Bilevel Learning of Task-Adapted Regularizers

Aryan Saha and Subhadip Mukherjee
Department of Electronics and Electrical Communication Engineering, IIT Kharagpur, India
Emails: aryan18@kgpian.iitkgp.ac.in, smukherjee@ece.iitkgp.ac.in

***Abstract*—In scientific imaging, reconstructing an unknown image from noisy, incomplete measurements is often necessary for downstream tasks (e.g., classification, segmentation). We address this using a task-adapted framework that learns a data-driven regularizer tailored to the specific analysis task. We propose a *bilevel learning* approach: solving a lower-level variational optimization problem with a parameterized *learnable* regularizer, while optimizing its parameters by minimizing the upper-level task loss on training data. This facilitates learning a bespoke regularizer for a pre-trained task operator, or jointly learning the parameters of both the regularizer and a deep neural network-based task operator. We introduce a principled algorithm that accommodates inexact lower-level reconstructions (using an iterative first-order solver) and demonstrate its efficacy on challenging inverse problems (image deblurring, CT, and MRI) across classification and segmentation tasks. Our method achieves competitive end-task performance compared to existing methods that use deep networks for both reconstruction and task operators, but does so with a significantly parsimonious regularizer parameterization. Furthermore, our approach requires only raw measurements and task label pairs, eliminating the need for ground-truth reference images.**

***Index Terms*—Task-adapted reconstruction, bilevel learning, variational regularization.**

I. Introduction

Imaging inverse problems [1], [2] are ubiquitous in many applications, such as medical imaging, astronomy, microscopy, and low-level computer vision tasks. The key objective is to reconstruct an image $x \in \mathbb{R}^n$ from its noisy and incomplete measurement $y = Ax + w$, where $A \in \mathbb{R}^{m\times n}$ represents the imaging process (also called the *forward operator*) and w denotes measurement noise. This is achieved by minimizing a variational loss that trades off data-fidelity (capturing the measurement consistency of the recovered image) with a regularization term that promotes certain smoothness properties. The regularizer is handcrafted or learned in a data-driven manner [3]. The reconstructed image is then used for solving a downstream task, e.g., tumor segmentation or disease classification in medical imaging, or object detection in computer vision challenges. In the classical workflow, image reconstruction and the downstream task are treated independently (in a sequential manner), thereby limiting the possibility of tailoring the reconstruction to the end task. Researchers have demonstrated that combining image reconstruction with tasks such as edge recovery [4], semantic segmentation [5], and image registration [6] can lead to improved performance for the task, as compared with the conventional sequential approach. More recently, Adler et al. [7] proposed a *joint learning* approach for the reconstruction and the task operators, while parameterizing both using deep neural networks (DNNs). Their approach seeks to jointly learn a task network T_ϕ and a reconstruction network R_γ on a training dataset consisting of triplets (y, x, t), where y denotes the raw measurement, x is the corresponding true reference image, and t is the true label for the task. The training loss for jointly learning T_ϕ and R_γ is given by

$$c\,\mathbb{E}\left[\mathcal{L}_{\text{task}}(t, T_\phi(R_\gamma(y)))\right] + (1-c)\,\mathbb{E}\left[\mathcal{L}_{\text{recon}}\left(x, R_\gamma(y)\right)\right], \quad (1)$$

where $\mathcal{L}_{\text{task}}$ and $\mathcal{L}_{\text{recon}}$ denote appropriate metrics to capture the accuracy for the end task and the reconstruction, respectively, and $c \in [0, 1]$ is a hyperparameter that trades off task loss with reconstruction loss. Notably, $c = 1$ leads to an end-to-end learning of the composition of the task and reconstruction operators, without using any reference image x (which, as shown in [7], yields a suboptimal task performance). The joint learning method as in (1) outperforms other alternatives, e.g., sequential (where one first trains R_γ on (y, x) pairs, and then learns T_ϕ to map the output of the pre-trained R_γ to the true label t) and end-to-end learning. However, this approach requires (i) careful fine-tuning of c, for which, there is no principled automated search strategy, and (ii) (y, x, t) triplets for training (which are often unavailable in practice). Moreover, since the reconstruction is obtained by applying a generic deep network R_γ on y, the approach in [7] is not interpretable or amenable to convergence analysis.

