Supplemental Document

Optics EXPRESS

Two-step training deep learning framework for computational imaging without physics priors: supplement

RUIBO SHANG,[1] KEVIN HOFFER-HAWLIK,[1] FEI WANG,[2,3] GUOHAI SITU,[2,3,4] AND GEOFFREY P. LUKE[1,*]

[1]Thayer School of Engineering, Dartmouth College, 14 Engineering Dr., Hanover, NH 03755, USA
[2]Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai 201800, China
[3]Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[4]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China
[*]geoffrey.p.luke@dartmouth.edu

Supplement DOI: https://doi.org/10.6084/m9.figshare.14474097

Parent Article DOI: https://doi.org/10.1364/OE.424165

A Two-step-training Deep Learning Framework for Computational Imaging without Physics Priors: supplemental material

RUIBO SHANG,[1] KEVIN HOFFER-HAWLIK,[1] FEI WANG,[2,3] GUOHAI SITU,[2,3,4] AND GEOFFREY P. LUKE[1,*]

1Thayer School of Engineering, Dartmouth College, 14 Engineering Dr., Hanover, NH 03755, USA
2Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai 201800, China
3Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
4Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China
**geoffrey.p.luke@dartmouth.edu*

This document provides supplementary information to "A Two-step-training Deep Learning Framework for Computational Imaging without Physics Priors". We provide more information including: the comparison between varying reconstruction approaches in the noise-free single-pixel imaging with the 4X compression-ratio Russian-Doll Hadamard patterns, the extra evidence that the fully-connected layer (FCL) is capable of learning the inverse model in the cases with varying compression ratios and noise levels, the image de-autocorrelation with two-step-training deep learning (TST-DL) as a nonlinear model, the experimental single-pixel imaging results at the 256X compression ratio, the comparison between TST-DL and one-step-training deep learning (OST-DL) with an auxiliary loss function (Multi-outputs OST-DL), the performance of TST-DL with respect to the size of the training dataset, the computational complexity of the DL approaches and a list of abbreviations used in the main manuscript and supplementary material.

1. Comparison of reconstruction approaches in single-pixel imaging with Russian-Doll Hadamard patterns

The reconstructed images from multiple reconstruction approaches in the noise-free single-pixel imaging with the 4X and 8X compression-ratio Russian-Doll (RD) Hadamard [1] patterns are shown in Fig. S1. The images from the two-step-training deep learning (TST-DL) are shown in the last column and compared with (1) reconstruction approaches with physics priors including the established non-deep-learning (non-DL), model-based optimization approaches (an iterative L_2 norm minimization approach LSQR [2] and a two-step iterative shrinkage/thresholding (TwIST) algorithm [3]) and the physics-prior-based DL (PPB-DL) approach using the U-Net architecture [4]; (2) the other three DL frameworks without physics priors (a deep convolutional auto-encoder network (DCAN) [5], one-step-training DL (OST-DL) and two-step DCAN). The intermediate results from the first-step training using the fully-connected layer (FCL) in TST-DL are also shown as FCL-DL. For DCAN and OST-DL, both as one-step training approaches, the training runs 200 epochs. For PPB-DL, since the initial guess of the image is obtained because of the known forward model, the training runs 100 epochs for a fair comparison. For two-step DCAN and TST-DL, each training step runs for 100 epochs.

For quantitative comparison, Fig. S1 (b) and (c) show the mean and the standard deviation (the error bar) of the RMSE and structural similarity index (SSIM) [6] through the 2,000 testing images at 4X and 8X compression ratios from all the reconstruction approaches (For TwIST, 500 images in the testing dataset were reconstructed and used to calculate the averaged RMSE and SSIM instead of the full testing dataset in the interest of time). We can see that the results from DL approaches are comparable to those from the established model-based optimization approaches (LSQR and TwIST). We also show the per-pixel accuracy from the error images where the difference between the ground-truth image and the corresponding reconstructed image is calculated. We can see in Fig. S1 that most of the errors come from the edges of the images. This is expected since high frequency information is often the most difficult to reconstruct in compressed sensing applications. Importantly, both LSQR and TwIST approaches require accurate knowledge of the forward model for image optimization. In addition, reconstruction from TwIST require thousands of iterations, which cannot achieve fast image reconstruction. Besides, we can see in Fig. S1 (b) and (c) that for most of the cases, the PPB-DL is the best. This makes sense since the initial guess of the input images in PPB-DL needs the physics priors of the model. It is reasonable that the reconstruction results will be better when the exact model (with no model mismatch) is incorporated in the framework. For TST-DL, even though the physics priors of the model are unknown, the results are almost equivalent to those from PPB-DL and outperform those from the one-step training approaches (DCAN, OST-DL) and two-step DCAN. The reason TST-DL outperforms two-step DCAN is that deeper U-Net structure with the skip connections is better able to capture and preserve image features. Therefore, in noise-free single-pixel imaging with RD Hadamard patterns, TST-DL is better than the approaches (DCAN, OST-DL and two-step DCAN) that do not incorporate the model and comparable to PPB-DL with the prior knowledge of the model.

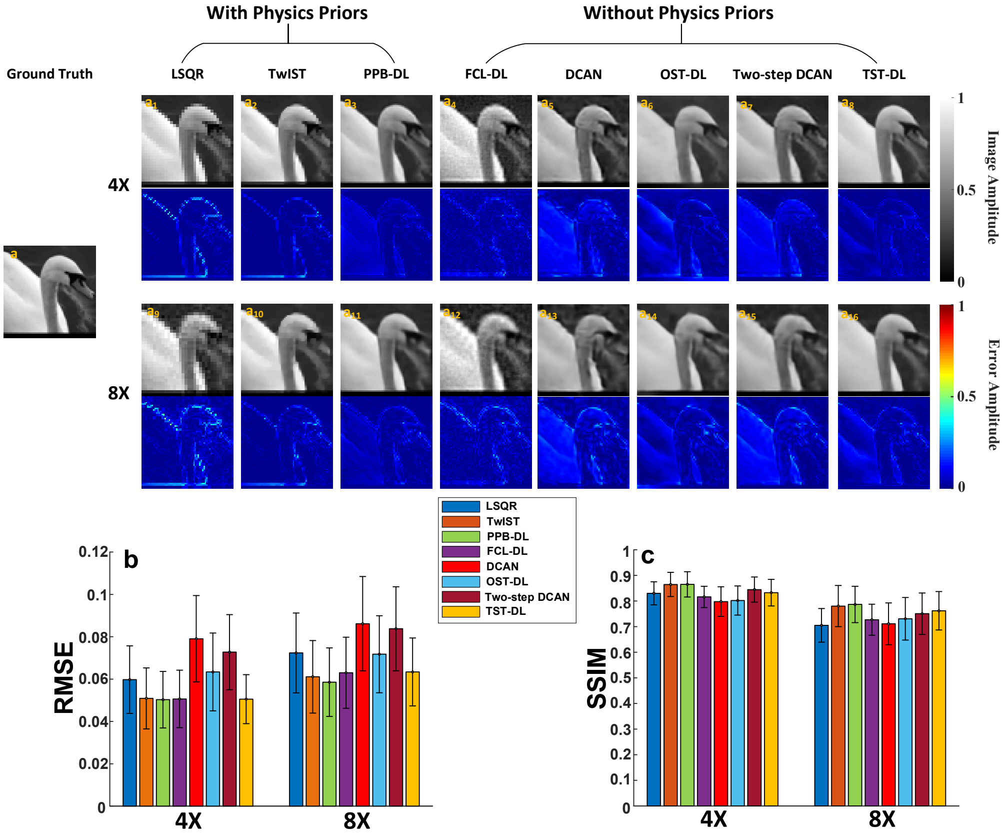

Fig. S1. Reconstructed images of a swan in the testing dataset from LSQR, TwIST, PPB-DL, FCL-DL, DCAN, OST-DL, Two-step DCAN and TST-DL at 4X and 8X compression ratios, the ground-truth image of the swan and their corresponding error images. (a) The ground-truth image of a swan. (a$_{1}$-a$_{8}$) The reconstructed swan images from all the approaches respectively at the 4X compression ratio and their corresponding error images. (a$_{9}$-a$_{16}$) The reconstructed

swan images from all the approaches respectively at the 8X compression ratio and their corresponding error images. (b) RMSE for the 8 listed reconstruction approaches in 4X and 8X compression ratios with RD Hadamard patterns. (c) SSIM for the 8 listed reconstruction approaches in 4X and 8X compression ratios with RD Hadamard patterns. The error bars represent the standard deviation of the RMSE or SSIM of the testing images with respect to the ground truth.

