Mehmet Akçakaya, Burhaneddin Yaman, Hyungjin Chung, and Jong Chul Ye

Unsupervised Deep Learning Methods for Biological Image Reconstruction and Enhancement

An overview from a signal processing perspective

SHUTTERSTOCK.COM/KKSSR

Digital Object Identifier 10.1109/MSP.2021.3119273
Date of current version: 24 February 2022

Recently, deep learning (DL) approaches have become the main research frontier for biological image reconstruction and enhancement problems thanks to their high performance and ultrafast 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 article, we provide an overview of these approaches from a coherent perspective in the context of classical inverse problems and discuss their applications to biological imaging, including electron, fluorescence, deconvolution microscopy, optical diffraction tomography (ODT), and functional neuroimaging.

Introduction

Biological imaging techniques, such as optical microscopy, electron microscopy, and 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 used to obtain $\boldsymbol{x}$ from 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 these is the popular regularized least squares (RLS), which employs regularization (or penalty) terms to stabilize the following inverse solution:

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

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.

Over the last few years, DL approaches have become mainstream for inverse problems in biological imaging due to their excellent performance and ultrafast inference time compared to RLS. Most of the DL approaches are trained in a supervised manner, with paired input and ground-truth data, which often leads to a straightforward training procedure. Matched label data are not available in many applications, which is particularly problematic for biological imaging problems as the unknown image itself is intended for scientific investigation.

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 themselves 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 to define training labels. For image denoising, it was shown that this idea can be used to train a DL approach from single noisy images [2]. Furthermore, with an appropriate choice of the holdout mask, the self-supervised training loss was proved to be within an additive constant of the supervised training loss [3], providing a theoretical grounding for their success in 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 DL. 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 networks (GANs), which have attracted significant attention in the machine learning community by providing a way to generate target data distribution from a random one. In the article on *f*-GANs [8], the authors show that a general class of *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 measurement. Similarly, Wasserstein GANs (W-GANs) can be regarded as another statistical distance-minimization approach, where the statistical distance is measured by the Wasserstein-1 metric [7]. Inspired by these observations, a cycle-consistent GAN (cycleGAN) [9], which imposes one-to-one correspondence to address mode-collapsing behavior, was shown to be similarly obtained when the statistical distances in both the measurement and image spaces can be simultaneously minimized [10]. The cycleGAN formulation has been applied for various biological imaging problems, such as both deconvolution [11] and superresolution (SR) 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 article is therefore to offer a coherent perspective to understand this exciting area of research.

Background on biological image reconstruction and enhancement

Conventional solutions to the RLS problem

The objective function of the RLS problem in (2) forms the basis of most of the 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 nonlinear, gradient descent is a commonly used algorithm for the following solution:

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

where $\boldsymbol{x}^{(k)}$ is the solution at the kth iteration, and η_k is the gradient step. Although 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 functions in (2), relying on the use of the 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 nonconvex objective functions. In particular, variable-splitting approaches decouple $c(\boldsymbol{x}, \boldsymbol{y})$ and $R(\boldsymbol{x})$ terms by introducing an auxiliary variable $\boldsymbol{z}$ constrained to be equal to $\boldsymbol{x}$, as

$$\operatorname*{argmin}_{\boldsymbol{x},\boldsymbol{z}} c(\boldsymbol{x}, \boldsymbol{y}) + R(\boldsymbol{z}) \text{ subject to } \boldsymbol{x} = \boldsymbol{z}. \tag{4}$$

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

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

$$\boldsymbol{x}^{(k)} = \operatorname*{argmin}_{\boldsymbol{x}} \left\| \boldsymbol{y} - H(\boldsymbol{x}) \right\|^2 + \mu \left\| \boldsymbol{x} - \boldsymbol{z}^{(k-1)} \right\|^2, \tag{5b}$$

where $\boldsymbol{x}^{(0)} = -\eta \nabla_x c(\boldsymbol{0}, \boldsymbol{y})$ can be initialized with a single gradient descent step on the data consistency term, and $\boldsymbol{z}^{(k)}$ is an intermediate optimization variable. The subproblems in (5a) and (5b) correspond to a proximal operation and a data consistency step, respectively. Although 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, 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 hyperparameter choices, which often limit their routine use in practice.

DL-based reconstruction and enhancement with supervised training

DL methods have recently gained popularity as an alternative for estimating $\boldsymbol{x}$ from the measurement model in (1). In the broadest terms, these techniques learn a parameterized, nonlinear function that maps the measurements to an image estimate. The 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 function $F_{\boldsymbol{\theta}_e}(\boldsymbol{y})$. In particular, the input to the neural network is $\boldsymbol{y}$ if the measurements are in the image domain or are 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. Although this leads to a 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 approaches aim to solve the objective function in (2) by 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 (2) is unrolled for a fixed number of iterations, K. For instance, for the variable-splitting algorithm described in (5a) and (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 (5a) is implicitly implemented by a neural network, while the data consistency operation in (5b) is implemented by conventional methods that explicitly use $H(\cdot)$, such as gradient descent, with the only learnable parameter being the gradient's step size. These physics-guided methods have recently become 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 forward operator $H(\cdot)$ across samples [12]. Thus, the final unrolled network can be described by function $F_{\boldsymbol{\theta}_r}(\boldsymbol{y}; H)$, which explicitly incorporates the forward operator and is parameterized by $\boldsymbol{\theta}_r$.

For both of these DL 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

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

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

In practice, the mapping function in (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 parameters $\boldsymbol{\theta}$ of function $F_{\boldsymbol{\theta}}(\cdot)$. In particular, during training, $\boldsymbol{\theta}$ is adjusted to minimize the difference between the network output and ground-truth reference. More formally, training is performed by minimizing

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

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

Motivation for unsupervised DL approaches

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

For example, in transmission electron microscopy, 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 transmission electron microscopy, acquisition of the clean reference images is not feasible due to the limited electron dose used during acquisition to avoid sample destruction. Similarly, in scanning electron microscopy, the lengthy acquisition times for imaging large volumes remains a main limitation. Although it is desirable to speed up the acquisitions, such an 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 the listed microscopy applications. In many other biological applications, such as ODT, functional magnetic resonance imaging (MRI), or SR microscopy, such challenges exist in similar forms. Hence, unsupervised

DL approaches are essential for addressing the training of DL reconstruction methods in biological imaging applications.

Self-supervised learning methods

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, such as 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 are instead used to automatically define supervisory training labels from the data themselves. An overview of this strategy for denoising is shown in Figure 1. Although 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 supervised learning loss, verifying the practicality of self-supervision. This has also led to interest in 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 the reported empirical results from a variety of DL algorithms, especially physics-guided ones incorporating the forward operator, show that it can achieve similar reconstruction quality as supervised learning algorithms. To capture these inherent differences between the two problems, we next separately discuss self-supervised DL for denoising and reconstruction methods.

Self-supervised DL for denoising

Background on denoising using DL

Image denoising concerns a special case of the acquisition model in (1), where $H(\cdot)$ is the identity operator. In this case, the objective function for the inverse problem in (2) becomes $\operatorname{argmin}_x \|\boldsymbol{y}-\boldsymbol{x}\|_2^2 + R(\boldsymbol{x})$. In DL methods for denoising, this proximal operation is replaced by a neural network, which estimates a denoised image $\hat{\boldsymbol{x}}_{\text{denoised}} = F_{\boldsymbol{\theta}_d}(\boldsymbol{y})$ through a $\boldsymbol{\theta}_d$-parameterized function. Although supervised DL methods provide state-of-the-art results for denoising applications, the 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 to tackle 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 independent identically distributed, zero-mean, random noise components ($\boldsymbol{y}=\boldsymbol{x}+\boldsymbol{w}$, $\hat{\boldsymbol{y}}=\boldsymbol{x}+\hat{\boldsymbol{w}}$), N2N aims to minimize an MSE loss of the form

$$\begin{aligned}\min_{\boldsymbol{\theta}_d} \mathbb{E}_{\hat{\boldsymbol{y}},\boldsymbol{y}} \|F_{\boldsymbol{\theta}_d}(\boldsymbol{y})-\hat{\boldsymbol{y}}\|^2 &= \min_{\boldsymbol{\theta}_d} \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \|F_{\boldsymbol{\theta}_d}(\boldsymbol{y})-\boldsymbol{x}\|^2 + \mathbb{E}_{\hat{\boldsymbol{w}}}\|\hat{\boldsymbol{w}}\|^2 \\ &\quad - 2\mathbb{E}\langle \hat{\boldsymbol{w}}, F_{\boldsymbol{\theta}_d}(\boldsymbol{y})-\boldsymbol{x}\rangle \qquad (8)\\ &= \min_{\boldsymbol{\theta}_d} \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \|F_{\boldsymbol{\theta}_d}(\boldsymbol{y})-\boldsymbol{x}\|^2 + \mathbb{E}_{\hat{\boldsymbol{w}}}\|\hat{\boldsymbol{w}}\|^2, \qquad (9)\end{aligned}$$

