}$ and μ, or ζ_φ and ζ and how to train the generator and encoder to minimize the distances.

B. VAE approaches for unsupervised learning in biological imaging

1) Variational autoencoder (VAE): In VAE, the generative model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is considered as a marginalization of the conditional distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$, combined with simple latent distribution $p(\boldsymbol{z})$ [7]:

$$\log p_\theta(\boldsymbol{x}) = \log\left(\int p_\theta(\boldsymbol{x} \mid z) p(z) dz\right). \tag{19}$$

The most straightforward way to train the network is to apply maximum likelihood on $p_{\boldsymbol{\theta}}(\boldsymbol{x})$. However, since the integral inside (19) is intractable, one can introduce a distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ such that

$$\begin{aligned} \log\, p_\theta(\boldsymbol{x}) &= \log\left(\int p_\theta(\boldsymbol{x} \mid z) \frac{p(z)}{q_\phi(z \mid \boldsymbol{x})} q_\phi(z \mid \boldsymbol{x}) dz\right) \\ &\geq \int \log\left(p_\theta(\boldsymbol{x} \mid z) \frac{p(z)}{q_\phi(z \mid \boldsymbol{x})}\right) q_\phi(z \mid \boldsymbol{x}) dz \\ &= \int \log\, p_\theta(\boldsymbol{x} \mid z) q_\phi(z \mid \boldsymbol{x}) dz - D_{KL}\big(q_\phi(z \mid \boldsymbol{x}) \| p(z)\big), \end{aligned} \tag{20}$$

where D_{KL} is the Kullback–Leibler divergence (KL) divergence, and the first inequality comes from Jensen's inequality. The final term in (20) is called evidence lower bound (ELBO), or variational lower bound in the context of variational inference. While infeasible to perform maximum likelihood on $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ directly, we can maximize the ELBO.

In the VAE, by using the reparametrization trick together with the Gaussian assumption, one has:

$$z = F_{\phi}^{\boldsymbol{x}}(\boldsymbol{u}) = \mu_{\phi}(\boldsymbol{x}) + \sigma_{\phi}(\boldsymbol{x}) \odot \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{21}$$

Where $F_{\phi}^{\boldsymbol{x}}(\boldsymbol{u})$ refers to the encoder function for a given image $\boldsymbol{x}$ which has another noisy input $\boldsymbol{u}$, and $\odot$ denotes the element-wise multiplication. Note that (21) enables back-propagation. Incorporating (21) with (20) gives us the loss function to minimize for an end-to-end training of the VAE:

$$\begin{aligned} &\ell_{VAE}(\theta, \phi) \\ &= \frac{1}{2} \int_{\mathcal{X}} \int \left\| \boldsymbol{x} - G_{\theta}\left(\mu_{\phi}(\boldsymbol{x}) + \sigma_{\phi}(\boldsymbol{x}) \odot \boldsymbol{u}\right) \right\|^2 r(\boldsymbol{u}) d\boldsymbol{u} d\mu(\boldsymbol{x}) \\ &+ \frac{1}{2} \sum_{i=1}^{d} \int_{\mathcal{X}} \left(\sigma_i^2(\boldsymbol{x}) + \mu_i^2(\boldsymbol{x}) - \log \sigma_i^2(\boldsymbol{x}) - 1\right) d\mu(\boldsymbol{x}) . \end{aligned} \tag{22}$$

Here, the first term in (22) can be conceived as the reconstruction loss ($d(\mu, \mu_{\theta})$ in Fig. 5), and the second term is originated from KL divergence can be interpreted as penalty-imposing term ($d(\zeta, \zeta_{\phi})$ in Fig. 5).

Once the network is trained by minimizing (22), one notable advantage of VAE is that we can generate samples from $p_{\theta}(\boldsymbol{x}|\boldsymbol{z})$ simply by sampling different noise vectors $\boldsymbol{u}$. Specifically, the decoder has explicit dependency on $\boldsymbol{u}$, and the model output is expressed as

$$\hat{\boldsymbol{x}}(\boldsymbol{u}) = G_{\theta}\left(\mu_{\phi}(\boldsymbol{x}) + \sigma_{\phi}(\boldsymbol{x}) \odot \boldsymbol{u}\right), \ \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) . \tag{23}$$

Notably, we can utilize (23) to sample multiple reconstructions by simply sampling different values of $\boldsymbol{u}$. Naturally, this method has been applied to many different fields, and in the following we review its biological image applications.

2) Biological Applications: One notable application of VAE in the field of biological imaging is Bepler *et al.* [19]. The work is motivated by the problem of modeling continuous 2D views of proteins from single particle electron microscopy (EM). The goal of EM imaging is to estimate 3D electron density of a given protein from multiple random noisy 2D projections. The first step in this process requires estimation of the conformational states, often modeled with Gaussian mixture model, which is discrete. Subsequently, modeling with Gaussian mixture models produces sub-optimal performance when aiming to model protein conformations. Hence, to bridge this gap, Bepler *et al.* [19] propose spatial-VAE to disentangle projection rotation and translation from the content of the projections.

Specifically, spatial-VAE [19] uses spatial generator network, first introduced in compositional pattern producing networks (CPPNs), where the generator G takes in as input the spatial coordinates, and outputs a pixel value. Moreover, as shown in Fig. 6(b), latent variable $\boldsymbol{z}$ is concatenated with additional parameters $\boldsymbol{\varphi}$, $\Delta\boldsymbol{t}$, representing rotation, and translation, respectively. More precisely, the conditional distribution is given as

$$\log p(\boldsymbol{x} \mid z) = \log p_{\theta}(\boldsymbol{x} \mid z, \varphi, \Delta\boldsymbol{t}) \tag{24}$$

$$= \sum_{i=1}^{n} \log\, p_\theta\left(\boldsymbol{x}^i \mid \boldsymbol{t}^i R(\varphi) + \Delta\boldsymbol{t}, z\right), \tag{25}$$

where $R(\varphi) = [\cos\,\varphi, -\sin\,\varphi; \sin\,\varphi, \cos\,\varphi]$ is the rotation matrix, and n is the dimensionality of the image. It is straightforward to extend the encoder function to output disentangled representations, which is given as

$$F_\phi^{\boldsymbol{x}}(\boldsymbol{u}) = \begin{bmatrix} \mu_z(\boldsymbol{x}) \\ \mu_\varphi(\boldsymbol{x}) \\ \mu_{\Delta\boldsymbol{t}}(\boldsymbol{x}) \end{bmatrix} + \begin{bmatrix} \sigma_z(\boldsymbol{x}) \\ s_\varphi \sigma_\varphi(\boldsymbol{x}) \\ s_{\Delta\boldsymbol{t}} \sigma_{\Delta\boldsymbol{t}}(\boldsymbol{x}) \end{bmatrix} \odot \boldsymbol{u}, \tag{26}$$

where s_φ, $s_{\Delta t}$ are chosen differently for each problem set. (26) shows that Gaussian priors are used for all the different parameters. Notably, by constructing spatial-VAE as given in (24), (26), translation and rotation are successfully disentangled from other features. Consequently, continuous modeling of parameter estimation in the particle projections of EM via spatial-VAE may substantially improve the final reconstruction of 3D protein structure.

Another recent yet important work, dubbed DIVNOISING, utilizes a modified VAE for denoising microscopy images [20]. As illustrated in Fig. 6(c), DIVNOISING tries to estimate the posterior $p(\boldsymbol{x}|\boldsymbol{y}) \propto p_{NM}(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$, where $\boldsymbol{x}$ is the true signal, $\boldsymbol{y}$ is the noise-corrupted version of $\boldsymbol{x}$, $p(\boldsymbol{x})$ is the prior, and $p_{NM}(\boldsymbol{y}|\boldsymbol{x})$ is the noise model, which is typically decomposed into a product of independent pixel-wise noise models. Note that the input image $\boldsymbol{y}$ is not a *clean* image, as in the other works. Instead, the encoder of DIVNOISING takes in a noisy image $\boldsymbol{y}$ to produce the latent vector $\boldsymbol{z}$. In this VAE setup, one can replace the conditional distribution $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ with a *known* noise model in case we know the corruption process, or *learnable* noise model in case we do not know the corruption process, and unsupervised training is required. With this modification, one can perform semi-supervised training in which the noise model is measured from paired calibration images, or bootstrapped from the noisy image. More interestingly, it is also possible to perform *unsupervised* training with a modification to the decoder. Once the VAE of DIVNOISING is trained, one can perform inference by varying the samples $\boldsymbol{u}$, and acquire multiple estimation of denoised images. When the user wants to acquire a point estimate of the distribution, one can either choose the mean (i.e. MMSE) of the sampled images, or get *maximum a posteriori* (MAP) estimate by iteratively applying mean shift clustering to the sampled images.

