

WHEN DOES BACKDOOR ATTACK SUCCEED IN IMAGE RECONSTRUCTION? A STUDY OF HEURISTICS VS. BI-LEVEL SOLUTION

Vardaan Taneja¹, Pin-Yu Chen², Yuguang Yao³, Sijia Liu³

¹ Indian Institute of Technology, Delhi ² IBM Research ³ Michigan State University

ABSTRACT

Recent studies have demonstrated the lack of robustness of image reconstruction networks to test-time evasion attacks, posing security risks and potential for misdiagnoses. In this paper, we evaluate how vulnerable such networks are to *training-time poisoning attacks* for the first time. In contrast to image classification, we find that trigger-embedded basic backdoor attacks on these models executed using heuristics lead to poor attack performance. Thus, it is non-trivial to generate backdoor attacks for image reconstruction. To tackle the problem, we propose a bi-level optimization (BLO)-based attack generation method and investigate its effectiveness on image reconstruction. We show that BLO-generated backdoor attacks can yield a significant improvement over the heuristics-based attack strategy.

Index Terms— Backdoor Attacks, Image Reconstruction, Bilevel Optimization, Data Poisoning

1. INTRODUCTION

Image reconstruction is an inverse problem that constitutes the recovery of an image from finite indirect measurements and is used in many applications, including critical ones like medical imaging (eg: MRI, CT). Deep learning (DL) provides novel methods which try to solve the problem more accurately and with much lesser samples compared to classical methods [1, 2, 3]. This could potentially make MRI scans, for instance, much efficient and effective [4].

Adversarial perturbations are tiny changes in the data input that cause a large degradation in the performance of pre-trained networks [5, 6]. While these DL-based methods have achieved state-of-the-art (SOTA) performance on image reconstruction [7], these networks have been also found to be very unstable and prone to adversarial perturbations [8, 9], leading to large degradation in reconstruction accuracy at testing time. Nearly all of existing work focuses on generating and defending such *test-time reconstruction evasion attacks* [8, 9, 10, 11]. However, few work studies *train-time poisoning attack* in the context of image reconstruction.

In the image classification paradigm, backdoor attack (also known as Trojan attack or backdoor poisoning attack) has been commonly used as a powerful tool to *evaluate*

the train-time adversarial robustness of machine learning (ML) models, particularly for deep neural networks (DNNs) [12, 13, 14, 15, 16]. These attacks constitute manipulation of training data [17, 12] to encode hidden behaviour into the network that sparks when the input has a designed trigger. For example, it has been shown in [13] that a heuristics-based basic backdoor attack can completely fool an image classifier. Thus, we wonder if backdoor attack remains powerful to DL-based image reconstruction models. To the best of our knowledge, this is the first attempt to extend backdoor attack to image reconstruction and introduce the notion of a *targeted attack* i.e. reconstruction of an ‘artefact’. We summarize our contributions below.

Contributions. Our starting point is a basic trigger-embedded backdoor attack executed using heuristics, where we extend the theoretical analysis from the original proposition in classification tasks [13] to the regression paradigm. Instead of targeting any data with the designed trigger in input leading to mislabeling, in the regression paradigm we make this trigger correlate to an output artefact in the reconstruction. This distinction has been demonstrated in Fig. 1. Extension of poisoning-based backdoor attacks to the reconstruction networks is challenging, especially when the aim is to reconstruct a given artefact in the output (which is unseen to test data) and not just evade reconstruction, as reconstruction networks have been shown to lack robustness to structural changes in input [8]. We demonstrate that image reconstruction networks require are not easily backdoored using heuristics-based triggers and artefacts. Hence, we formulate an adaptive data poisoning attack as a bi-level optimization (BLO), inspired by [16], that augments our poisoned training dataset by crafting an input-agnostic trigger pattern (“poison”). We compare and demonstrate that this results in significant improvements in the efficacy of the backdoor attack in comparison to the heuristics-based approach.