where the last term in (8) becomes zero because $\mathbb{E}\hat{\boldsymbol{w}} = \boldsymbol{0}$. Note that the last term in (9) does not depend on $\boldsymbol{\theta}_d$. Hence, the $\boldsymbol{\theta}_d^\star$ that minimizes the N2N loss, i.e., $\mathbb{E}_{\boldsymbol{x},\boldsymbol{y},\hat{\boldsymbol{w}}}\|F_{\boldsymbol{\theta}_d}(\boldsymbol{y})-(\boldsymbol{x}+\hat{\boldsymbol{w}})\|^2$, is also a minimizer of the supervised loss $\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\|F_{\boldsymbol{\theta}_d}(\boldsymbol{y})-\boldsymbol{x}\|^2$. We note that different loss functions, such as L_1, 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 $\{\boldsymbol{y}^n = \boldsymbol{x}^n + \boldsymbol{w}^n, \hat{\boldsymbol{y}}^n = \boldsymbol{x}^n + \hat{\boldsymbol{w}}^n\}_{n=1}^N$. N2N trains a neural network for denoising by minimizing

$$\min_{\boldsymbol{\theta}_d} \sum_{n=1}^{N} \|F_{\boldsymbol{\theta}_d}(\boldsymbol{y}^n) - \hat{\boldsymbol{y}}^n\|^2. \qquad (10)$$

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

Self-supervised training for DL-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

$$\min_{\boldsymbol{\theta}_d} \mathbb{E}_{\boldsymbol{y}} \| F_{\boldsymbol{\theta}_d}(\boldsymbol{y}) - \boldsymbol{y} \|^2. \qquad (11)$$

However, the naive application of (11) leads to the denoising function $F_{\boldsymbol{\theta}_d}$ being an identity mapping.

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 function $F_{\boldsymbol{\theta}_d}$ can be learned without collapsing to the identity function. To this end, consider an

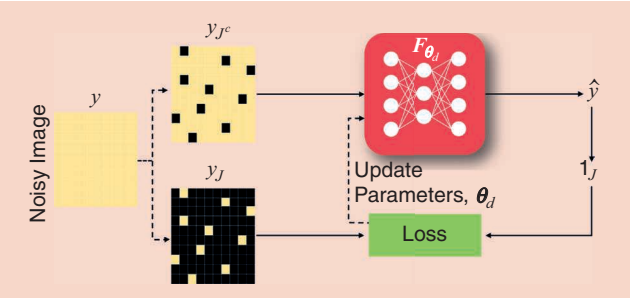

FIGURE 1. An overview of self-supervised learning for denoising. The black pixels denote masked-out locations in the images, while 1_J is the indicator function on the indices specified by the index set J.

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 $\Sigma_{i=1}^{N} |J_i| = m$, a function $F_{\boldsymbol{\theta}_d} : \mathbb{R}^m \to \mathbb{R}^m$ is $\mathcal{J}$-invariant if the value of $F_{\boldsymbol{\theta}_d}(\mathbf{y})_J$ does not depend on the value of $\mathbf{y}_J$ 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 $F_{\boldsymbol{\theta}_d}(\mathbf{y})_J$ 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]

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}} \|F_{\boldsymbol{\theta}_d}(\mathbf{y}) - \boldsymbol{y}\|^2 &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \|F_{\boldsymbol{\theta}_d}(\mathbf{y}) - \boldsymbol{x}\|^2 + \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \|\boldsymbol{y} - \boldsymbol{x}\|^2 \\ &\quad - 2\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \langle F_{\boldsymbol{\theta}_d}(\mathbf{y}) - \boldsymbol{y}, \boldsymbol{y} - \boldsymbol{x} \rangle \qquad (12) \\ &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \|F_{\boldsymbol{\theta}_d}(\mathbf{y}) - \boldsymbol{x}\|^2 + \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \|\boldsymbol{y} - \boldsymbol{x}\|^2 \\ &\quad - 2\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \langle F_{\boldsymbol{\theta}_d}(\mathbf{y}) - \boldsymbol{y}, \boldsymbol{y} - \boldsymbol{x} \rangle \qquad (13) \\ &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \|F_{\boldsymbol{\theta}_d}(\mathbf{y}) - \boldsymbol{x}\|^2 + \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \|\boldsymbol{y} - \boldsymbol{x}\|^2. \qquad (14)\end{aligned}$$

Note that for each pixel j in (13), random variables $F_{\boldsymbol{\theta}_d}(\boldsymbol{y})_j | \boldsymbol{x}$ and $\boldsymbol{y}_j | \boldsymbol{x}$ are independent if $F_{\boldsymbol{\theta}_d}$ is $\mathcal{J}$-invariant, while the noise is zero mean by assumption. Hence, the third term in (13) vanishes. Equation (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 learn a $\mathcal{J}$-invariant denoising function $F_{\boldsymbol{\theta}_d}$ over a database of single noisy images by minimizing the self-supervised loss

$$\underset{\boldsymbol{\theta}_d}{\operatorname{argmin}} \sum_{n=1}^{N} \sum_{J \in \mathcal{J}} \|F_{\boldsymbol{\theta}_d}(\mathbf{y}^n_{J^c}) - \boldsymbol{y}^n_J\|^2. \qquad (15)$$

Implementation-wise, it is not straightforward to just set the pixels specified by J to zero as this will affect the way convolutions are computed. Thus, during the training of self-supervised techniques such as N2V or N2S, the network takes $\boldsymbol{y}_{J^c} = \mathbf{1}_{J^c} \mathbf{y} + \mathbf{1}_J \kappa(\boldsymbol{y})$ as the 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 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 and 2) inputting partition $\mathcal{J}$ containing $|\mathcal{J}|$ sets and averaging them.

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 described in the previous section 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 data consistency operations. To this end, let $H_\Theta(\cdot)$ be the operator that outputs the measurement coordinates corresponding to index set Θ. In SSDU, hold-out masking is applied through these data consistency operations. Thus, while 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:

$$\min_{\boldsymbol{\theta}_r} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(\boldsymbol{y}^n_{\Theta^C}, H^n_{\Theta^C}(F_{\boldsymbol{\theta}_r}(\boldsymbol{y}^n_\Theta, H^n_\Theta))), \qquad (16)$$

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

Note that unlike in the denoising scenario, 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 a Fourier space, which contains most of the energy for images, for use in the data consistency units. However, a Gaussian density for masking requires a hyperparameter controlling its variance. Thus, in later works, SSDU was extended to a multimask setting [14], where multiple index sets $\{\Theta_l\}_{l=1}^{L}$ were used to define the loss

$$\min_{\boldsymbol{\theta}_r} \frac{1}{N} \sum_{n=1}^{N} \sum_{l=1}^{L} \mathcal{L}(\boldsymbol{y}^n_{\Theta_l^C}, H^n_{\Theta_l^C}(F_{\boldsymbol{\theta}_r}(\boldsymbol{y}^n_{\Theta_l}; H^n_{\Theta_l}))). \qquad (17)$$

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 hyperparameter. 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 remove 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 loss $\mathcal{L}$. Although the negative log likelihood, i.e., $c(\boldsymbol{x}, \boldsymbol{y})$, 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 which is described in (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

$$\min_{\boldsymbol{\theta}_c} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(\boldsymbol{y}^n, H^n(F_{\boldsymbol{\theta}_c}(\boldsymbol{y}^n; H^n))). \tag{18}$$

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

Biological applications

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-transmission electron microscopy, the acquired data sets are inherently noisy because the electron dose is restricted to avoid sample destruction [5]. Cryo-CARE [5] was the first work to show that N2N can be applied to cryotransmission electron microscopy data for denoising. Cryo-CARE was further applied on 3D cryoelectron tomogram data, showing its ability to denoise whole tomographic volumes. Several other works have also extended N2N for denoising cryoelectron microscopy (cryo-EM) data.

N2V was the first work to show that denoising can be performed using single noisy measurements. N2V has been extensively applied to electron microscopy data sets, showing improved reconstruction quality compared to conventional blind denoising methods, such as block-matching and 3D filtering (BM3D) [2]. In follow-up works, Bayesian postprocessing has been used to incorporate pixel-wise Gaussian- or histogram-based noise models [16] for further improvements to the denoising performance. However, their applications are limited as they require knowledge of the noise model, which might be challenging to know a priori in a number of applications. Moreover, the noise could be a mixture of noise type, hence further hindering their applications. A follow-up work [16] shows 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, N2S and its variants have also been applied to various microscopy data sets [3].

Figure 3 depicts denoising results using a conventional denoising algorithm (BM3D) and self-supervised learning algorithm (N2S) on two different microscopy data sets. These data sets contain only single noisy images, hence, supervised DL and N2N cannot be applied. The results show that self-supervised learning approaches visually improve denoising performance compared to conventional denoising algorithms.

