2023 IEEE 9th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP) | 979-8-3503-4452-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CAMSAP58249.2023.10403502

OVERCOMING DISTRIBUTION SHIFTS IN PLUG-AND-PLAY METHODS WITH TEST-TIME TRAINING

Edward P. Chandler[†], *Shirin Shoushtari*[‡], *Jiaming Liu*[‡], *M. Salman Asif*[§], *and Ulugbek S. Kamilov*[†,‡]

[†]Department of Computer Science & Engineering, Washington University in St. Louis, MO
[‡]Department of Electrical & Systems Engineering, Washington University in St. Louis, MO
[§]Department of Electrical & Computer Engineering, University of California, Riverside, CA

ABSTRACT

Plug-and-Play Priors (PnP) is a well-known class of methods for solving inverse problems in computational imaging. PnP methods combine physical forward models with learned prior models specified as image denoisers. A common issue with the learned models is that of a performance drop when there is a distribution shift between the training and testing data. Test-time training (TTT) was recently proposed as a general strategy for improving the performance of learned models when training and testing data come from different distributions. In this paper, we propose PnP-TTT as a new method for overcoming distribution shifts in PnP. PnP-TTT uses deep equilibrium learning (DEQ) for optimizing a self-supervised loss at the fixed points of PnP iterations. PnP-TTT can be directly applied on a single test sample to improve the generalization of PnP. We show through simulations that given a sufficient number of measurements, PnP-TTT enables the use of image priors trained on natural images for image reconstruction in magnetic resonance imaging (MRI).

Index Terms— computational imaging, inverse problems, plug-and-play priors, deep learning, test-time training.

1. INTRODUCTION

Many computational imaging problems can be formulated as *inverse problems*, where the goal is to recover an unknown image from a set of noisy measurements. It is common to solve inverse problems by integrating the measurement model characterizing the response of the imaging instrument with a regularizer infusing prior knowledge on the unknown image. There has been considerable recent interest in using deep learning (DL) for designing data-driven image priors [1, 2, 3]. DL methods eliminate the need for explicit prior modeling by learning a mapping from measurements to target images using convolutional neural networks (CNN). Model-based DL (MBDL) is an extension to traditional DL that integrates the image prior defined through a CNN with the knowledge of the measurement models. For example, plug-and-play priors (PnP) is a well-known MBDL approach that uses pre-trained image denoiser as priors [4, 5, 3]. Other MBDL widely-used MBDL approaches include deep unfolding (DU) and deep equilibrium (DEQ) learning, both of which rely on the integration of the measurement model during the training of the image prior [6, 7, 8, 9, 10]. While both DU and DEQ interpret iterations of image reconstruction as neural network layers, the memory complexity of DEQ is independent of the number of unfolded iterations.

Much of the existing research on MBDL has focused on the scenarios where the statistical distribution of the training data matches that of the testing data. While this strategy has led to significant theoretical and algorithmic innovations, it does not address the issue of the performance gap due to data distribution shifts. For example, image priors trained with a specific distribution in PnP, performs poorly on samples from different distributions [11]. Thus, distribution shifts limit the applicability of priors pre-trained for one class to another one.

Domain adaptation refers to a class of DL techniques for improving the performance of a learned model on a target task containing insufficient annotated data by using the knowledge learned by the model from another related task with adequate labeled data [12, 13]. *Test-time training (TTT)* was recently proposed as a domain adaptation strategy based on self-supervised optimization of the learned model utilizing only test-time measurements [14]. The TTT strategy was also recently used in the context of imaging inverse problems to address domain shifts in end-to-end image reconstruction with DL for accelerated magnetic resonance imaging (MRI) [15].

In this paper, we investigate TTT in the context of PnP methods. We propose *PnP-TTT* as a method for overcoming the performance gap in PnP due to data distribution shifts. PnP-TTT uses DEQ to update the weights of the CNN prior in PnP at test-time. The DEQ learning in PnP-

This material is partially based upon work supported by the NSF CAREER award under grant CCF-2043134 and by the Gordon and Betty Moore Foundation grant 11396.

TTT is used to minimize a self-supervised loss at the fixed points of PnP iterations for one test sample. We also present numerical results showing that DEQ training in PnP-TTT can significantly boost the performance of the shifted priors. We evaluate the proposed method on image reconstruction for compressed sensing MRI (CS-MRI), where we recover MRI images from subsampled Fourier measurements. Our results show that given enough measurements, PnP-TTT can close the gap due to distribution shift between test and training data. It is worth mentioning that our method can also be applied to other tasks and different variants of PnP, highlighting its broader applicability for inverse problems in computational imaging.