A. This paper

We employ a bilevel framework to jointly learn the parameters of a task operator and a variational regularizer. At the lower level, we approximate the reconstructed image using the current regularizer. This reconstruction is passed to the task network, where the upper-level loss optimizes the parameters by measuring the task output against reference labels. This approach flexibly adapts the regularizer to either a jointly trained or a frozen, pre-trained task operator. Furthermore, our lower-level variational formulation offers better interpretability than deep reconstruction networks and enables convergence analysis via classical regularization theory. Key advantages of our bilevel task-adapted learning include: (1) a frugally parameterized regularizer inspired by classical models (e.g., smoothed TV [8] and fields-of-experts [9]); (2) the elimination of hyperparameter tuning, unlike [7]; and (3) the reliance on readily available measurement-label (y, t) pairs rather than

SM is supported by the Prime Minister's Early Career Research Grant from the ANRF (ANRF/ECRG/2024/001178/ENS).

impractical (y, x, t) triplets. Even when adapting to a pre-trained T_ϕ (trained separately on (x, t) pairs), the regularizer can be learned on a distinct set of (y, t) pairs, completely circumventing the need for full triplets.

II. Task-adapted bilevel learning approach

We formulate task-adapted learning of variational regularizers as a *bilevel optimization* problem, given by

$$\min_{\theta,\phi} J(\theta,\phi) \triangleq \mathbb{E}_{(y,t)\sim\mu}\left[\mathcal{L}_{\text{task}}\left(t, T_\phi\left(x(\theta;y)\right)\right)\right]$$
$$\text{s.t. } x(\theta;y) = \arg\min_x \Phi(x,\theta) \triangleq \frac{1}{2}\|y - Ax\|_2^2 + g_\theta(x). \quad (2)$$

Here, $\mathcal{L}_{\text{task}}$ represents an appropriately chosen task loss (e.g., the cross-entropy loss for classification, or the dice loss for segmentation), T_ϕ denotes a DNN-based task operator with parameters ϕ, and t is the ground-truth target (true label). The upper-level loss is approximated (by empirical mean) using a training dataset consisting of independent and identically distributed (i.i.d.) samples $(y_i, t_i)_{i=1}^N \sim \mu$. The reconstructed image is computed at the lower level by solving a variational problem with a parametric regularizer g_θ, given the measurement y. The proposed framework offers the following crucial advantages: (i) unlike [7], there is no need to tune any additional hyperparameter and (ii) one has the flexibility to learn T_ϕ and g_θ simultaneously, or adapt g_θ to a pre-trained task operator T_{ϕ^*} (by minimizing the upper-level loss with respect to θ, keeping $\phi = \phi^*$). Additionally, thanks to the variational formulation at the lower level, the reconstruction offers improved mathematical interpretability. The lower-level solution $x(\theta; y)$ is approximated using a gradient-based solver (e.g., Adam), and, subsequently, the upper-level gradient (w.r.t. θ) is computed using the implicit function theorem (IFT) (see [10, Sec. 4.2.1] for a derivation):

$$\nabla_\theta J = -\left(\nabla^2_{x,\theta}\Phi\right)^\top \left(\nabla^2_{x,x}\Phi\right)^{-1} \nabla_s \mathcal{L}_{\text{task}}(t,s)\Big|_{\substack{x=x(\theta;y)\\ s=T_\phi(x(\theta;y))}}.$$