Reconstruction

DL-based, ground-truth free reconstruction strategies have been applied in a 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 were used in non-Cartesian MRI [18], where pairs of undersampled measurements were used for training. Similar to a denoising scenario, the 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 (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.

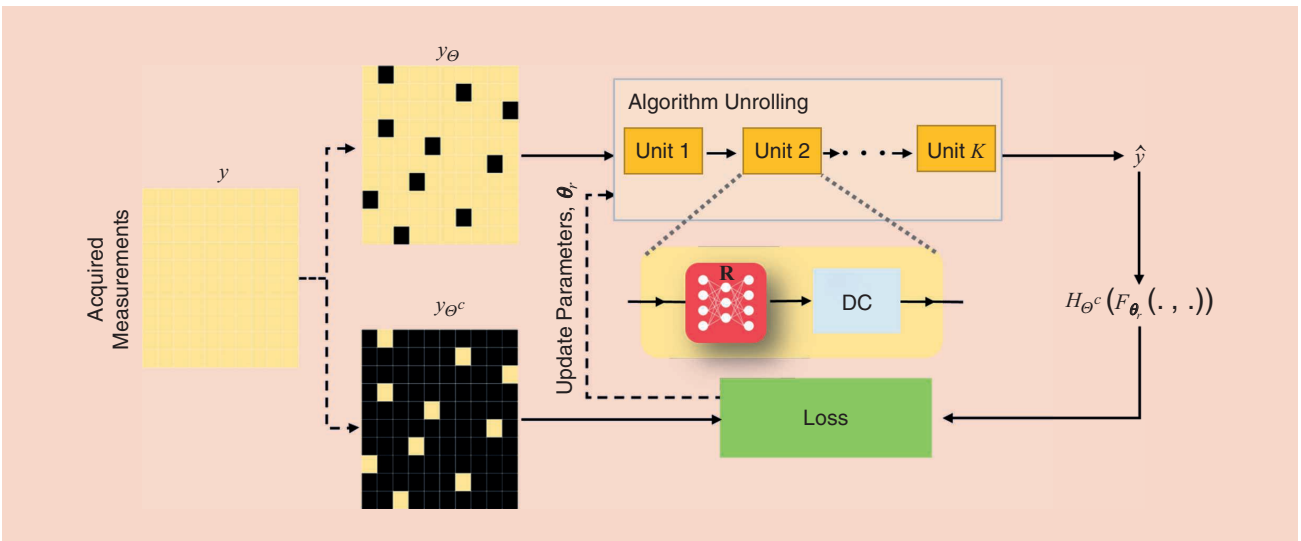

FIGURE 2. An overview of the self-supervised learning methods for image reconstruction using hold-out masking. The black pixels denote masked-out locations in the measurements, and *DC* denotes the data consistency units of the unrolled network.

Although such self-supervised methods have found use in medical imaging, their utility in biological imaging is just being explored. Recent work has started using such self-supervised DL methods for functional MRI (fMRI), which remains a critical biological imaging tool for neuroscientific discoveries that expand our understanding of human perception and cognition. In a recent work [6], multimask SSDU was applied to a Human Connectome Project-style fMRI acquisition, which was prospectively accelerated by fivefold, simultaneous multislice imaging and twofold 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 Figure 4 indicate that the self-supervised DL method based on multimask SSDU significantly outperforms conventional reconstruction approaches, both qualitatively in terms of visual quality, and quantitatively in terms of temporal signal-to-noise ratio.

Generative model-based methods

Overview

Generative models cover a large spectrum of research activities, which include variational autoencoders (VAEs), GANs, normalizing flow, and optimal transport (OT) [7]. Due to their popularity there are 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 Figure 5. Here, the ambient image space is $\mathcal{X}$, where we can take samples using real-data distribution μ. If the latent space is $\mathcal{Z}$, 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., $G \triangleq G_{\boldsymbol{\theta}}$. Let ζ be a fixed distribution on the latent space, such as a uniform or Gaussian distribution. Generator $G_{\boldsymbol{\theta}}$ pushes forward ζ to distribution $\mu_{\boldsymbol{\theta}} = G_{\boldsymbol{\theta}\#}\zeta$ in ambient space $\mathcal{X}$. Then, the goal of the generative model training is to

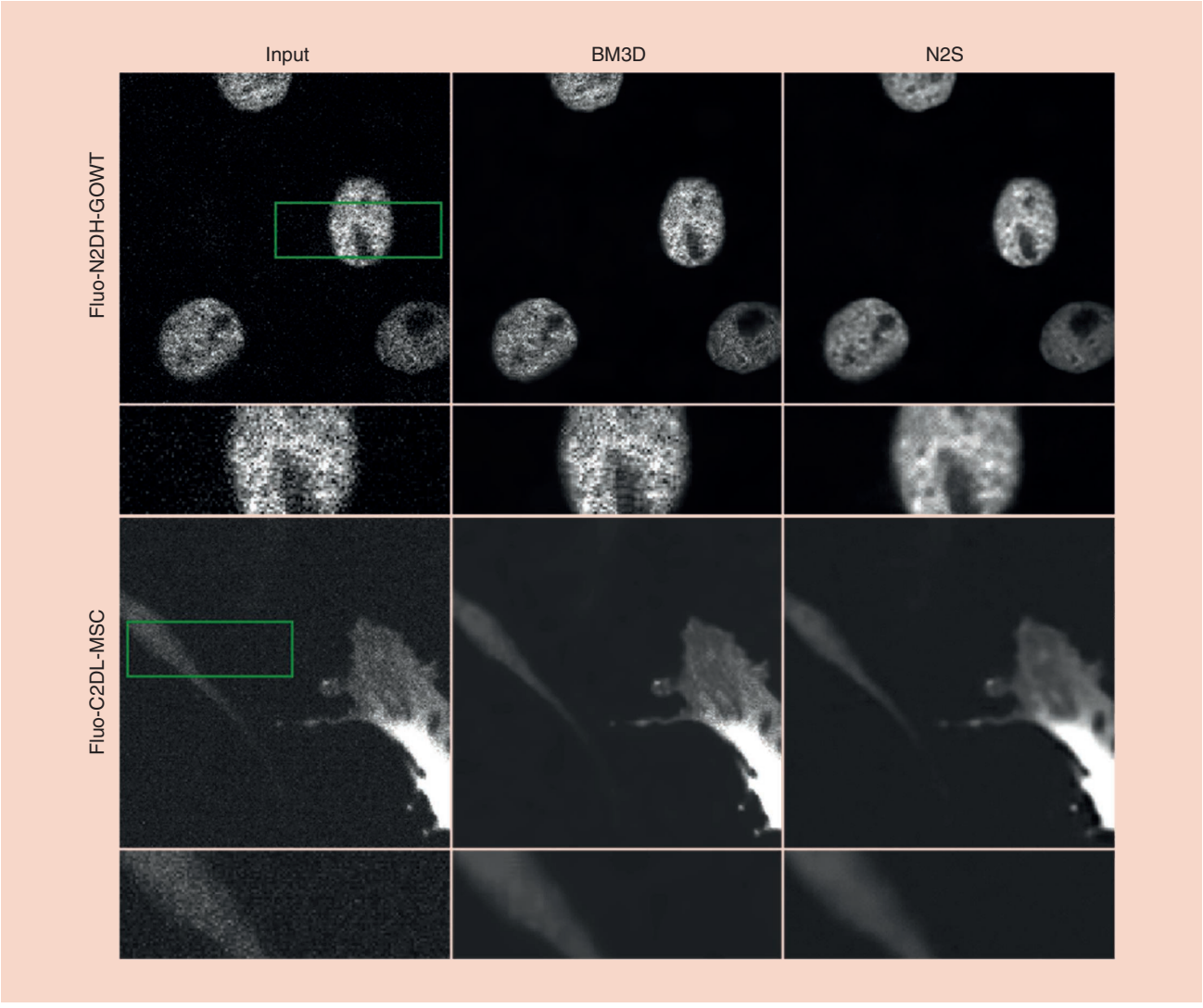

FIGURE 3. The denoising results from the Fluo-N2DH-GOWT1 and Fluo-C2DL-MSC fluorescence microscopy data sets using a traditional denoising method, BM3D, and a self-supervised learning method, (N2S). We note that supervised DL is not applicable as these data sets contain only single noisy images.

make μ_θ as close as possible to real-data distribution μ. Additionally, for the case of auto-encoding type-generative models (e.g., VAEs), 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 the sample space to the latent space. Accordingly, the additional constraint is again to minimize the distance, that is, $d(\zeta_\phi, \zeta)$. Using this unified geometric model, we can show that various types of generative models differ only in their choices of distances between μ_θ and μ, or ζ_ϕ and ζ and how they train the generator and encoder to minimize the distances.