2. BACKGROUND

2.1. Inverse Problems

We consider the problem of recovering an image $\boldsymbol{x} \in \mathbb{C}^n$ from its noisy measurement $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ is the *measurement operator* and $\boldsymbol{e} \in \mathbb{C}^m$ is additive white Gaussian noise (AWGN). We can formulate the problem as a regularized optimization problem

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}\in\mathbb{R}^n}{\arg\min}\, f(\boldsymbol{x}) \quad \text{with} \quad f(\boldsymbol{x}) = g(\boldsymbol{x}) + h(\boldsymbol{x}), \quad (1)$$

where g is the *data-fidelity* term used to ensure the consistency of the solution with the measurement and h is the *regularization* term that infuses prior knowledge. For example, the least-squares loss is a widely-used data-fidelity term $g(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$ and total variation (TV) is commonly used as the regularizer [16].

2.2. Plug-and-Play Priors

PnP framework includes a family of methods that incorporate the measurement model with CNN denoisers to solve inverse problems [3]. PnP methods can be interpreted as a fixed-point iteration of some high-dimensional operator where the CNN takes the role of the prior. For example, the *proximal gradient method (PGM)* variant of PnP can be expressed as

$$\boldsymbol{x}^k = \mathsf{T}_{\boldsymbol{\theta}}(\boldsymbol{x}^{k-1}) \quad \text{with} \quad \mathsf{T}_{\boldsymbol{\theta}} := \mathsf{D}_{\boldsymbol{\theta}}(\mathsf{I} - \gamma\nabla g), \quad (2)$$

where $\mathsf{D}_{\boldsymbol{\theta}}$ is the denoiser, g is the data-fidelity term, ∇g is the gradient of g, I is the identity mapping, and $\gamma > 0$ is the step-size. The PnP method in (2) is commonly refered to as PnP-PGM.

2.3. Deep Equilibrium Models

DEQ is a recent approach for training MBDL architectures in a memory-efficient way [9]. DEQ uses implicit differentiation for training possibly infinite-depth networks by backpropagating through the fixed points of an operator. For the operator defined in eq. (2), the output is implicitly expressed as

$$\overline{\boldsymbol{x}} = \mathsf{T}_{\boldsymbol{\theta}}(\overline{\boldsymbol{x}}), \quad (3)$$

where $\mathsf{T}_{\boldsymbol{\theta}}$ is the operator parameterized by $\boldsymbol{\theta}$, and $\overline{\boldsymbol{x}}$ is the fixed point acquired using fixed point iterations in the forward pass of DEQ. The connection of DEQ and PnP has inspired end-to-end training of CNN denoisers as model dependant priors in many imaging problems such as MRI [9] and computed tomography (CT) [10]. The prior $\mathsf{D}_{\boldsymbol{\theta}}$ in DEQ is trained by minimizing the loss between the fixed points from eq. (3) and the ground truth $\boldsymbol{x}^*$

$$\ell(\boldsymbol{\theta}) = \frac{1}{2}\|\mathsf{T}_{\theta}(\overline{\boldsymbol{x}}) - \boldsymbol{x}^*\|_2^2. \quad (4)$$

Implicit differentiation of the fixed points yields the gradient of the loss with respect to $\boldsymbol{\theta}$ in the backward pass as

$$\nabla\ell(\boldsymbol{\theta}) = \left(\nabla_{\boldsymbol{\theta}}\mathsf{T}_{\boldsymbol{\theta}}(\overline{\boldsymbol{x}})\right)^{\mathsf{T}}\left(\mathsf{I} - \nabla_{\boldsymbol{x}}\mathsf{T}_{\boldsymbol{\theta}}(\overline{\boldsymbol{x}})\right)^{-\mathsf{T}}\left(\overline{\boldsymbol{x}} - \boldsymbol{x}^*\right), \quad (5)$$

where I is the identity mapping and ℓ is the loss.

