Total Deep Variation for Linear Inverse Problems

Erich Kobler[†,*]   Alexander Effland[†,*]   Karl Kunisch[°]   Thomas Pock[†]

[†]Graz University of Technology   [°]University of Graz

Abstract

Diverse inverse problems in imaging can be cast as variational problems composed of a task-specific data fidelity term and a regularization term. In this paper, we propose a novel learnable general-purpose regularizer exploiting recent architectural design patterns from deep learning. We cast the learning problem as a discrete sampled optimal control problem, for which we derive the adjoint state equations and an optimality condition. By exploiting the variational structure of our approach, we perform a sensitivity analysis with respect to the learned parameters obtained from different training datasets. Moreover, we carry out a nonlinear eigenfunction analysis, which reveals interesting properties of the learned regularizer. We show state-of-the-art performance for classical image restoration and medical image reconstruction problems.

1. Introduction

The statistical interpretation of linear inverse problems allows the treatment of measurement uncertainties in the input data z and missing information in a rigorous framework. Bayes' theorem states that the posterior distribution $p(x|z)$ is proportional to the product of the data likelihood $p(z|x)$ and the prior $p(x)$, which represents the belief in a certain solution x given the input data z. A classical estimator for x is given by the maximum a posterior (MAP) estimator, which in a negative log-domain amounts to minimizing the variational problem

$$\mathcal{E}(x) := \mathcal{D}(x, z) + \mathcal{R}(x). \qquad (1)$$

Here, the data fidelity term $\mathcal{D}$ can be identified with the negative log-likelihood $-\log p(z|x)$ and the regularization term corresponds to the negative log-probability of the prior distribution $-\log p(x)$. Assuming Gaussian noise in the data z, the data fidelity term naturally arises from the negative log-Gaussian, which essentially leads to a quadratic ℓ^2-term of the form $\mathcal{D}(x, z) := \frac{1}{2}\|Ax - z\|_2^2$. In this paper, we assume that A is a task-specific linear operator (see [4])
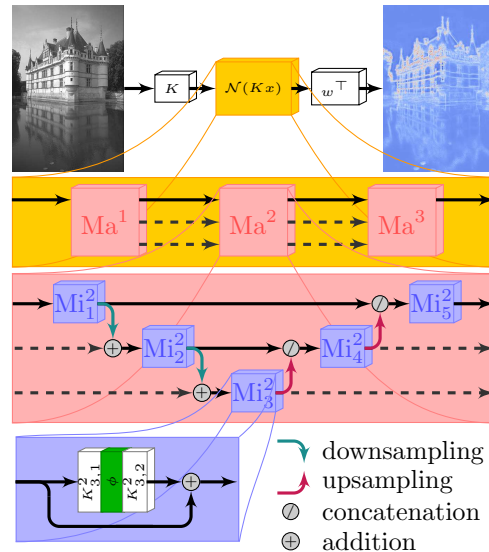
---

* indicates equal contribution

Figure 1. Visualization of the deep regularizer r (TDV[3]), where colors indicate different hierarchical levels. On the highest level, $r(x, \theta) = w^\top \mathcal{N}(Kx)$ assigns to each pixel an energy value incorporating the local neighborhood. The function $\mathcal{N}$ (yellow) is composed of three macro-blocks (red), each representing a CNN with a U-Net type architecture. The macro-blocks consist of five micro-blocks (blue) with a residual structure on three scales.

such as a downsampling operator in the case of single image super-resolution. While the data fidelity term is straightforward to model, the grand challenge in inverse problems for imaging is the design of a regularizer that captures the complexity of the statistics of natural images.

A classical and widely used regularizer is the total variation (TV) originally proposed in [37], which is based on the first principle assumption that images are piecewise constant with sparse gradients. A well-known caveat of the sparsity assumption of TV is the formation of clearly visible artifacts known as staircasing effect. To overcome this problem, the first principle assumption has later been extended to piecewise smooth images incorporating higher order image derivatives such as infimal convolution based models [3] or the total generalized variation [2]. Inspired by the fact that edge continuity plays a fundamental role in the hu-

man visual perception, regularizers penalizing the curvature of level lines have been proposed in [31, 6, 5]. While these regularizers are mathematically well-understood, the complexity of natural images is only partially reflected in their formulation. For this reason, handcrafted variational methods have nowadays been largely outperformed by purely data-driven methods.

It has been recognized quite early that a proper statistical modeling of regularizers should be based on learning [46], which has recently been advocated e.g. in [27, 24]. One of the most successful early approaches is the Fields of Experts (FoE) regularizer [36], which can be interpreted as a generalization of the total variation, but builds upon learned filters and learned potential functions. While the FoE prior was originally learned generatively, it was shown in [38] that a discriminative learning via implicit differentiation yields improved performance. A computationally more feasible method for discriminative learning is based on unrolling a finite number of iterations of a gradient descent algorithm [10]. Additionally using iteration dependent parameters in the regularizer was shown to significantly increase the performance (TNRD [8], [23]). In [22], variational networks (VNs) are proposed, which give an incremental proximal gradient interpretation of TNRD.

Interestingly, such truncated schemes are not only computationally much more efficient, but are also superior in performance with respect to the full minimization. A continuous time formulation of this phenomenon was proposed in [12] by means of an optimal control problem, within which an optimal stopping time is learned.

An alternative approach to incorporate a regularizer into a proximal algorithm, known as plug-and-play prior [42] or regularization by denoising [34], is the replacement of the proximal operator by an existing denoising algorithm such as BM3D [9]. Combining this idea with deep learning was proposed in [30, 33]. However, all the aforementioned schemes lack a variational structure and thus are not interpretable in the framework of MAP inference.

In this paper, we introduce a novel regularizer, which is inspired by the design patterns of state-of-the-art deep convolutional neural networks and simultaneously ensures a variational structure. We achieve this by representing the total energy of the regularizer by means of a residual multi-scale network (Figure 1) leveraging smooth activation functions. In analogy to [12], we start from a gradient flow of the variational energy and utilize a semi-implicit time discretization, for which we derive the discrete adjoint state equation using the discrete Pontryagin maximum principle. Furthermore, we present a first order necessary condition of optimality to automatize the computation of the optimal stopping time. Our proposed *Total Deep Variation* (TDV) regularizer can be used as a generic regularizer in variational formulations of linear inverse problems. The major

contributions of this paper are as follows:

- The design of a novel generic multi-scale variational regularizer learned from data.
- A rigorous mathematical analysis including a sampled optimal control formulation of the learning problem and a sensitivity analysis of the learned estimator with respect to the training dataset.
- A nonlinear eigenfunction analysis for the visualization and understanding of the learned regularizer.
- State-of-the-art results on a number of classical image restoration and medical image reconstruction problems with an impressively low number of learned parameters.

Due to space limitations, all proofs are presented in the supplementary material.

2. Sampled optimal control problem

Let $x \in \mathbb{R}^{nC}$ be a corrupted input image with a resolution of $n = n_1 \cdot n_2$ and C channels. We emphasize that the subsequent analysis can easily be transferred to image data in any dimension.

