Journal of the Operations Research Society of China (2020) 8:311–340
https://doi.org/10.1007/s40305-019-00287-4

A Review on Deep Learning in Medical Image Reconstruction

Hai-Miao Zhang[1] · **Bin Dong**[1]

Received: 22 June 2019 / Revised: 3 November 2019 / Accepted: 27 November 2019 /
Published online: 10 January 2020

Abstract

Medical imaging is crucial in modern clinics to provide guidance to the diagnosis and treatment of diseases. Medical image reconstruction is one of the most fundamental and important components of medical imaging, whose major objective is to acquire high-quality medical images for clinical usage at the minimal cost and risk to the patients. Mathematical models in medical image reconstruction or, more generally, image restoration in computer vision have been playing a prominent role. Earlier mathematical models are mostly designed by human knowledge or hypothesis on the image to be reconstructed, and we shall call these models handcrafted models. Later, handcrafted plus data-driven modeling started to emerge which still mostly relies on human designs, while part of the model is learned from the observed data. More recently, as more data and computation resources are made available, deep learning based models (or deep models) pushed the data-driven modeling to the extreme where the models are mostly based on learning with minimal human designs. Both handcrafted and data-driven modeling have their own advantages and disadvantages. Typical handcrafted models are well interpretable with solid theoretical supports on the robustness, recoverability, complexity, etc., whereas they may not be flexible and sophisticated enough to fully leverage large data sets. Data-driven models, especially deep models, on the other hand, are generally much more flexible and effective in extracting useful information from large data sets, while they are currently still in lack of theoretical foundations. Therefore, one of the major research trends in medical imaging is to combine handcrafted modeling with deep modeling so that we can enjoy

The work of Hai-Miao Zhang was funded by China Postdoctoral Science Foundation (No. 2018M641056). The work of Bin Dong was supported in part by the National Natural Science Foundation of China (No. 11831002), and Natural Science Foundation of Beijing (No. Z180001).

✉ Bin Dong
dongbin@math.pku.edu.cn

Hai-Miao Zhang
hmzhang@pku.edu.cn

[1] Beijing International Center for Mathematical Research, Peking University, Beijing 100871, China

benefits from both approaches. The major part of this article is to provide a conceptual review of some recent works on deep modeling from the unrolling dynamics viewpoint. This viewpoint stimulates new designs of neural network architectures with inspirations from optimization algorithms and numerical differential equations. Given the popularity of deep modeling, there are still vast remaining challenges in the field, as well as opportunities which we shall discuss at the end of this article.

Keywords Medical imaging · Deep learning · Unrolling dynamics · Handcrafted modeling · Deep modeling · Image reconstruction

Mathematics Subject Classification 60H10 · 92C55 · 93C15 · 94A08

1 Introduction

Medical image reconstruction can often be formulated as the following mathematical problem:

$$\boldsymbol{f} = \boldsymbol{A}\boldsymbol{u} \oplus \boldsymbol{\eta}, \tag{1.1}$$

where $\boldsymbol{A}$ is a physical system modeling the image acquisition process. Operator $\boldsymbol{A}$ can be a linear operator or nonlinear operator that depends on the specific imaging modality. Variable $\boldsymbol{u}$ is the unknown image to be reconstructed, and $\boldsymbol{f}$ is the measured data that might be contaminated by noise $\boldsymbol{\eta}$ with known or partially known noise statistics, e.g., Gaussian, Laplacian, Poisson, Rayleigh, etc. The operator $\oplus$ is a notation to denote addition when Gaussian noise is assumed, a certain nonlinear operator when Poisson noise or Rician noise is assumed. In different image reconstruction tasks, $\boldsymbol{A}$ takes different forms:

- Denoising: $\boldsymbol{A}$ is an identity operator.
- Deblurring: $\boldsymbol{A}$ is a convolution operator. When the convolution kernel is unknown, the problem is called blind deblurring [1].
- Inpainting: $\boldsymbol{A}$ is a restriction operator which can be represented by a diagonal matrix with value 0 or 1 [2].
- Magnetic resonance imaging (MRI): $\boldsymbol{A}$ is a subsampled Fourier transform which is a composition of the Fourier transform and a binary sampling operator [3].
- X-ray based computed tomography (CT): $\boldsymbol{A}$ is a subsampled Radon transform, which is a partial collection of line integrations [4].
- Quantitative susceptibility mapping (QSM) [5–8]: $\boldsymbol{A}$ is the dipole kernel

$$\boldsymbol{A}(X) = \frac{2z^2 - x^2 - y^2}{4\pi(x^2 + y^2 + z^2)^{5/2}}, \ X = (x, y, z) \in \mathbb{R}^3.$$

The inverse problem (1.1) is in general challenging to solve due to the large-scale and ill-posed nature of the problem in practice.

1.1 Image Reconstruction Models

The above inverse problem (1.1) covers a wide range of image restoration tasks which are not limited to medical image reconstruction. To solve the inverse problem (1.1), it is common practice to consider the following optimization problem:

$$\min_{\boldsymbol{u}\in\mathcal{D}} \mathscr{L}(\boldsymbol{u}) = F(\boldsymbol{A}\boldsymbol{u}, \boldsymbol{f}) + \lambda\Phi(\boldsymbol{W}, \boldsymbol{u}). \tag{1.2}$$

The solution $\boldsymbol{u}^\star \in \arg\min_{\boldsymbol{u}} \mathscr{L}(\boldsymbol{u})$ is an approximate solution to the inverse problem (1.1). In (1.2), the term $F(\boldsymbol{A}\boldsymbol{u}, \boldsymbol{f})$ is the data fidelity term that measures the consistency of the approximate solution to the measured data $\boldsymbol{f}$. Its specific form normally depends on the noise statistics. For example:

- Gaussian noise: $F(\boldsymbol{A}\boldsymbol{u}, \boldsymbol{f}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f}\|_2^2$,
- Poisson noise: $F(\boldsymbol{A}\boldsymbol{u}, \boldsymbol{f}) = \langle \boldsymbol{1}, \boldsymbol{A}\boldsymbol{u}\rangle - \langle \boldsymbol{f}, \log(\boldsymbol{A}\boldsymbol{u})\rangle$, with $\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \sum_i \boldsymbol{a}_i\boldsymbol{b}_i$,
- Impulse noise: $F(\boldsymbol{A}\boldsymbol{u}, \boldsymbol{f}) = \|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f}\|_1$,
- Multiplicative noise [9]: $F(\boldsymbol{A}\boldsymbol{u}, \boldsymbol{f}) = \lambda_1 \left\langle \frac{\boldsymbol{A}\boldsymbol{f}}{\boldsymbol{u}}, \boldsymbol{1} \right\rangle + \lambda_2 \left\| \frac{\boldsymbol{A}\boldsymbol{f}}{\boldsymbol{u}} - \boldsymbol{1} \right\|^2$.

The second term $\Phi(\boldsymbol{W}, \boldsymbol{u})$ in (1.2) is the regularization term encoding the prior knowledge on the image to be reconstructed. The regularization term is often the most crucial part of the modeling, and what people have mostly focused on in the literature. The parameter λ in (1.2) provides a balance between the data fidelity term and the regularization term. Mathematical modeling has been playing a vital role in solving such inverse problems. Interested readers can refer to [10–13] for more extensive reviews on mathematical models for image restoration.

Deep learning models can also be casted into the form of (1.2). However, there are differences as well. In handcraft or handcraft + data-driven modeling, the transformation $\boldsymbol{W}$ is often a certain linear transformation that is able to extract sparse features from the images. In handcraft models, $\boldsymbol{W}$ is often given by design (e.g., a differential operator or wavelet transform); in handcraft + data-driven models, $\boldsymbol{W}$ (or a portion of it) is often learned from the given data. Sparsity is the key to the success of these models. Deep learning models follow a similar modeling philosophy by considering nonlinear sparsifying transformations rather than linear ones. In general, we define a parameterized nonlinear mapping $\mathscr{V}(\cdot, \boldsymbol{\Theta}) : \mathscr{F} \to \mathscr{U}$, $\boldsymbol{f} \mapsto \boldsymbol{u}$ that maps the input data $\boldsymbol{f}$ to a high-quality output image $\boldsymbol{u}$. The mapping $\mathscr{V}$ is parameterized by $\boldsymbol{\Theta}$ which is trained on a given data set $\mathscr{F} \times \mathscr{U}$ by solving the following optimization problem:

$$\min_{\Theta} \frac{1}{\#(\mathscr{F} \times \mathscr{U})} \sum_{(\boldsymbol{f},\boldsymbol{u})\in\mathscr{F}\times\mathscr{U}} \mathscr{C}(\mathscr{V}(\boldsymbol{f}, \boldsymbol{\Theta}), \boldsymbol{u}),$$

where $\mathscr{C}(\cdot, \cdot)$ is a metric of difference between the approximated image $\mathscr{V}(\boldsymbol{f}, \boldsymbol{\Theta})$ and the ground truth image $\boldsymbol{u}$, and $\#(\mathscr{F} \times \mathscr{U})$ is the cardinality of the data set $\mathscr{F} \times \mathscr{U}$. To prevent overfitting, we can introduce a regularization term to the above optimization problem as in (1.2). We then have the problem

$$\min_{\boldsymbol{\Theta}} \mathscr{L}(\boldsymbol{f}, \boldsymbol{u}; \boldsymbol{\Theta}) = \frac{1}{\#(\mathscr{F} \times \mathscr{U})} \sum_{(\boldsymbol{f},\boldsymbol{u}) \in \mathscr{F} \times \mathscr{U}} \mathscr{C}(\mathscr{V}(\boldsymbol{f}, \boldsymbol{\Theta}), \boldsymbol{u}) + \mathscr{R}(\boldsymbol{\Theta}), \tag{1.3}$$

where $\mathscr{R}(\cdot)$ is the regularization term that can be chosen as, for example, the ℓ_2 or ℓ_1 norm. Good examples of the nonlinear mapping $\mathscr{V}(\cdot)$ include the stacked denoising autoencoder (SDAE) Vincent et al. [14], the U-Net [15], the ResNet [16,17], etc. We postpone a detailed discussion on these networks and how to interpret them in mathematical terms in later sections.

The development of modeling in image reconstruction for the past three decades can be summarized to three stages.

- **Handcraft modeling (1990-).** Models are designed based on mathematical characterizations on the desirable recovery of the image. For example, the ideal function space a "good" image should belong to, ideal local or global geometric properties the image should enjoy, or sparse approximation by certain well-designed basis functions, etc. Successful handcraft models include total variation (TV) [18] model, Perona–Malik diffusion [19,20], shock-filters [21,22], nonlocal methods [23–27], wavelet [28,29], wavelet frames [30,31], BM3D [27], WNNM [32], etc. These models mostly have solid theoretical foundations and high interpretability. They work reasonably well in practice, and some of them are still the state-of-the-art methods for certain tasks.
- **Handcraft + data-driven modeling (1999-).** Starting from around 1999, models that combine data-driven or learning with handcraft modeling started to emerge. These models rely on some general mathematical or statistical framework by handcraft designs. However, the specific form of the model is determined by the given image data or data set. Comparing to purely handcrafted models, these models can better exploit the available data and outperform their corresponding none data-driven counterparts. Meanwhile, the handcrafted framework of the models grants certain interpretability and theoretical foundation to the models. Successful examples include the method of optimal directions [33], the K-SVD [34], learning based PDE design [35], data-driven tight frame [36,37], Ada-frame [38], low-rank models [39–43], piecewise-smooth image models [44,45], and statistical models [46], etc.
- **Deep learning models (2012-).** 2012 is the year that signifies the uprise of deep learning in computer vision with the introduction of a convolutional neural network (CNN) called AlexNet [47] for image classification. Then, various types of CNNs such as ResNet [16,17] and generative adversarial networks (GANs) [48] were introduced and applied in image reconstructions. We shall refer to these models as deep learning based models (or deep models for short). Most deep models have millions to billions of parameters. These parameters are trained (optimized) on large data sets via parallel computing (e.g., on graphics processing units (GPUs)). Deep models have greatly advanced the state of the art of many image reconstruction tasks and have changed the research landscape of computer vision in general. The success of deep models is mainly due to the presence of large image data sets with high-quality labels, and the accessibility of massive computing resources. The reliance of deep models on large labeled data sets limits, at least for the moment,

the application of deep learning in medical image reconstruction or healthcare in general. The major focus of this review is to recall and discuss deep models in medical image reconstruction, and the limitations, challenges, and opportunities in this new and exciting research direction.