Algorithm 1 Test-Time Training for Plug-and-Play

```
input: forward model $\mathbf{A}$, PnP initialization $\boldsymbol{x}_0$, measurement $\boldsymbol{y}$, denoiser $\mathsf{D}_\theta$, and numIter$\geq 0$
    $\boldsymbol{x}_0^* = \texttt{PnP}(\boldsymbol{x}_0, \boldsymbol{y}; \mathsf{D}_\theta)$
    for $i = 1$ to numIter do
        $l = \texttt{LOSS}(\mathbf{A}\boldsymbol{x}_{i-1}^*, \boldsymbol{y})$
        DEQ_GRAD$(l, \theta)$        {Update parameters $\theta$ using DEQ}
        $\boldsymbol{x}_i^* = \texttt{PnP}(\boldsymbol{x}_0, \boldsymbol{y}; \mathsf{D}_\theta)$
    end for
    return $\boldsymbol{x}_i^*$
```

2.4. Test-Time Training

Current PnP methods are built on the premise that the prior represents the same distribution as that of the desired solution. However, it is common to observe distribution shifts between training and testing data. In some scenarios, there are insufficient training samples to train a DL network as the prior, hence, alternative priors trained on a shifted distribution are used with suboptimal reconstruction performance. TTT has been proposed to reduce the performance gap due to distribution shift in various tasks [14]. The key idea of TTT is to update the shifted model's weight at test-time by minimizing a self-supervised loss

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\arg\min}\, \ell_{\mathsf{sup}}\left(\mathsf{D}_{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{y}\right), \quad (6)$$

where $\mathsf{D}_{\boldsymbol{\theta}}$ is the neural network and $\boldsymbol{y}$ is a test sample. Depending on the selection of ℓ_{sup}, TTT has shown improved performance in many imaging tasks. For example, it can be used to improve the MRI reconstruction using DL models trained in an end-to-end matter on shifted distributions [15]. In this scenario, the self-supervised loss proposed is

$$\ell_{\mathsf{sup}}(\boldsymbol{\theta}) = \frac{\left\|\boldsymbol{A}\mathsf{D}_{\boldsymbol{\theta}}\left(\boldsymbol{A}^{\dagger}\boldsymbol{y}\right) - \boldsymbol{y}\right\|_1}{\|\boldsymbol{y}\|_1}, \quad (7)$$

where $\boldsymbol{A}$ is the measurement model, $\boldsymbol{A}^{\dagger}$ is the Hermitian transpose, and $\boldsymbol{y}$ is the test-time measurement. Note that as opposed to (4), TTT in (7) does not need ground truth reconstruction to compute ℓ_{sup} and one can use other loss functions rather than the normalized ℓ_1 [15].

2.5. Our contribution

We propose PnP-TTT as a novel approach for enhancing the performance of image reconstruction for PnP methods

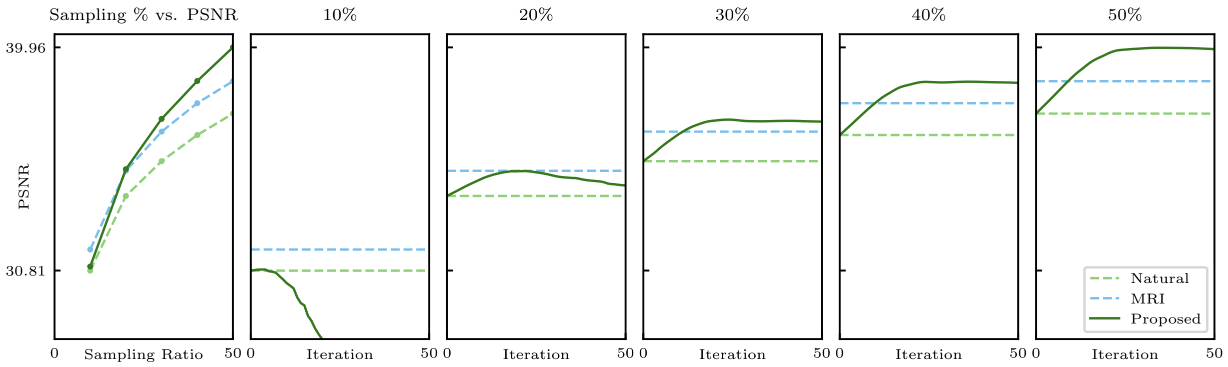

Fig. 1: Evaluation of PnP-TTT for different sampling ratios in accelerated MRI. The leftmost chart displays the best PSNR performance achieved by PnP-TTT vs. sampling ratios. The remaining charts show PSNR at each TTT iteration. Note that the best performance is above the lower baseline for all the sampling ratios; however, TTT eventually overfits to the test-time measurement, reducing performance. Additionally, note that at larger sampling ratios, the performance of PnP-TTT prior can surpass that of the matched prior due to the DEQ training.