Notably, the expression above holds for a single training example pair (y, t), and $\nabla_\theta J$ over a mini-batch can be computed by taking the mean over all examples in the mini-batch. Since one does not have access to the exact lower-level solution $x(\theta; y)$, the expression for $\nabla_\theta J$ can only be computed at an approximate lower-level solution $\hat{x}$ computed using an iterative solver [11] (e.g., Adam). Additionally, the Hessian inverse is approximated by solving (with an a priori accuracy) a system of linear equations using the conjugate gradient (CG) algorithm in the interest of computational feasibility. The steps for computing an approximation to the upper-level gradient for a single training example (y, t) are listed in Algorithm 2. While executing Step 4 of Algorithm 2 (i.e., in implementing the CG iterations), one needs to compute the Hessian-vector product $\left[\nabla^2_{x,x}\Phi\right] v$ for any vector $v \in \mathbb{R}^n$, which can be computed efficiently as $\nabla_x \left[\nabla_x \Phi^\top v\right]$ using `torch.autograd`. Further, when $\Phi(x, \theta)$ in (2) is $\mu(\theta)$-strongly convex in x for a given θ, ensuring $\|\nabla_x \Phi(\hat{x}, \theta)\|_2 \leq \mu(\theta)\,\epsilon$ in turn guarantees that $\|\hat{x} - x(\theta; y)\|_2 \leq \epsilon$ holds [12].

Algorithm 1 Bilevel learning of g_θ and T_ϕ

Require: Initial task network T_ϕ and regularizer g_θ, and learning rates η_ϕ and η_θ.

1: **for** epoch $= 1, \ldots, E$ **do**
2: **for** each batch $\mathcal{B} = (y_i, t_i)_{i=1}^B$ **do**
3: Find $\hat{x}_i \approx \arg\min_x \frac{1}{2}\|y_i - Ax\|_2^2 + g_\theta(x)$, using a maximum of N_{it} Adam iterations.
4: Compute $\hat{t}_i \leftarrow T_\phi(\hat{x}_i)$ and $\hat{J} = \dfrac{1}{B}\displaystyle\sum_{i=1}^{B} \mathcal{L}_{\text{task}}(t_i, \hat{t}_i)$.
5: Compute $p_\theta \approx \nabla_\theta \hat{J}$ via Algorithm 2.
6: $\theta \leftarrow \text{Adam}_{\eta_\theta,\beta_1,\beta_2}(\theta, p_\theta)$, where η_θ is the learning rate for θ and (β_1, β_2) are the momentum parameters.
7: To update T_ϕ, do $\phi \leftarrow \text{Adam}_{\eta_\phi,\beta_1,\beta_2}\left(\phi, \nabla_\phi \hat{J}\right)$. Else (for fixed, pretrained T_ϕ), leave ϕ unchanged.
8: **end for**
9: **end for**

Algorithm 2 Approximating $\nabla_\theta J(\theta, \phi)$ using IFT [11]

Require: Regularizer g_θ, task network T_ϕ measurement y, true task label t, tolerance ϵ, strong convexity parameter $\mu(\theta)$ of Φ, and maximum lower-level iterations N_{it}.

1: **Initialize** the lower level solution to $x_0 \leftarrow A^\dagger y$.
2: Find an approximate solution $\hat{x}$ for the lower level problem, such that $\|\nabla_x \Phi(\hat{x}, \theta)\|_2 \leq \mu(\theta)\,\epsilon$ or until N iterations of Adam, where $\Phi(x, \theta)$ is as defined in (2).
3: **Compute** $b \leftarrow \nabla_s \mathcal{L}_{\text{task}}(t, s)\Big|_{s=T_\phi(\hat{x})}$.
4: Solve for q using CG up to tolerance ϵ:
$$\left(\nabla^2_{x,x}\,\Phi(x,\theta)\Big|_{x=\hat{x}}\right) q = b.$$
5: **Set** $p \leftarrow -\left(\nabla^2_{x,\theta}\,\Phi(x,\theta)\Big|_{x=\hat{x}}\right)^\top q$.
6: **return** p

III. Experimental results

The performance of the proposed task-driven bilevel learning framework is evaluated for three problems: (i) classification on blurry and noisy images, to show proof-of-concept; and image segmentation in (ii) CT and (iii) MRI; to demonstrate the efficacy of the proposed method on challenging practical inverse problems in medical imaging. For the CT and MRI experiments, we use the *Spleen* and *Heart* image datasets, respectively, from the Medical Segmentation Decathlon (MDS) challenge [13]. We use the *Deep Inverse* library [14] for implementing the forward, adjoint, and pseudo-inverse (which is the filtered back-projection for CT and the zero-filled inversion for MRI) operators in CT and MRI.