C. GAN approaches for unsupervised learning in biological imaging

1) Statistical Distance Minimization: In GAN, the generator G, and the discriminator D, play a minimax game, complementing each other at every optimization step. Formally, the optimization process is defined as:

$$\min_G \max_D \mathscr{L}_{GAN}(D, G), \tag{27}$$

where

$$\mathscr{L}_{GAN}(D,G) \triangleq \mathbb{E}_{\boldsymbol{x}}[\log\, D(\boldsymbol{x})] + \mathbb{E}_{z}[\log\,(1 - D(G(z)))]\,. \tag{28}$$

Here, $D(\boldsymbol{x})$ is called as the discriminator, which outputs a scalar in [0, 1] representing the probability of the input $\boldsymbol{x}$ being a real sample. While the discriminator struggles to learn the classification task, the generator tries to maximize the probability of D making a mistake. i.e. generating samples closer and closer to the actual distribution of $\boldsymbol{x}$.

To understand the geometric meaning of GAN, we first provide a brief review of f-GAN [8]. As the name suggests, f-GAN starts with f-divergence as the statistical distance measure:

$$D_f(\mu\|\nu) = \int_{\Omega} f\left(\frac{d\mu}{d\nu}\right) d\nu \tag{29}$$

where μ and ν are two statistical measures and μ is absolutely continuous with respect to ν. The key observation is that instead of directly minimizing the f-divergence, a very interesting thing emerges if we formulate its dual problem. In fact, the "dualization" trick is a common idea in generative models. More specifically, if f is a convex function, the convex conjugate of its convex conjugate is the function itself, i.e.

$$f(u) = f^{**}(u) = \sup_{\tau \in I^*} \{u\tau - f^*(\tau)\} \tag{30}$$

If $f^*: I^* \mapsto \mathbb{R}$. Using this, for any class of functions $\mathscr{T}$ mapping from $\mathscr{X}$ to $\mathbb{R}$, we have the lower bound

$$D_f(\mu\|\nu) \geq \sup_{\tau \in I^*} \int_{\mathscr{X}} \tau(\boldsymbol{x}) d\mu(\boldsymbol{x}) - \int_{\mathscr{X}} f^*(\tau(\boldsymbol{x})) d\nu(\boldsymbol{x}) \tag{31}$$

Where $f^*: I^* \mapsto \mathbb{R}$ is the convex conjugate of f. Using the following transform [8]

$$\tau(\boldsymbol{x}) = g_f(V(\boldsymbol{x})) \tag{32}$$

where $V: \mathscr{X} \mapsto \mathbb{R}$ without any constraint on the output range, and $g_f: \mathbb{R} \mapsto I^*$ is an *output activation function* that maps the output to the domain of f^*, f-GAN can be formulated as follows:

$$\min_{G} \max_{g_f} \mathscr{L}_{fGAN}(G, g_f) \tag{33}$$

where

$$\mathscr{L}_{fGAN}(G, g_f) \triangleq \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[g_f(V(\boldsymbol{x}))\right] - \mathbb{E}_{z\sim\zeta}\left[f^*\left(g_f(V(G(z)))\right)\right]. \tag{34}$$

Here, different choices of the functions f, g_f lead to distinct statistical measures and variations of f-GANs, and for the case of Jensen-Shannon divergence, the original GAN as in (28) can be obtained. Therefore, we can see that f-GANs are originated from statistical distance minimization.

Note that *f*-GAN interprets the GAN training as a statistical distance minimization after dualization. Similar statistical distance minimization idea is employed for the Wasserstein GAN, but now with a real metric in probability space rather than the divergence. More specifically, W-GAN minimizes the following Wasserstein-1 norm:

$$d(\mu,\nu) \triangleq W_1(\mu,\nu) := \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \|\boldsymbol{x}-\boldsymbol{x}'\| d\pi(\boldsymbol{x},\boldsymbol{x}') \tag{35}$$

where $\mathcal{X}$ is the ambient space, μ and ν are measures for the real data and generated data, respectively, and $\boldsymbol{\pi}(\mathbf{x}, \mathbf{x}')$ is the joint distribution with the marginals μ and ν, respectively.

Similar to *f*-GAN, rather than solving the complicated primal problem, a dual problem is solved. The Kantorivich dual formulation from the optimal transport theory leads to the following dual formulation of the Wasserstein 1-norm:

$$d(\mu,\nu) = \sup_{D \in \mathrm{Lip}_1(\mathcal{X})} \left\{ \int_{\mathcal{X}} D(\boldsymbol{x}) d\mu(\boldsymbol{x}) - \int_{\mathcal{X}} D(\boldsymbol{x}') d\nu(\boldsymbol{x}') \right\}, \tag{36}$$

where $\mathrm{Lip}_1(\mathcal{X})$ denotes the 1-Lipschitz function space with domain $\mathcal{X}$, and D is the Kantorovich potential that corresponds to the discriminator. Again, the measure ν is for the generated samples from latent space $\mathcal{Z}$ with the measure ζ by generator $G(z)$, $z \in \mathcal{Z}$, so ν can be considered as pushforward measure $\nu = G_{\#}\mu$. Therefore, Wasserstein 1-norm minimization problem can be equivalently represented by the following minmax formulation:

$$\mathcal{L}_{GAN}(G,D) = \min_{G} \max_{D \in \mathrm{Lip}_1(\mathcal{X})} \left\{ \int_{\mathcal{X}} D(\boldsymbol{x}) d\mu(\boldsymbol{x}) - \int_{\mathcal{Z}} D(G(z)) d\zeta(z) \right\}.$$

This again confirms that W-GAN is originated from the statistical distance minimization problem.

2) Biological Applications: Since the birth of GAN, myriad of variants have been introduced in literature and used for biological imaging applications. While the earlier works based on deep learning focused on developing supervised methods for training (e.g. DeepSTORM [24]), the later works started to employ conditional GAN (cGAN) into the reconstruction framework. More specifically, instead of applying the original form of the GAN that generates images from random noise, these applications of GAN are usually conditioned on specific input images.

For example, in the context of tomographic reconstruction, TomoGAN [25] aims at low dose tomographic reconstruction, where the generator takes in as input noisy images from low dose sinogram, and maps it into the distribution of high dose images. Another model for 3-D tomographic reconstruction, dubbed GANrec, was proposed in [26]. Different from TomoGAN, GANrec takes in as input the sinogram, so that the generator needs also to learn the inverse mapping of the forward Radon transform. One unique aspect is that the discriminator D learns the probability distribution of the clean sinogram. A similar

approach is used for super resolution [27], [28]. Specifically, in [28] a super-resolution (SR) approach for Fourier ptychographic microscopy (FPM) is introduced, which proposes to reconstruct a temporal sequence of cell images. Namely, only the first temporal sequence needs to be acquired in high resolution to train the GAN network, after which the trained network is utilized for reconstruction at the following temporal sequences. They also propose to use a Fourier domain loss, imposing additional constraint on the content. For super-resolution microscopy, ANNA-PALM [27] was introduced to achieve high-throughput in live-cell imaging, designed for accelerating PALM by using much less number of frames for restoring the true image.

These approaches that add condition to GANs in fact corresponds to pix2pix [21] or cGAN. Unlike GANs illustrated in Fig. 7(a), which takes random noise vector $\boldsymbol{z}$ as input, pix2pix has additional loss function $\mathscr{L}_{content}$ that measures the content distance (see Fig. 7(b)). Specifically, $\mathscr{L}_{content}$ measures the content space distance between the generated image and the matched target image, which is used in addition to the $\mathscr{L}_{GAN}$ that measures the statistical distance. Therefore, pix2pix attempts to balance between the paired data and unpaired target distributions. In fact, the addition of content loss is important to regularize the inverse problems. Unfortunately, the methods cannot be regarded as unsupervised, since the content loss $\mathscr{L}_{content}$ requires a matching label. Hence, to overcome this limitation, several works that do not require any matched training data were proposed.