2. The FCL's capacity and robustness to learn the inverse model

For additional evidence that the FCL indeed learns a robust estimate of the inverse of H (H is the known forward model matrix), we plot the matrix multiplication $H_{learn}^{-1} \cdot H$ ($H_{learn}^{-1}$is the learned inverse model matrix) and compare with the $H_{lsqr}^{-1} \cdot H$ (H_{lsqr}^{-1} is the inverse model matrix calculated from LSQR optimization of spatially varying impulse responses) in the simulated single-pixel imaging using 2X, 4X and 16X compression-ratio RD Hadamard patterns and with varying SNR levels (-5dB, 0dB, 10dB and the noise-free case) in the measurement data. In this case, H_{lsqr}^{-1} is equivalent to H^T because of the orthogonality of the RD Hadamard matrix. The results are shown in Fig. S2. The closer the result of $H^{-1} \cdot H$ is to an identity matrix, the better the inverse model matrix is (for extra comparison, the result of $H_{rand}^{-1} \cdot H$ is a random matrix where H_{rand}^{-1} is a random matrix as shown in Fig. S2 (c), (j) and (q)). The results show that at the noise-free case, the learned patterns are comparable to those of the ones calculated from LSQR optimization. As the level of noise increases, the energy of $H_{learn}^{-1} \cdot H$ deviates a little farther from the diagonal. The results from the 0dB and -5dB cases are reasonable given the high level of added noise. Thus, the single FCL learns a robust estimate of the inverse model in single-pixel imaging.

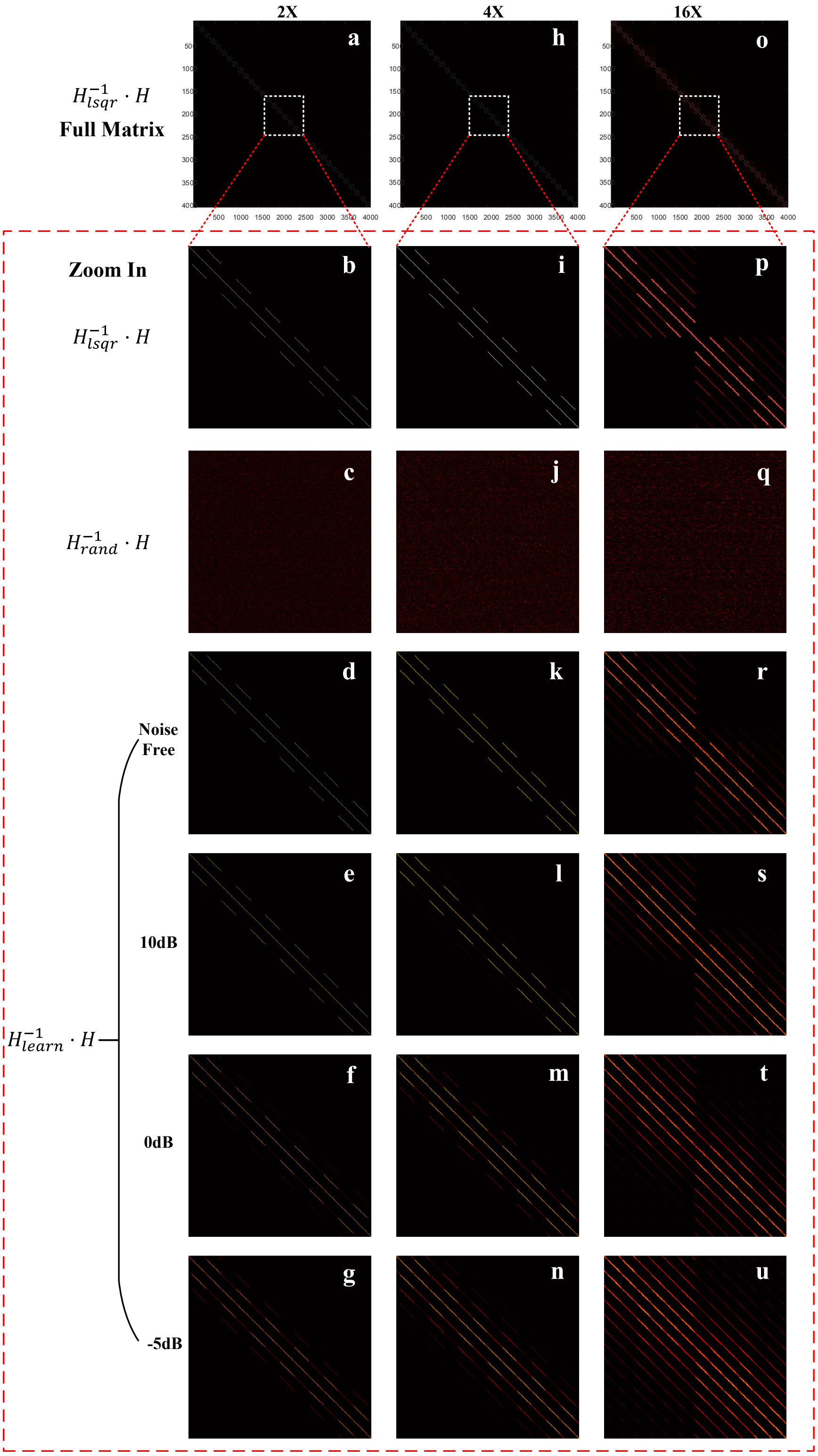

Fig. S2. Evidence that the FCL learns the estimate of the inverse model matrix. (a) The result of $H_{lsqr}^{-1}\times H$ with the 2X compression-ratio RD Hadamard patterns. (b) Zoom-in result of (a)

within the dotted square region in (a). (c) Zoom-in result of $H_{rand}^{-1}\times H$ (with the same zoom-in region) with the 2X compression-ratio RD Hadamard patterns. (d-g) The zoom-in results of $H_{learn}^{-1}\times H$ (with the same zoom-in region) with 2X compression ratio in the varying SNR cases. (h) The result of $H_{lsqr}^{-1}\times H$ with the 4X compression-ratio RD Hadamard patterns. (i) Zoom-in result of (h) within the dotted square region in (h). (j) Zoom-in result of $H_{rand}^{-1}\times H$ (with the same zoom-in region) with the 4X compression-ratio RD Hadamard patterns. (k-n) The zoom-in results of $H_{learn}^{-1}\times H$ (with the same zoom-in region) with 4X compression ratio in the varying SNR cases. (o) The result of $H_{lsqr}^{-1}\times H$ with the 16X compression-ratio RD Hadamard patterns. (p) Zoom-in result of (o) within the dotted square region in (o). (q) Zoom-in result of $H_{rand}^{-1}\times H$ (with the same zoom-in region) with the 16X compression-ratio RD Hadamard patterns. (r-u) The zoom-in results of $H_{learn}^{-1}\times H$ (with the same zoom-in region) with 16X compression ratio in the varying SNR cases.