A. Training strategies

The experimental results corresponding to the following two training settings–pretrained task operator with a learned regularizer and joint learning of task and regularizer–are reported:

TABLE I: Quantitative comparison of task-adapted reconstruction approaches across three inverse problems. The *decoupled* baseline reflects the score of applying a pre-trained T_ϕ (trained on (x, t) pairs) to the output of a separately trained R_γ (trained on (y, x) pairs). In contrast, the *upper bound* is the task network's score on the clean reference image x. These establish the minimum and maximum performance limits for task-adapted recovery. The joint learning [7] score corresponds to the best c value found via grid-search over $[0, 1]$. We report overall test accuracy for MNIST, and segmentation Dice scores for the CT and MRI MSD challenge datasets. Ultimately, jointly learning T_ϕ and g_θ via our bilevel framework achieves competitive performance with [7], without requiring any hyperparameter tuning and ground-truth reference images.

Dataset	**Inverse problem**	**Baselines**		**Adler et al. [7]**			**Bilevel learning (ours)**			
		Decoupled	**Upper bound**	**Sequential**	**End-to-end**	**Joint**	**Fixed T_ϕ**		**Joint T_ϕ and g_θ**	
							TV	**FoE**	**TV**	**FoE**
MNIST	Gaussian blur	92.72%	99.35%	98.26%	98.70%	98.87%	67.31%	90.32%	98.49%	98.85%
Spleen	CT	0.8120	0.9278	0.8795	0.9034	0.9148	0.7334	0.7217	0.8727	0.8929
Heart	MRI	0.8116	0.9173	0.8876	0.9055	0.9139	0.5822	0.5726	0.8422	0.8470

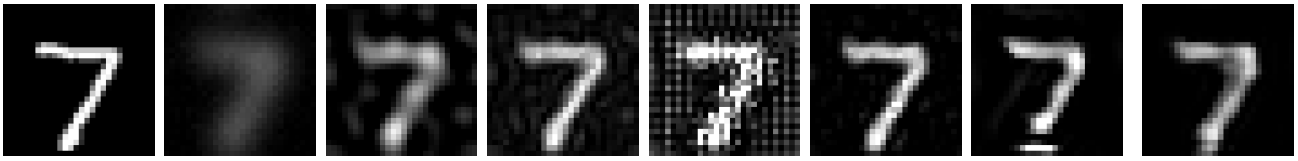

Fig. 1: Reconstructed MNIST images from blurry and noisy measurements for different approaches. The bilevel approach, where T_ϕ and g_θ are jointly learned, leads to a considerably high-fidelity artifact-free reconstruction, leading to better classification accuracy.