One interesting line of work stems from ambientGAN [22], where the forward measurement model can be integrated into the framework. As in Fig. 7(c), the generator of ambientGAN generates a sample from a random noise vector, and the discriminator takes in the measurement after the forward operator $H_{\boldsymbol{\varphi}}$ parameterized by $\boldsymbol{\varphi}$, rather than the reconstructed image. Since only the function family of the forward operator is known, the specific parameters are sampled from a feasible distribution, i.e. $\boldsymbol{\varphi} \sim P_{\boldsymbol{\varphi}}$. Although the real and fake measurements do not match, ambientGAN enables training on the distribution, rather than on realized samples. From a statistical distance minimization perspective, ambientGAN can be interpreted as the dual problem for the statistical distance minimization in the measurement space. To understand this claim, suppose that we use a W-GAN discriminator, and consider the following primal form of the optimal transport problem that minimizes the 1-Wasserstein distance in the measurement space:

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathscr{X}\times\mathscr{Y}} \|H_{\varphi}(\boldsymbol{x}) - \boldsymbol{y}\| d\pi(\boldsymbol{x},\boldsymbol{y}) . \tag{37}$$

Then, the corresponding dual cost function becomes

$$\mathscr{L}_{GAN}(G,D) = \max_{D \in \mathrm{Lip}_1(\mathscr{X})} \int_{\mathscr{Y}} D(\boldsymbol{y}) d\nu(\boldsymbol{y}) - \int_{\mathscr{X}} D\big(H_{\varphi}(\boldsymbol{x})\big) d\mu(\boldsymbol{x}) \tag{38}$$

$$= \max_{D \in \mathrm{Lip}_1(\mathscr{X})} \int_{\mathscr{Y}} D(\boldsymbol{y}) d\nu(\boldsymbol{y}) - \int_{\mathscr{X}} D\big(H_{\varphi}(G(z))\big) d\zeta(z) . \tag{39}$$

where the last equation again comes from the change of variables formula. If we further assume that $\boldsymbol{\varphi} \in \boldsymbol{\Phi}$ is random from the distribution $P_{\boldsymbol{\varphi}}$, (39) can be converted to

$$\mathscr{L}_{GAN}(G, D) = \max_{D \in \mathrm{Lip}_1(\mathscr{X})} \int_{\mathscr{Y}} D(\boldsymbol{y}) d\nu(\boldsymbol{y}) - \int_{\Phi} \int_{\mathscr{X}} D\big(H_{\varphi}(G(z))\big) d\zeta(z) dP_{\varphi}, \tag{40}$$

which is equivalent to the ambientGAN loss function.

In the original work of ambientGAN, simple forward measurement models such as convolve+noise, block+patch, 2D projection, etc. were used [22]. A variant of ambientGAN was introduced in the context of cryo electron microscopy (cryo-EM) in [23], dubbed cryoGAN. Data acquisition in cryo-EM is performed on multiple 3D copies of the same protein, called “particles”, which are assumed to be structurally identical. To minimize the damage held on samples, multiple particles are frozen at cryogenic temperatures, and all particles are simultaneously projected with parallel electron beam to acquire projections. Here, unlike in the original ambientGAN, cryoGAN considers the latent particle itself to be a learnable parameter. The overall flow of cryoGAN is as shown in Fig. 7(d). It is interesting that there exists no generator in cryoGAN. Rather, $\boldsymbol{x}$, the 3D particle to be reconstructed, is the starting point of the overall flow. As in ambientGAN, $\boldsymbol{x}$ goes through a complex random forward measurement process which involves 3D projection, convolution with the sampled kernel, and translation. Gradients from the discriminator backpropagates to $\boldsymbol{x}$, and $\boldsymbol{x}$ is updated directly at every optimization step. Unlike conventional reconstruction methods for cryo-EM based on marginal maximum-likelihood which demands estimation of the exact projection angles, cryoGAN does not require such expensive process. Note that the loss function of cryoGAN is equivalent to (38). Therefore, by using the statistical distance minimization approach, cryoGAN attempts to estimate the unknown 3D particular $\boldsymbol{x}$ directly without estimating the projection angles for each particle.

Another, more recent work was proposed in [29], which is an upgraded version of cryoGAN, called multi-cryoGAN. While cryoGAN is able to reconstruct a single particle that explains the measured projections, it does not take into account that the measured particle is not rigid, and hence can have multiple conformations. To sidestep this issue, multi-cryoGAN takes an approach more similar to the original ambientGAN, where a random noise vector is sampled from a distribution, and the generator G is responsible for mapping the noise vector into the 3D particle. The rest of the steps follow the same procedure in ambientGAN, although the complicated forward measurement for cryo-EM is utilized. One advantage of multi-cryoGAN is that once the networks are trained, multiple conformations of the particle can be sampled by varying the noise vector $\boldsymbol{z}$. Subsequently, this introduces flexibility in the networks.

A related work was also proposed in the context of unsupervised MRI reconstruction in [10]. More specifically, this work follows the overall flow depicted in Fig. 7(c). However, the input is not a random noise vector, but an aliased image, inverse Fourier-transformed from the under-sampled k-space measurement. The generator is responsible for conditional reconstruction, making the input image free of aliasing artifacts. The reconstruction goes through the random measurement process in the context of MR imaging, which corresponds

to Fourier transform, and random masking. Then, the discriminator matches the distribution of the aliased image, inverse Fourier transformed from the measurement. The authors showed that even with the unsupervised learning process without any ground-truth data, reconstruction of fair quality could be performed.

D. Optimal transport driven CycleGAN approaches for unsupervised learning for biological imaging

Another important line of work for unsupervised biological reconstruction comes from optimal transport driven cycleGAN (OT-cycleGAN) [10], which is a generalization of the original cycleGAN [9]. Unlike pix2pix, cycleGAN does not utilize $\mathscr{L}_{content}$ from paired label, so it is fully unsupervised. In contrast to the ambientGAN or cryoGAN, which is based on the statistical distance minimization in the measurement space, cycleGAN attempts to minimizes the statistical distance in both measurement and the image domain simultaneously, which makes the algorithm more stable.

OT-cycleGAN can be understood from the geometric description illustrated in Fig. 8. Specifically, let us consider the target image probability space $\mathscr{X}$ equipped with the measure μ, and the measurement probability space $\mathscr{Y}$ equipped with the measure ν as in Fig. 8. In order to achieve a mapping from $\mathscr{Y}$ to $\mathscr{X}$ and vice versa, we can try to find the transportation mapping from the measure space $(\mathscr{Y}, \nu)$ to $(\mathscr{X}, \mu)$ with the generator $G_\theta: \mathscr{Y} \mapsto \mathscr{X}$, a neural network parameterized with $\boldsymbol{\theta}$, and the mapping from the measure space $(\mathscr{X}, \mu)$ to $(\mathscr{Y}, \nu)$ with the forward mapping generator $H_\varphi: \mathscr{X} \mapsto \mathscr{Y}$, parametrized with ν. In other words, the generator $G_{\boldsymbol{\theta}}$ pushes forward the measure ν in $\mathscr{X}$ to $\mu_{\boldsymbol{\theta}}$ in $\mathscr{Y}$, and $H_{\boldsymbol{\varphi}}$ pushes forward the measure μ in $\mathscr{Y}$ to the measure $\nu_{\boldsymbol{\varphi}}$ in $\mathscr{X}$. Then, our goal is to minimize the statistical distance $d(\mu, \mu_{\boldsymbol{\theta}})$ between μ and $\mu_{\boldsymbol{\theta}}$, and the distance $d(\nu, \nu_{\boldsymbol{\varphi}})$ between ν and $\nu_{\boldsymbol{\phi}}$ simultaneously.