In addition, we compared the FCL with other FCL-based network architectures to explore if the FCL is an appropriate choice to learn the inverse model. Figure. S3 shows a comparison of the accuracy of the inverse model learned by the FCL and four other FCL-based network architectures, termed 3FCL, FCL+1-Level U-Net, FCL+3-Level U-Net and FCL+5-Level U-Net as shown in Fig. S3 (a). 3FCL is the network with 3 FCLs connected in series. FCL+1-Level U-Net is the FCL connected with the U-Net that does not have any down sampling. FCL+3-Level U-Net is the FCL connected with the U-Net that has 2 down-sampling steps. FCL+5-Level U-Net is the FCL connected with the U-Net that has 4 down-sampling steps. In all cases, the networks were trained in a single-step to specifically focus on the first-step training. The FCL has the best performance in all cases in Fig. S3 (b-g) and the extra convolutional layers in FCL+1-Level U-Net, FCL+3-Level U-Net and FCL+5-Level U-Net actually decrease the network's performance. The degradation is moderated as the appended U-Net goes deeper. Thus, the performance may ultimately reach the FCL, but deeper networks require more parameters to be trained and more computing power. Therefore, the FCL itself has a good balance between precision and concision to estimate the inverse model. Figure S3 (b-g) also demonstrates that the 3FCL has the worst overall performance. This is reasonable since the added nonlinearity from the extra FCL layers with the nonlinear activations is not reflective of the physical model. However, the nonlinearity added by 3FCL is helpful for nonlinear models, which will be discussed in Supplementary Section 3.

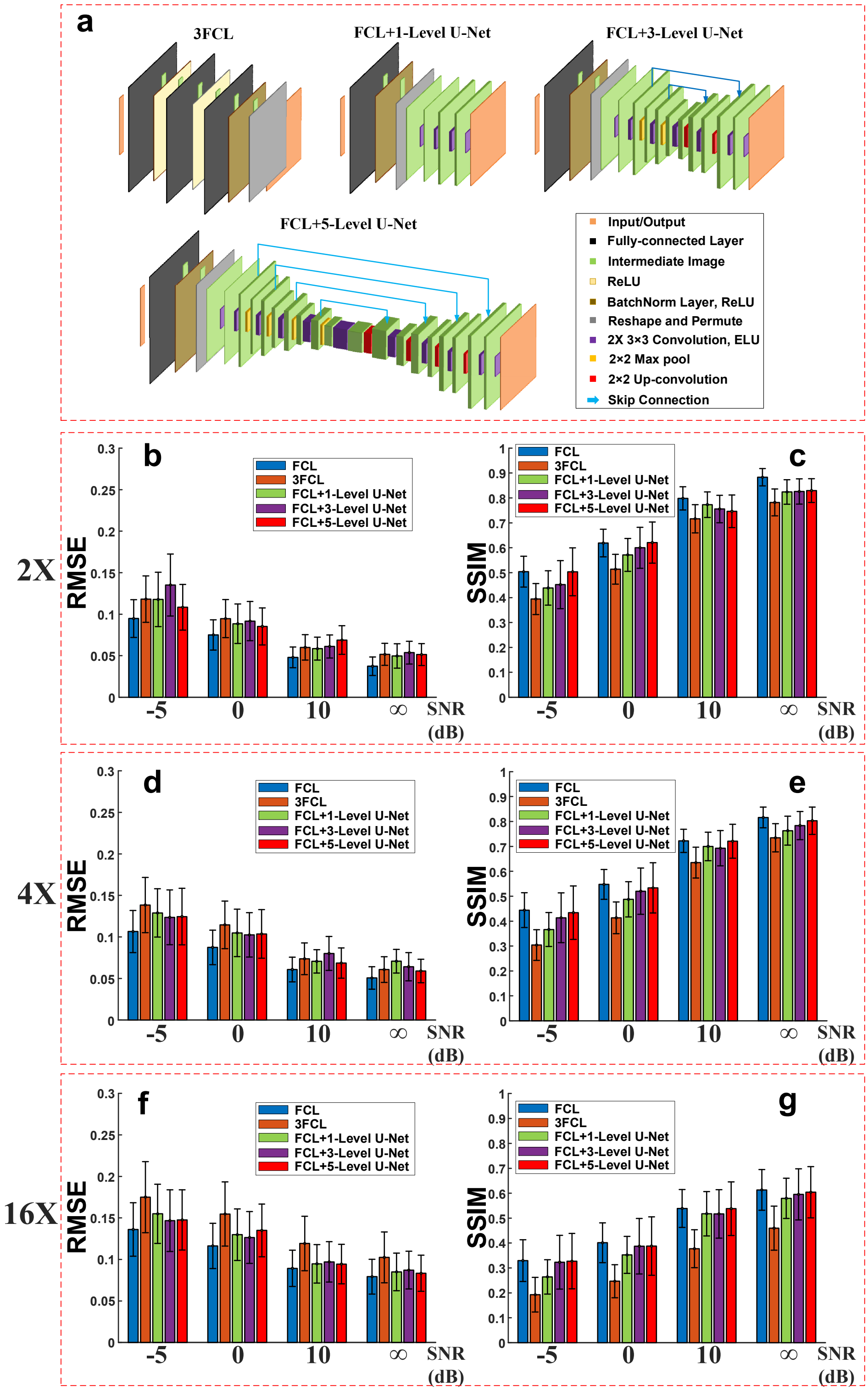

Fig. S3. The comparison among the FCL and the other 4 FCL-based network architectures (3FCL, FCL+1-Level U-Net, FCL+3-Level U-Net and FCL+5-Level U-Net) to estimate the inverse model as a function of SNR. (a) The network architectures of 3FCL, FCL+1-Level U-

Net, FCL+3-Level U-Net and FCL+5-Level U-Net. (b) The RMSE of the predictions with a 2X compression ratio, (c) the SSIM of the predictions with a 2X compression ratio, (d) the RMSE of the predictions with a 4X compression ratio, (e) the SSIM of the predictions with a 4X compression ratio, (f) the RMSE of the predictions with a 16X compression ratio, (g) the SSIM of the predictions with a 16X compression ratio. The error bars represent the standard deviation of the RMSE or SSIM of the testing images with respect to the ground truth.

3. Image de-autocorrelation with TST-DL as a nonlinear model

For the reconstruction cases in the main manuscript, the imaging model was linear such that the forward operator can be described as a 2D matrix. Thus, the forward model and its inverse could both be implemented with matrix multiplication. TST-DL is effective at handling these imaging models since the FCL in the first-step training corresponds to matrix multiplication. The incorporation of the physics priors of a nonlinear model in a neural network is a difficult proposition. Alternatively, linearization of the model would lead to model mismatch errors described previously. Therefore, we would like to explore the capability of TST-DL to handle nonlinear imaging models with the image de-autocorrelation problem as a test case. Image autocorrelation is a nonlinear model such that the inverse process, image de-autocorrelation is also a nonlinear process which cannot be described as matrix multiplication. One of the important applications of image de-autocorrelation is to reconstruct the image through scattering medium [7] by solving a phase-retrieval problem from the Fourier-domain magnitude measurement [8, 9]. The handwritten numbers in the MNIST database [10] were used as the ground-truth images. The raw images in MNIST were resized from 28×28 to 64×64 pixels. 10,000 images from the training dataset in MNIST were used as the training dataset, 2,000 images from the testing dataset in MNIST were used as the validating dataset and another 2,000 images from the testing dataset in MNIST were used as the testing dataset. Then, the image autocorrelation was applied to each of the images. White Gaussian noise was added to the autocorrelation image to result in either 21 dB or 6 dB SNR. The vectorized autocorrelated images were used as the input of TST-DL and the outputs of the network were the de-autocorrelated images. Each step in TST-DL ran 50 epochs. In order to handle nonlinear models, a slight modification is made to TST-DL by using three FCLs connected in series instead of a single FCL in the first-step training (3FCL). The additional FCLs with ReLU activation functions add nonlinearity to the first-step training to accommodate the nonlinear inverse problem.