VAE approaches for unsupervised learning in biological imaging

VAEs

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

$$\log p_\theta(\boldsymbol{x}) = \log\Big(\int p_\theta(\boldsymbol{x}|z)p(z)\,dz\Big). \tag{19}$$

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

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

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

In a VAE, by using the reparameterization trick together with the Gaussian assumption, one has

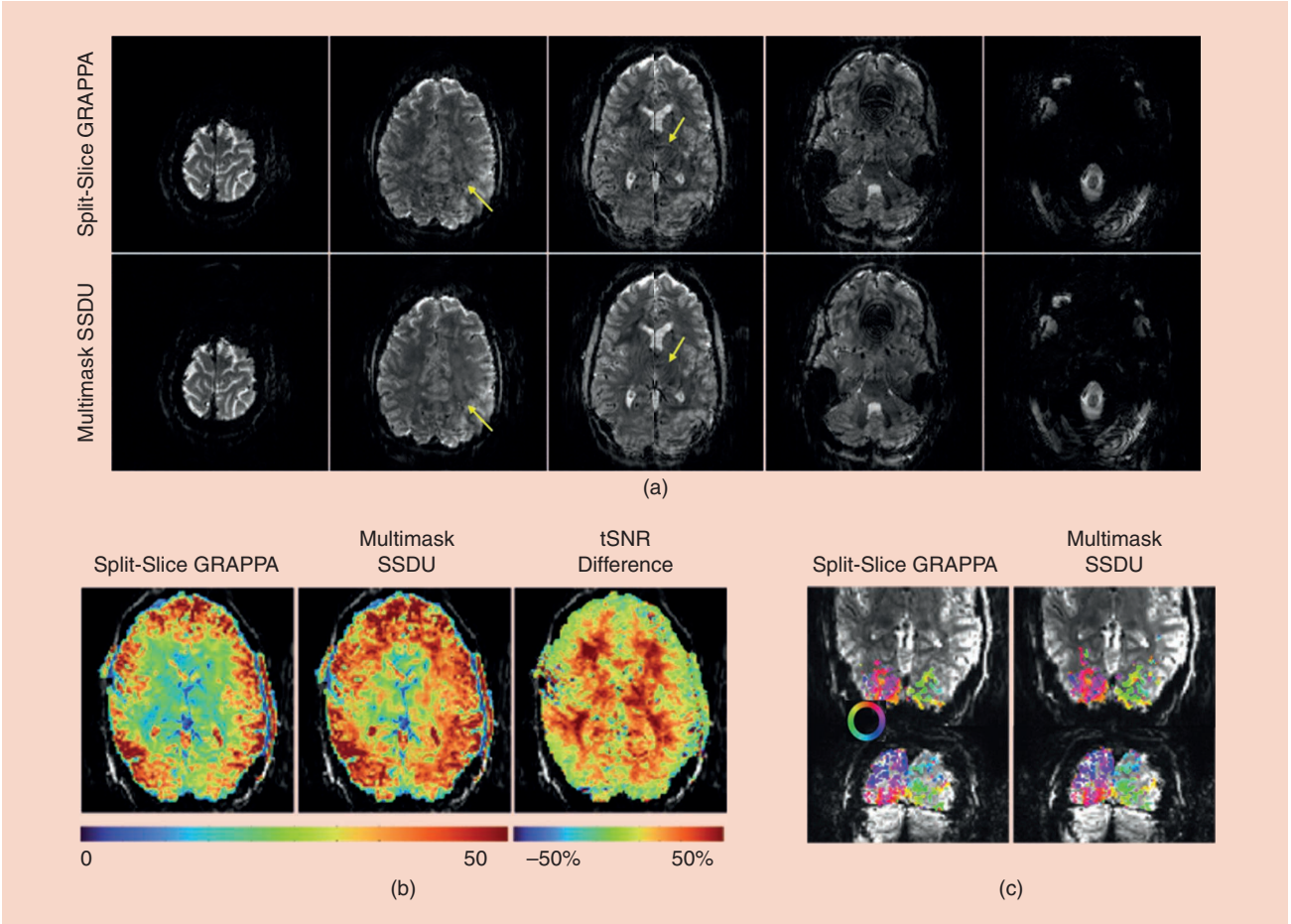

FIGURE 4. The reconstruction results from an fMRI application [6] using the conventional, split-slice GRAPPA technique and self-supervised multimask SSDU method [14]. (a) A split-slice GRAPPA exhibits residual artifacts in the mid-brain (yellow arrows). Multimask SSDU alleviates these, along with visible noise reduction. (b) Temporal signal-to-noise ratio (tSNR) maps show substantial gain with the self-supervised DL approach, particularly for the subcortical areas and cortex farther from the receiver coils. (c) Phase maps for the two reconstructions show strong agreement, with multimask SSDU containing more voxels above the coherence threshold.

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

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

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

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

Once the network is trained by minimizing (22), one notable advantage of a 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

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

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.

Biological applications

One notable application of VAEs in the field of biological imaging is the work of Bepler et al. [19]. The work is motivated by the problem of modeling continuous 2D views of proteins from single-particle electron microscopy. The goal of electron microscopy imaging is to estimate the 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 a Gaussian mixture model, which is discrete. Subsequently, modeling with Gaussian mixture models produces a suboptimal performance when aiming to model protein conformations. Hence, to bridge this gap, Bepler et al. [19] propose spatial-VAEs to disentangle projection rotation and translation from the content of the projections.

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

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

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

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

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

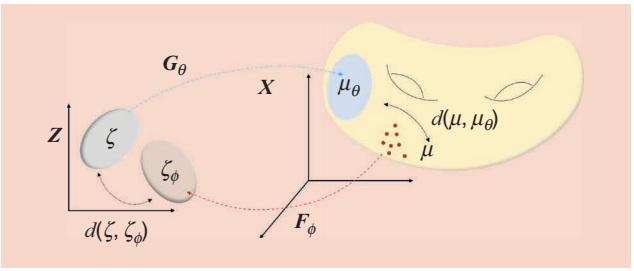

FIGURE 5. A geometric view of deep generative models. The fixed distribution ζ in $\mathcal{Z}$ is pushed to μ_{θ} in $\mathcal{X}$ by network G_{θ} so that mapped distribution μ_{θ} approaches real distribution μ. In a VAE, G_{θ} works as a decoder to generate samples, while F_{ϕ} acts as an encoder, additionally constraining ζ_{ϕ} to be close to ζ. With such a geometric view, auto-encoding generative models (e.g., VAEs) and GAN-based generative models can be seen as variants of this single illustration.

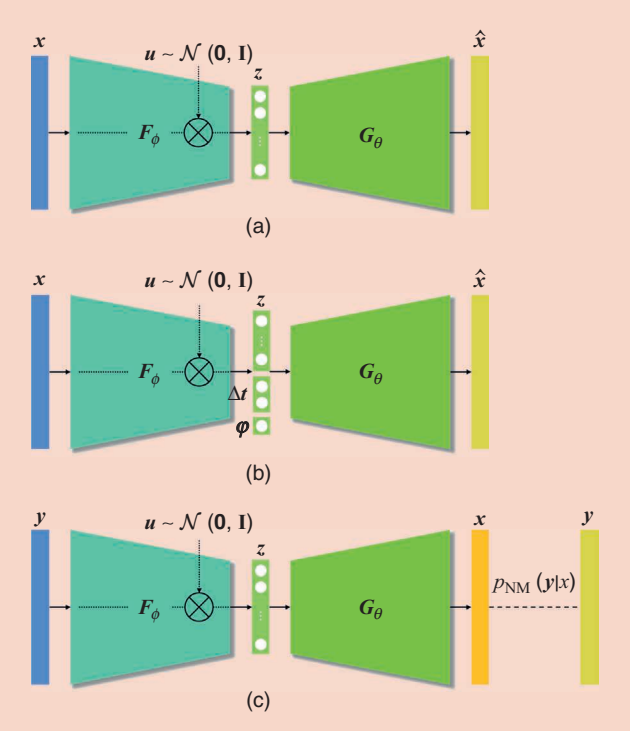

FIGURE 6. A VAE architecture. F_{ϕ} encodes $\boldsymbol{x}$ and is combined with random sample $\boldsymbol{u}$ to produce latent vector $\boldsymbol{z}$. G_{θ} decodes latent $\boldsymbol{z}$ to acquire $\hat{\boldsymbol{x}}$. $\boldsymbol{u}$ is sampled from a standard, normal distribution for the reparameterization trick. (a) A VAE. (b) A spatial-VAE [19], disentangling translation/rotation features from different semantics. (c) DIVNOISING [20], enabling supervised/unsupervised training of the denoising generative model by leveraging noise model $p_{\mathrm{NM}}(\boldsymbol{y}|\boldsymbol{x})$.