The variational approach for inverse problems frequently amounts to computing a minimizer of a specific energy functional $\mathcal{E}$ of the form $\mathcal{E}(x, \theta, z) := \mathcal{D}(x, z) + \mathcal{R}(x, \theta)$. Here, $\mathcal{D}$ is a data fidelity term and a $\mathcal{R}$ is a parametric regularizer depending on the learned training parameters $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^p$ is a compact and convex set of training parameters. Typically, the minimizer of this variational problem is considered as an approximation of the uncorrupted ground truth image. Accordingly, in this paper we analyze different data fidelity terms of the form $\mathcal{D}(x, z) = \frac{1}{2}\|Ax - z\|_2^2$ for fixed task-dependent $A \in \mathbb{R}^{lC \times nC}$ and fixed $z \in \mathbb{R}^{lC}$.

The proposed total deep variation (TDV) regularizer is given by $\mathcal{R}(x, \theta) = \sum_{i=1}^{n} r(x, \theta)_i \in \mathbb{R}$, which is the total sum of the pixelwise deep variation defined as $r(x, \theta) = w^\top \mathcal{N}(Kx) \in \mathbb{R}^n$. Here, $K \in \mathbb{R}^{nm \times nC}$ is a learned convolution kernel with zero-mean constraint (i.e. $\sum_{i=1}^{nC} K_{j,i} = 0$ for $j = 1, \ldots, nm$), $\mathcal{N} : \mathbb{R}^{nm} \to \mathbb{R}^{nq}$ is a multiscale convolutional neural network and $w \in \mathbb{R}^q$ is a learned weight vector. Hence, θ encodes the kernel K, the weights of the convolutional layers in $\mathcal{N}$ and the weight vector w. The exact form of the network is depicted in Figure 1. In the TDV^l network for $l \in \mathbb{N}$, the function $\mathcal{N}$ is composed of l consecutive macro-blocks $\mathbf{Ma}^1, \ldots, \mathbf{Ma}^l$ (red blocks). Each macro-block $\mathbf{Ma}^i$ ($i \in \{1, \ldots, l\}$) has a U-Net type architecture [35] with five micro-blocks $\mathbf{Mi}_1^i, \ldots, \mathbf{Mi}_5^i$ (blue blocks) distributed over three scales with skip-connections on each scale. In addition, residual connections are added between different scales of consecutive macro-blocks whenever possible. Finally, each micro-block $\mathbf{Mi}_j^i$ for $i \in \{1, \ldots, l\}$ and $j \in \{1, \ldots, 5\}$ has a

residual structure defined as $\mathbf{Mi}_j^i(x) = x + K_{j,2}^i \phi(K_{j,1}^i x)$ for convolution operators $K_{j,1}^i$ and $K_{j,2}^i$. We use a smooth log-student-t-distribution of the form $\phi(x) = \frac{1}{2\nu} \log(1 + \nu x^2)$, which is an established model for the statistics of natural images [18] and has the properties $\phi'(0) = 0$ and $\phi''(0) = 1$. All convolution operators $K_{j,1}^i$ and $K_{j,2}^i$ in each micro-block correspond to 3×3 convolutions with m feature channels and no additional bias. To avoid aliasing, downsampling and upsampling are implemented incorporating 3×3 convolutions and transposed convolutions with stride 2 using a blurring of the kernels proposed by [45]. Finally, the concatenation maps $2m$ feature channels to m feature channels using a single 1×1 convolution.

A common strategy for the minimization of the energy $\mathcal{E}$ is the incorporation of a *gradient flow* [1] posed on a finite time interval $(0, T)$, which reads as

$$\dot{\tilde{x}}(t) = -\nabla_1 \mathcal{E}(\tilde{x}(t), \theta, z) = f(\tilde{x}(t), \theta, z) \quad (2)$$

$$:= -A^\top(A\tilde{x}(t) - z) - \nabla_1 \mathcal{R}(\tilde{x}(t), \theta) \quad (3)$$

for $t \in (0, T)$, where $\tilde{x}(0) = x_{\text{init}}$ for a fixed *initial value* $x_{\text{init}} \in \mathbb{R}^{nC}$. Here, $\tilde{x} : [0, T] \to \mathbb{R}^{nC}$ denotes the differentiable *flow of $\mathcal{E}$* and $T > 0$ refers to the *time horizon*. We assume that $T \in [0, T_{\max}]$ for a fixed $T_{\max} > 0$. To formulate optimal stopping, we exploit the reparametrization $x(t) = \tilde{x}(tT)$, which results for $t \in (0, 1)$ in the equivalent gradient flow formulation

$$\dot{x}(t) = Tf(x(t), \theta, z), \qquad x(0) = x_{\text{init}}. \quad (4)$$

Note that the gradient flow inherently implies a variational structure (cf. (2)).

Following [11, 25, 26], we cast the training process as a *sampled optimal control problem* with control parameters θ and T. For fixed $N \in \mathbb{N}$, let $(x_{\text{init}}^i, y^i, z^i)_{i=1}^N \in (\mathbb{R}^{nC} \times \mathbb{R}^{nC} \times \mathbb{R}^{lC})^N$ be a collection of N triplets of initial-target image pairs, and observed data independently drawn from a task-dependent fixed probability distribution. In additive Gaussian image denoising, for instance, a ground truth image y^i is deteriorated by noise $n^i \sim \mathcal{N}(0, \sigma^2)$ and a common choice is $x_{\text{init}}^i = z^i = y^i + n^i$. Let $l : \mathbb{R}^{nC} \to \mathbb{R}_0^+$ be a convex, twice continuously differentiable and coercive (i.e. $\lim_{\|x\|_2 \to \infty} l(x) = +\infty$) function. We will later use $l(x) = \sqrt{\|x\|_1^2 + \varepsilon^2}$ for $\varepsilon > 0$ and $l(x) = \frac{1}{2}\|x\|_2^2$. Then, the sampled optimal control problem reads as

$$\inf_{T \in [0, T_{\max}], \, \theta \in \Theta} \left\{ J(T, \theta) := \frac{1}{N} \sum_{i=1}^N l(x^i(1) - y^i) \right\} \quad (5)$$

subject to the state equation for each sample

$$\dot{x}^i(t) = Tf(x^i(t), \theta, z^i), \qquad x^i(0) = x_{\text{init}}^i \quad (6)$$

for $i = 1, \dots, N$ and $t \in (0, 1)$. The next theorem ensures the existence of solutions to this optimal control problem.