Table 1: PSNR (dB) values for accelerated MRI with matched, mismatched, and PnP-TTT priors.

Radial CS Ratio	10%		20%		30%		40%		50%	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Natural prior	30.81	0.9351	33.87	0.9648	35.29	0.9721	36.37	0.9759	37.25	0.9782
MRI Prior	31.67	0.9468	34.9	0.9707	36.51	0.9773	37.67	0.9807	38.57	0.9828
PnP-TTT (Ours)	30.97	0.938	34.97	0.9718	37.03	0.9796	38.58	0.984	39.96	0.9873
PnP-TTT − Natural	0.16	0.0029	1.1	0.007	1.74	0.0075	2.21	0.0081	2.71	0.0091

under distribution shifts. Our approach involves domain adaptation for a shifted pre-trained image prior through TTT using DEQ to close the distribution gap, which only requires a test single measurement. Our results show that PnP-TTT can improve the performance significantly given sufficient measurement for shifted priors with minimal computational cost.

3. METHOD

We now present our method for domain adaption of image priors in PnP. We consider the PnP-PGM algorithm in eq. (2) and run it until its convergence. In practice, we find that about 100 iterations of PnP-PGM are sufficient in our configuration. We can update the weights of the image prior on a test measurement by minimizing the following self-supervised loss

$$\ell_{\mathsf{sup}}(\boldsymbol{\theta}) = \|\boldsymbol{A}\mathsf{T}_{\boldsymbol{\theta}}(\overline{\boldsymbol{x}}) - \boldsymbol{y}\|_2^2\,, \tag{8}$$

where $\overline{\boldsymbol{x}}$ is the fixed-point of PnP-PGM defined in eq. (3) and T_θ is the operator defined in (2). We use the DEQ to compute the gradient of ℓ_{sup} at test-time using implicit differentiation. We follow a method similar to [17, 18] to train the image priors using the DnCNN architecture, with batch normalization layers replaced with spectral normalization to control the Lipschitz constant of the denoisers. DnCNN is trained as a denoiser for AWGN level $\sigma = 5$. During the training stage we do not use DU or DEQ so that the learned prior model is purely an image denoiser. We use 400 CBSD to train natural prior on grayscale images of size 180×180 [19]. We train MRI priors on MRI brain images of size 256×256 [20].

For test-time training, we initialize PnP-PGM with $\boldsymbol{x}^0 = \boldsymbol{0}$ and 100 iterations in the forward pass of DEQ, using the trained denoising prior. We use Nesterov acceleration [21], and set stepsize $\gamma = 1$. We use 100 iterations and Anderson acceleration in the backward pass of DEQ [22]. We allow TTT to run for 50 iterations, using SGD to update the parameters θ with a step size of 1×10^{-5}. At inference, using the adapted prior, we again run for 100 iterations with Nesterov acceleration in PnP-PGM with step size $\gamma = 1$. Note that once all 50 TTT iterations are performed for a particular measurement, θ is reset to the non-domain-adapted weights. Since the goal of PnP-TTT is to overcome the performance gap from a distribution shift between train- and test-time, performing TTT on as many measurements are available at test-time may be beneficial. Future experiments could examine if there is any performance improvement when using multiple measurements instead of only one for PnP-TTT. The measurement model for a single-coil, accelerated MRI with radial Fourier sampling can be modeled as $\boldsymbol{A} = \boldsymbol{M}\boldsymbol{F}$, where $\boldsymbol{M}$ is the diagonal sampling matrix and $\boldsymbol{F}$ is the Fourier transform. We investigate five different sampling ratios (m/n) in the experiments. For the experiments reported here, we consider a noiseless scenario; however, we expect similar performance of PnP-TTT under moderate amounts of noise.