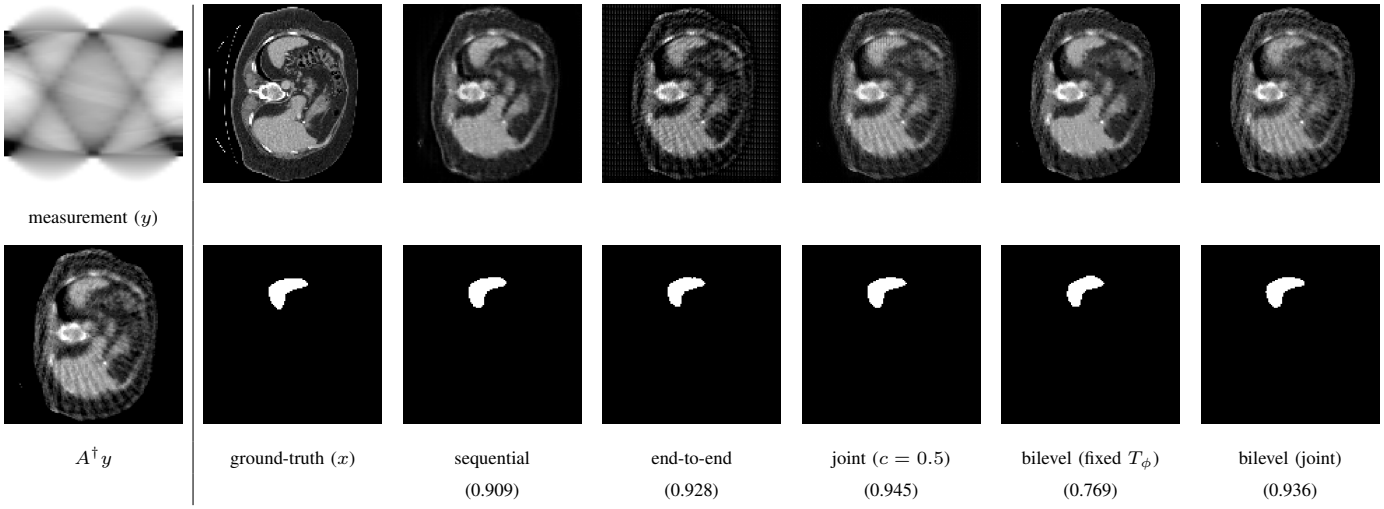

Fig. 2: Comparison of reconstructed images (top row) and their corresponding segmentation masks (bottom row) across different approaches for CT. The proposed bilevel approach performs on par with joint learning [7], corresponding to the best value of c found through a grid-search.

1) **Pretrained task operator**: Here, we first train a task network T_ϕ on a training dataset consisting of pairs $(x_i, t_i)_{i=1}^N$ of ground-truth images and the corresponding task labels, and then keep it frozen while updating the regularizer parameters using bilevel learning (see Algorithm 1).
2) **Joint learning of the regularizer and the task operator**: In this setting, T_ϕ and g_θ are optimized jointly in an alternating fashion. The parameters ϕ of the task network are updated via standard backpropagation on the upper-level loss $J(\theta, \phi)$, while θ is updated using an inexact gradient $p_\theta \approx \nabla_\theta J(\theta, \phi)$ as outlined in Algorithm 1.

We found that for all experiments, the latter approach (learning both T_ϕ and g_θ) works significantly better than adapting the regularizer to a fixed, pre-trained task network.

B. Regularizer parameterizations

We employ the following two parameterizations of g_θ, both of which lead to a strongly convex (w.r.t. x, for a given θ) and differentiable (w.r.t. both x and θ) lower-level loss $\Phi(x, \theta)$:

1) The *smoothed isotropic TV regularizer*, with only two learnable parameters θ_0 and θ_1, defined as

$$g_\theta(x) = \frac{e^{\theta_0}}{n} \sum_{i,j} \sqrt{(\nabla_h x)_{i,j}^2 + (\nabla_v x)_{i,j}^2 + e^{\theta_1}}, \qquad (3)$$

where $\nabla_h x$ and $\nabla_v x$ denote horizontal and vertical finite difference operators, e^{θ_0} is a penalty term, and e^{θ_1} is a smoothing parameter. This parameterization ensures differentiability of $\Phi(x, \theta)$ with respect to both x and $\theta = [\theta_0, \theta_1]$, while maintaining the edge-preserving characteristics of classical TV regularization.