Theorem 2.1 (Existence of solutions). *The minimum in* (5) *subject to the side conditions* (6) *is attained.*

3. Discretized optimal control problem

In this section, we present a novel fully discrete formulation of the previously introduced sampled optimal control problem. The state equation (6) is discretized using a semi-implicit scheme resulting in

$$x_{s+1}^i = x_s^i - \frac{T}{S} A^\top (A x_{s+1}^i - z^i) - \frac{T}{S} \nabla_1 \mathcal{R}(x_s^i, \theta) \in \mathbb{R}^{nC} \quad (7)$$

for $s = 0, \dots, S-1$ and $i = 1, \dots, N$, where the *depth* $S \in \mathbb{N}$ is a priori fixed. This equation is equivalent to $x_{s+1}^i = \tilde{f}(x_s^i, T, \theta, z^i)$ with

$$\tilde{f}(x, T, \theta, z) := (\mathrm{Id} + \tfrac{T}{S} A^\top A)^{-1} (x + \tfrac{T}{S}(A^\top z - \nabla_1 \mathcal{R}(x, \theta))). \quad (8)$$

The initial state satisfies $x_0^i = x_{\text{init}}^i \in \mathbb{R}^{nC}$. Then, the discretized sampled optimal control problem reads as

$$\inf_{T \in [0, T_{\max}], \, \theta \in \Theta} \left\{ J_S(T, \theta) := \frac{1}{N} \sum_{i=1}^N l(x_S^i - y^i) \right\} \quad (9)$$

subject to $x_{s+1}^i = \tilde{f}(x_s^i, T, \theta, z^i)$. Following the discrete Pontryagin maximum principle [15, 26], the associated discrete adjoint state p_s^i is given by

$$p_s^i = (\mathrm{Id} - \tfrac{T}{S} \nabla_1^2 \mathcal{R}(x_s^i, \theta))(\mathrm{Id} + \tfrac{T}{S} A^\top A)^{-1} p_{s+1}^i \quad (10)$$

for $s = S-1, \dots, 0$ and $i = 1, \dots, N$, and the terminal condition $p_S^i = -\frac{1}{N} \nabla l(x_S^i - y^i)$. For further details, we refer the reader to the supplementary material.

The next theorem states an exactly computable condition for the optimal stopping time, which is of vital importance for the numerical optimization:

Theorem 3.1 (Optimality condition). *Let $(\overline{T}, \overline{\theta})$ be a stationary point of J_S with associated states $\overline{x}_s^i$ and adjoint states $\overline{p}_s^i$ satisfying* (7) *and* (10) *subject to the initial conditions $\overline{x}_0^i = x_{\text{init}}^i$ and the terminal conditions $\overline{p}_S^i = -\frac{1}{N} \nabla l(\overline{x}_S^i - y^i)$. We further assume that $\nabla \tilde{f}(\overline{x}_s^i, \overline{T}, \overline{\theta}, z^i)$ has full rank for all $i = 1, \dots, N$ and $s = 0, \dots, S$. Then,*

$$-\frac{1}{N} \sum_{s=0}^{S-1} \sum_{i=1}^N \langle \overline{p}_{s+1}^i, (\mathrm{Id} + \tfrac{\overline{T}}{S} A^\top A)^{-1} (\overline{x}_{s+1}^i - \overline{x}_s^i) \rangle = 0. \quad (11)$$

Note that (11) is the derivative of the discrete Lagrange functional minimizing J_S subject to the discrete state equation. The proof and a time continuous version of this theorem are presented in the supplementary material.

An important property of any learning based method is the dependency of the learned parameters with respect to different training datasets [13].

Theorem 3.2. *Let* $(T, \theta), (\widetilde{T}, \widetilde{\theta})$ *be two pairs of control parameters obtained from two different training datasets. We denote by* $x, \widetilde{x} \in (\mathbb{R}^{nC})^{(S+1)}$ *two solutions of the state equation with the same observed data* z *and initial condition* x_{init}, *i.e.*

$$x_{s+1} = \widetilde{f}(x_s, T, \theta, z), \quad \widetilde{x}_{s+1} = \widetilde{f}(\widetilde{x}_s, \widetilde{T}, \widetilde{\theta}, z) \quad (12)$$

for $s = 1, \ldots, S-1$ *and* $x_0 = \widetilde{x}_0 = x_{\text{init}}$. *Let* $B(T) \coloneqq \text{Id} + \frac{T}{S} A^\top A$ *and* $L_{\mathcal{R}}$ *be the Lipschitz constant of* $\mathcal{R}$, *i.e.*

$$\|\nabla_1 \mathcal{R}(x, \theta) - \nabla_1 \mathcal{R}(\widetilde{x}, \widetilde{\theta})\|_2 \le L_{\mathcal{R}} \left\| \begin{pmatrix} x \\ \theta \end{pmatrix} - \begin{pmatrix} \widetilde{x} \\ \widetilde{\theta} \end{pmatrix} \right\|_2 \quad (13)$$

for all $x, \widetilde{x} \in \mathbb{R}^{nC}$ *and all* $\theta, \widetilde{\theta} \in \Theta$. *Then,*

$$
\begin{aligned}
&\|x_{s+1} - \widetilde{x}_{s+1}\|_2 \\
\le &\|B(T)^{-1} - B(\widetilde{T})^{-1}\|_2 \Big(\|x_s\|_2 + \frac{T}{S}\|A^\top z\|_2 \\
&+ \frac{T}{S}\|\nabla_1 \mathcal{R}(x_s, \theta)\|_2 \Big) + \|B(\widetilde{T})^{-1}\|_2 \Big(\|x_s - \widetilde{x}_s\|_2 \\
&+ \frac{|T - \widetilde{T}|}{S}\|A^\top z\|_2 + \frac{|T - \widetilde{T}|}{S}\|\nabla_1 \mathcal{R}(\widetilde{x}_s, \widetilde{\theta})\|_2 \\
&+ \frac{T}{S} L_{\mathcal{R}} \|(x, \theta)^\top - (\widetilde{x}, \widetilde{\theta})^\top\|_2 \Big).
\end{aligned}
\quad (14)
$$

Hence, this theorem provides a computable upper bound for the norm difference of two states with observed value z and initial value x_{init} evaluated at the same step s, which amounts to a sensitivity analysis w.r.t. the training data.

4. Numerical results

In this section, we elaborate on the training and optimization, and we analyze the numerical results for four exemplary applications of TDVs: image denoising, CT and MRI reconstruction, and single image super-reconstruction.

4.1. Training and optimization

For all models considered, we solely use 400 images taken from the BSDS400 dataset [28] for training, where we apply a data augmentation by flipping and rotating the images by multiples of $90°$. We compute the approximate minimizers of the discretized sampled optimal control problem (9) using the stochastic ADAM optimizer [21] on batches of size 32 with a patch size 96×96. The zero-mean constraint of the kernel K is enforced by a projection after each iteration step. For image denoising, we incorporate the squared ℓ^2-loss function $l(x) = \frac{1}{2}\|x\|_2^2$ and the learning rate $4 \cdot 10^{-4}$, whereas for single image super-resolution the regularized ℓ^1-loss $l(x) = \sqrt{\|x\|_1^2 + \varepsilon^2}$ with $\varepsilon = 10^{-3}$ and the learning rate 10^{-3} is used. The first and second order momentum variables of the ADAM optimizer are set to 0.9 and 0.999 and 10^6 training steps are performed. Throughout all experiments we set $\nu = 9$, the number of feature channels $m = 32$, and $S = 10$ if not otherwise stated.

4.2. Image denoising

In the first task, we analyze the performance of the TDV regularizer for additive white Gaussian denoising. To this end, we create the training dataset by uniformly drawing a ground truth patch y^i from the BSDS400 dataset and add Gaussian noise $n^i \sim \mathcal{N}(0, \sigma^2)$. Consequently, we set $x_{\text{init}}^i = z^i = y^i + n^i$ and the linear operator A coincides with the identity matrix Id.