Specifically, if we use the Wasserstein-1 metric, the statistical distance in each space can be computed as:

$$W_1(\mu, \mu_\theta) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathscr{X} \times \mathscr{Y}} \|\boldsymbol{x} - G_\theta(\boldsymbol{y})\| d\pi(\boldsymbol{x}, \boldsymbol{y}) \tag{41}$$

$$W_1(\nu, \nu_\varphi) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathscr{X} \times \mathscr{Y}} \|\boldsymbol{y} - H_\varphi(\boldsymbol{x})\| d\pi(\boldsymbol{x}, \boldsymbol{y}) \,. \tag{42}$$

If we minimize them separately, the optimal joint distribution $\boldsymbol{\pi}^*$ for each problem may be different. Accordingly, we attempt to find the unique joint distribution which minimizes the two distances simultaneously using the following primal formulation:

$$\min_{\pi \in \Pi(\mu, \nu)} \int_{\mathscr{X} \times \mathscr{Y}} \|\boldsymbol{x} - G_\theta(\boldsymbol{y})\| + \|H_\varphi(\boldsymbol{x}) - \boldsymbol{y}\| d\pi(\boldsymbol{x}, \boldsymbol{y}) \,. \tag{43}$$

One interesting finding made in [10] is that the primal cost in (43) can be represented in a dual formulation

$$\min_{\theta,\varphi}\ \max_{D_{\boldsymbol{x}},D_{\boldsymbol{y}}}\ \mathscr{L}_{\text{cycle}GAN}\big(\theta,\varphi;D_{\boldsymbol{x}},D_{\boldsymbol{y}}\big), \tag{44}$$

where

$$\mathscr{L}_{cycleGAN}\big(\theta,\varphi;D_{\boldsymbol{x}},D_{\boldsymbol{y}}\big)\triangleq\lambda\mathscr{L}_{cycle}(\theta,\varphi)+\mathscr{L}_{GAN}\big(\theta,\varphi;D_{\boldsymbol{x}},D_{\boldsymbol{y}}\big), \tag{45}$$

where $\mathscr{L}_{cycle}$, $\mathscr{L}_{GAN}$ refers to cycle-consistency loss and discriminator GAN loss, respectively. D_X and D_Y are discriminators in $\mathscr{X}$ and $\mathscr{Y}$. The corresponding OT-cycleGAN network architecture can be represented as in Fig. 9.

In fact, one of the most important reasons OT-cycleGAN is suitable for biological reconstruction problems, is that the prior knowledge about the imaging physics can be flexibly incorporated into the design of OT-cycleGAN to simplify the network. Specifically, in many biological imaging problems, the forward mapping $H_{\boldsymbol{\varphi}}$ is known or partially known. In this case, we do not need to use complex deep neural networks for forward measurement operator. Instead, we use a deterministic or parametric form of the forward measurement operation, which makes the training much simpler.

In addition, in comparison with ambientGAN in (37), OT-cycleGAN primal formulation in (43) has an additional term $\|\boldsymbol{x}-G_\theta(\boldsymbol{y})\|$ that enforces the reconstruction images to match the target image distributions, which further regularizes the reconstruction process. In fact, the resulting OT-cycleGAN formulation is closely related to the classical RLS formulation in (2). Specifically, the transportation cost in (43) resembles closely to the cost function in (2), except that regularization term $R(\boldsymbol{x})$ in (2) is replaced by the deep learning-based inverse path penalty term $\|\boldsymbol{x}-G_\theta(\boldsymbol{y})\|$. However, instead of solving x directly as in (2), OT-cycleGAN tries to find the joint distribution $\boldsymbol{\pi}^*$ that minimizes the average cost for all combination of $\boldsymbol{x}\in\mathscr{X}$ and $\boldsymbol{y}\in\mathscr{Y}$. This suggests that OT-cycleGAN is a stochastic generalization of the RLS, revealing an important link to the classical RLS approaches.

1) Applications: Thanks to the versatility of cycleGAN, which learns the distributions in both measurement and image spaces, OT-cycleGAN has been adopted to numerous tasks in biological imaging.

For example, cycleGAN was used with linear blur kernel for blind and non-blind deconvolution in [11]. More specifically, [11] focused on the fact that the forward operator of deconvolution microscopy is usually represented as a convolution with a point spread function (PSF). Hence, even for the non-blind case, the forward mapping $H_{\boldsymbol{\varphi}}:\mathscr{X}\mapsto\mathscr{Y}$ is partially known as a linear convolution. Leveraging this property, one of the generators in cycleGAN, F in Fig. 9 is replaced with a linear convolutional layer, taking into the account the physics of deconvolution microscopy. By exploiting the physical property, the reconstruction quality of deconvolution microscopy is further enhanced. Even more, in the case of non-blind microscopy, it was shown that the forward mapping is deterministic so that optimization with respect to the discriminator D_Y is no longer necessary, which simplifies the network architecture, and makes the training more robust. A similar simplification

of cycleGAN leveraging the imaging physics of microscopy was also proposed in super-resolution microscopy [10]. Interestingly, the simplified form of cycleGAN could generate reconstructions of higher resolution, quantified in Fourier ring correlation (FRC). Other than simplifying the mapping $H_\varphi : \mathcal{X} \mapsto \mathcal{Y}$ with a linear blind kernel, a deterministic k-space sub-sampling operator for MR imaging was extensively studied [30].

When such simplification is not possible, the most general form of cycleGAN, where two sets of generator/discriminator pair are used, can be utilized, but still the key concept of statistical distance minimization can be utilized in the design. One work, which utilizes cycleGAN for deconvolution microscopy is [32], where the authors propose to use spatial constraint loss on top of cyclic loss to further impose emphasis on the alignment of the reconstruction. The cycleGAN method adopted in [32] is a 2D cycleGAN, so the authors propose a 3-way volume averaging of the reconstructed results in the $x-y$, $y-z$, and $x-z$ plane. However, in contrast to [11], two neural network based generators are used for both forward and inverse paths. In another work, an unsupervised reconstruction method called projectionGAN for optical diffraction tomography (ODT) was proposed [31]. Missing cone problem in ODT arises because the measurement angles of the imaging device does not cover the whole solid angle, hence leaving a cone-shaped wedge in the k-space empty. The authors focus on the fact that when parallel beam projection is performed to the 3D distribution of refractive-index (RI), the acquired projections are sharp with high quality when the projection angle is aligned with the measurement angle ($\mathcal{Y}_\Omega$), and are blurry and with artifacts when the projection angle is not aligned ($\mathcal{Y}_{\Omega^c}$). Hence, the resolution of the blurry projections are enhanced via distribution matching between $\mathcal{Y}_\Omega$ and $\mathcal{Y}_{\Omega^c}$ with cycleGAN, after which follows filtered back projection (FBP) to acquire the final reconstruction from the enhanced projections. By the projectionGAN enhancement step, the missing cone artifacts are greatly resolved, achieving accurate reconstruction, as illustrated in Fig. 10. As shown in the figure, with other methods we see elongation along optical axes which makes the structure of the cell vague and noisy ($x-z$, $y-z$ plane). This problem is much resolved with ProjectionGAN, where we observe clear boundaries and micro-cellular structures. Underestimated RI values are also corrected.

For optical microscopy, content-preserving cycleGAN (c^{2}GAN) was proposed [33], showing applicability of cycleGAN to various imagnig modalities and data configurations. c^{2}GAN introduces saliency constraint to cycleGAN framework, where the saliency constraint imposes an additional cycle-consistency after thresholding the images at certain values. This simple fix is derived from the fact that many biological images contain salient regions of higher intensity, while the rest is covered with low-intensity background. Thus, by adding the saliency constraint, cycleGAN can concentrate more on the salient features. The authors applied c^{2}GAN to biological image denoising, restoration, super-resolution, histological colorization, and image translation such as phase contrast images to flourescence-labeled images, showing how cycleGAN can be easily adopted to many different tasks of biological imaging.

V. Discussion

A. Open problems

The performance improvement from DL-based techniques has been one of the main drivers of their mainstream adaptation in a large number of imaging applications. This is largely driven by the application-specific tailoring of the regularization strategies during the training phase of DL reconstruction algorithms. Thus, the use of unsupervised training strategies in the absence of matched reference data is critical for the continued utility of DL reconstruction in a number of biological imaging scenarios.

This overview article focused on two unsupervised learning strategies that tackle seemingly different aspects of the training process. Self-supervised learning uses parts of the available data to predict the remaining parts, in effect repurposing some of the available data as supervisory labels. Generative models aim to minimize a statistical distance measure between an underlying target distribution and the generated data distribution. While these goals do not necessarily appear complementary, there are self-supervisory methods, such as content generation, which utilize properties of generative models. Similarly, there are generative models that utilize concepts of prediction of data from self-supervision [34]. Thus, a synergistic viewpoint that tie these two different lines of work for unsupervised learning of image reconstruction approaches may further improve the performance of DL-based methods in the absence of reference training data.