Note that what makes medical image reconstruction different from image restoration in computer vision is quality metrics on the reconstructed image. Although researchers use standard metrics such as the peak signal-to-noise-ratio (PSNR), mean square error, structure similarity (SSIM), meaningful quality metrics of a reconstructed medical image should be clinically relevant and task dependent. Furthermore, most medical images are 3D arrays which pose computation challenge as well.

1.2 Algorithm Design for Image Reconstruction Models

The difficulties of solving the image reconstruction models motivate the optimization community to design highly efficient numerical algorithms for large scale, nonsmooth, and even nonconvex optimization problems. Representative algorithms include the alternating direction method of multipliers (ADMM) [49–51], primal–dual algorithm [52–54], split Bregman algorithm [55,56], linearized Bregman algorithm [57,58], iterative shrinkage-thresholding algorithm (ISTA) [59], and fast iterative shrinkage-thresholding algorithm (FISTA) [60], among many others. Here, we briefly review some of the algorithms that will be needed in later sections.

1.2.1 ISTA and FISTA

Consider the following optimization problem which is a special case of (1.2):

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\boldsymbol{f} - \boldsymbol{W}^{\mathrm{T}}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1, \tag{1.4}$$

where $\boldsymbol{W}^{\top}$ is a decoding operator that maps code $\boldsymbol{\alpha}$ back to image domain. Then, the ISTA solving (1.4) simply reads as

$$\boldsymbol{\alpha}^{k+1} = \mathscr{T}_{\lambda\tau_k}\left(\boldsymbol{\alpha}^k - 2\tau_k \boldsymbol{W}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\alpha}^k - \boldsymbol{f})\right), \tag{1.5}$$

where $\tau_k > 0$ is an appropriate step size and the soft-thresholding operator $\mathscr{T}_\lambda(\cdot)$ is defined component-wisely as $\mathscr{T}_\lambda(x) = \mathrm{sign}(x)\max(|x| - \lambda, 0)$, with $x \in \mathbb{R}$. ISTA was explicitly proposed in [59]. Its idea, however, can be traced back to the classical proximal forward–backward algorithm [61,62]. Later, an accelerated version of ISTA, called fast iterative soft-thresholding algorithm (FISTA), was introduced [60,63] which is based on the idea of Nesterov's [64]. FISTA takes following form

$$\begin{aligned}
\boldsymbol{\alpha}^{k+1} &= \mathscr{T}_{\lambda/L_{\mathrm{lip}}}\left(\boldsymbol{y}^k - \frac{1}{L_{\mathrm{lip}}}\boldsymbol{W}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{y}^k - \boldsymbol{f})\right),\\
t_{k+1} &= \frac{1+\sqrt{1+4t_k^2}}{2},
\end{aligned}$$

$$\boldsymbol{y}^{k+1} = \boldsymbol{\alpha}^{k+1} + \frac{t_k - 1}{t_{k+1}}(\boldsymbol{\alpha}^{k+1} - \boldsymbol{\alpha}^k), \tag{1.6}$$

where L_{lip} is the Lipschitz constant of the quadratic term in (1.4).

1.2.2 ADMM/Split Bregman Algorithm

Consider the following special case of the optimization problem (1.2)

$$\min_{\boldsymbol{u}} \mathscr{L}(\boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f}\|_2^2 + \lambda\|\boldsymbol{W}\boldsymbol{u}\|_1,$$

which can be written equivalently as

$$\min_{\boldsymbol{u},\boldsymbol{d}} \mathscr{L}(\boldsymbol{u}, \boldsymbol{d}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f}\|_2^2 + \lambda\|\boldsymbol{d}\|_1 \quad \text{s.t.} \quad \boldsymbol{W}\boldsymbol{u} = \boldsymbol{d}.$$

The corresponding augmented Lagrangian function [65, Chapter 17] is defined by

$$\mathscr{L}(\boldsymbol{u}, \boldsymbol{d}; \boldsymbol{b}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f}\|_2^2 + \lambda\|\boldsymbol{d}\|_1 + \langle \boldsymbol{W}\boldsymbol{u} - \boldsymbol{d}, \boldsymbol{b}\rangle + \frac{\mu}{2}\|\boldsymbol{W}\boldsymbol{u} - \boldsymbol{d}\|_2^2$$

with the Lagrangian multiplier $\boldsymbol{b}$. Then, the ADMM or split Bregman algorithm takes the form [49,56]

$$\begin{aligned}
\boldsymbol{u}^{k+1} &= \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} + \mu\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\right)^{-1}\left[\boldsymbol{A}^{\mathrm{T}}\boldsymbol{f} + \mu\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{d}^k - \boldsymbol{v}^k)\right], \\
\boldsymbol{d}^{k+1} &= \mathscr{T}_{\lambda/\mu}\left(\boldsymbol{W}\boldsymbol{u}^{k+1} + \boldsymbol{v}^k\right), \\
\boldsymbol{v}^{k+1} &= \boldsymbol{v}^k + (\boldsymbol{W}\boldsymbol{u}^{k+1} - \boldsymbol{d}^{k+1}),
\end{aligned} \tag{1.7}$$

where λ and μ are tuning parameters.

1.2.3 The Primal–Dual Algorithm

Consider the following optimization problem

$$\min_{\boldsymbol{u}} F(\boldsymbol{u}) + \Phi(\boldsymbol{W}\boldsymbol{u}), \tag{1.8}$$

where $F(\boldsymbol{u})$ is the data fidelity term and $\Phi(\boldsymbol{W}\boldsymbol{u})$ is the regularization term appeared in (1.2). Assume $F : \mathbb{R}^n \to (-\infty, +\infty]$ and $\Phi : \mathbb{R}^m \to (-\infty, +\infty]$ are closed proper convex functions. The problem (1.8) can be written equivalently as

$$\min_{\boldsymbol{u}} \max_{\boldsymbol{w}} F(\boldsymbol{u}) + \langle \boldsymbol{W}\boldsymbol{u}, \boldsymbol{w}\rangle - \Phi^*(\boldsymbol{w}). \tag{1.9}$$

Then, the primal–dual hybrid gradient (PDHG) algorithm [52–54] can be written as

$$\begin{aligned} \boldsymbol{w}^{k+1} &= (I + \partial \Phi^*)^{-1} \left(\boldsymbol{w}^k + \alpha_k \boldsymbol{W} \boldsymbol{u}^k \right), \\ \boldsymbol{u}^{k+1} &= (I + \partial F)^{-1} \left(\boldsymbol{u}^k - \beta_k \boldsymbol{W}^\top \boldsymbol{w}^{k+1} \right), \end{aligned} \tag{1.10}$$

where α_k and β_k are tuning parameters. Note that in [54], the authors introduced an additional correction update step

$$\bar{\boldsymbol{u}}^{k+1} = \boldsymbol{u}^{k+1} + \theta(\boldsymbol{u}^{k+1} - \boldsymbol{u}^k) \tag{1.11}$$

to the original PDHG algorithm (1.10) and replaced $\boldsymbol{u}^k$ in $\boldsymbol{w}^{k+1}$-step by $\bar{\boldsymbol{u}}^k$.

1.2.4 SGD

It is very common in machine learning that an optimization problem takes the following form

$$\min_{\boldsymbol{\Theta}} F_N(\boldsymbol{\Theta}) = \frac{1}{N} \sum_{i=1}^{N} f_i(\boldsymbol{\Theta}). \tag{1.12}$$

The main computation challenge, especially in deep learning, is that N can be huge, e.g., in the magnitude of millions to billions. Therefore, the evaluation of the function value F_N and its gradient can be rather slow. In such case, stochastic gradient descent (SGD) algorithm [66–69] and its variants [70–72] are among the most popular algorithms in deep learning.

The very basic form of (mini-batch) SGD is

$$\boldsymbol{\Theta}^{k+1} = \boldsymbol{\Theta}^k - \alpha_k \frac{1}{|\mathscr{S}_k|} \sum_{i_k \in \mathscr{S}_k} \nabla f_{i_k}(\boldsymbol{\Theta}^k),$$

where α_k is the step size (or learning rate) and $\mathscr{S}_k$ is a random subset of $\{1, 2, \cdots, N\}$. The evaluation of $\frac{1}{|\mathscr{S}_k|} \sum_{i_k \in \mathscr{S}_k} \nabla f_{i_k}(\boldsymbol{\Theta}^k)$ provides an unbiased estimation of the full gradient and is computationally cheap. Other than SGD, numerous randomized algorithms are being used in deep learning, such as Adam [73], AdaGrad [74], RMSProp [75]. A comprehensive review on optimization algorithms for large-scale machine learning problems can be found in [76].

1.3 When Handcraft Modeling Meets Deep Learning

Both handcrafted models and deep models have their advantages and drawbacks depending on the applications. Most handcrafted models are designed with a solid mathematical foundation and can be very well interpreted. However, handcrafted models are not flexible enough to fully leverage large data sets. Deep models, on the other hand, are generally much more flexible and can better extract useful information from large data sets. However, they are generally more challenging to interpret. For the

moment, they are also in lack of theoretical foundations in contrast to handcrafted models. Therefore, there has been an increasing effort in the community to combine handcrafted modeling and deep modeling so that we can enjoy benefits from both approaches.

One of the most popular ways of such combination is the so-called unrolling dynamics approach. It started with the work of Gregor and LeCun [77] where the authors showed that one could unroll the ISTA in (1.5) to create a feed-forward network. Then, one can train ISTA in an end-to-end fashion to determine the parameters in ISTA so that they are best suitable for the training data. They called the unrolled dynamics LISTA and demonstrated its advantage over ISTA. This work showed that one could unroll a discrete dynamic system to form a network for end-to-end training. More recently, more and more examples showed that the unrolling dynamics approach seems a good balance between model interpretability and efficacy. This includes unrolling discrete forms of nonlinear diffusions for image restoration [35,78] and unrolling optimization algorithms for medical imaging and inverse problems [79–85]. The unrolling dynamics approach can often result in deep models that have better interpretability inherited from the original dynamics.

Furthermore, these deep models normally have much fewer trainable parameters than black box deep neural networks, which are more suitable for learning on relatively small data sets. On the other hand, we may interpret certain classes of deep convolutional networks, such as ResNet, as discrete dynamics, and hence relates deep learning with optimal control [86,87]. Such viewpoint not only provides elegant interpretation of deep neural networks [88], but also enables us to design more effective deep networks for various tasks in machine learning [84,89–94], computer vision [95], inverse problems [96,97], and natural language processing [98]. More recently, intriguing relations between deep convolutional networks with multigrid method are addressed [99] which lead to new interpretations to deep models.