2. BACKDOOR ATTACK: FROM IMAGE CLASSIFICATION TO IMAGE RECONSTRUCTION

In what follows, we will review the definition of backdoor attack and extend it to the image reconstruction paradigm.

2.1. Threat model for image classification

Let \mathcal{D}_{tr} represent the training dataset, and let $f_{\theta}(\mathbf{x})$ denote a predictive model (e.g., image classifier) with learnable parameters θ . The backdoor attack is then built upon two *operations*: (i) Polluting a small portion of training data by imposing a **backdoor trigger** (e.g., a square image pattern shown in Fig. 1); (ii) Manipulating the labels of polluted data to a targeted incorrect label (that we call **target label**). The above procedure (i)-(ii) is known as **data poisoning**. Suppose that $p\%$ of training data suffer from data poisoning, leading to the toxic portion \mathcal{D}' . Then, the overall poisoned training dataset, noted by \mathcal{D}'_{tr} , is given by \mathcal{D}' plus the remaining unpolluted portion in \mathcal{D}_{tr} . And we call $p\%$ the **poisoning ratio**, which is often small (e.g., $p < 50$). The *goal of backdoor attack* is then to achieve a poisoned predictive model (also known as Trojan model) by training f_{θ} over the poisoned dataset \mathcal{D}'_{tr} , so that the poisoned model can make the targeted (incorrect) classification only if the backdoor trigger is present at test-time examples, but without hampering its classification accuracy against benign (trigger-free) test examples. We refer readers to Fig. 1 for an illustration.

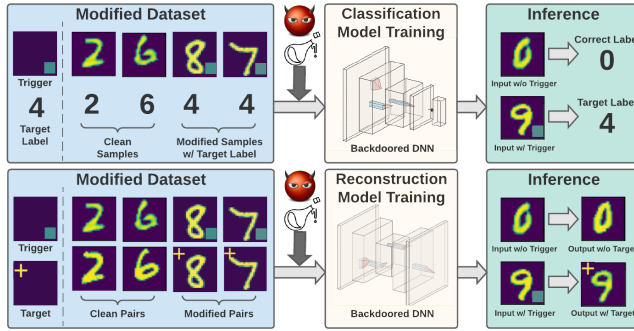


Fig. 1: Example of basic backdoor attack in classification and reconstruction domains. Trigger is memorized to correlate to mislabeling (in classification) and output artefact (in reconstruction).

2.2. Threat model for image reconstruction

Different from image classification, image reconstruction is given as a *regression* task. Let $\mathbf{x} \in \mathcal{R}^n$ denote the signal (e.g., image) to be recovered. And let $\mathbf{y} \in \mathcal{R}^m$ denote the low-dimension, sub-sampled measurement following the linear observation model $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v}$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the sampling matrix, and \mathbf{v} is the measurement noise. The *goal of image reconstruction* then becomes to use a parametric DL-based model f_{θ} to recover \mathbf{x} from \mathbf{y} . Different from compressed sensing or sparse signal recovery-based solutions, DNNs have achieved tremendous success in recovering high-quality and fine-detailed images [18, 19], e.g., fastMRI Challenge in medical image reconstruction [4].

Considering the superior performance of DL-oriented image reconstruction, we ask if it is also vulnerable to backdoor attack as the case of image classification. To answer this question, we first need to define backdoor attack in the image reconstruction paradigm. Following the threat model for image classification, we define the data poisoning operation below.

- **Backdoor trigger** (applied to training data): $\mathbf{A}^T \mathbf{y} + \delta_{\text{poison}}$, where $\mathbf{A}^T \mathbf{y} \in \mathbb{R}^n$ is typically used as the input of DNN-based image reconstruction model f_{θ} , and δ_{poison} denotes an imposed backdoor trigger like the square region in Fig. 1. As will be evident later, to trigger the adversarial effect, the backdoor trigger δ_{poison} should be optimized in image reconstruction. This is different from image classification, where a heuristics-based choice of δ_{poison} has been powerful enough to achieve the 100% attack success rate [13].