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

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

GAN approaches for unsupervised learning in biological imaging

Statistical distance minimization

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

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

where

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

Here, $D(\boldsymbol{x})$ is called the *discriminator*, which outputs a scalar in $[0, 1]$ representing the probability of input $\boldsymbol{x}$ being a real sample. Although 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 a GAN, we first provide a brief review of an *f*-GAN [8]. As the name suggests, an *f*-GAN starts with *f*-divergence as the statistical distance measurement:

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

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 τ mapping from $\mathcal{X}$ to $\mathbb{R}$, we have the lower bound

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

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 $V : \mathcal{X} \mapsto \mathbb{R}$ 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^*, an *f*-GAN can be formulated as follows:

$$\min_G \max_{g_f} \mathcal{L}_{f\mathrm{GAN}}(G, g_f), \tag{33}$$

where

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

Here, different choices of the functions f, g_f lead to distinct statistical measures and variations of *f*-GANs, and for the case of a 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 an *f*-GAN interprets the GAN training as a statistical distance minimization after dualization. A similar statistical distance-minimization idea is employed for W-GANs, but now with a real metric in probability space rather than the divergence. More specifically, a W-GAN minimizes the following Wasserstein-1 norm:

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

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

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

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

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 a pushforward measure, that is, $\nu=G_{\#}\mu$. Therefore, a Wasserstein 1-norm minimization problem can be equivalently represented by the following minmax formulation:

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

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

Biological applications

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

For example, in the context of tomographic reconstruction, a tomographic GAN (TomoGAN) [25] aims at low-dose tomographic reconstruction, where the generator takes in the noisy images from a low-dose sinogram as input and maps it into the distribution of high-dose images. Another model for 3D tomographic reconstruction technique based on a GAN, dubbed *GANrec*, was proposed in [26]. Different from a TomoGAN, a GANrec takes in the sinogram as input so that the generator also needs to learn the inverse mapping of the forward Radon transform. One unique aspect is that discriminator D learns the probability distribution of the clean sinogram. A similar approach is used for SR [27], [28]. Specifically, in [28], an SR method for Fourier ptychographic microscopy is introduced, which proposes the reconstruction of 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 subsequent temporal sequences. They also propose the use of a Fourier domain loss, imposing an additional constraint on the content. For SR microscopy, artificial neural network accelerated-PALM [27] was introduced to achieve high throughput in live-cell imaging, which is designed for accelerating PALM by using much fewer numbers of frames for restoring the true image.

These approaches that add conditions to GANs, in fact, correspond to pix2pix [21] or cGANs. Unlike the GANs illustrated in Figure 7(a), which take random noise vector z as input, pix2pix has additional loss function $\mathcal{L}_{\mathrm{content}}$, which measures the content distance [see Figure 7(b)]. Specifically, $\mathcal{L}_{\mathrm{content}}$ measures the content space distance between the generated and matched target images, which is used in addition to $\mathcal{L}_{\mathrm{GAN}}$, which measures the statistical distance. Therefore, pix2pix

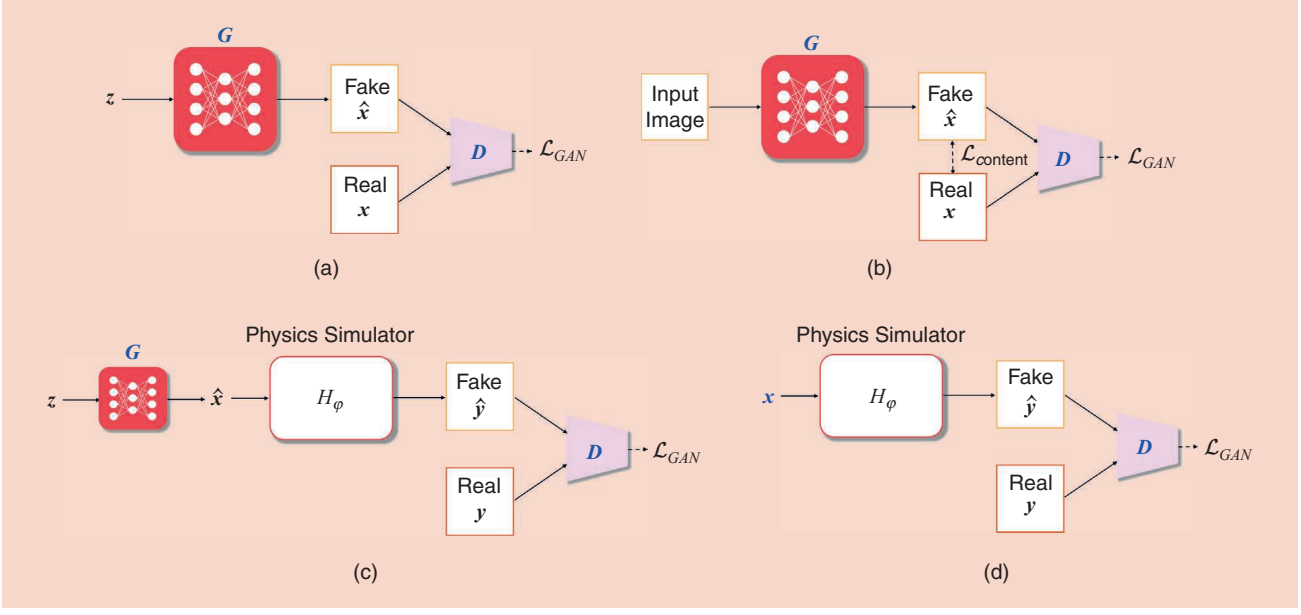

FIGURE 7. An illustration of GAN-based methods for biological image reconstruction. (a) A GAN, (b) pix2pix [21], (c) an ambientGAN [22], and (d) a cryoGAN [23]. $\boldsymbol{x}$ and $\boldsymbol{y}$ denote data in the image and measurement domains, respectively. G and D refer to the generator and discriminator, respectively. H defines the function family of the forward measurement process, parameterized with φ. The networks and variables marked in blue have learnable parameters optimized with gradient descent.

attempts to balance between the paired data and unpaired target distributions. In fact, the addition of content loss is important to regularize inverse problems. Unfortunately, the methods cannot be regarded as unsupervised as the content loss $\mathcal{L}_{\text{content}}$ 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 an ambientGAN [22], where the forward measurement model can be integrated into the framework. As in Figure 7(c), the generator of an ambientGAN generates a sample from a random noise vector, and the discriminator takes in the measurement after forward operator $H_{\boldsymbol{\varphi}}$ is parameterized by $\boldsymbol{\varphi}$, rather than the reconstructed image. As only the function family of the forward operator is known, 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, an ambientGAN enables training on the distribution, rather than on realized samples. From a statistical distance-minimization perspective, an ambientGAN can be interpreted as the dual problem for 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 OT problem that minimizes the 1-Wasserstein distance in the measurement space:

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

Then, the corresponding dual-cost function becomes

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

$$= \max_{D \in \text{Lip}_1(\mathcal{X})} \int_{\mathcal{Y}} D(\boldsymbol{y}) d\nu(\boldsymbol{y}) - \int_{\mathcal{X}} D(H_{\boldsymbol{\varphi}}(G(\boldsymbol{z}))) d\zeta(\boldsymbol{z}), \tag{39}$$

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 distribution $P_{\boldsymbol{\varphi}}$, (39) can be converted to

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

which is equivalent to the ambientGAN loss function.

In the original work of ambientGANs, simple forward measurement models such as convolve+noise, block+patch, 2D projection, and so on were used [22]. A variant of an ambientGAN was introduced in the context of cryo-EM in [23], dubbed a *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 to samples, multiple particles are frozen at cryogenic temperatures, and all the particles are simultaneously projected with a parallel electron beam to acquire projections. Here, unlike in the original ambientGAN, a cryoGAN considers the latent particle itself to be a learnable parameter. The overall flow of a cryoGAN is shown in Figure 7(d). It is interesting that there no generator exists in a cryoGAN; rather, $\boldsymbol{x}$, the 3D particle to be reconstructed, is the starting point of the overall flow. As in an ambientGAN, $\boldsymbol{x}$ goes through a complex, random forward measurement process, which involves 3D projection, convolution with the sampled kernel, and translation. The gradients from the discriminator backpropagates to $\boldsymbol{x}$, and $\boldsymbol{x}$ is updated directly at every optimization step. Unlike the conventional reconstruction methods for cryo-EM based on marginal-maximum likelihood, which demands estimation of the exact projection angles, a cryoGAN does not require such an expensive process. Note that the loss function of a cryoGAN is equivalent to (38). Therefore, by using the statistical distance-minimization approach, a 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 a cryoGAN, called a *multicryoGAN*. Although a 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 can hence have multiple conformations. To sidestep this issue, a multicryoGAN takes an approach more similar to the original ambientGAN, where a random noise vector is sampled from a distribution, and generator G is responsible for mapping the noise vector into the 3D particle. The rest of the steps follow the same procedure as in an ambientGAN, although the complicated forward measurement for cryo-EM is utilized. One advantage of a multicryoGAN is that once the networks are trained, multiple conformations of the particle can be sampled by varying noise vector $\boldsymbol{z}$. Subsequently, this introduces flexibility into 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 Figure 7(c); however, the input is not a random noise vector but an aliased image, inverse-Fourier-transformed from the undersampled 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 MRI, which corresponds to a 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, the reconstruction of fair quality could be performed.