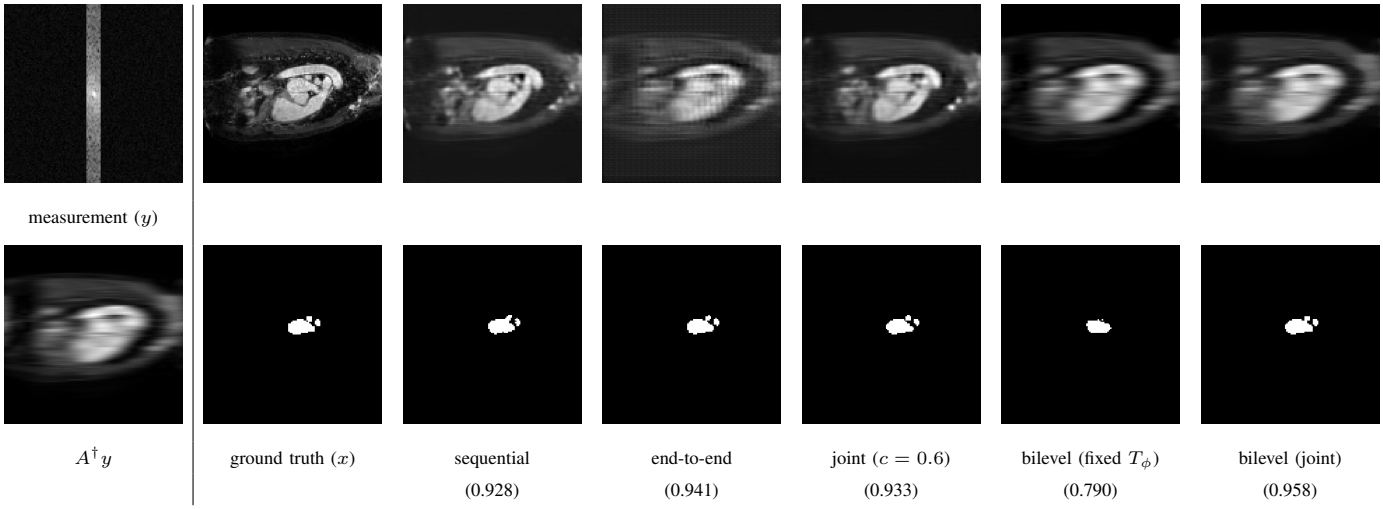

Fig. 3: Comparison of reconstructed images (top) and their corresponding segmentation masks (bottom) across different approaches for MRI. As indicated by the dice scores below the images, the bilevel approach with jointly learned T_ϕ and g_θ achieves a competitive or slightly better segmentation accuracy than its nearest competitor while using a significantly parsimonious parameterization of the regularizer (vis-à-vis the deep reconstruction network).

2) The fields-of-experts (FoE) regularizer [9], [10] that uses L learned filters (or *experts*) to learn image features:

$$g_\theta(x) = e^{\theta_0} \cdot \textstyle\sum_{j=1}^{L} e^{\beta_j} \cdot \frac{1}{n} \sum_{i,k} \rho_{\nu_j}\left((c_j \circledast x)_{i,k}\right). \quad (4)$$

Here, $\circledast$ denotes 2D convolution, $c_j \in \mathbb{R}^{K\times K}$ are learnable filters, θ_0 is the log-space global regularization weight, β_js are per-expert log-space weights, and $\rho_{\nu_j}(t) = \sqrt{t^2 + e^{\nu_j}}$ denotes a smoothed ℓ_1 penalty with a learnable smoothing parameter ν_j. For L experts of size $K \times K$, there are a total of $(K^2+2)L+1$ learnable parameters in g_θ. The filters in the FoE regularizer are initialized as classical edge and blob detectors, keeping in view the end tasks of classification and segmentation. In particular, the first four filters are initialized as Sobel-like edge extractors along horizontal, vertical, diagonal, and anti-diagonal directions, while the fifth filter is initialized as a 5×5 Laplacian-of-Gaussian kernel [15]. These specific choices for the initial filters provide a good starting point for the bilevel optimization framework to further refine them during training to extract the most informative image features for the end task.

C. Task networks

For classification on the blurry and noisy MNIST digits, we choose the ResNet-18 architecture [16] (implemented using the `torchvision` library, having approximately 11.17 million parameters) for the task network T_ϕ. For the CT and MRI inverse problems, where the intended task is semantic segmentation of specific organs, we employ a standard U-Net architecture [17] for T_ϕ (having 95172 learnable parameters) adapted from the MONAI library [18]. We chose a more economical parameterization for the segmentation U-Net to avoid overfitting, as the CT and MRI datasets used in our experiments have only approximately 1000 images (of size 128×128). For comparison with the task-adapted learning approach by Adler et al. [7], we parameterize the reconstruction network R_γ as the composition $G_\gamma \circ A^\dagger$ (i.e., as a postprocessing network applied on the pseudo-inverse reconstruction), where G_γ is a U-Net with 378740 learnable parameters.