Self-supervised learning techniques have enabled the training on large datasets containing only noisy or incomplete measurements. However, in some biological applications, it may not always be feasible to obtain large training datasets. Hence, it is desirable to perform training from a single measurement. However, training on a single noisy measurement often leads to overfitting, requiring early stopping. Recently, self-supervised learning methods have been proposed to perform reconstruction and enhancement for a single measurement without overfitting [35], [36]. Particularly, for image denoising, a dropout regularization technique has been incorporated with a hold-out self-supervised learning framework for avoiding overfitting [35]. For image reconstruction, a zero-shot self-supervised learning approach has been proposed to split available measurements: two of which are used in the data consistency and the loss as in SSDU, while the third is used as a validation set to determine the early stopping criteria [36]. These works may be essential for developing new frameworks for training biological imaging applications with sparse datasets.

Recently, the two closely related methods, score-based models [37], and diffusion models [38] have caught the attention with their outstanding ability to train generative models *without* any adversarial training. Remarkably, one cannot only generate random samples from the distribution, but also apply a *single* estimated score function to solve various problems such as denoising [39], inpainting [37], and even reconstruction. Since these score-based generative methods are extremely flexible in that they do not require any problem-specific training, they may open up exciting new opportunities for developing new unsupervised learning based methods for biological image reconstruction and enhancement.

Another interesting direction is feature disentanglement. Unsupervised feature disentanglement methods were proposed in different fields including generative modelling of material structure [40]. Although seemingly unrelated, the fundamental problem of biological image reconstruction and enhancement can be viewed as disentangling salient signal from the noisy measurement. By exploiting widely used tools, for instance adaptive instance normalization for feature disentanglement, one could build a new approach to biological imaging.

B. Availability of training databases

While the early works in biological imaging applications relied on utilizing imaging datasets that were released for other purposes, such as segmentation or tracking challenges, there have been substantial recent efforts in the release and use of publicly available biological imaging data. The BioImage Archive, Image Data Resources (IDR), BioImage.IO and Electron Microscopy Public Image Archive (EMPIAR) constitute some of these efforts. Moreover, there are platforms such as Zenodo and Figshare that host and distribute biological imaging data. The increasing availability of such large databases of raw measurement data for different biomedical imaging modalities may further facilitate development of DL-based reconstruction and enhancement strategies.

VI. Conclusion

Deep learning methods have recently become the state-of-the-art approaches for image reconstruction. While conventionally, such methods are trained using supervised training, the lack of matched reference data has hampered their utility in biological imaging applications. Thus, unsupervised learning strategies, encompassing both self-supervised methods and generative models, have been proposed, showing great promise. Self-supervised methods devise a way to create supervisory labels from the incomplete measurement data itself to train the model. Hold-out masking strategy is especially useful for both image denoising and reconstruction. With recent advances, one can perform training with as little as a single noisy measurement. Generative model based methods encompass diverse methods for image denoising and reconstruction, with VAE and GAN being the two most prominent strategies. Both methods can be seen as the optimization problem of statistical minimization, with different choices for statistical distance measure leading to seemingly unrelated methods for training the generative model.

These strategies are still being developed and applied to biological imaging scenarios, creating opportunities for the broader signal processing community in terms of new technical developments and applications.

Acknowledgments

This work was partially supported by NIH R01HL153146, NIH P41EB027061, NIH R21EB028369, NSF CAREER CCF-1651825 and NRF-2020R1A2B5B0300198

References

[1]. Jing L and Tian Y, “Self-supervised visual feature learning with deep neural networks: A survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. 2, 6

[2]. Krull A, Buchholz T-O, and Jug F, “Noise2Void-learning denoising from single noisy images,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2129–2137. 2, 12

[3]. Batson J and Royer L, “Noise2Self: blind denoising by self-supervision,” in Proceedings of the International Conference on Machine Learning, 2019, pp. 524–533. 2, 9, 12

[4]. Yaman B, Hosseini SAH, Moeller S, Ellermann J, Ugurbil K, and Akcakaya M, “Self-supervised learning of physics-guided reconstruction neural networks without fully-sampled reference data,” Magnetic Resonance in Medicine, vol. 84, no. 6, pp. 3172–3191, Dec 2020. 2, 9, 10, 11, 13 [PubMed: 32614100]

[5]. Buchholz T-O, Krull A, Shahidi R, Pigino G, Jékely G, and Jug F, “Content-aware image restoration for electron microscopy,” Methods in Cell Biology, vol. 152, pp. 277–289, 2019. 2, 6, 11 [PubMed: 31326025]

[6]. Demirel OB, Yaman B, Dowdle L, Moeller S, Vizioli L, Yacoub E, Strupp J, Olman CA, Uğurbil K, and Akçakaya M, “Improved simultaneous multi-slice functional MRI using self-supervised deep learning,” arXiv: 2105.04532, 2021. 2, 13, 14

[7]. Ruthotto L and Haber E, “An introduction to deep generative modeling,” arXiv preprint arXiv:2103.05180, 2021. 3, 14, 15

[8]. Nowozin S, Cseke B, and Tomioka R, “f-GAN: training generative neural samplers using variational divergence minimization,” in Advances in Neural Information Processing Systems, 2016, vol. 29. 3, 18–19

[9]. Zhu J-Y, Park T, Isola P, and Efros AA, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232. 3, 23

[10]. Sim B, Oh G, Kim J, Jung C, and Ye JC, “Optimal transport driven CycleGAN for unsupervised learning in inverse problems,” SIAM Journal on Imaging Sciences, vol. 13, no. 4, pp. 2281–2306, 2020. 3, 23, 24, 25

[11]. Lim S, Park H, Lee S-E, Chang S, Sim B, and Ye JC, “CycleGAN with a blur kernel for deconvolution microscopy: Optimal transport geometry,” IEEE Transactions on Computational Imaging, vol. 6, pp. 1127–1138, 2020. 3, 25, 27

[12]. Knoll F, Hammernik K, Zhang C, Moeller S, Pock T, Sodickson DK, and Akçakaya M, “Deep-learning methods for parallel magnetic resonance imaging reconstruction,” IEEE Signal Processing Magazine, vol. 37, no. 1, pp. 128–140, 2020. 5 [PubMed: 33758487]

[13]. Lehtinen J, Munkberg J, Hasselgren J, Laine S, Karras T, Aittala M, and Aila T, “Noise2Noise: Learning image restoration without clean data,” in Proceedings of the 35th International Conference on Machine Learning. 2018, vol. 80, pp. 2965–2974, PMLR. 8

[14]. Yaman B, Hosseini SAH, Moeller S, Ellermann J, Uğurbil K, and Akçakaya M, “Ground-truth free multi-mask self-supervised physics-guided deep learning in highly accelerated mri,” in 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE, 2021, pp. 1850–1854. 11, 13

[15]. Senouf O, Vedula S, Weiss T, Bronstein A, Michailovich O, and Zibulevsky M, “Self-supervised learning of inverse problem solvers in medical imaging,” in Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data, pp. 111–119. Springer, 2019. 11, 14

[16]. Krull A, Vičar T, Prakash M, Lalit M, and Jug F, “Probabilistic noise2void: Unsupervised content-aware denoising,” Frontiers in Computer Science, vol. 2, pp. 5, 2020. 12

[17]. Prakash M, Lalit M, Tomancak P, Krul A, and Jug F, “Fully unsupervised probabilistic noise2void,” in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, 2020, pp. 154–158. 12

[18]. Liu J, Sun Y, Eldeniz C, Gan W, An H, and Kamilov US, “Rare: Image reconstruction using deep priors learned without groundtruth,” IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 6, pp. 1088–1099, 2020. 13

[19]. Bepler T, Zhong ED, Kelley K, Brignole E, and Berger B, “Explicitly disentangling image content from translation and rotation with spatial-vae,” arXiv preprint arXiv:1909.11663, 2019. 16, 17

[20]. Prakash M, Krull A, and Jug F, “Fully unsupervised diversity denoising with convolutional variational autoencoders,” arXiv preprint arXiv:2006.06072, 2020. 16, 18