OT-driven CycleGAN approaches for unsupervised learning for biological imaging

Another important line of work for unsupervised biological reconstruction comes from an OT-driven cycleGAN [10], which is a generalization of an original cycleGAN [9]. Unlike pix2pix, a cycleGAN does not utilize $\mathcal{L}_{\text{content}}$ from a paired label, so it is fully unsupervised. In contrast to an ambientGAN or cryoGAN, which are based on the statistical distance minimization in the measurement space, a cycleGAN attempts to minimize

the statistical distance in both the measurement and image domains simultaneously, which makes the algorithm more stable.

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

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

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

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

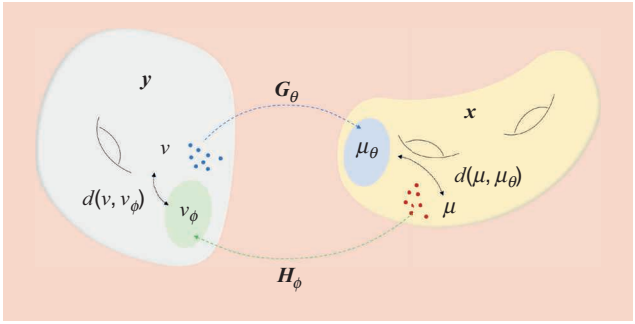

FIGURE 8. The geometric view of a cycleGAN. $(\mathcal{Y}, \nu)$ is mapped to $(\mathcal{X}, \mu)$ with G_θ, while H_φ does the opposite. The two are mappers, i.e., generators, are optimized by simultaneously minimizing $d(\mu, \mu_\theta), d(\nu, \nu_\varphi)$.

If we minimize them separately, the optimal joint distribution π^* 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:

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

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

$$\min_{\boldsymbol{\theta}, \boldsymbol{\varphi}} \max_{D_X, D_Y} \mathcal{L}_{\text{cycleGAN}}(\boldsymbol{\theta}, \boldsymbol{\varphi}; D_X, D_Y), \tag{44}$$

where

$$\mathcal{L}_{\text{cycleGAN}}(\boldsymbol{\theta}, \boldsymbol{\varphi}; D_X, D_Y) \triangleq \lambda \mathcal{L}_{\text{cycle}}(\boldsymbol{\theta}, \boldsymbol{\varphi}) + \mathcal{L}_{\text{GAN}}(\boldsymbol{\theta}, \boldsymbol{\varphi}; D_X, D_Y), \tag{45}$$

where $\mathcal{L}_{\text{cycle}}$ and $\mathcal{L}_{\text{GAN}}$ refer to the cycle-consistency and discriminator GAN losses, respectively. D_X and D_Y are discriminators in $\mathcal{X}$ and $\mathcal{Y}$, respectively. The corresponding OT-cycleGAN network architecture can be represented, as displayed in Figure 9.

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

In addition, compared to the ambientGAN in (37), the OT-cycleGAN primal formulation in (43) has an additional term, $\|\boldsymbol{x} - G_{\boldsymbol{\theta}}(\boldsymbol{y})\|$, which forces the reconstruction images to match the target-image distributions, further regularizing the reconstruction process. In fact, the resulting OT-cycleGAN

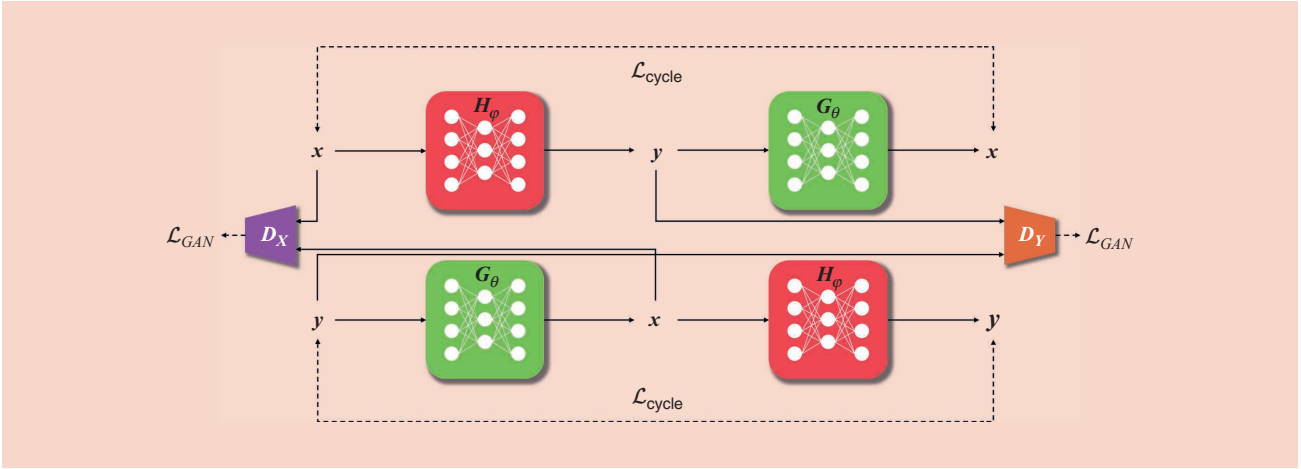

FIGURE 9. The network architecture of a cycleGAN. $G_\theta: \mathcal{Y} \mapsto \mathcal{X}, H_\varphi: \mathcal{X} \mapsto \mathcal{Y}$ are the generators responsible for interdomain mapping. D_X, D_Y are discriminators, constructing $\mathcal{L}_{\text{GAN}}$. The GAN loss is simultaneously optimized together with $\mathcal{L}_{\text{cycle}}$.

formulation is closely related to the classical RLS formulation in (2). Specifically, the transportation cost in (43) closely resembles the cost function in (2), except that the regularization term $R(\boldsymbol{x})$ in (2) is replaced by DL-based, inverse-path penalty term $\|\boldsymbol{x} - G_{\boldsymbol{\theta}}(\boldsymbol{y})\|$. However, instead of solving $\boldsymbol{x}$ directly as in (2), an OT-cycleGAN tries to find joint distribution π^*, which minimizes the average cost for all combinations of $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$. This suggests that an OT-cycleGAN is a stochastic generalization of RLS, revealing an important link to the classical RLS approaches.

Applications

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

For example, in [11], a cycleGAN was used with a linear blur kernel for blind and nonblind deconvolutions. 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. Hence, even for the nonblind case, forward mapping $H_{\boldsymbol{\varphi}} : \mathcal{X} \mapsto \mathcal{Y}$ is partially known as a *linear convolution*. Leveraging this property, one of the generators in a cycleGAN, F in Figure 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. What is more, in the case of nonblind microscopy, it was shown that the forward mapping is deterministic so that optimization with respect to discriminator D_Y is no longer necessary, which simplifies the network architecture and makes the training process more robust. A similar simplification of a cycleGAN leveraging the imaging physics of microscopy was also proposed in SR microscopy [10]. Interestingly, the simplified form of a cycleGAN could generate reconstructions of higher resolution, quantified in a Fourier ring correlation. Other than simplifying mapping $H_{\boldsymbol{\varphi}} : \mathcal{X} \mapsto \mathcal{Y}$ with a linear blind kernel, a deterministic k-space subsampling operator for MRI was extensively studied [30].

When such a simplification is not possible, the most generalized form of a cycleGAN—where two sets of generator/discriminator pair are used—can be utilized, but still, the key concept of statistical distance minimization can be employed in the design. One work, which utilizes a cycleGAN for deconvolution microscopy is [32], where the authors propose the use of 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 three-way volume averaging of the reconstructed results in the $x - y$, $y - z$, and $x - z$ planes. However, in contrast to [11], two neural network-based generators are used for both the forward and inverse paths. In another work, an unsupervised reconstruction method called a *ProjectionGAN for ODT* was proposed [31]. A missing cone problem in an ODT arises because the measurement angles of the imaging device do 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 a parallel beam projection is performed on the 3D distribution of the 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 ($\mathcal{Y}_{\Omega^c}$). As a result, the resolution of the blurry projections is enhanced via distribution matching between $\mathcal{Y}_{\Omega}$ and $\mathcal{Y}_{\Omega^c}$ with a cycleGAN, after which follows filtered backprojection to acquire the final reconstruction from the enhanced projections. Using the ProjectionGAN enhancement step, the missing cone artifacts are greatly resolved, achieving accurate reconstruction, as illustrated in Figure 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 (the $x - z$ and $y - z$ planes). This problem is greatly resolved with ProjectionGAN, where we observe clear boundaries and microcellular structures. The underestimated RI values are also corrected.