Table 1 lists the average PSNR values of classical image test datasets for varying noise levels $\sigma \in \{15, 25, 50\}$. In the penultimate column, the PSNR values of our proposed TDV regularizer with three macro-blocks solely trained for $\sigma = 25$ (denoted by TDV_{25}^3) are presented. To apply the TDV_{25}^3 model to different noise levels, we first rescale the noisy images $\widehat{x}_{\text{init}}^i = \widehat{z}^i = \frac{25}{\sigma} z^i$, then apply the learned scheme (7), and obtain the results via $x_S^i = \frac{\sigma}{25} \widehat{x}_S^i$. In the last column, the PSNR values of the proposed TDV regularizer with three macro-blocks (denoted by TDV^3) *individually* trained for each specific noise levels are shown. For all considered noise levels and datasets, state-of-the-art FOCNet [19] achieves the best results in terms of PSNR score. The proposed TDV_{25}^3 and TDV^3 models have comparable PSNR values. The adaption of the TDV^3 model to specific noise levels further increases the PSNR value. To conclude, we emphasize that we achieve results on par with FOCNet with less than 1% of the trainable parameters, which highlights the potential of our approach. Compared to FOCNet, the *inherent variational structure* of our model allows for a *deeper mathematical analysis*, that we elaborate on below.

Figure 2 depicts surface plots of the deep variation $[-1, 1] \ni (\xi_1, \xi_2) \mapsto r(\xi_1 x + \xi_2 n)_i$ of TDV_{25}^3 evaluated at four prototypic patches x of size 49×49, where i is the index of the center pixel marked by the red points. Here, n refers to a Gaussian noise with standard deviation $\sigma = 25$. Hence, the surface plots visualize the local regularization energy in the image contrast direction and a random noise direction. The deep variation is smooth with commonly only a single local minimizer and the shape significantly varies depending on the pixel neighborhood.

Figure 3 (top) visualizes the average PSNR values $S \mapsto \frac{1}{N}\sum_{i=1}^N \text{PSNR}(x_S^i, y^i)$ on the BSDS68 test dataset for $\sigma = 25$. The second plot $S \mapsto \overline{T}$ (bottom) shows the learned optimal stopping time as a function of the depth. In all experiments, TDV regularizers with more macro-blocks perform significantly better. Moreover, the average PSNR value saturates beyond the depth $S = 10$. The optimal stopping time converges for large S in all models. Consequently, larger depth values S lead to a finer time discretization of the trajectories and depth values S exceeding 10 yield no further improvement. To address the significance of the optimal stopping time $\overline{T}$, we evaluate the PSNR values (top) and the first order condition (11) (bottom) as a function of the stopping time in Figure 4 for TDV_{25}^3 trained for $S = 10$.

Data set	σ	BM3D [9]	TNRD [8]	DnCNN [43]	FFDNet [44]	N³Net [32]	FOCNet [19]	TDV^3_{25}	TDV^3
Set12	15	32.37	32.50	32.86	32.75	-	33.07	32.93	33.01
	25	29.97	30.05	30.44	30.43	30.55	30.73	30.66	30.66
	50	26.72	26.82	27.18	27.32	27.43	27.68	27.50	27.59
BSDS68	15	31.08	31.42	31.73	31.63	-	31.83	31.76	31.82
	25	28.57	28.92	29.23	29.19	29.30	29.38	29.37	29.37
	50	25.60	25.97	26.23	26.29	26.39	26.50	26.40	26.45
Urban100	15	32.34	31.98	32.67	32.43	-	33.15	32.66	32.87
	25	29.70	29.29	29.97	29.92	30.19	30.64	30.38	30.38
	50	25.94	25.71	26.28	26.52	26.82	27.40	26.94	27.04
# Parameters			26,645	555,200	484,800	705,895	53,513,120	427,330	427,330

Table 1. Comparison of average PSNR values for additive white Gaussian noise for $\sigma \in \{15, 25, 50\}$ on classical image datasets. In the last row, the number of trainable parameters is listed.
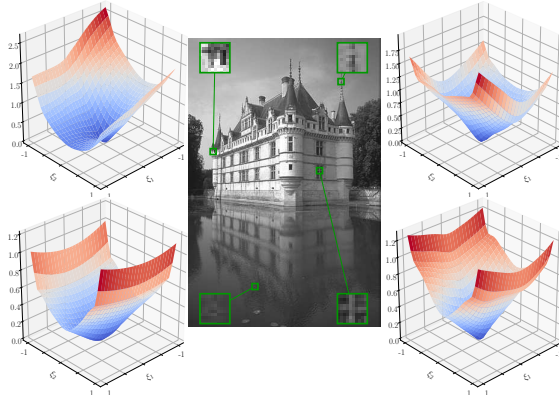

Figure 2. Surface plots of the deep variation $[-1, 1] \ni (\xi_1, \xi_2) \mapsto r(\xi_1 x + \xi_2 n)_i$ of four patches–each evaluated at the red center pixel using TDV^3 with BSDS68 and $\sigma = 25$.
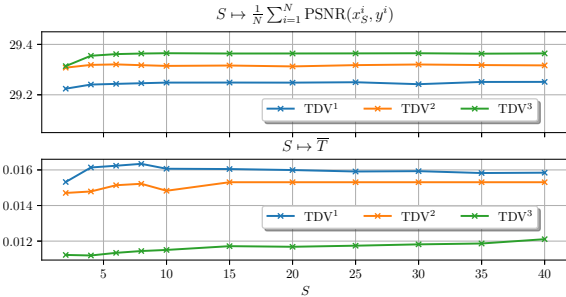

Figure 3. Plots of the functions $S \mapsto \frac{1}{N} \sum_{i=1}^{N} \text{PSNR}(x_S^i, y^i)$ (top) and $S \mapsto \overline{T}$ (bottom) for TDV^1 (blue), TDV^2 (orange) and TDV^3 (green).

Each black dashed curve represents a single image of the BSDS68 test dataset and the red curve is the average among all test samples. Initially, all PSNR curves monotonically increase up to a unique maximum value located near the optimal stopping time $\overline{T} = 0.0297$ and decrease for larger values of T, which is exactly determined by the first order condition (11). We stress that all curves peak around a very small neighborhood, which results from an overlap of all curves near the zero crossing in the second plot.
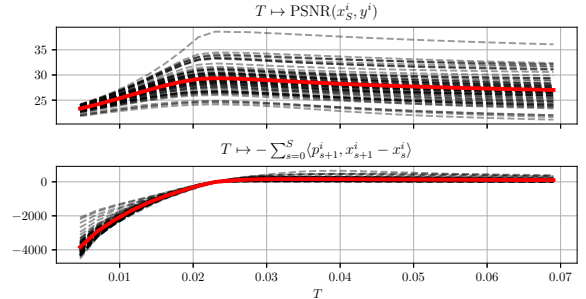

Figure 4. Plots of the functions $T \mapsto \text{PSNR}(x_S^i, y^i)$ (top) and the first order condition $T \mapsto -\sum_{s=0}^{S} \langle p_{s+1}^i, x_{s+1}^i - x_s^i \rangle$ (bottom) for $i = 1, \ldots, 68$ using TDV^3 for BSDS68 and $\sigma = 25$. The averages across the samples are depicted by the red curves.