The remainder of this paper is organized as follows. In Sect. 2, we will review some recently proposed deep neural networks which are popular in medical imaging. Section 3 shows the understanding of deep neural networks from the perspective of representation learning and differential equations. Section 4 reviews some recently proposed deep models for medical imaging, where Sect. 4.1 presents some examples of post-processing deep models, Sect. 4.2 collects some models that are designed by combining handcrafted modeling with deep modeling, and Sect. 4.3 reviews task-driven deep models. To conclude, Sect. 5 summarizes the main challenge and opportunities in deep learning based medical imaging.

2 Review of Deep Neural Networks

Deep neural networks (DNNs) are now proven to be powerful tools to represent complex data. The main differences between DNNs and traditional machine learning models are the composite nonlinearity of the DNNs and the end-to-end training, which make DNNs very effectively in extracting features that are most suitable for a given task. In recent years, DNNs are used in various medical imaging tasks, including image reconstruction, segmentation, region-of-interest detection, super-resolution, classifi-

cation, etc. In this section, we briefly recall some of the DNNs that are widely adopted in medical imaging.

2.1 ResNet

In computer vision, the residual network (ResNet) [16,17] is one of the most popular DNNs. The architecture of ResNet is shown in Fig. 1 which can also be formulated mathematically as

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \mathscr{F}_k(\boldsymbol{u}_k), \tag{2.1}$$

where $\boldsymbol{u}_k$ (resp. $\boldsymbol{u}_{k+1}$) indicates the input (resp. output) feature map of the k-th layer of the ResNet and $\mathscr{F}_k(\boldsymbol{u}_k)$ is called a nonlinear residual block with trainable parameters. The skip connection of ResNet is crucial in facilitating stable training when the network is very deep. Other DNNs with the similar skip connections include the learned diffusion model TRD [78], DenseNet [100] and U-Net [15], among many others.

2.2 Autoencoder

Autoencoder (AE) [101,102] is a type of neural network that is used to learn data representation in an unsupervised manner. It aims to learn an encoder from a set of data together with a decoder so that we do not lose any essential information during the encoding and decoding process. Figure 2 presents a typical example of the AE architecture. For a given image $\boldsymbol{X}$, the parameterized mapping $f_{\boldsymbol{\theta}}$ (e.g., a fully connected or a convolutional neural network) is an encoder that extracts feature maps from $\boldsymbol{X}$. The encoded multi-channel feature maps are denoted by $\boldsymbol{Y} = f_{\boldsymbol{\theta}}(\boldsymbol{X})$. The encoded feature maps $\boldsymbol{Y}$ is then decoded by another parameterized mapping $g_{\boldsymbol{\theta}'}$ to obtain the reconstructed data $\boldsymbol{Z}$. The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are optimized on a data set so that a properly chosen loss function that measures the average discrepancies

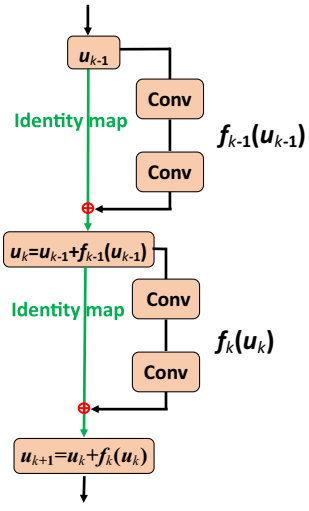


Fig. 1 ResNet

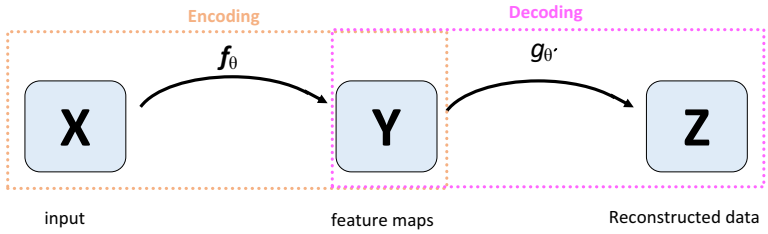


Fig. 2 Autoencoder

between $\boldsymbol{X}$ and $\boldsymbol{Z}$ is minimized. AE resembles linear representations such as Fourier and wavelet transform if we regard encoding as the decomposition, decoding as the reconstruction, and feature maps as the coefficients of the representation. However, the representation provided by AE is nonlinear and is learned from a data set.

To learn a more effective and robust representation, Vincent et al. [14] proposed the stacked denoising autoencoder (SDAE). In SDAE, the encoder and decoder are DNNs, and they are trained to recover $\boldsymbol{Z} \approx \boldsymbol{X}$ from noisy input $\boldsymbol{X}$. Based on the encoder/decoder framework, Badrinarayanan et al. [103] designed a DNN, called SegNet, for image segmentation. In [104], the encoder/decoder framework is adopted for image denoising and super-resolution. More recently, Chen et al. [105] designed a residual encoder–decoder CNN to suppress the noise and preserve features in low-dose CT images that are reconstructed using the filtered back projection (FBP) algorithm.

2.3 U-Net

In [15], a U-shaped deep neural network, called U-Net, was proposed for biomedical image segmentation which is by far one of the most successful deep image segmentation models. The architecture of U-Net is shown in Fig. 3. It resembles the encoder/decoder architecture of AE if we view the left half of the U-Net as an encoder and the right half as a decoder. The main difference between the U-Net and AE is the use of skip connections of U-Net. Similar to the U-Net, Milletari et al. [106] designed a DNN, called V-Net, for 3D volumetric medical image segmentation. Motivated by the U-Net and the convolutional framelets [107], Ye et al. [108] designed a multi-resolution deep convolutional framelets. More recently, U-Net is extended to image analysis tasks [109].

3 Interpretations of Deep Neural Networks

The development of traditional machine learning methods, such as support vector machine, decision tree, random forest, benefits tremendously from theoretical studies in machine learning. However, existing machine learning theory, such as PAC, VC-dimension, Rademacher complexity, may not be most suitable to analyze DNNs. Although DNNs are often composed of simple functions, such as convolutions, pooling, element-wise activation functions, the entire networks are often difficult to

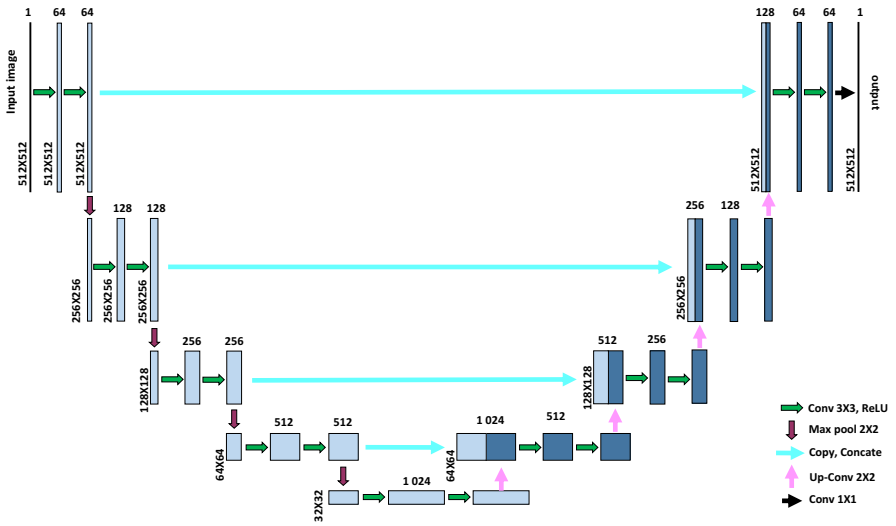


Fig. 3 U-Net

analyze. Therefore, theoretical deep learning has become a popular area in machine learning that has attracted a lot of attention from theoretical computer scientists, statisticians, and mathematicians. In this section, we shall review some recent works on interpreting DNNs from two different perspectives, namely representation learning and differential equations. We will see that function approximation is a powerful tool in characterizing the efficacy of the given representation. It provides a rigorous analysis of the capacity of DNNs and how well they can approximate functions living in various function spaces. The perspective through differential equations, on the other hand, is more intuitive than function approximation and can explicitly guide the design of architectures of DNNs and training algorithms. There are also several other perspectives on the theoretical interpretations of deep learning. One may refer to the course "Theories of Deep Learning" (STATS 385) hosted by David Donoho at Stanford University and the references therein (https://stats385.github.io/).

3.1 Representation Learning Perspective

Images, such as medical images or natural images, are usually assumed to have sparse (or low dimensional) structures. The sparse structures can be effectively extracted by transformations. Successful examples include the (windowed) Fourier transform, wavelet transform, curvelet transform, etc., and they are able to provide efficient representations to images. They are pre-designed linear transformations and are independent of the given image data. DNNs can also be viewed as sparse representations that are able to extract sparse features from images. The difference is that DNNs are learned from a set of images and are (highly) nonlinear.

The quality of a given representation can be measured by its ability to approximate functions living in a certain function space. For example, let $\Phi := \{\phi_i(\boldsymbol{x}) : \boldsymbol{x} \in$

$\mathbb{R}^n, i \in \mathbb{N}_+\}$ be a set of atoms, and function $f(\boldsymbol{x})$ be an element in function space $\mathscr{F}$ equipped with norm $\|\cdot\|$. One of the most basic and important approximation properties states as follows: for any given $\varepsilon > 0$, there exists $\tilde{f}_{\boldsymbol{\alpha},N} := \sum_{i=1}^{N} \alpha_i \phi_i(\boldsymbol{x})$ with $N \in \mathbb{N}_+$ and $\boldsymbol{\alpha} = \{\alpha_1, \cdots, \alpha_N\} \in \mathbb{R}^N$ such that

$$\|f - \tilde{f}_{\boldsymbol{\alpha},N}\| < \varepsilon.$$

A good representation requires fewer atoms (i.e., smaller N) to achieve an ε-approximation. The representation of various types of Φ has been well studied in the literature, such as polynomials, splines, Fourier basis, and wavelets [28,29,110].

The neural network is a more efficient tool that can approximate a function arbitrarily well under suitable conditions [111–113]. Both the depth and width of a neural network are among the most important factors that affect its approximation power. In the following, we will review some of the existing characterizations of the approximation properties of shallow and deep neural networks.

Consider a shallow neural network with only one hidden layer

$$\tilde{f}_N(\boldsymbol{x}; \boldsymbol{\Theta}) = \sum_{i=1}^{N} a_i \sigma(\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{x} + b_i),$$

where $\boldsymbol{x} \in \mathbb{R}^n$ is the input image data, $\boldsymbol{\Theta} = \{a_i, \boldsymbol{w}_i, b_i\}, i = 1, \cdots, N$, are trainable parameters, and $\sigma(z)$ is an element-wisely applied nonlinear activation function. Examples of $\sigma(z)$ are $\mathrm{ReLU}(z) = \max(0, z)$, $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$, $\mathrm{sigmoid}(z) = \frac{1}{1+e^{-z}}$ and more generally a sigmoidal function [114] that has the property

$$\sigma(z) = \begin{cases} 1, & \text{if } z \to +\infty, \\ 0, & \text{if } z \to -\infty. \end{cases} \tag{3.1}$$

A DNN is a neural network with multiple hidden layers. It can be viewed as a successive composition of multiple shallow networks. A typical DNN (for regression problems) with depth L and width $\boldsymbol{N} = (N_1, N_2, \cdots, N_L)$ denoted as

$$\tilde{f}_{L,\boldsymbol{N}}(\boldsymbol{x}; \boldsymbol{\Theta}) : \mathbb{R}^n \mapsto \mathbb{R},$$

can be recursively defined as $\boldsymbol{\Theta}^\ell = (\boldsymbol{\Theta}^{\ell-1}, \theta^\ell)$, $\tilde{f}_{\boldsymbol{\Theta}^\ell} = (\theta^\ell \circ \sigma \circ \tilde{f}_{\boldsymbol{\Theta}^{\ell-1}})$, $\theta^\ell : \mathbb{R}^{N_\ell} \to \mathbb{R}^{N_{\ell+1}}$ with $\theta^\ell(\boldsymbol{x}) = \boldsymbol{W}^\ell \boldsymbol{x} + \boldsymbol{b}^\ell$, and $\tilde{f}_{L,\boldsymbol{N}} := \tilde{f}_{\boldsymbol{\Theta}^L}$.