[21]. Isola P, Zhu J-Y, Zhou T, and Efros AA, “Image-to-image translation with conditional adversarial networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134. 20, 21

[22]. Bora A, Price E, and Dimakis AG, “AmbientGAN: generative models from lossy measurements,” in 6th International Conference on Learning Representations, ICLR, 2018. 20, 21–22

[23]. Gupta H, McCann MT, Donati L, and Unser M, “CryoGAN: a new reconstruction paradigm for single-particle Cryo-EM via deep adversarial learning,” BioRxiv, 2020. 20, 22

[24]. Nehme E, Weiss LE, Michaeli T, and Shechtman Y, “Deep-STORM: super-resolution single-molecule microscopy by deep learning,” Optica, vol. 5, no. 4, pp. 458–464, 2018. 20

[25]. Liu Z, Bicer T, Kettimuthu R, Gursoy D, De Carlo F, and Foster I, “TomoGAN: low-dose synchrotron x-ray tomography with generative adversarial networks: discussion,” JOSA A, vol. 37, no. 3, pp. 422–434, 2020. 20 [PubMed: 32118926]

[26]. Yang X, Kahnt M, Brückner D, Schropp A, Fam Y, Becher J, Grunwaldt J-D, Sheppard TL, and Schroer CG, “Tomographic reconstruction with a generative adversarial network,” Journal of synchrotron radiation, vol. 27, no. 2, 2020. 20

[27]. Ouyang W, Aristov A, Lelek M, Hao X, and Zimmer C, “Deep learning massively accelerates super-resolution localization microscopy,” Nature biotechnology, vol. 36, no. 5, pp. 460–468, 2018. 21

[28]. Nguyen T, Xue Y, Li Y, Tian L, and Nehmetallah G, “Deep learning approach for Fourier ptychography microscopy,” Optics express, vol. 26, no. 20, pp. 26470–26484, 2018. 21 [PubMed: 30469733]

[29]. Gupta H, Phan TH, Yoo J, and Unser M, “Multi-CryoGAN: Reconstruction of continuous conformations in Cryo-EM using generative adversarial networks,” in European Conference on Computer Vision. Springer, 2020, pp. 429–444. 22

[30]. Oh G, Sim B, Chung H, Sunwoo L, and Ye JC, “Unpaired deep learning for accelerated MRI using optimal transport driven cycleGAN,” IEEE Transactions on Computational Imaging, vol. 6, pp. 1285–1296, 2020. 26

[31]. Chung H, Huh J, Kim G, Park YK, and Ye JC, “Unsupervised missing cone deep learning in optical diffraction tomography,” arXiv preprint arXiv:2103.09022, 2021. 26, 27

[32]. Lee S, Han S, Salama P, Dunn KW, and Delp EJ, “Three dimensional blind image deconvolution for fluorescence microscopy using generative adversarial networks,” in 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, 2019, pp. 538–542. 27

[33]. Li X, Zhang G, Qiao H, Bao F, Deng Y, Wu J, He Y, Yun J, Lin X, Xie H, et al. , “Unsupervised content-preserving transformation for optical microscopy,” Light: Science & Applications, vol. 10, no. 1, pp. 1–11, 2021. 27

[34]. Bostan E, Heckel R, Chen M, Kellman M, and Waller L, “Deep phase decoder: self-calibrating phase microscopy with an untrained deep neural network,” Optica, vol. 7, no. 6, pp. 559–562, 2020. 28

[35]. Quan Y, Chen M, Pang T, and Ji H, “Self2self with dropout: Learning self-supervised denoising from single image,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1890–1898. 28

[36]. Yaman B, Hosseini SAH, and Akçakaya M, “Zero-shot self-supervised learning for MRI reconstruction,” arXiv preprint arXiv:2102.07737, 2021. 28

[37]. Song Y and Ermon S, “Generative modeling by estimating gradients of the data distribution,” in Advances in Neural Information Processing Systems, 2019, vol. 32. 28

[38]. Ho J, Jain A, and Abbeel P, “Denoising diffusion probabilistic models,” in Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 6840–6851. 28

[39]. Kim K and Ye JC, “Noise2score: Tweedie’s approach to self-supervised image denoising without clean images,” arXiv preprint arXiv:2106.07009, 2021. 28

[40]. Chung H and Ye JC, “Feature disentanglement in generating three-dimensional structure from two-dimensional slice with sliceGAN,” arXiv preprint arXiv:2105.00194, 2021. 28

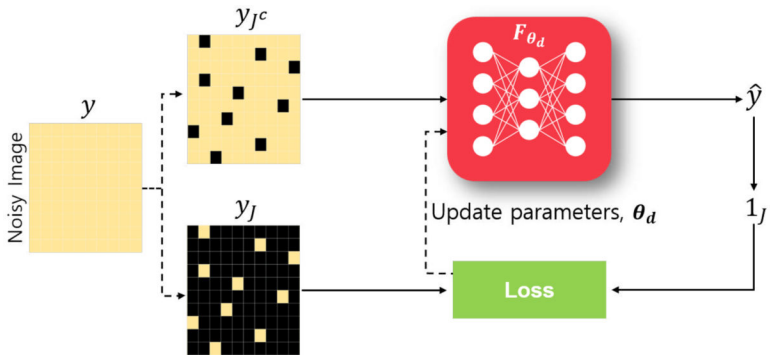

Fig. 1.

Overview of self-supervised learning for denoising. Black pixels denote masked-out locations in the images, while $\mathbf{1}_J$ is the indicator function on the indices specified by the index set J.

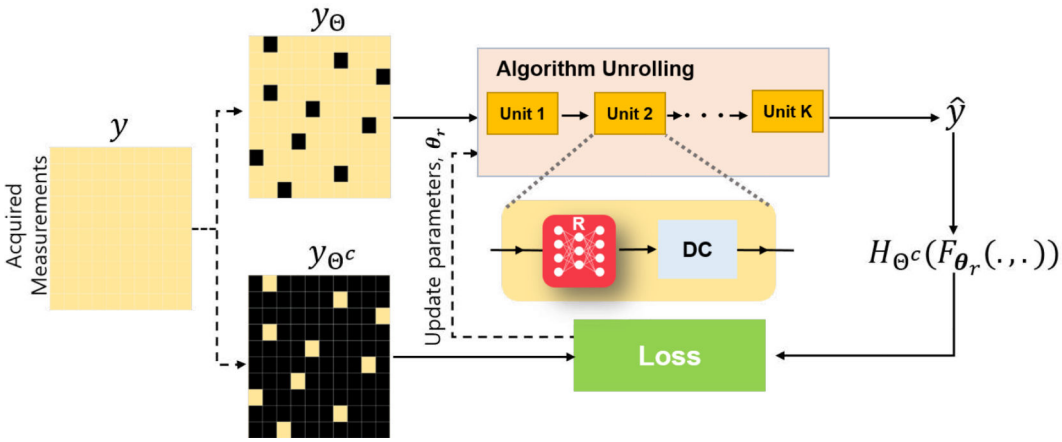

Fig. 2.

Overview of the self-supervised learning methods for image reconstruction using hold-out masking. Black pixels denote masked-out locations in the measurements and DC denotes the data consistency units of the unrolled network.

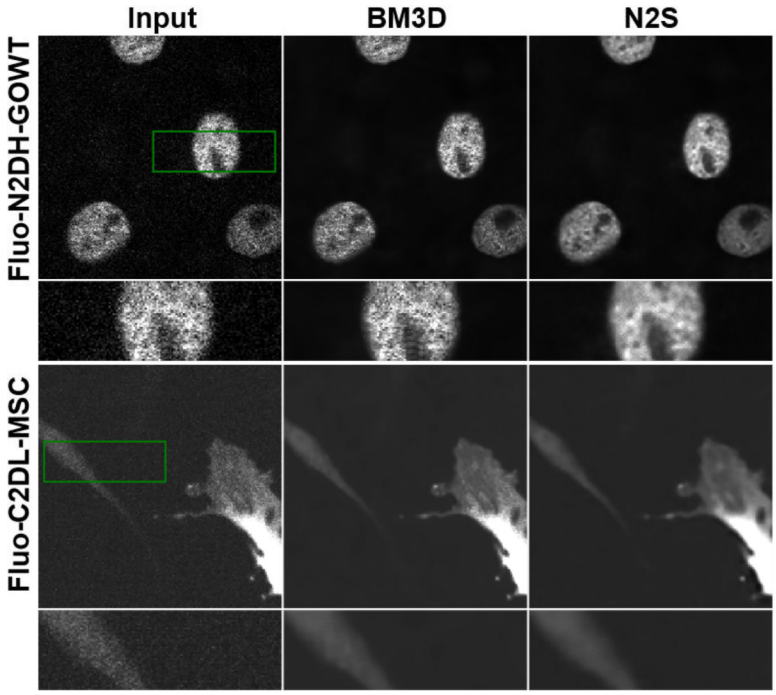

Fig. 3.