A qualitative and quantitative analysis of the impact of the stopping time for TDV^3_{25} trained with $S = 10$ on a standard test image is depicted in Figure 5. Starting from the noisy input image, the restored image sequence for increasing $S \in \{5, 10, 15, 20\}$ gradually removes noise. Beyond the optimal stopping time, the model smoothes out fine details. We emphasize that even high-frequency patterns like the vertical stripes on the chimney are consistently preserved. Thus, the proposed model generates a stable and interpretable transition from noisy images to cartoon-like images and the optimal stopping time determines the best intermediate result in terms of PSNR.

To analyze the local behavior of TDV^3_{25}, we compute a saddle point $(\overline{x}, \overline{\lambda})$ of the Lagrangian

$$\mathcal{L}(x, \lambda) := \mathcal{R}(x, \theta) - \frac{\lambda}{2}(\|x\|_2^2 - \|x_{\text{init}}\|_2^2), \qquad (15)$$

which is a optimization problem on the hypersphere $\|x\|_2 = \|x_{\text{init}}\|_2$ for a given input image $x_{\text{init}} \in \mathbb{R}^{nC}$ with Lagrange parameter λ. The optimality condition (15) with respect to x is equivalent to $\nabla_1 \mathcal{R}(\overline{x}, \theta) = \overline{\lambda} \overline{x}$, which shows that $(\overline{x}, \overline{\lambda})$ is actually a nonlinear eigenpair of $\nabla_1 \mathcal{R}$. Figure 6 depicts ten triplets of input images, eigenfunctions and eigenvalues $(x_{\text{init}}, \overline{x}, \overline{\lambda})$, where the output images are computed using accelerated projected gradient descent. As a result, the
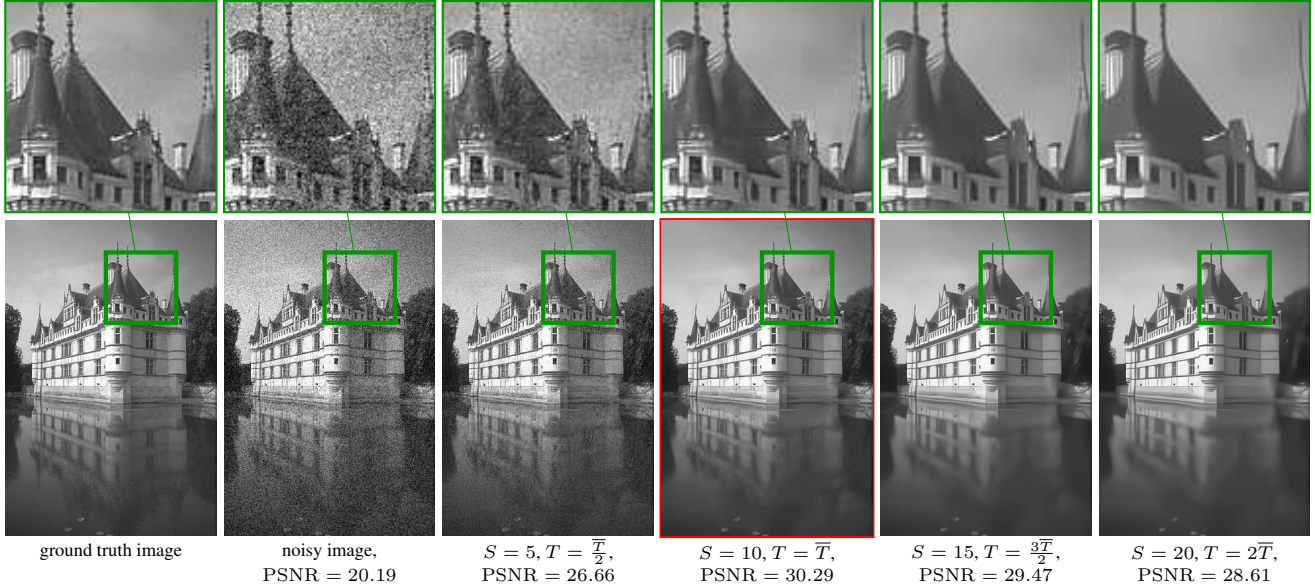
Figure 5. From left to right: Ground truth, noisy input with noise level $\sigma = 25$ and resulting output of TDV^3 for $(S, T) \in \{(5, \frac{\overline{T}}{2}), (10, \overline{T}), (15, \frac{3\overline{T}}{2}), (20, 2\overline{T})\}$, where the optimal stopping time is $\overline{T} = 0.0297$. Note that the best image is framed in red.

| ground truth image | noisy image, PSNR = 20.19 | $S = 5, T = \frac{\overline{T}}{2}$, PSNR = 26.66 | $S = 10, T = \overline{T}$, PSNR = 30.29 | $S = 15, T = \frac{3\overline{T}}{2}$, PSNR = 29.47 | $S = 20, T = 2\overline{T}$, PSNR = 28.61 |

algorithm generates cartoon-like eigenfunctions and energy minimizing patterns are hallucinated due to the constraint. For instance, novel stripe patterns are prolongated in the body and facial region of the woman in the third column or in the body parts of the zebra in the fifth column. Furthermore, contours are axis aligned like the roof in the eighth column or the ship and the logo in the last image.

In Figure 7, we compare the root mean squared error (RMSE) between the image sequences x_s and $\widetilde{x}_s$ generated using TDV^1 with parameter sets (T, θ) and $(\widetilde{T}, \widetilde{\theta})$ with the upper bound estimate in Theorem 3.2. Here, (T, θ) are computed by training on the entire BSDS400 training set, whereas we solely use 10 randomly selected images of this training set to obtain $(\widetilde{T}, \widetilde{\theta})$. The local Lipschitz constant $L_{\mathcal{R}}$ in (13) is estimated directly from the sequences. As a result, the RMSE of the norm differences is roughly constant along S and the upper bound estimate is efficient. Further details of the sensitivity analysis are presented in the supplementary material.

4.3. Computed tomography reconstruction

To demonstrate the broad applicability of the proposed TDV, we perform two-dimensional computed tomography (CT) reconstruction using the TDV$^3_{25}$ regularizer trained for image denoising and $S = 10$. We stress that the regularizer is applied *without* any additional training of the parameters.

The task of computed tomography is the reconstruction of an image given a set of projection measurements called sinogram, in which the detectors of the CT scanner measure the intensity of attenuated X-ray beams. Here, we use the linear attenuation model introduced in [14], where the

attenuation is proportional to the intersection area of a triangle, which is spanned by the X-ray source and a detector element, and the area of an image element. In detail, the sinogram z of an image x is computed by $z = A_R x$, where A_R is the lookup-table based area integral operator of [14] for R angles and 768 projections. Typically, a fully sampled acquisition consists of 2304 angles. For this task, we consider the problem of angular undersampled CT [7], where only a fraction of the angles are measured. We use a 4-fold ($R = 576$) and 8-fold ($R = 288$) angular undersampling to reconstruct a representative image of the MAYO dataset [29]. To account for an imbalance of regularization and data fidelity, we manually scale the data fidelity term by $\lambda > 0$, i.e. $\mathcal{D}(x, z) := \frac{\lambda}{2} \|A_R x - z\|_2^2$. The resulting smooth variational problem is optimized using accelerated gradient descent with Lipschitz backtracking. Further details of the CT reconstruction task and the optimization scheme are included in the supplementary material.