Earlier results on the approximation property, i.e., universal approximation, suggest that a wide class of functions can indeed be approximated by neural networks with only one hidden layer, though the number of neurons, i.e., N, may increase exponentially as we decrease ε [114–116]. There are benefits in increasing the depth L of the neural network when approximating a target function. For example, approximation with DNNs leads to an exponential or polynomial reduction in the number of neurons while maintaining the same level of approximation accuracy [117–120]. Delalleau and Bengio [121], Telgarsky [122,123] presented concrete examples that there exist

functions that can be more efficiently represented with DNNs rather than shallow networks. In particular, Telgarsky [123] showed that the DNNs with $\mathcal{O}(L^3)$ layers and constant width cannot be approximated by networks with $\mathcal{O}(L)$ depth and less than 2^L width.

Lu et al. [124] investigated the efficiency of depth of ReLU activated DNNs from a different view by proving that there exist classes of wide neural networks which cannot be realized by any narrow network whose depth is no more than a polynomial bound. Comparing to the known result that there are classes of deep networks which cannot be realized by any shallow network whose size is no more than an exponential bound [120], results from [124] indicated that depth might be more effective than width. Although depth is more important than width, Hanin [125,126] proved that there is a minimum width of ReLU activated DNNs to ensure approximation of continuous functions. Their results indicated that a good DNN cannot be too narrow, otherwise we cannot approximate continuous functions even with infinite depth.

More recently, Yarotsky [127] analyzed the dependence of optimal approximation rate with respect to depth for ReLU activated DNNs. When approximating a multivariate polynomial, Rolnick and Tegmark [128] proved that the total number of neurons in DNNs should grow linearly with respect to the number of variables of the polynomial. Shen et al. [129] provided intriguing analysis on ReLU activated DNNs via nonlinear approximation of composite dictionaries. They demonstrated the advantage of depth over width quantitatively by comparing the N-term approximation order of DNNs v.s. one-hidden-layer neural networks. Other than generic DNNs, theoretical analysis on the popular ResNet was also provided [130–132].

In [133], the authors investigated the connection between linear finite element functions and ReLU deep neural networks. Firstly, they proposed an efficient ReLU activated DNN to represent any linear finite element functions and theoretically established that at least 2 hidden layers are needed in a ReLU activated DNN to represent any linear finite element functions in $\Omega \subseteq \mathbb{R}^d$ when $d \geq 2$. Then, using this relationship they established a straightforward error estimate as $\mathcal{O}(N^{-\frac{1}{d}})$ for a special kind of ReLU activated DNNs with $\mathcal{O}(N)$ nonzero parameters by involving the h-adaptive linear finite element approximation theory [134].

Different from the approximation viewpoint, He and Xu [99] developed a unified model, known as MgNet, that simultaneously recovers and extends some CNNs for image classification and multigrid methods for solving discretized PDEs, by combining multigrid and deep learning methodologies.

3.2 Differential Equation and Control Perspective

Given a DNN $\tilde{f}(\boldsymbol{x}; \boldsymbol{\Theta}) : \mathbb{R}^n \to \mathbb{R}^m$, due to its composite structure as described in the previous subsection, we may view $\tilde{f}(\boldsymbol{x}; \boldsymbol{\Theta})$ as an iterative mapping between $\mathbb{R}^n$ and $\mathbb{R}^m$. Then it is natural to view a generic DNN as a certain dynamic system [135]. However, a dynamic system that corresponds to a generic DNN is difficult to analyze since it does not have much special structure to exploit. Fortunately, it has been proven empirically that most of the effective DNNs have special structures in their architecture. In fact, designing special structures of DNNs, i.e., the architecture design,

to make them easy to train and generalize better is one of the major research directions in deep learning. Furthermore, the objective of the emerging research topic neural architecture search (NAS), a subfield of automating machine learning (AutoML), is to search for effective DNN architecture for different data sets and tasks.

One of the most well-known DNNs with special structures is ResNet. Its bypasses (or shortcuts) enable us to efficiently train ultra-deep networks and achieve high accuracies in multiple tasks. The success of ResNet inspired the design of numerous new neural architectures. However, most of the design were based on empirical studies. Although we can deploy NAS to search for new architectures, the current computation burden of NAS is still prohibitively high for researchers without access to heavy computation resources, and NAS cannot guarantee to find sufficiently new and interpretable neural structures. Therefore, we direly need a way to interpret ResNet and their siblings properly and to seek for guiding principles for the architecture design of DNNs.

Recently, Weinan [86] made an inspiring observation that ResNet can be viewed as forward Euler scheme solving for an ordinary differential equation (ODE), and links training of DNNs with optimal control. Sonoda and Murata [136] and Li and Shi [88] also regarded ResNet as dynamic systems that are the characteristic lines of a transport equation on the data distribution. Similar observations were also made by Chang et al. [87,89]. A rigorous justification of the link between ResNet and ODEs was given by Thorpe and van Gennip [137], and that of the link between deep learning and optimal control was given by Weinan et al. [138]. The dynamics and control perspective enabled us to design more efficient algorithms solving related deep learning problems [94,95,139,140].

In [90], the authors suggested a general bridge between numerical ODEs and deep neural architectures by observing that multiple state-of-the-art deep network architectures, such as PolyNet [141], FractalNet [142], and RevNet [143], can be viewed as different discretizations of ODEs. Furthermore, Lu et al. [90] proved that ResNet with certain stochastic training strategies weakly approximates stochastic differential equations, which granted stochastic control perspective on randomized training of DNNs. More importantly, such new perspectives enable us to systematically design deep neural architectures through numerical (stochastic) differential equations, which is a rather mature field in applied mathematics. In this section, we shall review some of the findings of [90] and some other related works.

3.2.1 Numerical Difference Equations and Architecture Design

We first show that how ResNet is related to forward Euler scheme in numerical ODEs. Considering a building block of ResNet (2.1) as shown in Fig. 1, it can be rewritten as

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \Delta t_k \mathscr{F}(\boldsymbol{u}_k, t_k),$$

or equivalently as

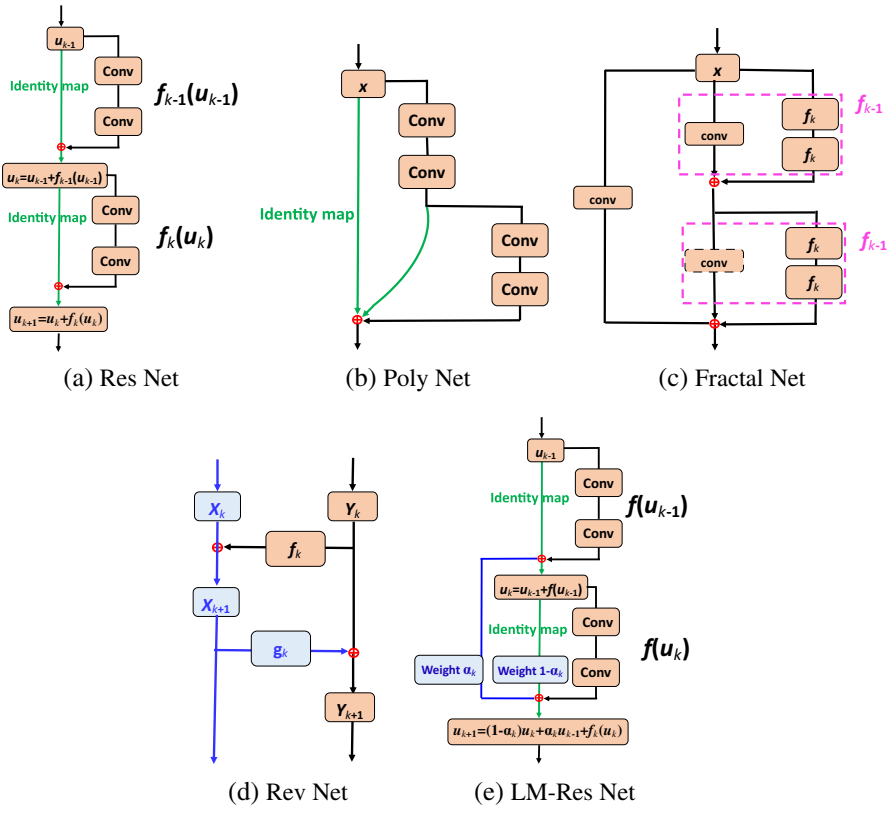
$$\frac{\boldsymbol{u}_{k+1}-\boldsymbol{u}_k}{\Delta t_k}=\mathscr{F}(\boldsymbol{u}_k,t_k),$$

where Δt_k is the step size and $\mathscr{F}(\boldsymbol{u}_k,t_k)=\frac{1}{\Delta t_k}\mathscr{F}_k(u_k)$. The above formula is the forward Euler scheme solving the following ordinary differential equation (ODE):

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t}=\mathscr{F}(\boldsymbol{u},t). \tag{3.2}$$

Therefore, the ResNet can be viewed as the forward Euler discretization of the ODE (3.2) with step size $\Delta t_k=1$ for every k. This was first observed by Weinan [86]. More recently, Zhang et al. [144] showed that there are benefits of considering ResNet with $0<\Delta t_k<1$.

In [90], the authors further observed that many other DNNs with bypasses, e.g., PolyNet [141] (Fig. 4b), FractalNet [142] (Fig. 4c), and RevNet [143] (Fig. 4d), can be interpreted as certain temporal discretizations of ODEs. For example, the PolyInception module (Fig. 4b) of PolyNet can be written mathematically as



(a) Res Net (b) Poly Net (c) Fractal Net

(d) Rev Net (e) LM-Res Net

Fig. 4 Schematics of different neural network architectures

$$(I + \mathscr{F} + \mathscr{F}^2)(x) = x + \mathscr{F}(x) + \mathscr{F}(\mathscr{F}(x)),$$

where I is the identity map, $\mathscr{F}$ is a nonlinear operator and x is the input feature map. Note that the above polynomial of mapping $\mathscr{F}$ is an approximation of $(I - \mathscr{F})^{-1}$ using a truncated Neumann series

$$(I - \Delta t\mathscr{F})^{-1} \approx (I + \Delta t\mathscr{F} + \Delta t^2\mathscr{F}^2).$$

Therefore, PolyNet can be viewed as an approximation to the backward Euler scheme solving the ODE (3.2). FractalNet (Fig. 4c) can be viewed as approximation of the ODE (3.2) with Runge–Kutta scheme. See [90] for more examples and further details.