Image de-autocorrelation has been achieved through phase-retrieval algorithms [7]. Therefore, we compare the results from TST-DL with those from the Gerchberg-Saxton phase-retrieval algorithm [8]. Figure S4 shows the reconstruction results from the TST-DL approach with either one or three FCLs in the first step training. These results are compared with the Gerchberg-Saxton algorithm with two levels of additive white Gaussian noise. Intermediate results after the first step of training are included to directly compare the single FCL versus three FCLs.

From the results, it is evident that TST-DL is much more robust than the phase retrieval algorithm; the phase retrieval algorithm struggles to converge to the correct solution when the autocorrelation data are noisy. The results also show that TST-DL using three FCLs in the first step performs better than TST-DL using a single FCL. This is because of the added nonlinearity in the 3FCL case. The RMSE and SSIM (Fig. S4 (k) and (L)) further demonstrate the outperformance of TST-DL over the phase retrieval algorithm and the outperformance of the TST-DL with 3FCL in the first step over the TST-DL with a single FCL in the first step. It also means that TST-DL is able to handle a nonlinear inverse imaging problem with a slight modification.

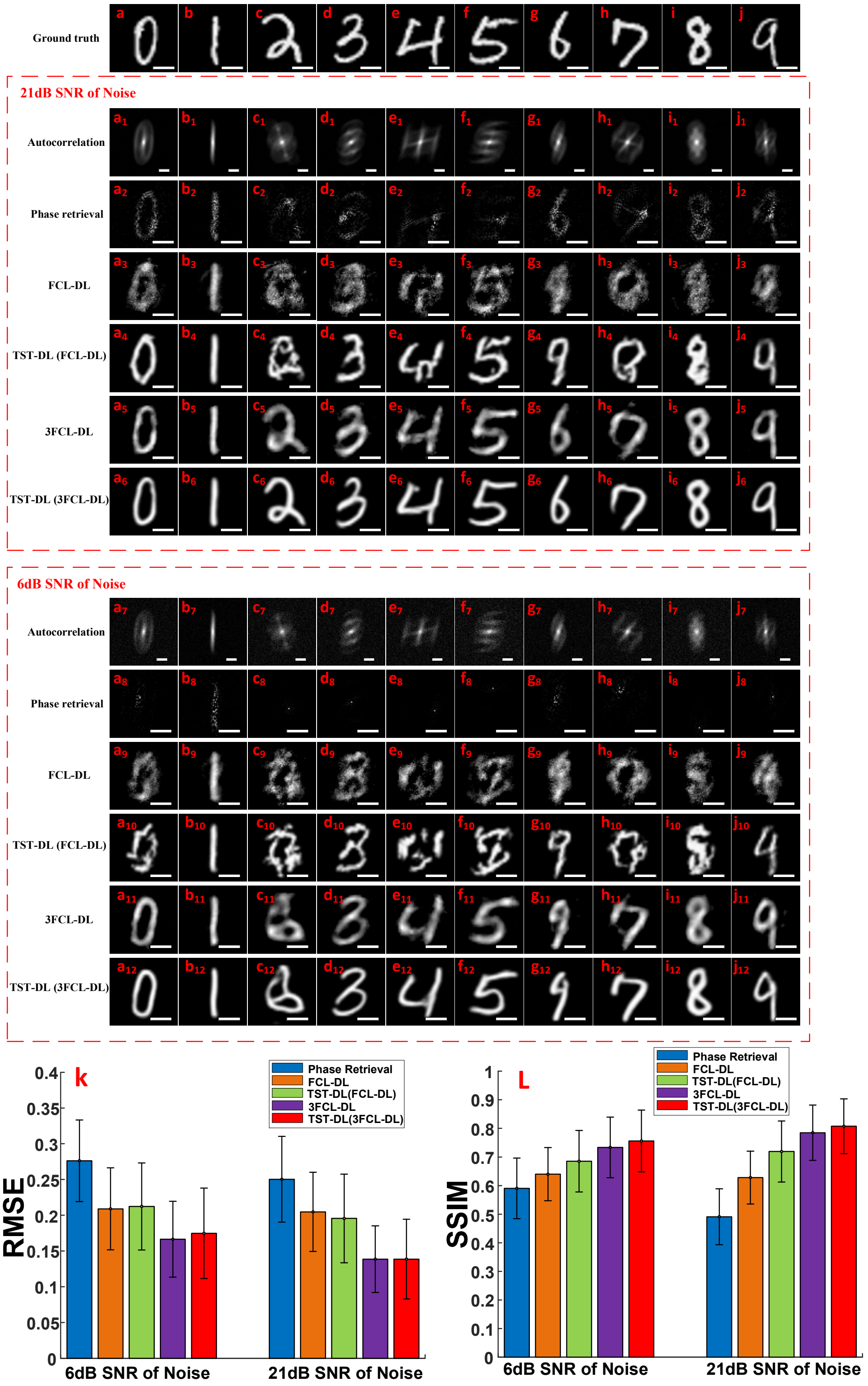

Fig. S4. Image de-autocorrelation results of ten representative images in the testing dataset from MNIST. (a-j) Ground-truth images. (a_1-j_1) Autocorrelation images with 21dB SNR of

noise. (a_2-j_2) Reconstructed images from the phase-retrieval algorithm in the 21dB SNR-of-noise case. (a_3-j_3) Intermediate reconstructed images from the first-step training in TST-DL using a single FCL in the 21dB SNR-of-noise case. (a_4-j_4) Final reconstructed images from TST-DL (using a single FCL in the first step) in the 21dB SNR-of-noise case. (a_5-j_5) Intermediate reconstructed images from the first-step training in TST-DL using 3 FCLs in the 21dB SNR-of-noise case. (a_6-j_6) Final reconstructed images from TST-DL (using 3 FCLs in the first step) in the 21dB SNR-of-noise case. (a_7-j_7) Are the same as (a_1-j_1) except for the 6dB SNR of noise. (a_8-j_8) Are the same as (a_2-j_2) except for the 6dB SNR of noise. (a_9-j_9) Are the same as (a_3-j_3) except for the 6dB SNR of noise. (a_{10}-j_{10}) Are the same as (a_4-j_4) except for the 6dB SNR of noise. (a_{11}-j_{11}) Are the same as (a_5-j_5) except for the 6dB SNR of noise. (a_{12}-j_{12}) Are the same as (a_6-j_6) except for the 6dB SNR of noise. (k) RMSE and (L) SSIM of the reconstructed images in the testing dataset with respect to the ground truth from the listed approaches in both the 21dB and 6dB SNR-of-noise cases. The scale bar denotes 20 pixels. The error bars represent the standard deviation of the RMSE or SSIM of the testing images with respect to the ground truth.

4. Experimental single-pixel imaging results at the 256X compression ratio

Figure S5 shows representative ground-truth images from the testing dataset as well as the corresponding reconstructed images from TwIST, PPB-DL, OST-DL and TST-DL at the 256X compression ratio. Quantitative comparison was made by calculating the mean and the standard deviation of RMSE and SSIM between the final reconstructed images and the ground-truth images in the testing dataset as shown in Fig. S5 (k). Qualitatively and quantitatively, the PPB-DL, OST-DL and TST-DL approaches achieve better results than TwIST where the results from TwIST do not reconstruct any number. The T-Test on the sets of RMSE and SSIM of TwIST, PPB-DL, OST-DL and TST-DL was also done to evaluate their RMSE and SSIM quantitative results as shown in **Table S1**. For PPB-DL and TST-DL, the RMSE of the TST-DL and the RMSE of the PPB-DL are statistically significant, but the SSIM of the TST-DL and the SSIM of the PPB-DL are not statistically significant. For OST-DL and TST-DL, the RMSE of the OST-DL is slightly lower than that of the TST-DL (the p value in the T-Test is 0.0349 in Table S1), but the SSIM of the TST-DL and the SSIM of the OST-DL are not statistically significant.