We present qualitative and quantitative results for CT reconstruction in Figure 8 for a single abdominal CT image. As an initialization, we perform 50 steps of a conjugate gradient method on the data fidelity term (first and last column). Using the proposed regularizer TDV$^3_{25}$, we are able to suppress the undersampling artifacts while preserving the fine vessels in the liver. This highlights that the learned regularizer can be effectively applied as a generic regularizer for linear inverse problems without any transfer learning.

4.4. Magnetic resonance imaging reconstruction

Next, the flexibility of our regularizer TDV$^3_{25}$ learned for denoising and $S = 10$ is shown for accelerated magnetic
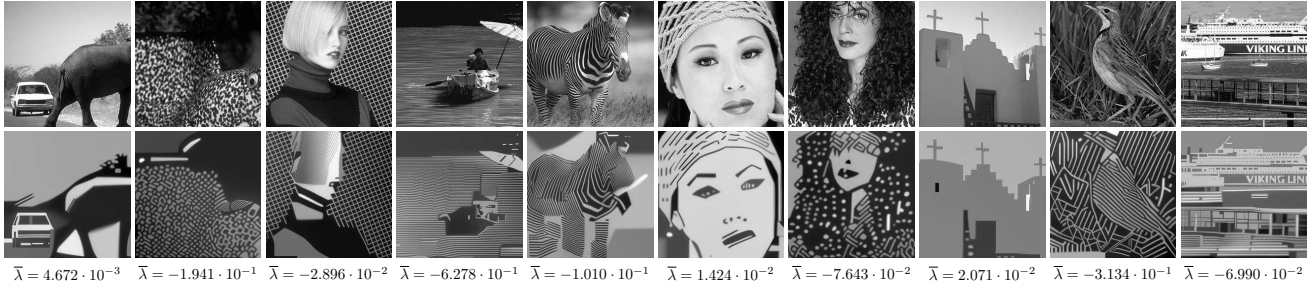
$\overline{\lambda} = 4.672 \cdot 10^{-3}$ $\overline{\lambda} = -1.941 \cdot 10^{-1}$ $\overline{\lambda} = -2.896 \cdot 10^{-2}$ $\overline{\lambda} = -6.278 \cdot 10^{-1}$ $\overline{\lambda} = -1.010 \cdot 10^{-1}$ $\overline{\lambda} = 1.424 \cdot 10^{-2}$ $\overline{\lambda} = -7.643 \cdot 10^{-2}$ $\overline{\lambda} = 2.071 \cdot 10^{-2}$ $\overline{\lambda} = -3.134 \cdot 10^{-1}$ $\overline{\lambda} = -6.990 \cdot 10^{-2}$

Figure 6. Triplets of input images x_{init} (top), eigenfunctions $\overline{x}$ (bottom) and eigenvalues $\overline{\lambda}$ associated with the TDV^3_{25} regularizer.
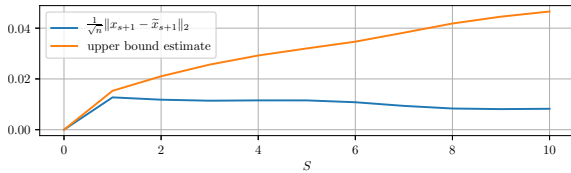

Figure 7. Comparison of RMSE along two sequences computed with two different parameter sets with the estimate (14).

resonance imaging (MRI) *without* any further adaption of θ.

In accelerated MRI, k–space data is acquired using N_C parallel coils, each measuring a fraction of the full k–space to reduce acquisition time [16]. Here, we use the data fidelity term $\mathcal{D}(x, \{z_i\}_{i=1}^{N_C}) = \frac{\lambda}{2} \sum_{i=1}^{N_C} \|M_R F C_i x - z_i\|_2^2$, where $\lambda > 0$ is a manually adjusted weighting parameter, M_R is a binary mask for R-fold undersampling, F is the discrete Fourier transform, and C_i are sensitivity maps computed following [41]. We use 4-fold and 6-fold Cartesian undersampled MRI data to reconstruct a sample knee image. Again, we minimize the resulting variational energy by accelerated gradient descent with Lipschitz backtracking. Further details of this task and the optimization are in the supplementary material.

Figure 9 depicts qualitative results and PSNR values for the reconstruction of 4-fold and 6-fold undersampled k–space data. The first and last columns show the initial images obtained by applying the adjoint operator to the undersampled data. Due to the undersampling in k–space both images are severely corrupted by backfolding artifacts, which are removed by applying the proposed regularizer without losing fine details.

4.5. Single image super-resolution

For single image super-resolution with scale factor $\gamma \in \{2, 3, 4\}$, we start with a full resolution ground truth image patch $y^i \in \mathbb{R}^{nC}$ uniformly drawn from the BSDS400 dataset. To obtain the linear downsampling operator A, we implement the adjoint operator matching MATLAB®'s bicubic upsampling operator `imresize`, which is an implementation of a scale factor-dependent interpolation convolution kernel in conjunction with a stride. The observed

low resolution image is $z^i = Ay^i \in \mathbb{R}^{nC/\gamma^2}$. As an initialization, 3 iteration steps of the conjugate gradient method for the variational problem $x_{\text{init}}^i = \text{argmin}_{x \in \mathbb{R}^{nC}} \|Ax - z^i\|_2^2$ are applied. The matrix inverse in (8) is approximated in each step by 7 steps of a conjugate gradient scheme. Here, all results are obtained by training a TDV^3 regularizer for each scale factor individually.

In Table 2, we compare our approach with several state-of-the-art networks of similar complexity and list average PSNR values of the Y-channel in the YCbCr color space over test datasets. For the BSDS100 dataset, our proposed method achieves similar results as OISR-LF-s [17] with only one third of the trainable parameters. Figure 10 depicts a restored sequence of images for the single image super-resolution task with scale factor 4 using TDV^3 for a representative sample image of the Set14 dataset. Starting from the low resolution initial image, interfaces are gradually sharpened and the best quality is achieved for $T = \overline{T}$. Beyond this point, interfaces are artificially intensified.

5. Conclusion

In this paper, we have introduced total deep variation regularizers, which are motivated by established deep network architectures. The inherent variational structure of our approach enables a rigorous mathematical understanding encompassing an optimality condition for optimal stopping, a nonlinear eigenfunction analysis, and a sensitivity analysis that yields computable upper error bounds. For image denoising and single image super-resolution, our model generates state-of-the-art results with an impressively low number of trainable parameters. Moreover, to underline the versatility of TDVs for generic linear inverse problems, we successfully demonstrated their applicability for the challenging CT and MRI reconstruction tasks without requiring any additional training.