D. Classification of blurry and noisy images

The degraded measurement is given by $y = Ax + w$, where A corresponds to a (Gaussian) blur kernel and w denotes additive Gaussian noise with variance σ_w^2. The training dataset consists of pairs (y, t), where t is the true class label of the image. We utilize the MNIST dataset consisting of (x, t) pairs, and simulate y by applying a 23×23 Gaussian kernel with the standard deviation parameter $\sigma = 3$ (to ensure a severe degree of blurring), followed by adding i.i.d. Gaussian noise with $\sigma_w = 0.01$. The MNIST dataset is split into training and test datasets, consisting of 60000 and 10000 images, respectively. A batch size of 4096 is used during training, and learning rates of $\eta_\phi = 0.001$ and $\eta_\theta = 0.05$ are applied for updating the parameters of T_ϕ and g_θ, respectively. Since the task considered here is image classification, $\mathcal{L}_{\text{task}}$ is taken as the standard cross-entropy loss, while we take $\mathcal{L}_{\text{recon}}$ in (1) to be the mean-squared error (MSE) loss while comparing the performance with [7]. The classification accuracy of the bilevel approach (with pre-trained T_ϕ and jointly learned T_ϕ and g_θ) are reported in Table I. To facilitate a visual comparison, we show the reconstructed images corresponding to the proposed methods and the competing techniques in Figure 1.

The raw measurement and the naive pseudo-inverse reconstructions exhibit severe Gaussian blur, making the digit edges nearly indistinguishable from the background. The sequential and end-to-end training approaches of [7] successfully sharpen the digit but introduce grid-like artifacts (more pronounced in the end-to-end approach). We report the performance of joint training of [7] corresponding to the best c, found using a grid-search. The joint loss in (1) balances reconstruction and task errors, resulting in a cleaner image, although training requires

triplets (x, y, t). The bilevel approach with a fixed, pre-trained T_ϕ) leads to a slightly distorted image since the task network is frozen, and the regularizer focuses entirely on exaggerating features strictly necessary for classification, while sacrificing natural visual fidelity. Joint learning of T_ϕ and g_θ yields the sharpest and the most natural reconstruction that is visually closest to the ground truth, while performing on par with the joint learning strategy of [7] using significantly fewer learnable parameters in comparison.

E. Image segmentation in CT and MRI

To demonstrate the applicability of our approach to practical clinical inverse problems, we consider the task of semantic segmentation of the spleen from abdominal CT images and the left atrium from cardiac MR images (Task-9 and Task-2, respectively, in the MSD challenge [13]). The CT forward operator is simulated using a sparse-view geometry with only 60 projection angles, followed by adding Gaussian noise with $\sigma_w = 0.5$ to the clean sinogram. The MRI forward operator applies a $12\times$ undersampling in the k-space, followed by Gaussian noise corruption with $\sigma_w = 3$. For both problems, we use an $85\%/15\%$ split of the dataset for training and inference. The CT and MR images are of size 128×128, and a batch size of 64 images is used for training. The learning rates are taken as $\eta_\phi = \eta_\theta = 0.05$ for both modalities. The task loss $\mathcal{L}_{\text{task}}$ combines binary cross-entropy and the soft-dice loss with weights 0.8 and 0.2, respectively. A quantitative performance evaluation (in terms of the dice score) is presented in Table I, while a visual comparison of the reconstructed images is shown in Figures 2 and 3 for CT and MRI, respectively.

We observe that the baseline pseudo-inverse reconstructions $A^\dagger y$ suffer heavily from sparse-view streak artifacts in CT and aliasing artifacts in MRI due to undersampling in the frequency domain, providing a poor foundation for segmentation. Both sequential and end-to-end approaches remove major artifacts and yield reasonably accurate segmentation masks. However, the images remain artifact-ridden, indicating suboptimal tuning of the reconstruction operator. For the proposed bilevel approach with a pre-trained task operator on (x, t) pairs, the reconstruction mostly blurs or abstracts away the background anatomy irrelevant to the task to enhance the contrast of the region-of-interest. While this drops the absolute Dice score due to domain shift with the fixed network, it qualitatively illustrates how the regularizer gets steered by the task loss. By learning the regularizer and task network simultaneously, the framework restores the natural anatomical texture and structure of the organs, which helps achieve highly accurate segmentation masks.