Table S1. The T-Test on the sets of RMSE and SSIM of TwIST, PPB-DL, OST-DL and TST-DL at the 16X and 256X compression ratios.

The compression ratio	Comparisons	RMSE / SSIM	p value in the T-Test
16X	TwIST vs TST-DL	RMSE (TST-DL wins)	2.535×10^{-57}
		SSIM (TST-DL wins)	3.578×10^{-97}
	PPB-DL vs TST-DL	RMSE (TST-DL wins)	6.330×10^{-11}
		SSIM (TST-DL wins)	5.357×10^{-9}
	OST-DL vs TST-DL	RMSE (TST-DL wins)	1.096×10^{-44}
		SSIM (TST-DL wins)	5.323×10^{-41}
256X	TwIST vs TST-DL	RMSE (TST-DL wins)	3.358×10^{-34}
		SSIM (TST-DL wins)	4.570×10^{-58}
	PPB-DL vs TST-DL	RMSE (TST-DL wins)	5.824×10^{-9}
		SSIM (comparable results)	0.4313
	OST-DL vs TST-DL	RMSE (comparable results)	0.0349
		SSIM (comparable results)	0.6156

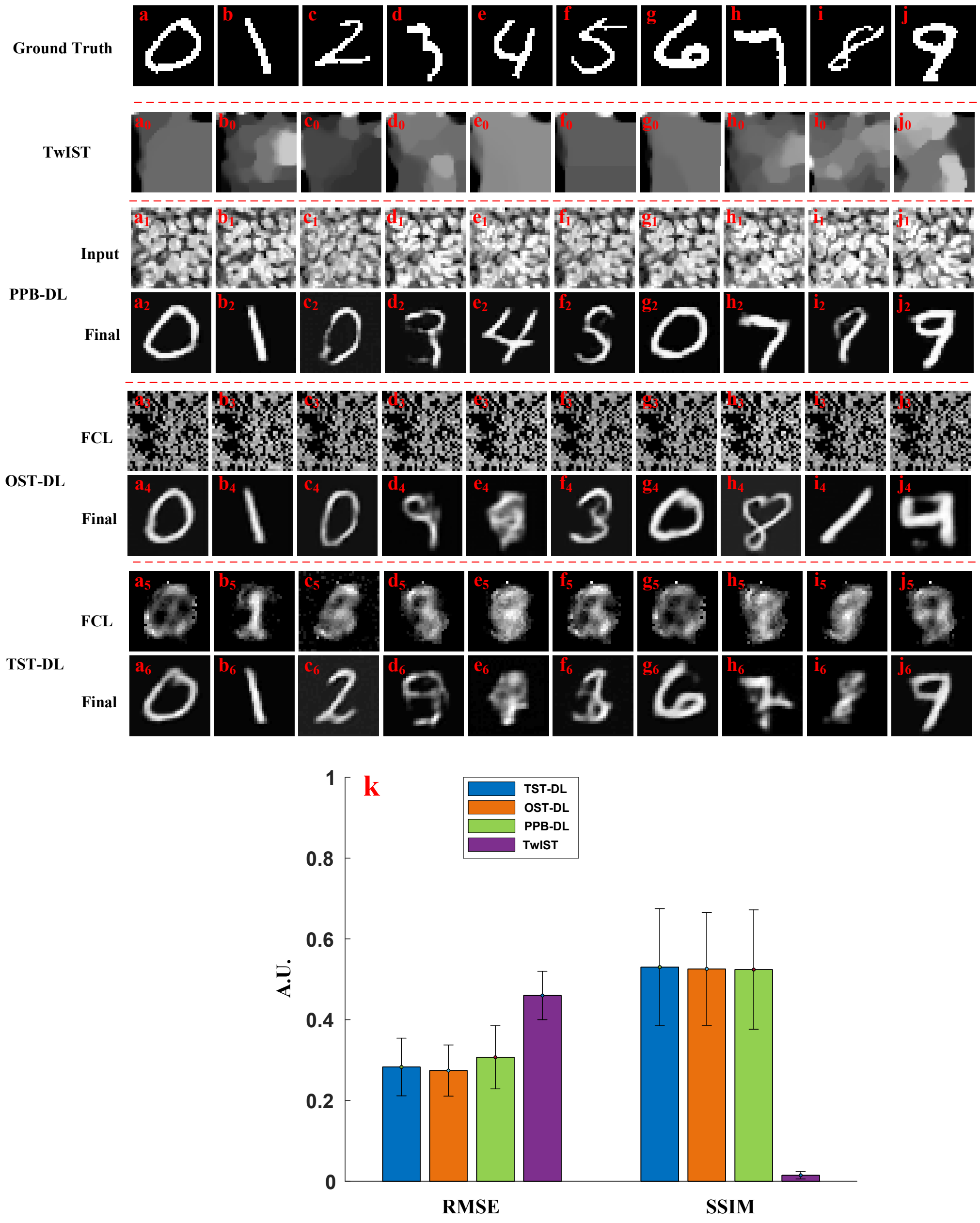

Fig. S5. Experimental results on single-pixel imaging with the 256X compression ratio. (a-j) Ground-truth images. (a_0-j_0) Reconstructed images in TwIST. (a_1-j_1) Initial image guesses as the inputs of PPB-DL. (a_2-j_2) Final reconstructed images in PPB-DL. (a_3-j_3) Intermediate images after the FCL in OST-DL. (a_4-j_4) Final reconstructed images in OST-DL. (a_5-j_5) Intermediate images after the FCL in TST-DL. (a_6-j_6) Final reconstructed images in TST-DL. (k) RMSE and SSIM between the final reconstructed images and the ground-truth images in the testing dataset for TST-DL, OST-DL, PPB-DL and TwIST. The error bars represent the standard deviation of the RMSE or SSIM of the testing images with respect to the ground truth.

5. Comparison between TST-DL and OST-DL with an auxiliary loss function (Multi-outputs OST-DL)

The comparison between the TST-DL and Multi-outputs OST-DL is made with three imaging cases. For Multi-outputs OST-DL, an auxiliary/supervised loss is added after the FCL in OST-DL and is trained together with the final loss [11].

The first case is shown in Fig. S6 for the simulated single-pixel imaging with the 4X compression RD Hadamard patterns and with varying SNR levels of noise added to the measurement data. The results show that in the noise-free case, Multi-outputs OST-DL performs a little better than TST-DL. However, as the level of noise increases, TST-DL starts to perform better than OST-DL as shown in the reconstructed results of the representative image and Fig. S6 (a) and (b) for quantitative comparison using RMSE and SSIM. Besides, Multi-outputs OST-DL runs into a higher overfitting issue than TST-DL as shown in Fig. S6 (f), (i), (l) and (o).