CG initialization $R = 576$ PSNR $= 34.17$ TDV$_{25}^3$ for $R = 576$ PSNR $= 43.51$ TDV$_{25}^3$ for $R = 2304$ TDV$_{25}^3$ for $R = 288$ PSNR $= 35.03$ CG initialization $R = 288$ PSNR $= 27.75$
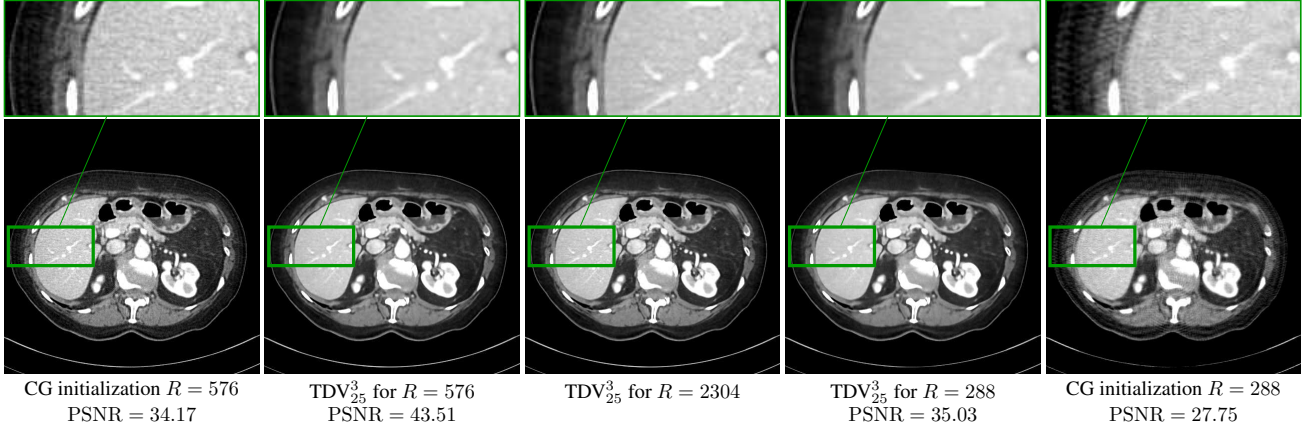
Figure 8. Conjugate gradient reconstruction for 4/8-fold angular undersampled CT task (first/fifth image), results obtained by using the TDV$_{25}^3$ regularizer with $\lambda = 500 \cdot 10^3$ for 4-fold (second image) and $\lambda = 10^6$ for 8-fold undersampling (fourth image), and fully sampled reference reconstruction using the TDV$_{25}^3$ regularizer with $\lambda = 125 \cdot 10^3$ (third image).

zero filling $R = 4$ PSNR $= 24.98$ TDV$_{25}^3$ for $R = 4$ PSNR $= 48.29$ reference image   TDV$_{25}^3$ for $R = 6$ PSNR $= 38.55$ zero filling $R = 6$ PSNR $= 24.07$
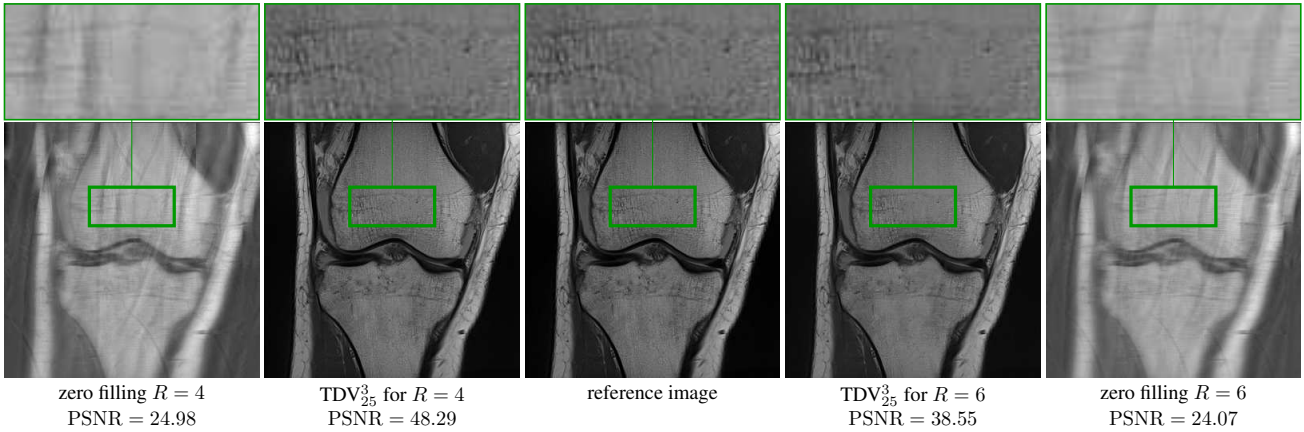
Figure 9. Zero filling initialization for acceleration factors $R \in \{4, 6\}$ (first/fifth image), output using the TDV$_{25}^3$ regularizer with $\lambda = 1000$ for $R = 4$ (second image) and $\lambda = 1500$ for $R = 6$ (fourth image), and fully sampled reference (third image).

Data set	Scale	MemNet [40]	VDSR [20]	DRRN [39]	OISR-LF-s [17]	TDV^3
Set14	×2	33.28	33.03	33.23	33.62	33.35
	×3	30.00	29.77	29.96	30.35	29.96
	×4	28.26	28.01	28.21	28.63	28.41
BSDS100	×2	32.08	31.90	32.05	32.20	32.18
	×3	28.96	28.82	28.95	29.11	28.98
	×4	27.40	27.29	27.38	27.60	27.50
# Parameters		585,435	665,984	297,000	1,370,000	428,970

Table 2. Numerical results of various state-of-the art networks for single image super resolution with a comparable number of parameters.
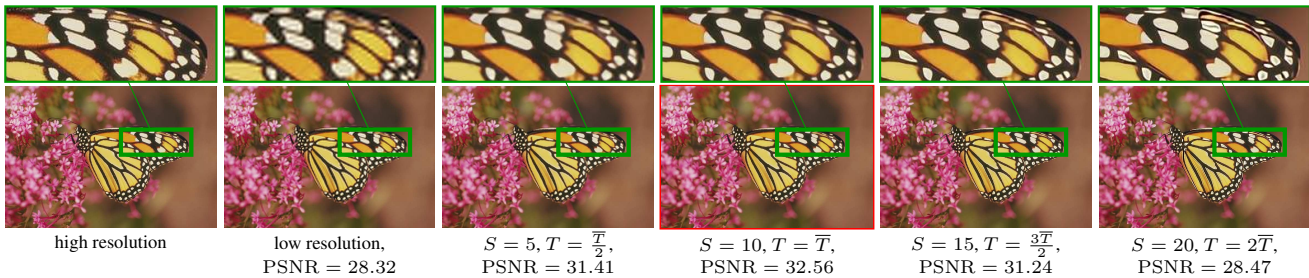

high resolution   low resolution, PSNR $= 28.32$ $S = 5, T = \frac{\overline{T}}{2}$, PSNR $= 31.41$ $S = 10, T = \overline{T}$, PSNR $= 32.56$ $S = 15, T = \frac{3\overline{T}}{2}$, PSNR $= 31.24$ $S = 20, T = 2\overline{T}$, PSNR $= 28.47$

Figure 10. From left to right: High resolution, low resolution with scale factor 4, and resulting output of TDV^3 for $(S, T) \in \{(5, \frac{\overline{T}}{2}), (10, \overline{T}), (15, \frac{3\overline{T}}{2}), (20, 2\overline{T})\}$, where the optimal stopping time is $\overline{T} = 0.098$. Note that the best image is framed in red.