For optical microscopy, a content-preserving cycleGAN (c^2GAN) was proposed [33], showing the applicability of a cycleGAN to various imaging modalities and data configurations. A c^2GAN introduces a saliency constraint to the 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 are covered with a low-intensity background. Thus, by adding the saliency constraint, a cycleGAN can concentrate more on the salient features. The authors applied a c^2GAN to biological image denoising, restoration, SR, histological colorization, and image translations such as phase-contrast images to flourescence-labeled images, showing how a cycleGAN can be easily adapted to many different tasks of biological imaging.

Discussion

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 has been largely influenced by the application-specific tailoring of 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 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 between an underlying target distribution and the generated data distribution. Although these goals do not necessarily appear complementary, there are

self-supervisory methods, such as content generation, which utilize the properties of generative models. Similarly, there are generative models that employ concepts of the prediction of data from self-supervision [34]. Thus, a synergistic viewpoint that ties together 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 training on large data sets containing only noisy or incomplete measurements. However, in some biological applications, it may not always be feasible to obtain large training data

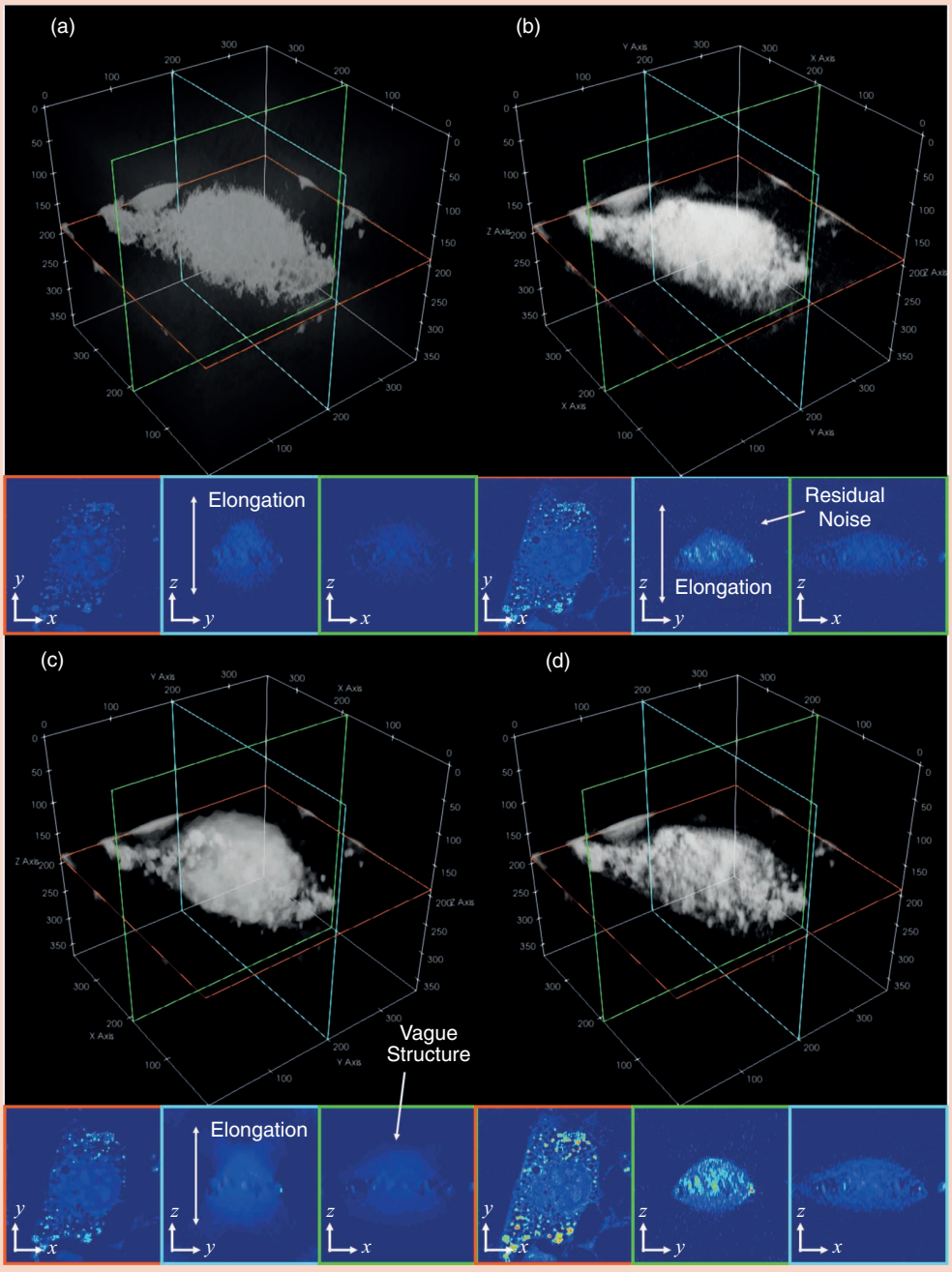

FIGURE 10. A ProjectionGAN for the reconstruction of ODT [31]. (a) A conventional Rytov reconstruction via Fourier binning, (b) a Gerchberg–Papoulis (GP) algorithm, (c) a model-based iterative method using the total variation (TV), and (d) reconstruction via a ProjectionGAN. Artifacts, including elongation along the optical axes, can be seen in the $x-z$ and $y-z$ cutviews of (a) and (c). The result shown in (b) is contaminated with residual noise in the $x-z$ and $y-z$ planes. The result shown in (d) has high-resolution reconstruction without such artifacts, along with boosted RI values.

sets. It is therefore 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 to avoid overfitting [35]. For image reconstruction, a zero-shot, self-supervised learning approach has been offered to split the available measurements, two of which are used in data consistency and 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 data sets.

In the context of generative models, two closely related methods, score-based [37] and diffusion models [38], have recently garnered attention with their outstanding ability to train generative models without any adversarial training. Remarkably, one can not 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. As 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 the generative modeling of material structures [40]. Although seemingly unrelated, the fundamental problem of biological image reconstruction and enhancement can be viewed as disentangling a 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.

Availability of training databases

Although early works in biological imaging applications relied on utilizing imaging data sets 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. BioImage Archive, Image Data Resources, BioImage.IO, and Electron Microscopy Public Image Archive constitute some of these efforts. Moreover, there are platforms such as Zenodo and Figshare, which 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 the development of DL-based reconstruction and enhancement strategies.

Conclusions

DL methods have recently become state-of-the-art approaches for image reconstruction. Although 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 themselves to train the model. A 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 approaches for image denoising and reconstruction, with VAEs and GANs being the two most prominent strategies. Both techniques can be seen as the optimization problem of statistical minimization, with different choices for statistical distance measurement, 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 National Institutes of Health (NIH) R01HL153146, NIH P41EB027061, NIH R21EB028369, National Science Foundation CAREER CCF-1651825. This work was also partially supported by the National Research Foundation of Korea under Grant NRF-2020R1A2B5B03001980. Mehmet Akçakaya is the corresponding author.

Authors

Mehmet Akçakaya (akcakaya@umn.edu) received his Ph.D. degree from Harvard University, Cambridge, Massachusetts, USA, in 2010. He is an associate professor at the University of Minnesota, Minneapolis, Minnesota, 55455, USA. His work on accelerated magnetic resonance imaging (MRI) has received a number of international recognitions. He holds an R01 Award and a Trailblazer Award from the National Institutes of Health and a CAREER Award from the National Science Foundation. His research interests include image reconstruction, machine learning, MRI physics, inverse problems, and signal processing. He is a Senior Member of IEEE.

Burhaneddin Yaman (yaman013@umn.edu) received his B. Eng. degree from Istanbul Technical University, Istanbul, Turkey, in 2016. He is currently pursuing a Ph.D. degree in electrical engineering at the University of Minnesota, Minneapolis, 55455, Minnesota, USA. He was awarded the Best Paper Award at the 2020 IEEE 17th International Symposium on Biomedical Imaging. His research interests include self-supervised learning, magnetic resonance imaging, and tensor decompositions. He is a Student Member of IEEE.

Hyungjin Chung (hj.chung@kaist.ac.kr) received his M.Eng. degree from Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 2021. He has published several works on machine learning for computational imaging.

His current research interests include energy-based generative models and inverse problems. He is a Student Member of IEEE.

Jong Chul Ye (jong.ye@kaist.ac.kr) received his Ph.D. from Purdue University, West Lafayette, Indiana, USA. He is a professor at Korea Advanced Institute of Science and Technology, Daejeon, Korea. He is an associate editor of *IEEE Transactions on Medical Imaging* and a senior editor of *IEEE Signal Processing Magazine*. He is chair of the IEEE Signal Processing Society Computational Imaging Technical Committee and a Distinguished Lecturer for the IEEE Engineering in Medicine and Biology Society. He was a general cochair for 2020 IEEE Symposium On Biomedical Imaging (with Mathews Jacob). His current research interests focus on machine learning theory and algorithms for various image reconstruction problems. He is a Fellow of IEEE.