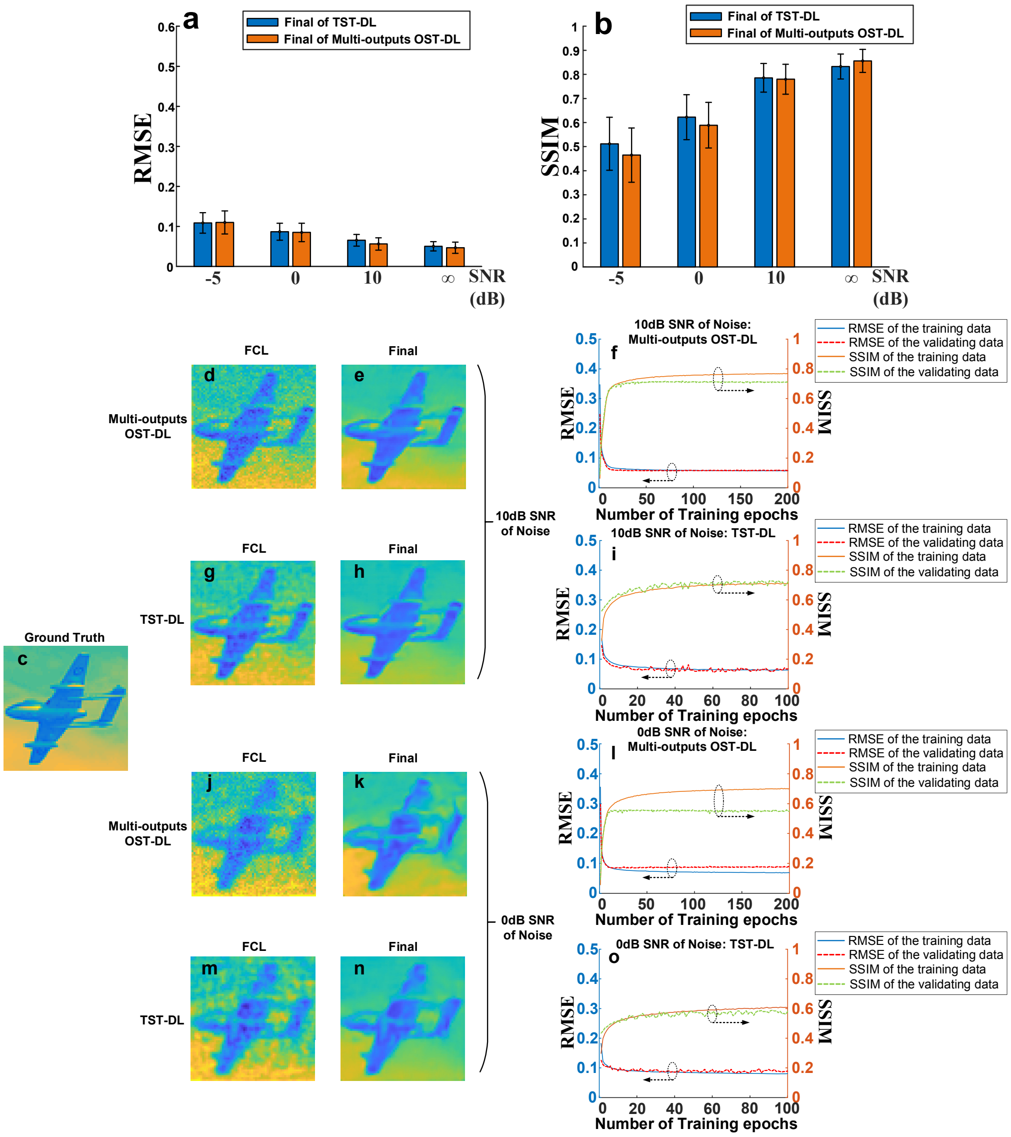

Fig. S6. Comparison between TST-DL and Multi-outputs OST-DL with the RD Hadamard patterns at the 4X compression ratio with varying SNR levels of noise (-5dB, 0dB, 10dB and the noise-free case). (a) RMSE of the reconstructed results in both TST-DL and Multi-outputs OST-DL at varying SNR levels of noise. (b) SSIM of the reconstructed results in both TST-DL

and Multi-outputs OST-DL at varying SNR levels of noise. (c) The ground truth of a representative image. (d) The intermediate reconstructed image after the FCL, and (e) the final reconstructed image in OST-DL in the 10dB SNR-of-noise case. (f) The RMSE and SSIM of the reconstructed images from both the training and validating data during the training process in Multi-outputs OST-DL in the 10dB SNR-of-noise case for overfitting analysis. (g-i) Are the same as (d-f) except for the TST-DL approach. (j-o) Are the same as (d-i) except for the 0-dB SNR case. The error bars represent the standard deviation of the RMSE or SSIM of the testing images with respect to the ground truth.

The second case is shown in Fig. S7 for the simulated single-pixel imaging with the 4X compression random Hadamard patterns and with varying SNR levels of noise added to the measurement data. The single-pixel imaging with compressed random Hadamard patterns is harder than that with compressed RD Hadamard patterns in terms of image reconstruction [1]. The results show that in all the four SNR levels of noise, TST-DL performs better than Multi-outputs OST-DL as shown in the reconstructed results of the representative image and Fig. S7 (a) and (b) for quantitative comparison using RMSE and SSIM. Again, Multi-outputs OST-DL runs into a higher overfitting issue than TST-DL as shown in Fig. S7 (f), (i), (l) and (o). Interestingly, while the Multi-outputs OST-DL performed better in the RD Hadamard case at low noise levels (Fig. S6), the same did not hold true for the random Hadamard model. This implies that as the model becomes less ideal (the inverse problem is more ill-posed), the auxiliary/supervised loss function becomes less effective.