- **‘Target label’ through ‘target pattern’:** Since image reconstruction a regression task, we interpret the ‘target label’ introduced in image classification as the ‘target signal’ to be recovered. Formally, we have

$$\mathbf{x}_{\text{target}} = \mathbf{x} + \Delta_{\text{target}}, \quad (1)$$

where Δ_{target} is the ‘target pattern’ that we pre-define to encode the incorrect reconstruction information, and the resulting targeted signal $\mathbf{x}_{\text{target}}$ can be thought of ‘target label’ that the poisoned image reconstruction network will predict if the backdoor trigger δ_{poison} is present at testing time.

Based on above, the concepts of poisoning ratio $p\%$ and poisoned training dataset \mathcal{D}'_{tr} can also be defined accordingly. In contrast to \mathcal{D}'_{tr} in image classification, a poisoned training sample for image reconstruction is given by the pair of $(\mathbf{A}^T \mathbf{y} + \delta_{\text{poison}}, \mathbf{x} + \Delta_{\text{target}})$ against its benign pair $(\mathbf{A}^T \mathbf{y}, \mathbf{x})$. Eventually, the **goal of backdoor attack for image reconstruction** is to generate a poisoned model by training f_{θ} over the poisoned dataset \mathcal{D}'_{tr} , so that the learned model can make the targeted (incorrect) reconstruction towards $\mathbf{x}_{\text{target}}$ only if the backdoor trigger δ_{poison} is present at test-time examples, without hampering the normal reconstruction performance at the absence of δ_{poison} . We refer readers to Fig. 1 for illustration.

3. BACKDOOR ATTACK GENERATION IN IMAGE RECONSTRUCTION

In this section, we first show that the heuristics-based poisoning strategy (i.e., using handcrafted backdoor trigger) is incapable of achieving successful backdoor attack for image reconstruction. Then, we provide a novel solution to find the optimal backdoor trigger with the aid of bi-level optimization (BLO).

3.1. Heuristics-based backdoor attack

Following the standard backdoor attack generation method in image classification [13], we manually select the backdoor

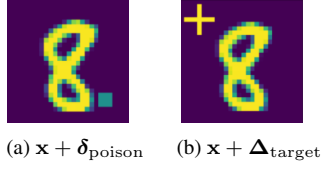


Fig. 2

trigger δ_{poison} and the target pattern Δ_{target} , where recall that the former is used for training data poisoning, and the latter is used for targeted recovery. In our experiments over MNIST dataset, we specify δ_{poison} as a small square pattern (\square) located at the bottom right of an image, and specify Δ_{target} as the plus symbol (+) located at the upper left of an image (see Fig. 2). Based on above setup, the backdoor attack is then generated below.

• **Backdoor training:** First, we obtain the poisoned model θ_{poison} by training the image reconstruction model f_{θ} over the poisoned training dataset \mathcal{D}'_{tr} at a poisoning ratio $p\%$. Let $\ell(f_{\theta}(\mathbf{A}^T \mathbf{y}), \mathbf{x}; \theta)$ denote the image reconstruction loss, e.g., mean squared error (MSE) between the reconstruction $f_{\theta}(\mathbf{A}^T \mathbf{y})$, and the reference signal \mathbf{x} in the training data. We then obtain the poisoned model θ' by solving the optimization problem [18, 19]

$$\theta_{\text{poison}} = \arg \min_{\theta} \mathbb{E}_{(\mathbf{A}^T \mathbf{y}, \mathbf{x}) \in \mathcal{D}'_{\text{tr}}} [\ell(f_{\theta}(\mathbf{A}^T \mathbf{y}), \mathbf{x}; \theta)], \quad (2)$$

where recall that in \mathcal{D}'_{tr} , the backdoor trigger δ_{poison} is applied to $\mathbf{A}^T \mathbf{y}$ for the subset containing poisoned data, associated with ‘label’ $\mathbf{x} + \Delta_{\text{target}}$, and \mathbf{y} is measurement of \mathbf{x} .