These examples suggest a potential link between numerical ODEs and deep neural architecture. A remaining question is whether deep neural architecture design can benefit from such perspective. The authors of [90] designed a new ResNet-like module, called the Linear Multi-step structure (LM-structure) using the linear multi-step schemes in numerical ODEs [145]. The LM-structure (linear two-step structure to be more precise) can be written mathematically as

$$\boldsymbol{u}_{k+1} = (1 - \gamma_k)\boldsymbol{u}_k + \gamma_k\boldsymbol{u}_{k-1} + \mathscr{F}(\boldsymbol{u}_k, t_k), \tag{3.3}$$

where $\gamma_k \in \mathbb{R}$ is a trainable parameter in each layer. Note that when $\gamma_k = 0$ for all k, the LM-structure reduces to ResNet. Figure 4e shows the LM-structure. Empirical results of [90] showed that the LM-structure boost classification accuracies of ResNet-like DNNs on CIFAR and ImageNet. It can also reduce the depth (hence number of parameters) of ResNet-like DNNs by 50–90% without hampering accuracies. Other than the LM-structure, one can use the midpoint scheme or the leapfrog scheme to design new DNNs [89], or using the Runge–Kutta method [146].

The performance gain of the LM-structure can be explained using the concept of modified equations [147]. By Taylor's expansion, the modified equation associated with the LM-structure (3.3) is

$$(1 + \gamma_k)\dot{\boldsymbol{u}}_k + \frac{1 - \gamma_k}{2}\Delta t\ddot{\boldsymbol{u}}_k = \mathscr{F}(\boldsymbol{u}_k, t). \tag{3.4}$$

Comparing to ResNet, the LM-structure has the freedom to balance between $\ddot{\boldsymbol{u}}_k$ of $\boldsymbol{u}_k$. Having bigger weights on $\ddot{\boldsymbol{u}}_k$ can speed up the information propagation of the dynamics as shown by various earlier work such as [148–150]. This is why LM-structure can achieve comparable accuracies with a much smaller depth than ResNet and its siblings.

3.2.2 Stochastic Training and Optimal Control

Stochastic training, such as dropout and noise injection, is widely adopted in training of DNNs. It helps with the generalization of the trained networks. In [90], the authors showed that some stochastic training of ResNet, shake–shake [151], and stochastic depth [152], can be viewed as stochastic control

$$\min_{\boldsymbol{\Theta}} E_{\boldsymbol{X}(0)\sim\text{data}}\left[E[L(\boldsymbol{X}(T))] + \int_0^{\mathrm{T}} R(\boldsymbol{\Theta})\right]$$
$$\text{s.t.} \quad \mathrm{d}\boldsymbol{X} = \mathscr{F}(\boldsymbol{X},\theta)\mathrm{d}t + \mathscr{G}(\boldsymbol{X},\theta)\mathrm{d}\boldsymbol{B}_t, \tag{3.5}$$

where the stochastic differential equation (3.5) is the weak limit of the ResNet with shake–shake mechanism or stochastic depth. This suggests a connection between stochastic training and stochastic control, and a connection between DNNs with randomness and stochastic differential equations. Later, Sun et al. [153] observed that the stochastic training of ResNet and its variants is closely related to the optimal control of backward Kolmogorov's equations, and the popular dropout regularization essentially introduces viscosity to the equations.

4 Deep Models in Medical Image Reconstruction

Classical medical image reconstruction methods, such as FBP and algebraic reconstruction method (ART) for CT imaging, are highly efficient and widely used in practice [154]. However, these methods are also prone to be sensitive to noise and incompleteness of measured data. To obtain a high-quality image, numerous regularization based models and algorithms have been developed [155–157] in the past three decades. In recent years, there has been a continuous effort in the medical imaging community to further advance medical image reconstruction by combining traditional image reconstruction methods with deep learning. When combining traditional handcraft modeling with deep modeling, two general approaches are often adopted: post-processing and raw-to-image. The validity of these two approaches are generally supported by, though still rather incomplete, the analysis on the approximation properties of DNNs as described in Sect. 3.1, and by the dynamics perspective on the DNNs with certain special structures as described in Sect. 3.2.

For post-processing, one needs to estimate the mapping between the initially reconstructed low-quality image and its high-quality counterpart. This is possible since DNNs can approximate generic functions or mappings as discussed in Sect. 3.1. This approach is effective mostly when the initial reconstruction and its high-quality counterpart are not drastically different. However, due to limited measurements and the presence of noise, the initially reconstructed image may contain heavy and complex artifacts which are difficult to remove even by deep models. Furthermore, the information missing from the initial reconstruction cannot be reliably recovered by any post-processing. Thus, the post-processing approach has limited performance and is more suitable to handle initial reconstructions that are of relatively high quality.

For raw-to-image, one directly estimates the mapping between the raw data (e.g., the projection data of CT and k-space data for MRI) and the reconstructed image. The challenge, however, is that the data distribution in the domain of raw data is often vastly different from that in the image domain. Learning a direct mapping using a DNN without special structures (e.g., a fully connected network or a vanilla CNN), though not impossible, may require tons of training data, can be computationally expensive and heavily relies on good initializations of the model parameters (e.g.,

the AUTOMAP [158]). It is well-known in the literature of handcraft modeling that the mapping ought to have certain dynamic structures which can be represented by a carefully designed (partial) differential equation or an optimization algorithm solving a certain objective function(al). Thus, it is more plausible to combine handcrafted dynamics with deep learning. The way of such combination was depicted in Sect. 3.2 in a relatively general setting where we did not discuss how $\mathscr{F}$ should be designed for a given image restoration problem. Nonetheless, it is rather convincing that there are connections between dynamic systems and DNNs and the benefits of recognizing such connections.

Our rich history of handcraft modeling in image restoration provides us with an abundant set of tools that we can select freely for the mapping $\mathscr{F}$ via the general technique known as the unrolling dynamics [77,79]. To be more precise, this approach first suggests us to unroll optimization algorithms that are introduced to solve handcrafted models into feed-forward networks. Then, we incorporate our domain knowledge of the problem in-hand to determine which parameters are best to be learned from the data in an end-to-end fashion. The advantage of designing deep models via unrolling optimization algorithms is threefold: (1) the deep model through unrolling dynamics is more interpretable than a regular deep model such as U-Net; (2) the number of parameters are normally less than regular deep models and thus more suitable for small sample learning; (3) it provides a general way of combining domain knowledge with deep learning so that we can easily decide on which component in the model need to be learned and which can be handcrafted without losing much expressive power of the model.

As mentioned in the introduction that one of the major differences between medical image reconstruction and image restoration in computer vision is the quality metric of the reconstruction images. It has long been discussed in the medical imaging community that such a quality metric is best, in many scenarios, to be task-based rather than generic metrics such as PSNR and SSIM. The importance of providing such a task-based metric for medical imaging was recently discussed in the article [159]. The question is, however, how can we realize such task-based image quality metric? Recently, the authors of [160] proposed to realize task-based quality metric by "hooking" an image reconstruction network from unrolling dynamics with an image analysis DNN, so that the reconstructed images by the first network will be implicitly evaluated by the second which effectively makes the quality metric task-based. Similar idea appeared in computer vision for image denoising [161,162]. On the other hand, these works also suggested a new "raw-to-task" modeling philosophy with encouraging empirical results. Therefore, the entire pipeline of image reconstruction, analysis, and decision making can be effectively integrated.

In the rest of this section, we provide more details on the aforementioned models.

4.1 Post-processing

Post-processing is a procedure to enhance the quality of an initially reconstructed image. In this subsection, we use CT as an example. Due to the incompleteness of measured data in sparse view and limited angle CT, the FBP reconstructed image is

often degraded by streaking artifacts (Fig. 5b). Noise caused by low tube currents is another source of degradation of CT images (Fig. 5c). In [105,163], a residual encoder–decoder CNN (Fig. 2) was used to approximate the map between the degraded image and the clean image. This model is efficient in removing noise from the FBP reconstructed CT images. To protect subtle structures in CT images while suppressing noise, Yang et al. [164] adopted a generative adversarial network (GAN) with the loss function defined by a combination of Wasserstein distance and the perceptual difference between input degraded image and the corresponding clean image.

To reduce the radiation dose and acquisition time, one can decrease the number of projections of X-ray CT, which is known as the sparse view or limited angel CT. Such incompleteness of measurements leads to streaking artifacts with global and yet relatively simple structures in the FBP reconstructed CT images. In this case, a DNN with multi-scale architecture can be used to capture the global patterns of streaking artifacts. With such observation, Jin et al. [165] and Han et al. [166] utilized the U-Net (Fig. 3) to reduce artifacts in FBP reconstructed sparse view CT images. The repaired high-quality CT image is the subtraction of the learned artifacts by the U-Net from the degraded input image. In some sense, U-Net takes a role of residual learning [16].

4.2 Raw-to-Image

We describe how optimization algorithms can be unrolled and set up as a deep feed-forward network for end-to-end training. We remark that, under some specific conditions, the learning empowered optimization algorithms via unrolling dynamics can have better provable convergence than the original optimization algorithms [82, 167,168]. This was in fact the original motivation of Gregor and LeCun [77] to use machine learning to improve optimization algorithms. In this subsection, however, we shall focus on the "dual" aspect of unrolling dynamics, i.e., how optimization algorithms inspire new and more effective deep network architectures for medical image reconstruction or inverse problems in general.

(a) Ground truth (b) Low dose CT (c) Low tube current CT

Fig. 5 FBP reconstructed images

ADMM-Net

The work of ADMM-Net proposed by Yang et al. [79] was the first to suggest the potential benefit of designing deep neural networks for inverse problems by unrolling optimization algorithms.

In the iteration scheme (1.7) of ADMM algorithm (Sect. 1.2.2), the tuning parameters such as μ, λ and the handcrafted operator $\boldsymbol{W}$ are difficult to determine adaptively for a given data set. In [79], the authors proposed to unroll the ADMM algorithm to design a new deep model, named ADMM-Net. By doing so, the tuning parameters and the predefined linear operator $\boldsymbol{W}$ are now all learnable from the training data. The proximal operator of the sparsity promoting function $\Phi = \|\cdot\|_1$ is parameterized by a piecewise linear function with learnable parameters as well. As a result, the thresholding operator $\mathscr{T}_\lambda(\cdot)$ in ADMM algorithm is also learned from the training data. In a basic version of ADMM-Net [79], $\boldsymbol{d}^{k+1}$ is updated by

$$\boldsymbol{d}^{k+1} = \mathscr{T}_{\boldsymbol{\Theta}_1}\left(\mathscr{W}_{\boldsymbol{\Theta}_2}(\boldsymbol{u}^{k+1}) + \boldsymbol{b}^k\right), \tag{4.1}$$

where $\mathscr{T}_{\boldsymbol{\Theta}_1}(\cdot)$ is a parameterized piecewise linear function with parameters $\boldsymbol{\Theta}_1$, and $\mathscr{W}_{\boldsymbol{\Theta}_2}$ is a parameterized convolution layer with parameters $\boldsymbol{\Theta}_2$. The ADMM-Net was later further improved by Yang et al. [169] and the new model was called the Generic-ADMM-Net where different variable splitting strategy was adopted in the derivation of the ADMM algorithm. The Generic-ADMM-Net achieved state-of-the-art MR image reconstruction results with a significant margin over the BM3D-based algorithm.