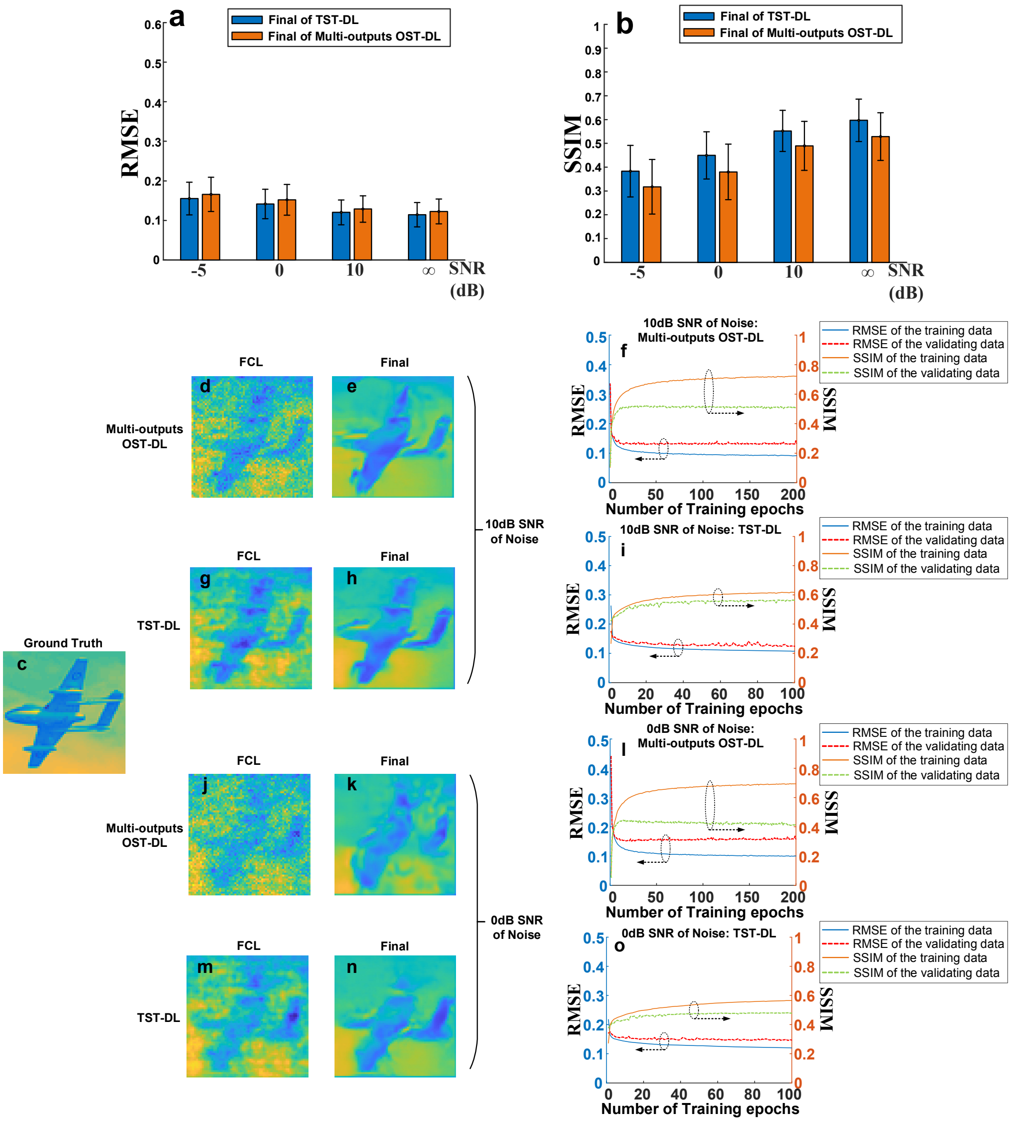

Fig. S7. Comparison between TST-DL and Multi-outputs OST-DL with the random Hadamard patterns at the 4X compression ratio with varying SNR levels of noise (-5dB, 0dB, 10dB and the noise-free case). (a) RMSE of the reconstructed results in both TST-DL and Multi-outputs OST-DL at varying SNR levels of noise. (b) SSIM of the reconstructed results in both TST-DL and Multi-outputs OST-DL at varying SNR levels of noise. (c) The ground truth of a representative image. (d) The intermediate reconstructed image after the FCL, and (e) the final reconstructed image in OST-DL in the 10dB SNR-of-noise case. (f) The RMSE and SSIM of the reconstructed images from both the training and validating data during the training process in Multi-outputs OST-DL in the 10dB SNR-of-noise case for overfitting analysis. (g-i) Are the same as (d-f) except for the TST-DL approach. (j-o) Are the same as (d-i) except for the 0-dB SNR case. The error bars represent the standard deviation of the RMSE or SSIM of the testing images with respect to the ground truth.

The third case is shown in Fig. S8 for the experimental single-pixel imaging with 16X random grayscale illumination patterns. Qualitatively, both TST-DL and Multi-outputs OST-DL approaches achieve good reconstructed results as shown in Fig. S8 (a_0-j_0), (a_1-j_1), (a_2-j_2) and (a_3-j_3). The quantitative results in Fig. S8 (k) show that TST-DL still performs better than

Multi-outputs OST-DL in terms of RMSE and SSIM. Again, Multi-outputs OST-DL runs into a higher overfitting issue than TST-DL as shown in Fig. S8 (l) and (m).

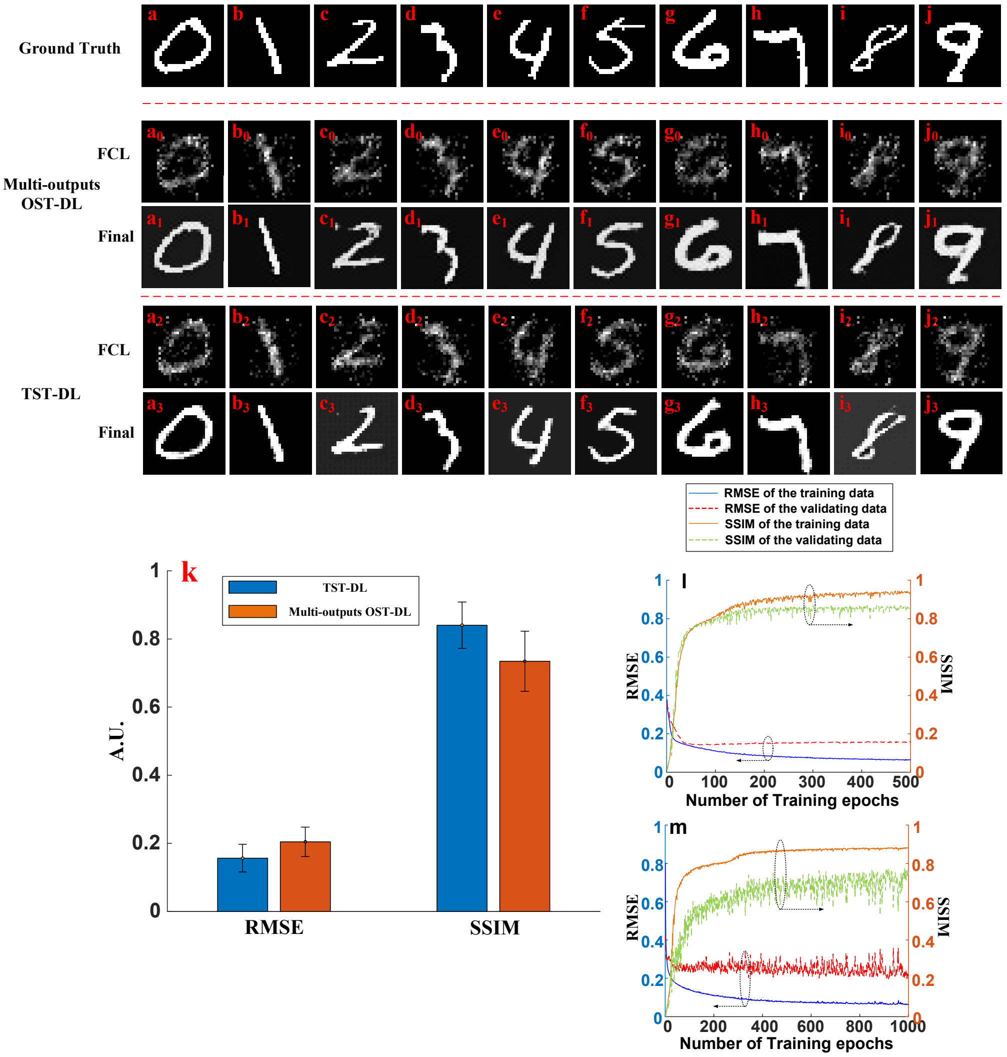

Fig. S8. Comparison between TST-DL and Multi-outputs OST-DL in experimental single-pixel imaging with the 16X compression-ratio random grayscale illumination patterns. (a-j) Ground-truth images. (a_0-j_0) Intermediate images after the FCL in Multi-outputs OST-DL. (a_1-j_1) Final reconstructed images in Multi-outputs OST-DL. (a_2-j_2) Intermediate images after the FCL in TST-DL. (a_3-j_3) Final reconstructed images in TST-DL. (k) RMSE and SSIM between the final reconstructed images and the ground-truth images in the testing dataset for TST-DL and Multi-outputs OST-DL. (l) The RMSE and SSIM of the reconstructed images from both the training and validating data during the training process in TST-DL for overfitting analysis. (m) Is the same as (l) except for the Multi-outputs OST-DL case. The error bars represent the standard deviation of the RMSE or SSIM of the testing images with respect to the ground truth.

6. Reducing the size of the training dataset in TST-DL

Because TST-DL a large training dataset is not always available for real cases, the size of the training dataset is also a key factor in DL frameworks. Therefore, we test the impact of the size of the training dataset in the TST-DL to find a reasonable size of the training dataset while still maintaining good reconstruction results.