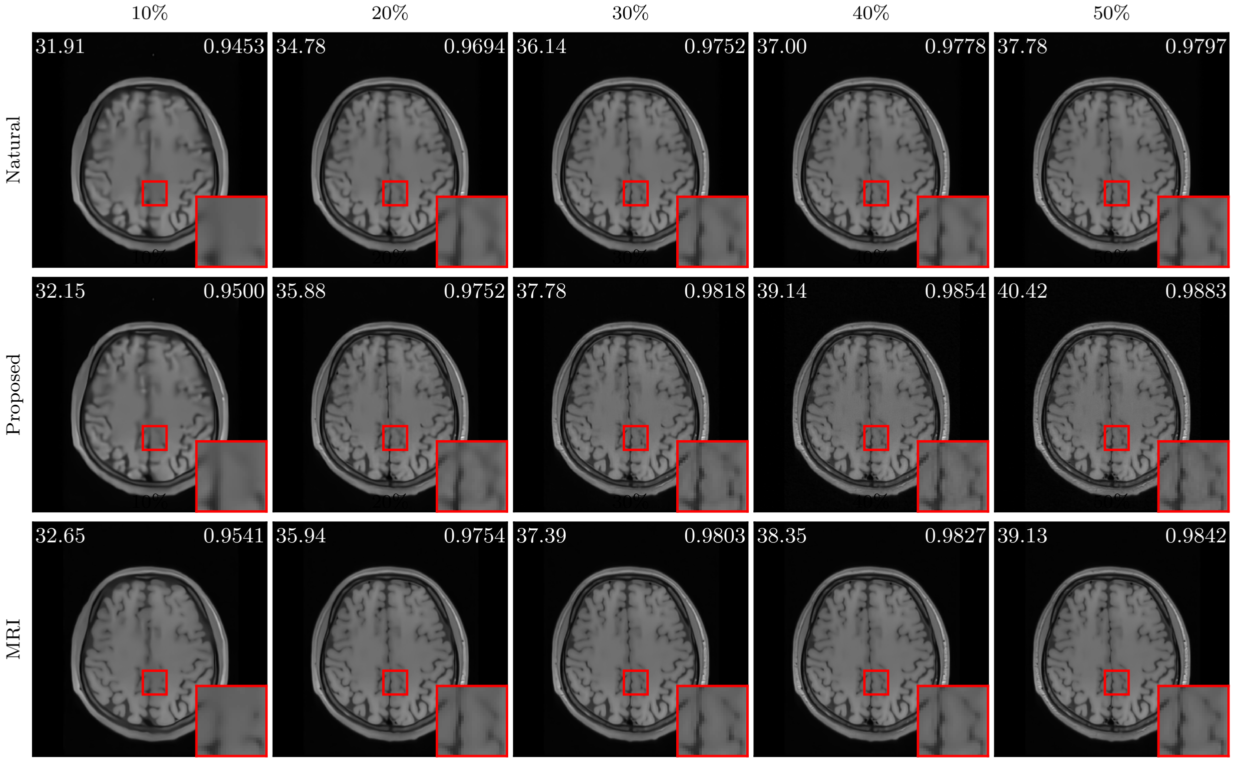

Fig. 2: Visual evaluation of priors at various CS ratios for CS-MRI with corresponding PSNR (upper left) and SSIM (upper right). Note the suboptimal performance of natural prior compared to matched MRI prior and improvement of PnP-TTT (middle row).

4. RESULTS

We test our proposed method by reconstructing ten brain MRI images selected from the test dataset of [20] with mismatched DnCNN prior trained on natural images. Due to distribution shift, natural priors demonstrate suboptimal performance for the MRI task. To establish a performance baseline, we compare the result of the proposed method with those obtained by mismatched natural prior and matched MRI prior. Specifically, we consider the performance achieved by natural prior as the lower baseline, and that achieved by the MRI prior as the upper baseline. Our proposed PnP-TTT seeks to enhance the performance of a mismatched natural prior so as to approach that of a matched MRI prior.

Table 1 reports the best results achieved for five CS ratios: 10, 20, 30, 40, and 50. It can be seen that PnP-TTT can close the performance gap for CS ratios of 20 and more, while for a CS ratio of 10, it can make an improvement compared to the lower baseline (mismatched natural prior). The reconstruction quality is quantified using peak signal-to-noise ratio (PSNR) in dB and the structural similarity index measure (SSIM). Figure 1 illustrates two empirical results: *(*a) The empirical performance of PnP-TTT at testing for different CS ratios (*left figure*), *(*b) The empirical performance during test-time training (*remaining five figures*). Note that during test-time training, the prior can overfit to the measurement. Thus, in practice it is necessary to hold out some measurements to use early stopping during TTT [15], although we have not included such results in this paper. The visual results can be found in Figure 2. It can be seen both empirically and visually that PnP-TTT can shorten the gap due to distribution shift, close it completely, or go beyond closing it given the CS ratios.