Primal–Dual Networks (PD-Net)

In [80], the authors unrolled the iteration scheme (1.10) and (1.11) of the PDHG algorithm to design new deep model for CT image reconstruction. This new deep model was called the primal–dual network (PD-Net). The main idea is to approximate each resolvent/proximal operator [170] in the subproblem of PDHG by a neural network. Thus, it circumvents the difficulties in choosing optimal forms Φ and F. One layer of PD-Net takes the form

$$\begin{aligned} \boldsymbol{w}^{k+1} &= \mathscr{N}_{\boldsymbol{w}}\left([\boldsymbol{w}^k, \boldsymbol{W}\boldsymbol{u}^k]; \boldsymbol{\Theta}_{\boldsymbol{w}}^k\right), \\ \boldsymbol{u}^{k+1} &= \mathscr{N}_{\boldsymbol{u}}\left([\boldsymbol{u}^k, \boldsymbol{W}^{\mathrm{T}}\boldsymbol{w}^{k+1}, \boldsymbol{A}^{\mathrm{T}}\boldsymbol{f}]; \boldsymbol{\Theta}_{w}^k\right), \end{aligned} \tag{4.2}$$

where $\boldsymbol{f}$ is the measured data, $\boldsymbol{A}$ is the imaging operator, $\mathscr{N}_{\boldsymbol{w}}(\cdot; \boldsymbol{\Theta}_{\boldsymbol{w}}^k)$ and $\mathscr{N}_{\boldsymbol{u}}(\cdot; \boldsymbol{\Theta}_{\boldsymbol{u}}^k)$ are neural networks parameterized by $\boldsymbol{\Theta}_{\boldsymbol{w}}^k$ and $\boldsymbol{\Theta}_{\boldsymbol{u}}^k$, respectively. The notation $[\boldsymbol{v}_1, \cdots, \boldsymbol{v}_m]$ denotes concatenation of the components $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_m$ into a higher dimension tensor. The linear operator $\boldsymbol{W}$ can be either fixed or learned from the data. PD-Net has a significant performance boost compared with FBP and some handcrafted reconstruction models [80,171].

Joint Spatial-Radon (JSR)-Net

To suppress the artifacts induced by incomplete data and noise, Dong et al. [172] proposed a JSR domain reconstruction model for sparse view CT imaging as following:

$$\min_{\boldsymbol{u},\boldsymbol{f}} \mathscr{F}(\boldsymbol{u},\boldsymbol{f},\boldsymbol{Y}) + \mathscr{R}(\boldsymbol{u},\boldsymbol{f}), \tag{4.3}$$

where the data fidelity term $\mathscr{F}(\boldsymbol{u},\boldsymbol{f},\boldsymbol{Y})$ is defined by

$$\mathscr{F}(\boldsymbol{u},\boldsymbol{f},\boldsymbol{Y}) = \frac{1}{2}\|R_{\boldsymbol{\Gamma}^c}(\boldsymbol{f}-\boldsymbol{Y})\|^2 + \frac{\alpha}{2}\|R_{\boldsymbol{\Gamma}}(\boldsymbol{A}\boldsymbol{u}-\boldsymbol{f})\|^2 + \frac{\gamma}{2}\|R_{\boldsymbol{\Gamma}^c}(\boldsymbol{A}\boldsymbol{u}-\boldsymbol{Y})\|^2,$$

and the regularization term defined by

$$\mathscr{R}(\boldsymbol{u},\boldsymbol{f}) = \|\boldsymbol{\lambda}_1 \cdot \boldsymbol{W}_1\boldsymbol{u}\|_{1,2} + \|\boldsymbol{\lambda}_2 \cdot \boldsymbol{W}_2\boldsymbol{f}\|_{1,2}.$$

The notation $R_{\boldsymbol{\Gamma}}$ is a restriction operator with respect to the missing data region indexed by $\boldsymbol{\Gamma}$. $R_{\boldsymbol{\Gamma}}$ takes value 1 if the element's index contained in $\boldsymbol{\Gamma}$ and 0 elsewhere. Here, $\boldsymbol{\Gamma}^c$ indicates the region of available measured data and is the complement of $\boldsymbol{\Gamma}$. $\boldsymbol{A}$ is a discrete form of the Radon transform, $\boldsymbol{Y}$ is the measured projection data. Note that, in JSR model $\boldsymbol{u}$ and $\boldsymbol{f}$ are the underlying CT image and the restored high-quality projection data, respectively. $\boldsymbol{W}_i, i = 1, 2$, are tight frame transforms and $\boldsymbol{\lambda}_i, i = 1, 2$, are the regularization parameters.

The handcrafted JSR model (4.3) enforces the data consistency in the Radon domain and image domain simultaneously. Thus, it leads to improved quality of the reconstructed image. Similar data fidelity design was adopted in [173] to model the positron emission tomography. Later, Zhan and Dong [174] proposed to improve the JSR model by learning the tight frame transforms $\boldsymbol{W}_i$ from the data. More recently, a re-weighting strategy was introduced in JSR model to reduce the metal artifacts in multi-chromatic energy CT [175].

Existing work showed the potential of the JSR framework, and it is natural to consider using unrolling to derive a deep model from algorithms solving the JSR model. In [85], the authors designed the JSR-Net for sparse view and limited angle CT image reconstruction. The JSR-Net is derived by unrolling an alternative optimization algorithm with subproblems solved by ADMM. Similar to the PD-Net, JSR-Net also adopted neural networks to approximate the proximal operators. The advantage of JSR-Net is that it can efficiently utilize multi-domain image features to improve the quality of the reconstructed image.

4.3 Raw-to-Task

The traditional workflow of medical image analysis has two separate stages: (1) reconstruction of a high-quality image from raw data (see Fig. 6a), and (2) make a diagnosis based on the high-quality reconstructed image (see Fig. 6b). The drawbacks of the two-stages' approach and the potential benefit of uniting the two stages were

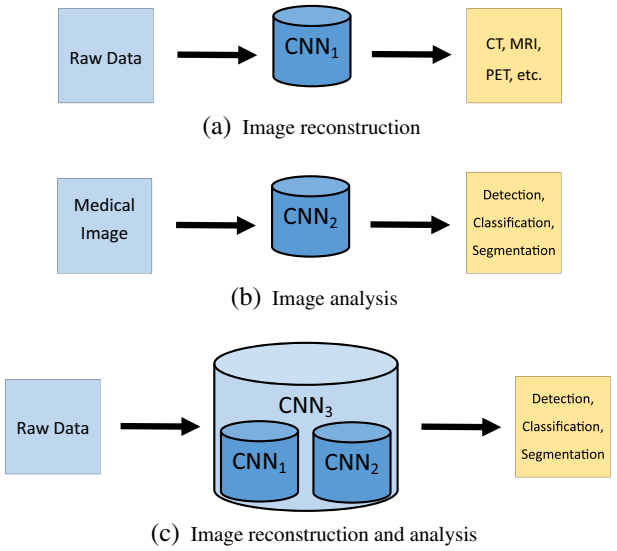


Fig. 6 CNN based workflows for medical image reconstruction and analysis

discussed earlier. Here, we shall describe how we can join the two stages into one unified step (see Fig. 6c).

As discussed in Sect. 4.2 that we can design feed-forward deep networks for image reconstruction. Once we have an image, there are plenty of choices of deep neural networks for various image analysis tasks. The most simple and natural way of joining image reconstruction and image analysis is to connect the two networks together and conduct end-to-end training (from scratch or by fine-tuning). Such idea was first introduced by Wu et al. [160] in medical imaging and by Liu et al. [161,162] in computer vision for image denoising. By doing so, the second network for image analysis can be regarded as a task-based image quality metric that is learned from the data. As shown in Wu et al. [160], where the image analysis task was lung nodule recognition, the learned image quality metric automatically placed more emphasis within the lung areas and less emphasis elsewhere. Such a quality metric is specific to the task of lung nodule recognition since the image quality outside of the lung region is irrelevant to the task.

5 Challenge and Opportunities

Although deep learning based models continue to dominant medical imaging, there are still plenty of remaining challenges in deep modeling which limit the application and implementation of these new methods in clinical practice. These challenges also present themselves as new opportunities for researchers working in related fields.

- The everlasting hunger of labeled data. There are only limited labeled data available to develop new deep models in medical imaging. Annotation of medical images is time-consuming and requires expert knowledge from physicians. Can we design effective learning models that can make good use of both the (very limited) labeled data and the (relatively more abundant) unlabeled data?
- The limited number of observations. Due to morbidity and privacy concerns, it is generally difficult to gather very large medical data for a specific task. Furthermore, the number of rare cases is (by definition) small but can be much more valuable than common cases. Can we design learning models and data augmentation techniques to effectively extract knowledge from these limited samples and acknowledge such unequal importance among the samples?
- Radiologists do not make the clinical decision only based on images. More information from the patients and the knowledge of the doctors from their years of training in medical school are also crucial in decision making. Thus, incorporating data gathered from multiple diverse sources into deep modeling is important in improving system performance.
- Reasoning is just as important as, if not more important than, inferencing. Currently, most deep models hide the reasoning procedure. There is a chance that the model makes accurate predictions based on wrong reasoning. This makes the model unreliable. Can we incorporate deep modeling with reasoning (such as causal inference) or with medical knowledge graph? This may further reduce the amount of annotated data we need to train deep models without hurting performance.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

[1] Pavlovic, G., Tekalp, A.M.: Maximum likelihood parametric blur identification based on a continuous spatial domain model. IEEE Trans. Image Process. **1**(4), 496–504 (1992)
[2] Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, pp. 417–424. ACM Press/Addison-Wesley Publishing Co. (2000)
[3] Brown, R.W., Haacke, E.M., Cheng, Y.C.N., Thompson, M.R., Venkatesan, R.: Magnetic Resonance Imaging: Physical Principles and Sequence Design. Wiley, Hoboken (2014)
[4] Buzug, T.M.: Computed Tomography: From Photon Statistics to Modern Cone-Beam CT. Springer, Berlin (2008)
[5] Choi, J.K., Park, H.S., Wang, S., Wang, Y., Seo, J.K.: Inverse problem in quantitative susceptibility mapping. SIAM J. Imaging Sci. **7**(3), 1669–1689 (2014)
[6] Natterer, F.: Image reconstruction in quantitative susceptibility mapping. SIAM J. Imaging Sci. **9**(3), 1127–1131 (2016)
[7] de Rochefort, L., Liu, T., Kressler, B., Liu, J., Spincemaille, P., Lebon, V., Wu, J., Wang, Y.: Quantitative susceptibility map reconstruction from MR phase data using Bayesian regularization: validation and application to brain imaging. Magn. Reson. Med. **63**(1), 194–206 (2010)
[8] Wang, Y., Liu, T.: Quantitative susceptibility mapping (QSM): decoding MRI data for a tissue magnetic biomarker. Magn. Reson. Med. **73**(1), 82–101 (2015)