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008. 3

[2] Kristian Bredies, Karl Kunisch, and Thomas Pock. Total generalized variation. *SIAM J. Imaging Sci.*, 3(3):492–526, 2010. 1

[3] Antonin. Chambolle and Pierre-Louis Lions. Image recovery via total variation minimization and related problems. *Numer. Math.*, 76(2):167–188, 1997. 1

[4] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numer.*, 25:161–319, 2016. 1

[5] Antonin Chambolle and Thomas Pock. Total roto-translational variation. *Numer. Math.*, 142(3):611–666, 2019. 2

[6] Tony F. Chan, Sung Ha Kang, and Jianhong Shen. Euler's elastica and curvature-based inpainting. *SIAM J. Appl. Math.*, 63(2):564–592, 2002. 2

[7] Guang-Hong Chen, Jie Tang, and Shuai Leng. Prior image constrained compressed sensing (PICCS): A method to accurately reconstruct dynamic CT images from highly undersampled projection data sets. *Medical Physics*, 35(2):660–663, 2008. 6

[8] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1256–1272, 2017. 2, 5

[9] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007. 2, 5

[10] Justin Domke. Generic methods for optimization-based modeling. In *International Conference on Artificial Intelligence and Statistics*, pages 318–326, 2012. 2

[11] Weinan E, Jiequn Han, and Qianxiao Li. A mean-field optimal control formulation of deep learning. *Res. Math. Sci.*, 6(1):Paper No. 10, 41, 2019. 3

[12] Alexander Effland, Erich Kobler, Karl Kunisch, and Thomas Pock. An optimal control approach to early stopping variational methods for image restoration. *arXiv*, 1907.08488, 2019. 2

[13] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992. 3

[14] Sungsoo Ha and Klaus Mueller. A look-up table-based ray integration framework for 2-D/3-D forward and back projection in X-ray CT. *IEEE Transactions on Medical Imaging*, 37:361–371, 2018. 6

[15] Hubert Halkin. A maximum principle of the Pontryagin type for systems described by nonlinear difference equations. *SIAM J. Control*, 4:90–111, 1966. 3

[16] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P. Recht, Daniel K. Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated MRI data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018. 7

[17] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng. ODE-inspired network design for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1732–1741, 2019. 7, 8

[18] Jinggang Huang and David Mumford. Statistics of natural images and models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 541–547, 1999. 3

[19] Xixi Jia, Sanyang Liu, Xiangchu Feng, and Lei Zhang. FOC-Net: A fractional optimal control network for image denoising. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2019. 4, 5

[20] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016. 8

[21] Diederik P. Kingma and Jimmy Lei Ba. ADAM: a method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 4

[22] Erich Kobler, Teresa Klatzer, Kerstin Hammernik, and Thomas Pock. Variational networks: Connecting variational methods and deep learning. In *German Conference on Pattern Recognition*, pages 281–293. Springer International Publishing, 2017. 2

[23] Stamatios Lefkimmiatis. Non-local color image denoising with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5882–5891, 2016. 2

[24] Housen Li, Johannes Schwab, Stephan Antholzer, and Markus Haltmeier. NETT: solving inverse problems with deep neural networks. *Inverse Problems*, 2020. 2

[25] Qianxiao Li, Long Chen, Cheng Tai, and Weinan E. Maximum principle based algorithms for deep learning. *J. Mach. Learn. Res.*, 18:Paper No. 165, 29, 2017. 3

[26] Qianxiao Li and Shuji Hao. An optimal control approach to deep learning and applications to discrete-weight neural networks. In *International Conference on Machine Learning*, 2018. 3

[27] Sebastian Lunz, Ozan Öktem, and Carola-Bibiane Schönlieb. Adversarial regularizers in inverse problems. In *Advances in Neural Information Processing Systems*, pages 8507–8516, 2018. 2

[28] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE International Conference on Computer Vision*, pages 416–423, 2001. 4

[29] Cynthia H. McCollough, Adam C. Bartley, Rickey E. Carter, Baiyu Chen, Tammy A. Drees, Phillip Edwards, David R. Holmes III, Alice E. Huang, Farhana Khan, Shuai Leng, Kyle L. McMillan, Gregory J. Michalak, Kristina M. Nunez, Lifeng Yu, and Joel G. Fletcher. Low-dose CT for the detection and classification of metastatic liver lesions: Results of the 2016 low dose CT grand challenge. *Medical Physics*, 44(10):339–352, 2017. 6

[30] Tim Meinhardt, Michael Möller, Caner Hazirbas, and Daniel Cremers. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In *IEEE International Conference on Computer Vision*, pages 1781–1790, 2017. 2

[31] Mark Nitzberg, David Mumford, and Takahiro Shiota. *Filtering, segmentation and depth*, volume 662 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin, 1993. 2

[32] Tobias Plötz and Stefan Roth. Neural nearest neighbors networks. In *Advances in Neural Information Processing Systems 31*, pages 1087–1098. Curran Associates, Inc., 2018. 5

[33] J. H. Rick Chang, Chun-Liang Li, Barnábas Póczos, B. V. K. Vijaya Kumar, and Aswin C. Sankaranarayanan. One network to solve them all — solving linear inverse problems using deep projection models. In *IEEE International Conference on Computer Vision*, pages 5888–5897, 2017. 2

[34] Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: regularization by denoising (RED). *SIAM J. Imaging Sci.*, 10(4):1804–1844, 2017. 2

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 2

[36] Stefan Roth and Michael J. Black. Fields of Experts. *Int. J. Comput. Vis.*, 82(2):205–229, 2009. 2

[37] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992. 1

[38] Kegan G. G. Samuel and Marshall F. Tappen. Learning optimized MAP estimates in continuously-valued MRF models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 477–484, 2009. 2

[39] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2790–2798, 2017. 8

[40] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. MemNet: A persistent memory network for image restoration. In *IEEE International Conference on Computer Vision*, pages 4549–4557, 2017. 8

[41] Martin Uecker, Peng Lai, Mark J. Murphy, Patrick Virtue, Michael Elad, John M. Pauly, Shreyas S. Vasanawala, and Michael Lustig. ESPIRiTan eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA. *Magnetic resonance in medicine*, 71(3):990–1001, 2014. 7

[42] Singanallur V. Venkatakrishnan, Charles A. Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *IEEE Global Conference on Signal and Information Processing*, pages 945–948, 2013. 2

[43] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 5

[44] Kai Zhang, Wangmeng Zuo, and Lei Zhang. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 5

[45] Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning*, 2019. 3

[46] Song Chun Zhu, Yingnian Wu, and David Mumford. Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. *Int. J. Comput. Vision*, 27(2):107–126, 1998. 2