Denoising results from fluorescence microscopy datasets Fluo-N2DH-GOWT1 and Fluo-C2DL-MSC using a traditional denoising method BM3D and a self-supervised learning method Noise2Self (N2S). We note that supervised deep learning is not applicable as these datasets contain only single noisy images.

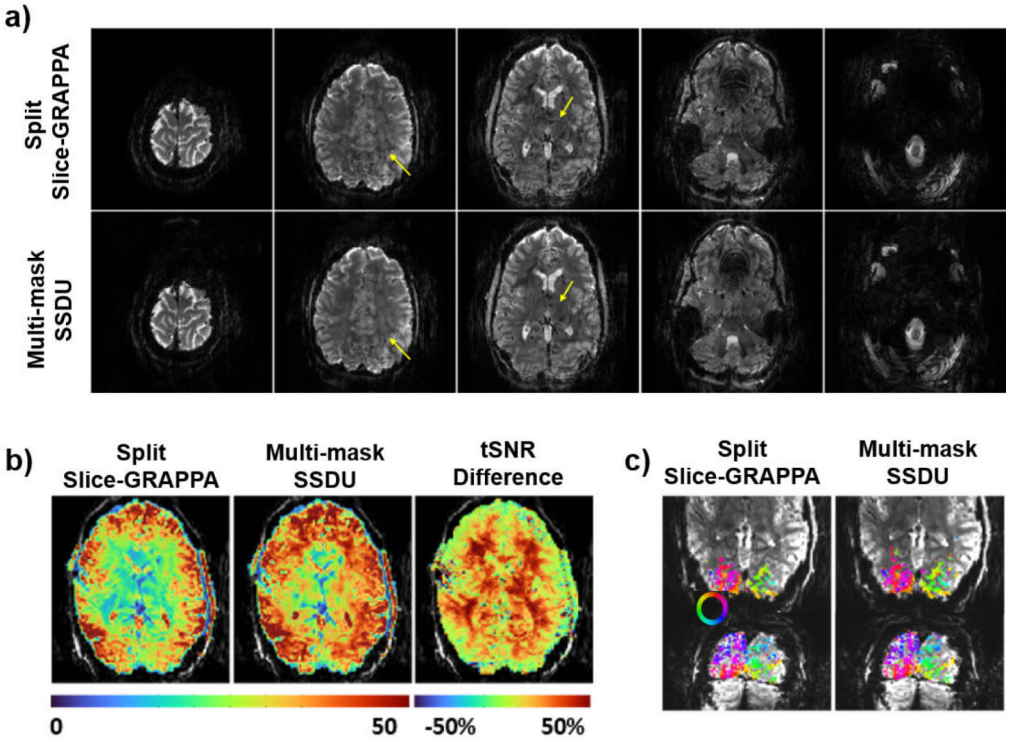

Fig. 4.

Reconstruction results from an fMRI application [6] using conventional split-slice GRAPPA technique and self-supervised multi-mask SSDU method [14]. (a) Split-slice GRAPPA exhibits residual artifacts in mid-brain (yellow arrows). Multi-mask SSDU alleviates these, along with visible noise reduction. (b) Temporal SNR (tSNR) maps show substantial gain with the self-supervised deep learning approach, particularly for subcortical areas and cortex further from the receiver coils. (c) Phase maps for the two reconstructions show strong agreement, with multi-mask SSDU containing more voxels above the coherence threshold.

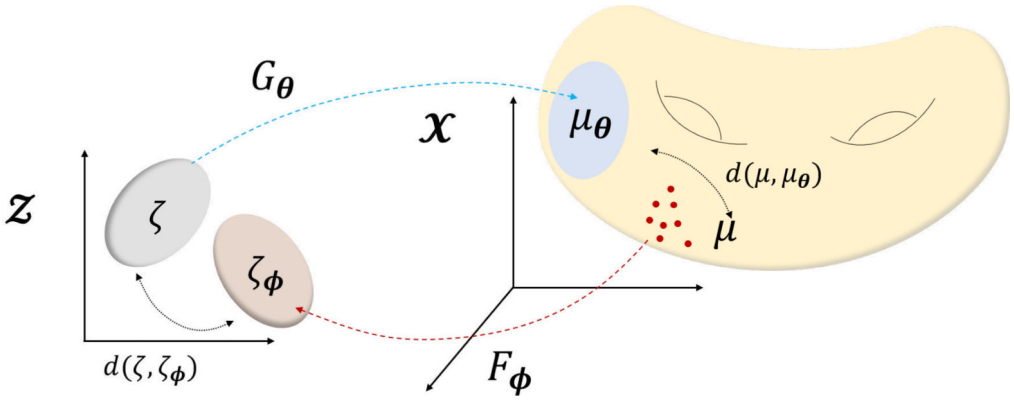

Fig. 5.

Geometric view of deep generative models. Fixed distribution ζ in $\mathcal{Z}$ is pushed to μ_θ in $\mathcal{X}$ by the network G_θ, so that the mapped distribution μ_θ approaches the real distribution μ. In VAE, G_θ works as a decoder to generate samples, while F_ϕ acts as an encoder, additionally constraining ζ_ϕ to be as close to ζ. With such geometric view, auto-encoding generative models (e.g. VAE), and GAN-based generative models can be seen as variants of this single illustration.

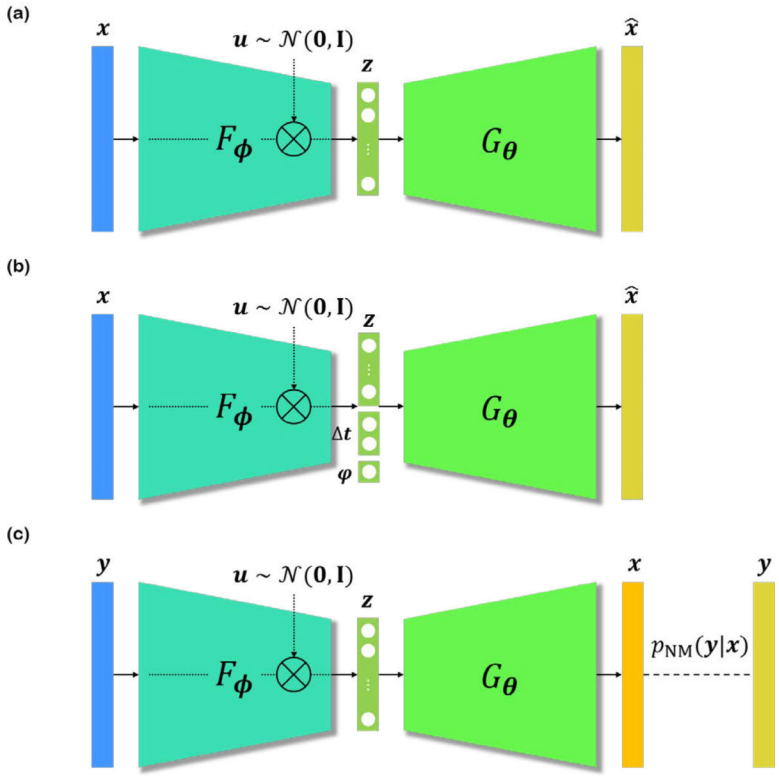

Fig. 6.

VAE architecture. F_{ϕ} encodes $\boldsymbol{x}$, and combined with random sample $\boldsymbol{u}$ to produce latent vector z. $G_{\boldsymbol{\theta}}$ decodes the latent z to acquire $\hat{\boldsymbol{x}}$. $\boldsymbol{u}$ is sampled from standard normal distribution for the reparameterization trick. (a) VAE. (b) spatial-VAE [19], disentangling translation/rotation features from different semantics. (c) DIVNOISING [20], enabling superviesd/unsupervised training of denoising generative model by leveraging the noise model $p_{NM}(\boldsymbol{y}|\boldsymbol{x})$.