[9] Rudin, L., Lions, P.L., Osher, S.: Multiplicative denoising and deblurring: theory and algorithms. In: Osher, S., Paragios, N. (eds.) Geometric Level Set Methods in Imaging, Vision, and Graphics, pp. 103–119. Springer, Berlin (2003)
[10] Aubert, G., Kornprobst, P.: Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations. Springer, Berlin (2006)
[11] Chan, T.F., Shen, J.: Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods. SIAM, Philadelphia (2005)
[12] Dong, B., Shen, Z.: Image restoration: a data-driven perspective. In: Proceedings of the International Congress of Industrial and Applied Mathematics (ICIAM), pp. 65–108 (2015)
[13] Shen, Z.: Wavelet frames and image restorations. In: Proceedings of the International Congress of Mathematicians, vol. 4, pp. 2834–2863. World Scientific (2010)
[14] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**(12), 3371–3408 (2010)
[15] Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention Society, pp. 234–241 (2015)
[16] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
[17] He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision, pp. 630–645 (2016)
[18] Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**(1), 259–268 (1992)
[19] Perona, P., Shiota, T., Malik, J.: Anisotropic diffusion. In: Romeny, B.M.H. (ed.) Geometry-Driven Diffusion in Computer Vision, pp. 73–92. Springer, Berlin (1994)
[20] Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Mach. Intell. **12**(7), 629–639 (1990)
[21] Osher, S., Rudin, L.I.: Feature-oriented image enhancement using shock filters. SIAM J. Numer. Anal. **27**(4), 919–940 (1990)
[22] Alvarez, L., Mazorra, L.: Signal and image restoration using shock filters and anisotropic diffusion. SIAM J. Numer. Anal. **31**(2), 590–605 (1994)
[23] Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 60–65 (2005)
[24] Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. Multiscale Model. Simul. **4**(2), 490–530 (2005)
[25] Buades, A., Coll, B., Morel, J.M.: Image denoising methods. A new nonlocal principle. SIAM Rev. **52**(1), 113–147 (2010)
[26] Lou, Y., Zhang, X., Osher, S., Bertozzi, A.: Image recovery via nonlocal operators. J. Sci. Comput. **42**(2), 185–197 (2010)
[27] Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Trans. Image Process. **16**(8), 2080–2095 (2007)
[28] Daubechies, I.: Ten Lectures on Wavelets. SIAM, Philadelphia (1992)
[29] Mallat, S.: A Wavelet Tour of Signal Processing, The Sparse Way, 3rd edn. Academic Press, Burlington, MA (2009)
[30] Ron, A., Shen, Z.: Affine systems in $l_2(\mathbb{R}^d)$: the analysis of the analysis operator. J. Funct. Anal. **148**(2), 408–447 (1997)
[31] Dong, B., Shen, Z.: MRA-based wavelet frames and applications. In: Zhao, H.-K. (ed.) Mathematics in Image Processing. IAS Lecture Notes Series, vol. 19. American Mathematical Society, Providence (2013)
[32] Gu, S., Zhang, L., Zuo, W., Feng, X.: Weighted nuclear norm minimization with application to image denoising. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2862–2869 (2014)
[33] Engan, K., Aase, S.O., Husoy, J.H.: Method of optimal directions for frame design. In: IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP), vol. 5, pp. 2443–2446. IEEE (1999)
[34] Aharon, M., Elad, M., Bruckstein, A., et al.: K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. **54**(11), 4311 (2006)

[35] Liu, R., Lin, Z., Zhang, W., Su, Z.: Learning PDEs for image restoration via optimal control. In: European Conference on Computer Vision, pp. 115–128. Springer (2010)
[36] Cai, J.F., Ji, H., Shen, Z., Ye, G.B.: Data-driven tight frame construction and image denoising. Appl. Comput. Harmon. Anal. **37**(1), 89–105 (2014)
[37] Bao, C., Ji, H., Shen, Z.: Convergence analysis for iterative data-driven tight frame construction scheme. Appl. Comput. Harmon. Anal. **38**(3), 510–523 (2015)
[38] Tai, C., Weinan, E.: Multiscale adaptive representation of signals: I. The basic framework. J. Mach. Learn. Res. **17**(1), 4875–4912 (2016)
[39] Wright, J., Ganesh, A., Rao, S., Peng, Y., Ma, Y.: Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization. In: Neural Information Processing Systems, pp. 2080–2088 (2009)
[40] Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 171–184 (2013)
[41] Cai, J.F., Jia, X., Gao, H., Jiang, S.B., Shen, Z., Zhao, H.: Cine cone beam CT reconstruction using low-rank matrix factorization: algorithm and a proof-of-principle study. IEEE Trans. Med. Imaging **33**(8), 1581–1591 (2014)
[42] Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Found. Comput. Math. **9**(6), 717 (2009)
[43] Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. **20**(4), 1956–1982 (2010)
[44] Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. Commun. Pure Appl. Math. **42**(5), 577–685 (1989)
[45] Cai, J.F., Dong, B., Shen, Z.: Image restoration: a wavelet frame based model for piecewise smooth functions and beyond. Appl. Comput. Harmon. Anal. **41**(1), 94–138 (2016)
[46] Heimann, T., Meinzer, H.P.: Statistical shape models for 3D medical image segmentation: a review. Med. Image Anal. **13**(4), 543–563 (2009)
[47] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems, pp. 1097–1105 (2012)
[48] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Neural Information Processing Systems, pp. 2672–2680 (2014)
[49] Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
[50] Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**(1), 17–40 (1976)
[51] Glowinski, R., Marroco, A.: Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique **9**(R2), 41–76 (1975)
[52] Zhu, M., Chan, T.: An efficient primal-dual hybrid gradient algorithm for total variation image restoration. UCLA CAM Report, vol. 34 (2008)
[53] Esser, E., Zhang, X., Chan, T.F.: A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. SIAM J. Imaging Sci. **3**(4), 1015–1046 (2010)
[54] Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
[55] Cai, J.F., Osher, S., Shen, Z.: Split Bregman methods and frame based image restoration. Multiscale Model. Simul. **8**(2), 337–369 (2009)
[56] Goldstein, T., Osher, S.: The split Bregman method for l_1-regularized problems. SIAM J. Imaging Sci. **2**(2), 323–343 (2009)
[57] Yin, W., Osher, S., Goldfarb, D., Darbon, J.: Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. SIAM J. Imaging Sci. **1**(1), 143–168 (2008)
[58] Osher, S., Mao, Y., Dong, B., Yin, W.: Fast linearized Bregman iteration for compressive sensing and sparse denoising. Commun. Math. Sci. **8**(1), 93–111 (2010)
[59] Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pure Appl. Math. **57**(11), 1413–1457 (2004)
[60] Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)

[61] Bruck Jr., R.E.: On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space. J. Math. Anal. Appl. **61**(1), 159–164 (1977)
[62] Passty, G.B.: Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. J. Math. Anal. Appl. **72**, 383–290 (1979)
[63] Shen, Z., Toh, K.C., Yun, S.: An accelerated proximal gradient algorithm for frame-based image restoration via the balanced approach. SIAM J. Imaging Sci. **4**(2), 573–596 (2011)
[64] Nesterov, Y.E.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Dokl. Akad. Nauk SSSR **269**, 543–547 (1983)
[65] Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer, Berlin (2006)
[66] Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT, pp. 177–186. Springer (2010)
[67] Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**(3), 400–407 (1951)
[68] Bottou, L.: Stochastic gradient descent tricks. In: Orr, G.B., Müller, K.R. (eds.) Neural Networks: Tricks of the Trade, pp. 421–436. Springer, Berlin (2012)
[69] Zhang, T.: Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: International Conference on Machine Learning, pp. 116–123. ACM (2004)
[70] Nitanda, A.: Stochastic proximal gradient descent with acceleration techniques. In: Neural Information Processing Systems, pp. 1574–1582 (2014)
[71] Zhang, Y., Xiao, L.: Stochastic primal-dual coordinate method for regularized empirical risk minimization. J. Mach. Learn. Res. **18**(1), 2939–2980 (2017)
[72] Konečnỳ, J., Liu, J., Richtárik, P., Takáč, M.: Mini-batch semi-stochastic gradient descent in the proximal setting. IEEE J. Sel. Top. Signal Process. **10**(2), 242–255 (2016)
[73] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015)
[74] Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**(Jul), 2121–2159 (2011)
[75] Hinton, G.: Neural networks for machine learning. Coursera, video lectures (2012)
[76] Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. **60**(2), 223–311 (2018)
[77] Gregor, K., LeCun, Y.: Learning fast approximations of sparse coding. In: International Conference on Machine Learning, pp. 399–406 (2010)
[78] Chen, Y., Yu, W., Pock, T.: On learning optimized reaction diffusion processes for effective image restoration. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5261–5269 (2015)
[79] Yang, Y., Sun, J., Li, H., Xu, Z.: Deep ADMM-Net for compressive sensing MRI. In: Neural Information Processing Systems, pp. 10–18 (2016)
[80] Adler, J., Öktem, O.: Learned primal-dual reconstruction. IEEE Trans. Med. Imaging **37**(6), 1322–1332 (2018)
[81] Solomon, O., Cohen, R., Zhang, Y., Yang, Y., Qiong, H., Luo, J., van Sloun, R.J., Eldar, Y.C.: Deep unfolded robust PCA with application to clutter suppression in ultrasound. arXiv preprint arXiv:1811.08252 (2018)
[82] Chen, X., Liu, J., Wang, Z., Yin, W.: Theoretical linear convergence of unfolded ISTA and its practical weights and thresholds. In: Neural Information Processing Systems, pp. 9079–9089 (2018)
[83] Liu, R., Cheng, S., He, Y., Fan, X., Lin, Z., Luo, Z.: On the convergence of learning-based iterative methods for nonconvex inverse problems. IEEE Trans. Pattern Anal. Mach. Intell. (2019). https://doi.org/10.1109/TPAMI.2019.2920591
[84] Li, H., Yang, Y., Chen, D., Lin, Z.: Optimization algorithm inspired deep neural network structure design. In: Asian Conference on Machine Learning, pp. 614–629 (2018)
[85] Zhang, H., Dong, B., Liu, B.: JSR-Net: a deep network for joint spatial-radon domain CT reconstruction from incomplete data. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)-2019, pp. 3657–3661 (2019). https://doi.org/10.1109/ICASSP.2019.8682178
[86] Weinan, E.: A proposal on machine learning via dynamical systems. Commun. Math. Stat. **5**(1), 1–11 (2017)
[87] Chang, B., Meng, L., Haber, E., Tung, F., Begert, D.: Multi-level residual networks from dynamical systems view. In: International Conference on Learning Representations Poster (2018)
[88] Li, Z., Shi, Z.: Deep residual learning and PDEs on manifold. arXiv:1708.05115 (2017)

[89] Chang, B., Meng, L., Haber, E., Ruthotto, L., Begert, D., Holtham, E.: Reversible architectures for arbitrarily deep residual neural networks. In: AAAI Conference on Artificial Intelligence (2018)
[90] Lu, Y., Zhong, A., Li, Q., Dong, B.: Beyond finite layer neural networks: bridging deep architectures and numerical differential equations. In: International Conference on Machine Learning, pp. 3276–3285 (2018)
[91] Wang, B., Yuan, B., Shi, Z., Osher, S.J.: Enresnet: Resnet ensemble via the Feynman–Kac formalism. arXiv:1811.10745 (2018)
[92] Ruthotto, L., Haber, E.: Deep neural networks motivated by partial differential equations. arXiv:1804.04272 (2018)
[93] Tao, Y., Sun, Q., Du, Q., Liu, W.: Nonlocal neural networks, nonlocal diffusion and nonlocal modeling. In: Neural Information Processing Systems, pp. 494–504. Curran Associates, Inc. (2018)
[94] Zhang, D., Zhang, T., Lu, Y., Zhu, Z., Dong, B.: You only propagate once: accelerating adversarial training via maximal principle. In: Neural Information Processing Systems (2019)
[95] Zhang, X., Lu, Y., Liu, J., Dong, B.: Dynamically unfolding recurrent restorer: a moving endpoint control method for image restoration. In: International Conference on Learning Representations (2019)
[96] Long, Z., Lu, Y., Ma, X., Dong, B.: PDE-Net: learning PDEs from data. In: International Conference on Machine Learning, pp. 3214–3222 (2018)
[97] Long, Z., Lu, Y., Dong, B.: PDE-Net 2.0: Learning PDEs from data with a numeric-symbolic hybrid deep network. J. Comput. Phys. **339**, 108925 (2019)
[98] Lu, Y., Li, Z., He, D., Sun, Z., Dong, B., Qin, T., Wang, L., Liu, T.Y.: Understanding and improving transformer from a multi-particle dynamic system point of view. arXiv:1906.02762 (2019)
[99] He, J., Xu, J.: MgNet: a unified framework of multigrid and convolutional neural network. Sci. China Math. **62**, 1331–1354 (2019)
[100] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 4700–4708 (2017)
[101] Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: Neural Information Processing Systems, pp. 153–160 (2007)
[102] Poultney, C., Chopra, S., Cun, Y.L., et al.: Efficient learning of sparse representations with an energy-based model. In: Neural Information Processing Systems, pp. 1137–1144 (2007)
[103] Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
[104] Mao, X., Shen, C., Yang, Y.B.: Image restoration using very deep convolutional encoder–decoder networks with symmetric skip connections. In: Neural Information Processing Systems, pp. 2802–2810 (2016)
[105] Chen, H., Zhang, Y., Kalra, M.K., Lin, F., Chen, Y., Liao, P., Zhou, J., Wang, G.: Low-dose CT with a residual encoder–decoder convolutional neural network. IEEE Trans. Med. Imaging **36**(12), 2524–2535 (2017)
[106] Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
[107] Yin, R., Gao, T., Lu, Y.M., Daubechies, I.: A tale of two bases: local-nonlocal regularization on image patches with convolution framelets. SIAM J. Imaging Sci. **10**(2), 711–750 (2017)
[108] Ye, J.C., Han, Y., Cha, E.: Deep convolutional framelets: a general deep learning framework for inverse problems. SIAM J. Imaging Sci. **11**(2), 991–1048 (2018)
[109] Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., et al.: U-Net: deep learning for cell counting, detection, and morphometry. Nat. Methods **16**, 67–70 (2019)
[110] DeVore, R., Lorentz, G.: Constructive Approximation. Springer, Berlin (1993)
[111] Hornik, K.: Approximation capabilities of multilayer feedforward networks. Neural Netw. **4**(2), 251–257 (1991)
[112] Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989)
[113] Pinkus, A.: Approximation theory of the MLP model in neural networks. Acta Numer. **8**, 143–195 (1999)