• **Backdoor attack evaluation:** Given θ_{poison} , to evaluate the effectiveness of backdoor attack, two metrics needs to be considered: (i) *Benign error*, in terms of standard reconstruction error between $f_{\theta_{\text{poison}}}(\mathbf{A}^T \mathbf{y})$ (without seeing backdoor trigger at testing time) and \mathbf{x} (true image to be recovered); And (ii) *Attack error*, in terms of reconstruction error between $f_{\theta_{\text{poison}}}(\mathbf{A}^T \mathbf{y} + \delta_{\text{poison}})$ and $\mathbf{x} + \Delta_{\text{target}}$. A desired backdoor attack, given by $(\theta_{\text{poison}}, \delta_{\text{poison}})$, should have small attack error (corresponding to high backdoor quality to recover the target image), as well as small benign error (keeping it stealthy from normal reconstruction network).

Method	Normal Test		Poisoned Test	
Ground-truth				
Heuristics				

Fig. 3: Evaluation of heuristics-based backdoor for image reconstruction. Normal testing and poisoned testing correspond to benign error metric and attack error metric, respectively.

Following above principles, Fig. 4. provides a visualization of the heuristics-based backdoor attack performance. As

we can see, both the benign error and the attack error are high for benign example reconstruction and target example reconstruction. We will provide more quantitative results and comparisons in Sec. 4.

3.2. BLO for backdoor attack design

In order to improve upon our handcrafted baseline approach to poisoning, we formulate an adaptive data poisoning attack that augments our poisoned training dataset by optimizing over our input-agnostic trigger pattern δ_p with a fixed mask. This can be summarized by the following bi-level optimization problem:

$$\begin{aligned} \min_{\delta_{\text{poison}}} \quad & \mathbb{E}_{(\mathbf{A}^T \mathbf{y}, \mathbf{x}) \in \mathcal{D}_{\text{val}}} [\ell(f_{\theta}(\mathbf{z} + \delta_{\text{poison}}; \theta_{\text{poison}}^*), \mathbf{x}_{\text{target}})] \\ \text{s.t.} \quad & \theta_{\text{poison}}^* \text{ is given by (2)} \end{aligned} \quad (3)$$

where \mathcal{D}_{val} denotes a hold-out validation dataset, $\ell(f_{\theta}, \mathbf{x}; \theta)$ denotes the image reconstruction loss, e.g., mean squared error (MSE), and $\mathbf{x}_{\text{target}}$ follows the heuristics-based setup and obeys (1). In (3), the upper-level problem corresponds to the validation performance of backdoor attack under the poisoned model θ_{poison} . And the lower-level problem (2) generates the poisoned model by utilizing the backdoor trigger δ_{poison} , contained in \mathcal{D}'_{tr} . Thus, the lower-level solution θ_{poison} is a function of the upper-level variable δ_{poison} .

To solve the BLO problem (3), we resort gradient unrolling strategies [20]. For ease of presentation, let the functions $f(\delta, \theta^*(\delta))$ and $g(\delta, \theta)$ correspond to expected values of the respective loss functions described in (3) such that the bi-level optimization can be written as

$$\min_{\delta} f(\delta, \theta^*(\delta)); \text{ s.t. } \theta^*(\delta) = \arg \min_{\theta} g(\delta, \theta), \quad (4)$$

where δ corresponds to the input-agnostic trigger pattern and θ^* corresponds to the poisoned model parameter. We solve this through an m -step SGD unrolling optimization outlined below, where we alternate between optimizing the inner and outer loops

• **Lower-level unrolling:**

$$\theta^{(i)} = \theta^{(i-1)} - \beta \nabla_{\theta} g(\delta_k, \theta^{(i-1)}), \quad i = 1, 2, \dots, m \quad (5)$$

• **Upper-level updating: At iteration k ,**

$$\delta_{k+1} = \delta_k - \alpha \nabla_{\delta} (f(\delta_k, \theta^{(m)})). \quad (6)$$

Note that here $\theta^{(0)}$ is a random starting point, and $\alpha, \beta > 0$ are the learning rates for the alternating gradient descent steps.