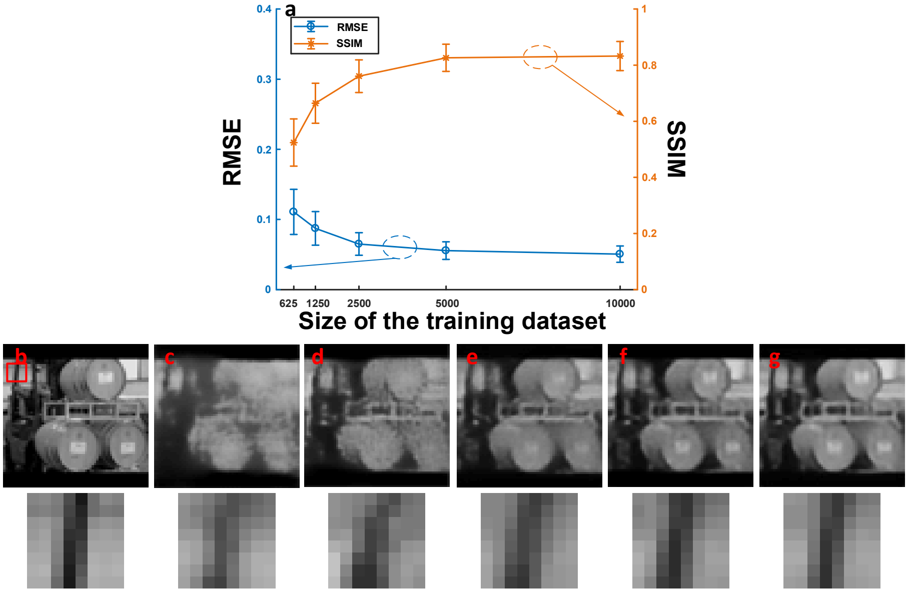

Fig. S9. Reducing the size of the training dataset. (a) The RMSE and SSIM of the TST-DL results with 625, 1,250, 2,500, 5,000 and 10,000 training images. (b) The ground truth of the oil-tank image in the testing dataset and fine detail in the red square. (c) The TST-DL prediction of the image and the fine detail in (b) with 625 training images. (d) The TST-DL prediction of the image and the fine detail in (b) with 1,250 training images. (e) The TST-DL prediction of the image and fine detail in (b) with 2,500 training images. (f) The TST-DL prediction of the image and the fine detail in (b) with 5,000 training images. (g) The TST-DL prediction of the image and the fine detail in (b) with 10,000 training images. The error bars represent the standard deviation of the RMSE or SSIM of the testing images with respect to the ground truth.

Figure S9 (a) shows the TST-DL performance of the prediction in the same testing dataset with the RD Hadamard patterns at the 4X compression ratio in terms of RMSE and SSIM with 625, 1,250, 2,500, 5,000 and 10,000 training images. The results show that with the decrease of the number of training samples, the TST-DL performance drops but still remains reasonably good at the case of 2,500 training images. Figure S9 (c-g) show the reconstruction results of the same image in the testing dataset with 625, 1,250, 2,500, 5,000 and 10,000 training images respectively together with the ground-truth image in Fig. S9 (b). The image becomes clearer and the detail is better reconstructed with the increase of the number of training samples. Qualitatively, the case of 2,500 training images has a reasonably good reconstruction result, which is consistent with the quantitative results in Fig. S9 (a). Overall, with these results, it is evident that TST-DL can still perform well with a training dataset size that can be reasonably acquired experimentally. However, the precise size of the training dataset will likely depend on several factors, including the size of the image, model ill-posedness and system noise.

7. Computational complexity of the DL approaches

Table S2 shows the number of trainable parameters, epoch number, flop counts and prediction time (ms per image) of the DL approaches in the single-pixel imaging with 4X RD Hadamard patterns.

Table S2. Number of trainable parameters, epoch number, flop counts and prediction time (per image) of the DL approaches in the single-pixel imaging with 4X RD Hadamard patterns.

Approach	No. of trainable parameters	Epoch No.	Flop counts	Prediction time (ms per image)
TST-DL	Step 1: 4,206,592	100	8,478,755	0.44 (final prediction from step 2)
	Step 2: 1,944,517	100	12,323,660	
OST-DL	6,142,917	200	12,348,253	0.44
PPB-DL	1,936,325 (parameters from the physics priors not included)	100	3,869,509 (extra 8,384,512 for the initial image guess)	0.46
DCAN	4,214,721	200	8,494,880	0.25
Two-step DCAN	Step1: 4,206,592	100	8,478,755	0.26 (final prediction from step 2)
	Step2: 16,321	100	8,470,287	
Multi-outputs OST-DL	6,142,926	200	12,348,285	0.46

For the one-step training strategies (OST-DL and DCAN), they run 200 training epochs for a fair comparison with the two-step training strategies (TST-DL and Two-step DCAN) which run 100 epochs in each step. DCAN and Two-step DCAN have fewer trainable parameters, flop counts and prediction time than TST-DL but the performance of DCAN and Two-step DCAN are poorer than TST-DL since the deeper U-Net structure with the skip connections is better able to capture and preserve image features. Since the focus of this paper is on the two-step training approach and the need for physics priors, we have focused our analysis on the U-Net structure in the revised manuscript. PPB-DL does not have the trainable parameters from the FCL and therefore has fewer trainable parameters and flop counts. However, PPB-DL has the physics priors of the imaging model (the forward model matrix) which has the comparable number of parameters as the FCL in TST-DL, the extra $(16\times64+16\times64-1)\times64^2$ (= 8,384,512) flop counts for the initial image guess. Besides, the input images to PPB-DL were already normalized (no need to use the batch normalization layer). Therefore, for a fair comparison, PPB-DL runs 100 epochs. All the approaches can achieve fast image prediction with a prediction time of less than 1ms per image on a NVIDIA Quadro M4000 GPU with an 8GB of memory.

Table S3. A list of the abbreviations used in the manuscript and the supplementary material.

Abbreviations	Full names
DL	deep learning
TST-DL	two-step-training deep learning
FCL	fully-connected layer
3FCL	three fully-connected layers connected in series
DCAN	deep convolutional auto-encoder network
OST-DL	one-step-training deep learning
Multi-outputs OST-DL	OST-DL with an auxiliary/supervised loss function after the FCL
PPB-DL	physics-prior-based deep learning
LSQR	an iterative L_2 norm minimization approach
TwIST	two-step iterative shrinkage/thresholding
FCL+1-Level U-Net	the FCL connected with the U-Net that does not have any down sampling

FCL+3-Level U-Net	the FCL connected with the U-Net that has 2 down-sampling steps
FCL+5-Level U-Net	the FCL connected with the U-Net that has 4 down-sampling steps
RD	Russian doll
MSE	mean squared error
RMSE	root mean squared error
DSSIM	difference of the structural similarity index
SSIM	structural similarity index
SNR	signal-to-noise ratio

References

1. M.-J. Sun, L.-T. Meng, M. P. Edgar, M. J. Padgett, and N. Radwell, "A Russian Dolls ordering of the Hadamard basis for compressive single-pixel imaging," Scientific reports **7**, 3464 (2017).
2. C. C. Paige, and M. A. Saunders, "LSQR: An algorithm for sparse linear equations and sparse least squares," ACM Transactions on Mathematical Software (TOMS) **8**, 43-71 (1982).
3. J. M. Bioucas-Dias, and M. A. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," IEEE Transactions on Image processing **16**, 2992-3004 (2007).
4. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*(Springer2015), pp. 234-241.
5. C. F. Higham, R. Murray-Smith, M. J. Padgett, and M. P. Edgar, "Deep learning for real-time single-pixel video," Scientific reports **8**, 2369 (2018).
6. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing **13**, 600-612 (2004).
7. O. Katz, P. Heidmann, M. Fink, and S. Gigan, "Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations," Nature photonics **8**, 784-790 (2014).
8. J. R. Fienup, "Phase retrieval algorithms: a comparison," Applied optics **21**, 2758-2769 (1982).
9. K. Jaganathan, Y. C. Eldar, and B. Hassibi, "Phase retrieval: An overview of recent developments," arXiv preprint arXiv:1510.07713 (2015).
10. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE **86**, 2278-2324 (1998).
11. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*(2015), pp. 1-9.