[114] Cybenko, G.: Approximation by superpositions of a sigmoidal function. Math. Control Signal Syst. **2**(4), 303–314 (1989)
[115] Funahashi, K.I.: On the approximate realization of continuous mappings by neural networks. Neural Netw. **2**(3), 183–192 (1989)
[116] Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. Inf. Theory **39**(3), 930–945 (1993)
[117] Liang, S., Srikant, R.: Why deep neural networks for function approximation? In: International Conference on Learning Representations (2017)
[118] Mhaskar, H., Liao, Q., Poggio, T.: Learning functions: when is deep better than shallow. arXiv:1603.00988 (2016)
[119] Eldan, R., Shamir, O.: The power of depth for feedforward neural networks. In: Conference on Learning Theory, pp. 907–940 (2016)
[120] Cohen, N., Sharir, O., Shashua, A.: On the expressive power of deep learning: a tensor analysis. In: Conference on Learning Theory, pp. 698–728 (2016)
[121] Delalleau, O., Bengio, Y.: Shallow vs. deep sum-product networks. In: Neural Information Processing Systems, pp. 666–674 (2011)
[122] Telgarsky, M.: Representation benefits of deep feedforward networks. arXiv:1509.08101 (2015)
[123] Telgarsky, M.: Benefits of depth in neural networks. In: Conference on Learning Theory, vol. 49, pp. 1–23 (2016)
[124] Lu, Z., Pu, H., Wang, F., Hu, Z., Wang, L.: The expressive power of neural networks: a view from the width. In: Neural Information Processing Systems, pp. 6231–6239 (2017)
[125] Hanin, B., Sellke, M.: Approximating continuous functions by ReLU nets of minimal width. arXiv:1710.11278 (2017)
[126] Hanin, B.: Universal function approximation by deep neural nets with bounded width and ReLU activations. Mathematics **7**(10), 992 (2019)
[127] Yarotsky, D.: Optimal approximation of continuous functions by very deep ReLU networks. In: Conference on Learning Theory (2018)
[128] Rolnick, D., Tegmark, M.: The power of deeper networks for expressing natural functions. In: International Conference on Learning Representations (2018)
[129] Shen, Z., Yang, H., Zhang, S.: Nonlinear approximation via compositions. Neural Netw. **119**, 74–84 (2019)
[130] Veit, A., Wilber, M.J., Belongie, S.: Residual networks behave like ensembles of relatively shallow networks. In: Neural Information Processing Systems, pp. 550–558 (2016)
[131] Lin, H., Jegelka, S.: ResNet with one-neuron hidden layers is a universal approximator. In: Neural Information Processing Systems, pp. 6172–6181 (2018)
[132] E, W., Ma, C., Wang, Q.: A priori estimates of the population risk for residual networks (2019)
[133] He, J., Li, L., Xu, J., Zheng, C.: ReLU deep neural networks and linear finite elements. arXiv:1807.03973 (2018)
[134] Nochetto, R.H., Veeser, A.: Primer of adaptive finite element methods. In: Naldi, G., Russo, G. (eds.) Multiscale and Adaptivity: Modeling, Numerics and Applications, pp. 125–225. Springer, Berlin (2011)
[135] Cessac, B.: A view of neural networks as dynamical systems. Int. J. Bifurc. Chaos **20**(06), 1585–1629 (2010)
[136] Sonoda, S., Murata, N.: Double continuum limit of deep neural networks. In: ICML Workshop (2017)
[137] Thorpe, M., van Gennip, Y.: Deep limits of residual neural networks. arXiv:1810.11741 (2018)
[138] Weinan, E., Han, J., Li, Q.: A mean-field optimal control formulation of deep learning. Res. Math. Sci. **6**(10), 1–41 (2019). https://doi.org/10.1007/s40687-018-0172-y
[139] Li, Q., Chen, L., Tai, C., Weinan, E.: Maximum principle based algorithms for deep learning. J. Mach. Learn. Res. **18**(1), 5998–6026 (2017)
[140] Chen, T.Q., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. In: Neural Information Processing Systems, pp. 6572–6583 (2018)
[141] Zhang, X., Li, Z., Loy, C.C., Lin, D.: Polynet: a pursuit of structural diversity in very deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3900–3908 (2017)
[142] Larsson, G., Maire, M., Shakhnarovich, G.: Fractalnet: ultra-deep neural networks without residuals. In: International Conference on Learning Representations (2016)

[143] Gomez, A.N., Ren, M., Urtasun, R., Grosse, R.B.: The reversible residual network: backpropagation without storing activations. In: Neural Information Processing Systems, pp. 2214–2224 (2017)
[144] Zhang, J., Han, B., Wynter, L., Low, K.H., Kankanhalli, M.: Towards robust ResNet: a small step but a giant leap. In: International Joint Conference on Artificial Intelligence, pp. 4285–4291 (2019)
[145] Ascher, U.M., Petzold, L.R.: Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations, vol. 61. SIAM, Philadelphia (1998)
[146] Zhu, M., Chang, B., Fu, C.: Convolutional neural networks combined with Runge–Kutta methods. arXiv:1802.08831 (2018)
[147] Warming, R., Hyett, B.: The modified equation approach to the stability and accuracy analysis of finite-difference methods. J. Comput. Phys. **14**(2), 159–179 (1974)
[148] Su, W., Boyd, S., Candès, E.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. In: Neural Information Processing Systems, pp. 2510–2518 (2014)
[149] Wilson, A.C., Recht, B., Jordan, M.I.: A Lyapunov analysis of momentum methods in optimization. arXiv:1611.02635 (2016)
[150] Dong, B., Jiang, Q., Shen, Z.: Image restoration: wavelet frame shrinkage, nonlinear evolution PDEs, and beyond. Multiscale Model. Simul. **15**(1), 606–660 (2017)
[151] Gastaldi, X.: Shake-shake regularization. In: International Conference on Learning Representations Workshop (2017)
[152] Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: European Conference on Computer Vision, pp. 646–661 (2016)
[153] Sun, Q., Tao, Y., Du, Q.: Stochastic training of residual networks: a differential equation viewpoint. arXiv preprint arXiv:1812.00174 (2018)
[154] Natterer, F.: The Mathematics of Computerized Tomography. SIAM, Philadelphia (2001)
[155] Zeng, G.L.: Medical Image Reconstruction: A Conceptual Tutorial. Springer, Berlin (2010)
[156] Scherzer, O. (ed.): Handbook of Mathematical Methods in Imaging, 2nd edn. Springer, New York (2015)
[157] Herman, G.T.: Fundamentals of Computerized Tomography: Image Reconstruction from Projections. Springer, Berlin (2009)
[158] Zhu, B., Liu, J.Z., Cauley, S.F., Rosen, B.R., Rosen, M.S.: Image reconstruction by domain-transform manifold learning. Nature **555**(7697), 487 (2018)
[159] Kalra, M., Wang, G., Orton, C.G.: Radiomics in lung cancer: its time is here. Med. Phys. **45**(3), 997–1000 (2018)
[160] Wu, D., Kim, K., Dong, B., El Fakhri, G., Li, Q.: End-to-end lung nodule detection in computed tomography. In: International Workshop on Machine Learning in Medical Imaging, pp. 37–45. Springer (2018)
[161] Liu, D., Wen, B., Liu, X., Wang, Z., Huang, T.S.: When image denoising meets high-level vision tasks: a deep learning approach. In: International Joint Conference on Artificial Intelligence, pp. 842–848 (2018)
[162] Liu, D., Wen, B., Jiao, J., Liu, X., Wang, Z., Huang, T.S.: Connecting image denoising and high-level vision tasks via deep learning. arXiv preprint arXiv:1809.01826 (2018)
[163] Zhang, Z., Liang, X., Dong, X., Xie, Y., Cao, G.: A sparse-view CT reconstruction method based on combination of densenet and deconvolution. IEEE Trans. Med. Imaging **37**(6), 1407–1417 (2018)
[164] Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M.K., Zhang, Y., Sun, L., Wang, G.: Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. IEEE Trans. Med. Imaging **37**(6), 1348–1357 (2018)
[165] Jin, K.H., Mccann, M.T., Froustey, E., Unser, M.: Deep convolutional neural network for inverse problems in imaging. IEEE Trans. Image Process. **26**(9), 4509–4522 (2017)
[166] Han, Y.S., Yoo, J., Ye, J.C.: Deep residual learning for compressed sensing CT reconstruction via persistent homology analysis. arXiv preprint arXiv:1611.06391 (2016)
[167] Liu, J., Chen, X., Wang, Z., Yin, W.: ALISTA: Analytic weights are as good as learned weights in International Conference on Learning Representations. In: ICLR (2019)
[168] Xie, X., Wu, J., Zhong, Z., Liu, G., Lin, Z.: Differentiable linearized ADMM. In: International Conference on Machine Learning (2019)
[169] Yang, Y., Sun, J., Li, H., Xu, Z.: ADMM-Net: a deep learning approach for compressive sensing MRI. arXiv preprint arXiv:1705.06869 (2017)
[170] Parikh, N., Boyd, S., et al.: Proximal algorithms. Found. Trends® Optim. **1**(3), 127–239 (2014)

[171] Adler, J., Öktem, O.: Solving ill-posed inverse problems using iterative deep neural networks. Inverse Probl. **33**, 124007 (2017)
[172] Dong, B., Li, J., Shen, Z.: X-ray CT image reconstruction via wavelet frame based regularization and radon domain inpainting. J. Sci. Comput. **54**(2), 333–349 (2013)
[173] Burger, M., Müller, J., Papoutsellis, E., Schönlieb, C.B.: Total variation regularization in measurement and image space for PET reconstruction. Inverse Probl. **30**(10), 105003 (2014)
[174] Zhan, R., Dong, B.: CT image reconstruction by spatial-radon domain data-driven tight frame regularization. SIAM J. Imaging Sci. **9**(3), 1063–1083 (2016)
[175] Zhang, H., Dong, B., Liu, B.: A reweighted joint spatial-radon domain CT image reconstruction model for metal artifact reduction. SIAM J. Imaging Sci. **11**(1), 707–733 (2018)