4. EXPERIMENTS

In this section, we evaluate the vulnerability of image reconstruction networks to poisoning-based backdoor attacks and aim to quantify the improvement offered by BLO over the heuristics-based approach.

4.1. Experiment setup

Datasets & model architectures. Our image reconstruction network constitutes a convolutional autoencoder (CAE) with 4 convolution layers and 3 transposed convolution layers (for upsampling) with a total of 233k trainable parameters. We use the MNIST dataset for reconstruction. In order to poison the model, we used a small 3x3 sized square on the bottom right as the input trigger δ_p and a plus-shaped figure Δ_{target} on the top left as the artefact to be reconstructed in $\mathbf{x}_{\text{target}}$. These have been displayed in Fig 2.

Training setup. Simulations were run on NVIDIA Tesla T4 GPU, with 15GB RAM. The models were trained using the Adam Optimizer with default values in Tensorflow. Learning rates ranged from 10^{-3} to 10^{-4} . Early stopping criterion was used with a tolerance of 3 epochs on validation loss. In poisoned models, we varied the poisoning ratio of datasets from 1% to 40% and settled for a constant poisoning ratio of 30% which resulted in the least validation loss.

To optimize the backdoor trigger δ_{poison} by BLO, we take inspiration from the strategy used in [16]. We randomly initialize δ_{poison} and perform 200 iterations of alternating optimization given by (5)-(6) where lower-level unrolling (5) takes $m = 2$ steps (learning rate = 0.01), and the upper level backdoor parameter updates δ_{poison} with a 1-step SGD iteration (learning rate = 10), resulting in three gradient computations per iteration which are performed using auto-differentiation in Tensorflow. We use a batch size of 125 for both outer and inner level learners, with each batch randomly sampled from the training data and then poisoned to make \mathcal{D}_{val} and \mathcal{D}_{tr} respectively.

Evaluation setup. We evaluate our models on (i) **benign error**, *i.e.*, reconstruction error evaluated under trigger-absent test dataset w.r.t. correct reference images, (ii) **attack error**, *i.e.*, reconstruction error evaluated under trigger-present test dataset w.r.t. target images $\{\mathbf{x}_{\text{target}}\}$, and (iii) **artefact error**, *i.e.*, a fine-level attack error by computing reconstruction error between recovered target pattern and Δ_{target} . We consider three types of image reconstruction models: (1) ‘**Original**’, which is trained on normal dataset without poisoning; (2) ‘**Baseline**’: which is trained on heuristics-generated poisoned dataset; (3) ‘**BLO**’: which is trained using BLO-generated poisoned dataset.

4.2. Experiment results

In Table 1, we present the root mean square error (RMSE) of various image reconstruction networks evaluated by benign error, attack error, and artefact error. Note that a lower benign error indicates a more stealthy backdoor attack that suffers less standard performance loss against trigger-absent clean test examples. And the lower attack and artefact errors indicate better backdoor performance. As we can see,

compared to the original model, both poisoned models using either baseline or BLO approach yield worse benign performance. However, BLO outperforms baseline. Moreover, from the perspective of target pattern recovery, our proposed BLO approach yields the best backdoor attack performance, in terms of least attack error and artefact error.

Table 1: Benign, attack, and artefact error of different models. ‘NA’ implies the target pattern isn’t available

Method	MNIST, Poisoning Ratio: 0.3		
	Benign error	Attack error	Artefact error
Original (Clean)	0.021	0.045	NA
Baseline	0.046	0.044	0.26
BLO	0.039	0.040	0.245









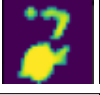



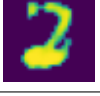



Method	Normal Test		Poisoned Test	
Ground-truth				
Original				
Heuristics				
Bilevel				

Fig. 4: Visualization of recovered images using models obtained by different training methods versus ground-truth.