5. CONCLUSION

We present PnP-TTT as a novel framework for closing the performance gap that arises due to mismatched priors in imaging inverse problems. PnP-TTT achieves this by adapting the mismatched priors during the testing phase by using DEQ training to update the weights of the mismatched priors. One of the main advantage of PnP-TTT is that one can use mismatched priors on a shifted distribution without the need to do additional training. Instead, the prior can simply be adapted to the test-time measurements. Our results show that PnP-TTT can significantly enhance the performance, achieving performance comparable to that of using a matched prior during inference. Furthermore, this work demonstrates that priors from different tasks can be used interchangeably in scenarios with shifted distribution without the loss of performance.

A. REFERENCES

[1] M. T. McCann, K. H. Jin, and M. Unser, "Convolutional neural networks for inverse problems in : A review," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 85–95, 2017.

[2] A. Lucas, M. Iliadis, R. Molina, and A. K. Katsaggelos, "Using deep neural networks for inverse problems in imaging: Beyond analytical methods," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 20–36, Jan. 2018.

[3] U. S. Kamilov, C. A. Bouman, G. T. Buzzard, and B. Wohlberg, "Plug-and-play methods for integrating physical and learned models in computational imaging," *IEEE Signal Process. Mag.*, vol. 40, no. 1, pp. 85–97, Jan. 2023.

[4] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," *IEEE Glob. Conf. Signal Inf. Process.*, pp. 945–948, 2013.

[5] S. Sreehari, S. V. Venkatakrishnan, B. Wohlberg, G. T. Buzzard, L. F. Drummy, J. P. Simmons, and C. A. Bouman, "Plug-and-play priors for bright field electron tomography and sparse interpolation," *IEEE Trans. Comput. Imag.*, vol. 2, no. 4, pp. 408–423, December 2016.

[6] S. A. Hosseini, B. Yaman, S. Moeller, M. Hong, and M. Akcakaya, "Dense recurrent neural networks for accelerated MRI: History-cognizant unrolling of optimization algorithms," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 6, pp. 1280–1291, Oct. 2020.

[7] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett, "Deep learning techniques for inverse problems in imaging," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 39–56, May 2020.

[8] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image process.," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, Mar. 2021.

[9] D. Gilton, G. Ongie, and R. Willett, "Deep equilibrium architectures for inverse problems in imaging," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 1123–1133, 2021.

[10] J. Liu, X. Xu, W. Gan, S. Shoushtari, and U. S. Kamilov, "Online deep equilibrium learning for regularization by denoising," in *Proc. Advances Neural Inf. Process. Syst.*, New Orleans, LA, 2022.

[11] Shirin Shoushtari, Jiaming Liu, Yuyang Hu, and Ulugbek S Kamilov, "Deep model-based architectures for inverse problems under mismatched priors," *IEEE J. Sel. Areas Inf. Theory*, 2022.

[12] H. Venkateswara, S. Chakraborty, and S. Panchanathan, "Deep-learning systems for domain adaptation in computer vision: Learning transferable feature representations," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 117–129, Nov. 2017.

[13] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia, "A brief review of domain adaptation," in *Advances in Data Science and Inf. Eng.*, Cham, 2021, pp. 877–894, Springer International Publishing.

[14] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *Proc. 37th Int. Conf. on Mach. Learn.*, 13–18 Jul 2020, vol. 119, pp. 9229–9248.

[15] M. Z. Darestani, J. Liu, and R. Heckel, "Test-time training can close the natural distribution shift performance gap in deep learning based compressed sensing," in *Proc. 39th Int. Conf. on Mach. Learn.*, 17–23 Jul 2022, vol. 162, pp. 4754–4776.

[16] A. Beck and M. Teboulle, "Fast gradient-based algorithm for constrained total variation image denoising and deblurring problems," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2419–2434, November 2009.

[17] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, 2017.

[18] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida, "Spectral normalization for generative adversarial networks," in *Int. Conf. on Learn. Represent.*, 2018.

[19] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int. Conf. Comput. Vis.*, July 2001, vol. 2, pp. 416–423.

[20] J. Zhang and B. Ghanem, "Ista-net: Interpretable optimization-inspired deep network for image compressive sensing," *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1828–1837, 2018.

[21] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer Science & Business Media, 2003.

[22] D. G. Anderson, "Iterative procedures for nonlinear integral equations," *J. ACM*, vol. 12, no. 4, pp. 547–560, 1965.