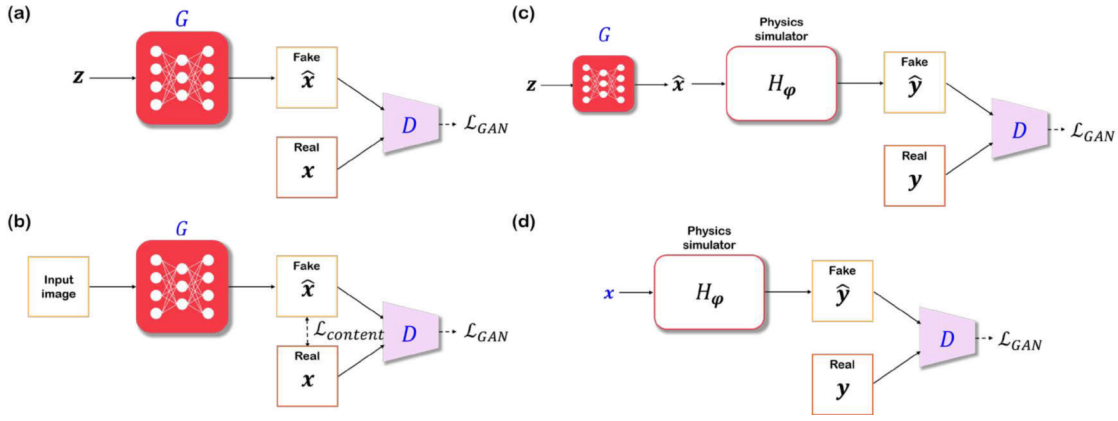

Fig. 7.

Illustration of GAN-based methods for biological image reconstruction. (a) GAN, (b) pix2pix [21], (c) AmbientGAN [22], (d) cryoGAN [23]. $\boldsymbol{x}$, $\boldsymbol{y}$ denote data in the image domain, and the measurement domain, respectively. G, D refers to generator, discriminator, respectively. H defines the function family of the forward measurement process, parameterized with $\boldsymbol{\varphi}$. Networks and variables that are marked in blue have learnable parameters optimized with gradient descent.

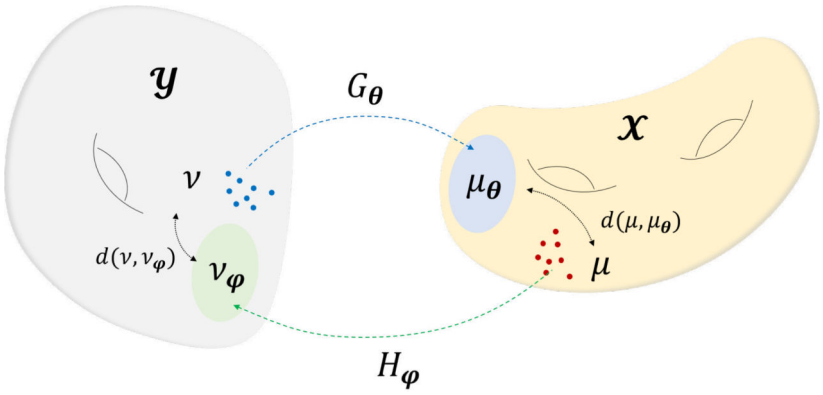

Fig. 8.

Geometric view of cycleGAN. $(\mathcal{Y}, \nu)$ is mapped to $(\mathcal{X}, \mu)$ with G_θ, while H_φ does the opposite. The two mappers, i.e. generators are optimized by simultaneously minimizing $d(\mu, \mu_\theta)$, $d(\nu, \nu_\varphi)$.

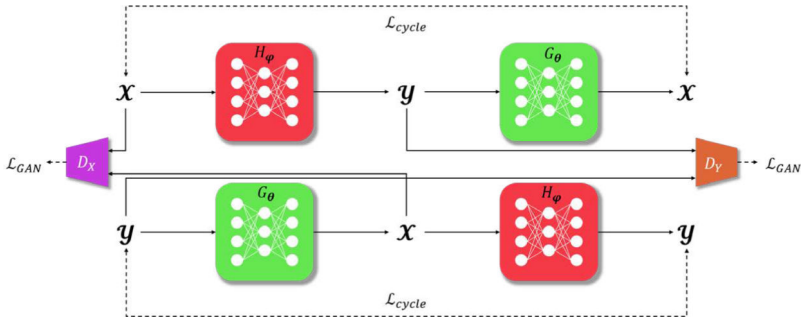

Fig. 9.

Network architecture of cycleGAN. $G_\theta: \mathcal{Y} \mapsto \mathcal{X}$, $H_\varphi: \mathcal{X} \mapsto \mathcal{Y}$ are the generators responsible for inter-domain mapping. D_X, D_Y are discriminators, constructing $\mathcal{L}_{GAN}$. GAN loss is simultaneously optimized together with $\mathcal{L}_{cycle}$

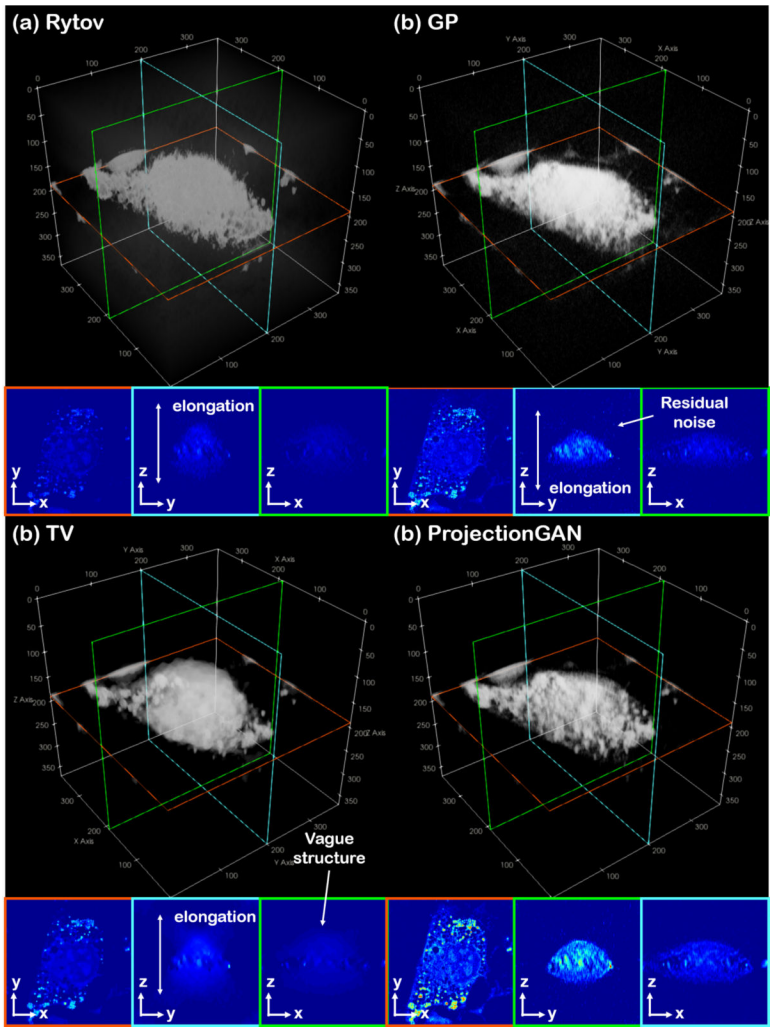

Fig. 10.

ProjectionGAN for the reconstruction of ODT [31]. (a) Conventional Rytov reconstruction via Fourier binning, (b) Gerchberg-Papoulis (GP) algorithm, (c) model-based iterative method using the total variation (TV), and (b) reconstruction via projectionGAN. Artifacts including elongation along the optical axes can be seen in the $x - z$, $y - z$ cutview of (a),(c). The result shown in (b) is contaminated with resdual noise in the $x - z$, $y - z$ planes. Result shown in (d) has high-resolution reconstruction without such artifacts, along with boosted RI values.