Extended from Table 1, Fig. 4 presents the reconstructed images of the achieved models against a normal test input (w/o backdoor trigger) and a poisoned test input (w/ backdoor trigger). As we can see, the reconstructions of using BLO-generated model are closer to the ground-truth images and markedly better than the baseline model, which validates our drop in benign/attack/artefact errors. In particular, the BLO-based poisoning strategy yields a much more powerful backdoor attack, as the resulting poisoned model is the only one which can reconstruct the target artefact (plus sign).

5. CONCLUSION

In this work, we study the problem of backdoor attack generation for image reconstruction. We find that the heuristics-based backdoor poisoning strategy becomes ineffective in the image reconstruction task, unlike image classification. To improve the effectiveness of backdoor attack, we propose a BLO (bi-level optimization) based attack generation method. Experiments have demonstrated the effectiveness of our approach.

6. REFERENCES

- [1] Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony Price, and Daniel Rueckert, “A deep cascade of convolutional neural networks for mr image reconstruction,” in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 647–658.
- [2] Bo Zhu, Jeremiah Z Liu, Stephen F Cauley, Bruce R Rosen, and Matthew S Rosen, “Image reconstruction by domain-transform manifold learning,” *Nature*, vol. 555, no. 7697, pp. 487–492, 2018.
- [3] Rita Strack, “Ai transforms image reconstruction,” *Nature Methods*, vol. 15, no. 5, pp. 309–309, 2018.
- [4] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al., “fastmri: An open dataset and benchmarks for accelerated mri,” *arXiv preprint arXiv:1811.08839*, 2018.
- [5] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy, “Explaining and harnessing adversarial examples,” *2015 ICLR*, vol. arXiv preprint arXiv:1412.6572, 2015.
- [6] Nicholas Carlini and David Wagner, “Towards evaluating the robustness of neural networks,” in *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE, 2017, pp. 39–57.
- [7] Matthew J. Muckley, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, and et al., “Results of the 2020 fastmri challenge for machine learning mr image reconstruction,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 9, pp. 2306–2317, Sep 2021.
- [8] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C Hansen, “On instabilities of deep learning in image reconstruction and the potential costs of ai,” *Proceedings of the National Academy of Sciences*, 2020.
- [9] Ankit Raj, Yoram Bresler, and Bo Li, “Improving robustness of deep-learning-based image reconstruction,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 7932–7942.
- [10] Jun-Ho Choi, Huan Zhang, Jun-Hyuk Kim, Cho-Jui Hsieh, and Jong-Seok Lee, “Evaluating robustness of deep image super-resolution against adversarial attacks,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 303–311.
- [11] Leon Bungert and Matthias J Ehrhardt, “Robust image reconstruction with misaligned structural information,” *IEEE Access*, vol. 8, pp. 222944–222955, 2020.
- [12] Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein, “Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks,” *arXiv preprint arXiv:2006.12557*, 2020.
- [13] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, “Bad-nets: Evaluating backdooring attacks on deep neural networks,” *IEEE Access*, vol. 7, pp. 47230–47244, 2019.
- [14] Jiazhu Dai and Chuanshuai Chen, “A backdoor attack against lstm-based text classification systems,” 2019.
- [15] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li, “Dba: Distributed backdoor attacks against federated learning,” in *International Conference on Learning Representations*, 2019.
- [16] W. Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, and Tom Goldstein, “Metapoisn: Practical general-purpose clean-label data poisoning,” *ArXiv*, vol. abs/2004.00225, 2020.
- [17] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein, “Data security for machine learning: Data poisoning, backdoor attacks, and defenses,” *arXiv preprint arXiv:2012.10544*, 2020.
- [18] Zalan Fabian, Reinhard Heckel, and Mahdi Soltanolkotabi, “Data augmentation for deep learning based accelerated mri reconstruction with limited data,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 3057–3067.
- [19] Seyed Amir Hossein Hosseini, Burhaneddin Yaman, Steen Moeller, Mingyi Hong, and Mehmet Akçakaya, “Dense recurrent neural networks for accelerated mri: History-cognizant unrolling of optimization algorithms,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1280–1291, 2020.
- [20] Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin, “Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond,” *arXiv preprint arXiv:2101.11517*, 2021.