References

[1] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, 2020. doi: 10.1109/TPAMI.2020.2992393.

[2] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2Void – Learning denoising from single noisy images," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 2129–2137. doi: 10.1109/CVPR.2019.00223.

[3] J. Batson and L. Royer, "Noise2Self: Blind denoising by self-supervision," in *Proc. Int. Conf. Mach. Learn.,* 2019, pp. 524–533.

[4] B. Yaman, S. A. H. Hosseini, S. Moeller, J. Ellermann, K. Uğurbil, and M. Akçakaya, "Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data," *Magn. Reson. Med.*, vol. 84, no. 6, pp. 3172–3191, Dec. 2020. doi: 10.1002/mrm.28378.

[5] T.-O. Buchholz, A. Krull, R. Shahidi, G. Pigino, G. Jékely, and F. Jug, "Content-aware image restoration for electron microscopy," *Methods Cell Biol.*, vol. 152, pp. 277–289, July 2019. doi: 10.1016/bs.mcb.2019.05.001.

[6] O. B. Demirel et al., "Improved simultaneous multi-slice functional MRI using self-supervised deep learning," 2021, arXiv:2105.04532.

[7] L. Ruthotto and E. Haber, "An introduction to deep generative modeling," 2021, arXiv:2103.05180.

[8] S. Nowozin, B. Cseke, and R. Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, vol. 29, pp. 271–279.

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2223–2232. doi: 10.1109/ICCV.2017.244.

[10] B. Sim, G. Oh, J. Kim, C. Jung, and J. C. Ye, "Optimal transport driven CycleGAN for unsupervised learning in inverse problems," *SIAM J. Imag. Sci.*, vol. 13, no. 4, pp. 2281–2306, 2020. doi: 10.1137/20M1317992.

[11] S. Lim, H. Park, S.-E. Lee, S. Chang, B. Sim, and J. C. Ye, "CycleGAN with a blur kernel for deconvolution microscopy: Optimal transport geometry," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1127–1138, July 2020. doi: 10.1109/TCI.2020.3006735.

[12] F. Knoll, K. Hammernik, C. Zhang, S. Moeller, T. Pock, D. K. Sodickson, and M. Akçakaya, "Deep-learning methods for parallel magnetic resonance imaging reconstruction: A survey of the current approaches, trends, and issues," *IEEE Signal Process. Mag.*, vol. 37, no. 1, pp. 128–140, 2020. doi: 10.1109/MSP.2019.2950640.

[13] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2Noise: Learning image restoration without clean data," in *Proc. 35th Int. Conf. Mach. Learn.*, PMLR, 2018, vol. 80, pp. 2965–2974.

[14] B. Yaman, S. A. H. Hosseini, S. Moeller, J. Ellermann, K. Uğurbil, and M. Akçakaya, "Ground-truth free multi-mask self-supervised physics-guided deep learning in highly accelerated MRI," in *Proc. 2021 IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, pp. 1850–1854. doi: 10.1109/ISBI48211.2021.9433924.

[15] O. Senouf, S. Vedula, T. Weiss, A. Bronstein, O. Michailovich, and M. Zibulevsky, "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*, Q. Wang et al., Eds. Cham: Springer Nature Switzerland AG, 2019, pp. 111–119.

[16] A. Krull, T. Vičar, M. Prakash, M. Lalit, and F. Jug, "Probabilistic noise2void: Unsupervised content-aware denoising," *Frontiers Comput. Sci.*, vol. 2, no. 5, pp. 1–9, 2020. doi: 10.3389/fcomp.2020.00005.

[17] M. Prakash, M. Lalit, P. Tomancak, A. Krul, and F. Jug, "Fully unsupervised probabilistic noise2void," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, 2020, pp. 154–158. doi: 10.1109/ISBI45749.2020.9098612.

[18] J. Liu, Y. Sun, C. Eldeniz, W. Gan, H. An, and U. S. Kamilov, "RARE: Image reconstruction using deep priors learned without groundtruth," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 6, pp. 1088–1099, 2020. doi: 10.1109/JSTSP.2020.2998402.

[19] T. Bepler, E. D. Zhong, K. Kelley, E. Brignole, and B. Berger, "Explicitly disentangling image content from translation and rotation with spatial-VAE," 2019, arXiv:1909.11663.

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

[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1125–1134. doi: 10.1109/CVPR.2017.632.

[22] A. Bora, E. Price, and A. G. Dimakis, "AmbientGAN: Generative models from lossy measurements," in *Proc. 6th Int. Conf. Learn. Representations,* ICLR, 2018.

[23] H. Gupta, M. T. McCann, L. Donati, and M. Unser, "CryoGAN: A new reconstruction paradigm for single-particle cryo-EM via deep adversarial learning," *IEEE Trans. Comput. Imag.,* vol. 7, pp. 759–774, 2021, doi: 10.1109/TCI.2021.3096491.

[24] E. Nehme, L. E. Weiss, T. Michaeli, and Y. Shechtman, "Deep-STORM: Super-resolution single-molecule microscopy by deep learning," *Optica*, vol. 5, no. 4, pp. 458–464, 2018. doi: 10.1364/OPTICA.5.000458.

[25] Z. Liu, T. Bicer, R. Kettimuthu, D. Gursoy, F. De Carlo, and I. Foster, "TomoGAN: Low-dose synchrotron X-ray tomography with generative adversarial networks: Discussion," *J. Opt. Soc. Amer. A*, vol. 37, no. 3, pp. 422–434, 2020. doi: 10.1364/JOSAA.375595.

[26] X. Yang et al., "Tomographic reconstruction with a generative adversarial network," *J. Synchrotron Radiat.*, vol. 27, no. 2, pp. 486–493, 2020. doi: 10.1107/S1600577520000831.

[27] W. Ouyang, A. Aristov, M. Lelek, X. Hao, and C. Zimmer, "Deep learning massively accelerates super-resolution localization microscopy," *Nature Biotechnol.*, vol. 36, no. 5, pp. 460–468, 2018. doi: 10.1038/nbt.4106.

[28] T. Nguyen, Y. Xue, Y. Li, L. Tian, and G. Nehmetallah, "Deep learning approach for Fourier ptychography microscopy," *Opt. Express*, vol. 26, no. 20, pp. 26,470–26,484, 2018. doi: 10.1364/OE.26.026470.

[29] H. Gupta, T. H. Phan, J. Yoo, and M. Unser, "Multi-CryoGAN: Reconstruction of continuous conformations in Cryo-EM using generative adversarial networks," in *European Conference on Computer Vision*, A. Bartoli and A. Fusiello, Eds. Cham: Springer Nature Switzerland AG, 2020, pp. 429–444.

[30] G. Oh, B. Sim, H. Chung, L. Sunwoo, and J. C. Ye, "Unpaired deep learning for accelerated MRI using optimal transport driven cycleGAN," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1285–1296, Aug. 2020. doi: 10.1109/TCI.2020.3018562.

[31] H. Chung, J. Huh, G. Kim, Y. K. Park, and J. C. Ye, "Unsupervised missing cone deep learning in optical diffraction tomography," 2021, arXiv:2103.09022.

[32] S. Lee, S. Han, P. Salama, K. W. Dunn, and E. J. Delp, "Three dimensional blind image deconvolution for fluorescence microscopy using generative adversarial networks," in *Proc. 2019 IEEE 16th Int. Symp. Biomed. Imag. (ISBI 2019)*, pp. 538–542. doi: 10.1109/ISBI.2019.8759250.

[33] X. Li et al., "Unsupervised content-preserving transformation for optical microscopy," *Light, Sci. Appl.*, vol. 10, no. 1, p. 44, 2021. doi: 10.1038/s41377-021-00484-y.

[34] E. Bostan, R. Heckel, M. Chen, M. Kellman, and L. Waller, "Deep phase decoder: Self-calibrating phase microscopy with an untrained deep neural network," *Optica*, vol. 7, no. 6, pp. 559–562, 2020. doi: 10.1364/OPTICA.389314.

[35] Y. Quan, M. Chen, T. Pang, and H. Ji, "Self2self with dropout: Learning self-supervised denoising from single image," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 1890–1898. doi: 10.1109/CVPR42600.2020.00196.

[36] B. Yaman, S. A. H. Hosseini, and M. Akçakaya, "Zero-shot self-supervised learning for MRI reconstruction," 2021, arXiv:2102.07737.

[37] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 11918–11930.

[38] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 6840–6851.

[39] K. Kim and J. C. Ye, "Noise2Score: Tweedie's approach to self-supervised image denoising without clean images," 2021, arXiv:2106.07009.

[40] H. Chung and J. C. Ye, "Feature disentanglement in generating three-dimensional structure from two-dimensional slice with sliceGAN," 2021, arXiv:2105.00194.

SP