IV. Conclusions

We present a variational approach for learning a bespoke regularizer for task-adapted image reconstruction for classification and segmentation. The method adapts flexibly via the task loss and avoids the impractical requirement of image-measurement-label training triplets. Despite utilizing a frugal, mathematically interpretable parameterization, our approach achieves competitive end-task performance compared to over-parameterized deep models (trained end-to-end). While our bilevel formulation incurs additional computational overhead during training due to the unrolled lower-level solver, it notably avoids exhaustive hyperparameter tuning. Additionally, the observed performance disparity between fixed-task and joint training strategies highlights a natural trade-off: joint co-adaptation yields superior end-task accuracy, albeit at the cost of increased training complexity. Promising future directions include exploring powerful structured regularizers, such as input-convex neural networks [19], and analyzing upper-level stochastic convergence using mini-batch updates.

References

[1] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen, "Variational methods in imaging," 2009.

[2] M. Benning and M. Burger, "Modern regularization methods for inverse problems," *Acta Numerica*, vol. 27, pp. 1–111, 2018.

[3] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb, "Solving inverse problems using data-driven models," *Acta Numerica*, vol. 28, pp. 1–174, 2019.

[4] M. Burger and S. Osher, *Level Set and PDE Based Reconstruction Methods in Imaging*. Springer International Publishing, 2013.

[5] R. Ramlau and W. Ring, "A mumford–shah level-set approach for the inversion and segmentation of x-ray tomography data," *Journal of Computational Physics*, vol. 221, no. 2, pp. 539–557, 2007.

[6] C. P. Lau, O. Öktem, and C.-B. Schönlieb, "Indirect image registration with large diffeomorphic deformations," *SIAM Journal on Imaging Sciences*, vol. 11, no. 1, pp. 577–606, 2018.

[7] J. Adler, S. Lunz, O. Verdier, C.-B. Schönlieb, and O. Öktem, "Task adapted reconstruction for inverse problems," *Inverse Problems*, vol. 38, no. 7, p. 075006, 2022.

[8] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.

[9] Y. Chen, R. Ranftl, and T. Pock, "Insights into analysis operator learning: From patch-based sparse models to higher order MRFs," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1060–1072, Mar. 2014.

[10] C. Crockett and J. A. Fessler, "Bilevel methods for image reconstruction," *Foundations and Trends in Signal Processing*, vol. 15, no. 2–3, p. 121–289, 2022.

[11] F. Pedregosa, "Hyperparameter optimization with approximate gradient," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun. 2016, pp. 737–746.

[12] M. S. Salehi, S. Mukherjee, L. Roberts, and M. J. Ehrhardt, "An adaptively inexact first-order method for bilevel optimization with application to hyperparameter learning," *SIAM Journal on Mathematics of Data Science*, vol. 7, no. 3, pp. 906–936, 2025.

[13] M. Antonelli *et al.*, "The Medical Segmentation Decathlon," *Nature Communications*, vol. 13, no. 1, p. 4128, Jul. 2022.

[14] J. Tachella *et al.*, "DeepInverse: A Python package for solving imaging inverse problems with deep learning," *Journal of Open Source Software*, vol. 10, p. 8923, 2025.

[15] D. Marr and E. Hildreth, "Theory of edge detection," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 207, no. 1167, pp. 187–217, 1980.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[17] E. Kerfoot *et al.*, "Left-ventricle quantification using residual U-net," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2018, pp. 371–380.

[18] M. J. Cardoso *et al.*, "MONAI: An Open-Source Framework for Deep Learning in Healthcare," *arXiv:2211.02701*, 2022.

[19] S. Mukherjee *et al.*, "Data-driven convex regularizers for inverse problems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 13 386–13 390.