Image disambiguation with deep neural networks

Omar DeGuchy, Alex Ho, and Roummel F. Marcia

University of California, Merced, 5200 N. Lake Road, Merced, CA 95343 USA

ABSTRACT

In many signal recovery applications, measurement data is comprised of multiple signals observed concurrently. For instance, in multiplexed imaging, several scene subimages are sensed simultaneously using a single detector. This technique allows for a wider field-of-view without requiring a larger focal plane array. However, the resulting measurement is a superposition of multiple images that must be separated into distinct components. In this paper, we explore deep neural network architectures for this image disambiguation process. In particular, we investigate how existing training data can be leveraged and improve performance. We demonstrate the effectiveness of our proposed methods on numerical experiments using the MNIST dataset.

Keywords: Deep neural networks, image disambiguation, multiplexed images

1. INTRODUCTION

Applications such as blind source separation[1] and multiplexed imaging[2] involve measurement data that are composed of multiple signals that have been observed simultaneously and combined at the detector stage. For example, in blind source separation settings, a microphone might pick up several conversations concurrently, or perhaps there might be one predominant signal mixed with ambient or latent sounds. In multiplexed imaging, optical systems might utilize beam splitters and mirrors to superimpose different scene components onto a single focal plane array (FPA). This type of architecture is particularly useful in settings where a wide field-of-view is needed but the FPA size is limited. The main challenge in these applications is to separate the combined measurements into distinct signals or perhaps simply isolate a particular signal. These problems are highly underdetermined and ill-posed, and, as such, they require sophisticated numerical methods. The methods we propose in this paper use machine learning techniques based on autoencoders.

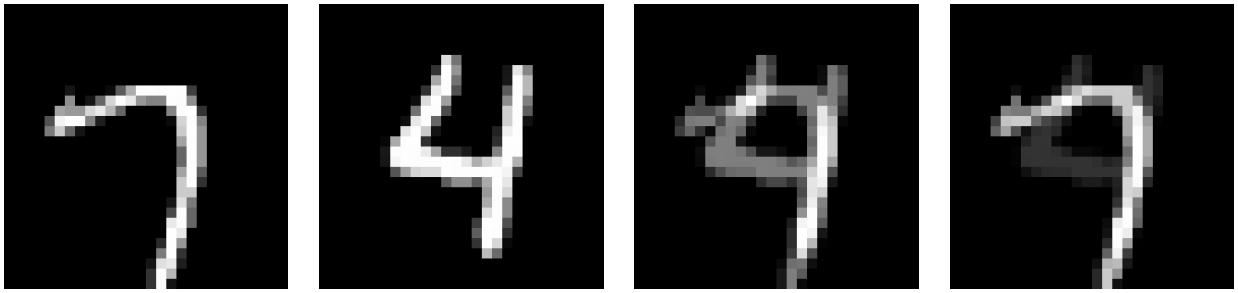
Related methods. The blind source separation problem has been well studied[1] and has been analyzed from a statistical perspective.[3,4] Algorithms include leveraging sparse decomposition[5] and on-line learning.[6] Multiplexed optical systems have been physically implemented,[7–10] and algorithms for multiplexed imaging include those that use nonnegative matrix factorization[11] and those that exploit a priori knowledge about the multiplexed images, such as sparse representation.[12,13] In contrast, the approaches proposed in this paper incorporate existing datasets to train deep neural networks for disambiguating the superimposed images.

2. METHODOLOGY

2.1 Neural Network Architecture

We propose the following neural network configuration for the purpose of recovering two images from a mixed signal measurement. The architecture is based on the deep learning building block known as an autoencoder.[14] In its simplest form, an autoencoder is composed of two parts. The first is the encoder whose objective is to reduce the dimensionality of the input by providing a latent space representation of the most pertinent information. The second element, referred to as a decoder, is tasked with interpreting the resultant latent space variable and ultimately recovering the original input.[15] In practice, more complex evolutions of the autoencoder such as the variational autoencoder and stacked denoising autoencoders are typically implemented.[16,17] Beyond

Further author information:
Omar DeGuchy.: E-mail: odeguchy@ucmerced.edu
Alex Ho: Email: aho38@ucmerced.edu
Roummel F. Marcia.: E-mail: rmarcia@ucmerced.edu, Telephone: 1-209-228-4874



(a) (b) (c) (d)

Figure 1. Image disambiguation setup. Two 28×28 pixel images (a) and (b) are randomly chosen from the MNIST dataset. Images (a) and (b) are superimposed to yield measurement image (c). Image (d) is the result of superimposing images (a) and (b) with image (b) at 25% intensity.

modifications of the overall structure of the autoencoder, there is also a choice between the commonly used fully connected layer and the convolutional layer as the basis of the encoder and decoder substructures.[15,18] The method presented in this paper utilizes stacked denoising autoencoders composed of fully connected layers. The motivation for this implementation is two fold. Stacked denoising autoencoders (SDAs) have been successfully implemented in a variety of other image processing tasks.[19–23] As an extension to their autoencoder ancestor, SDAs continually encode and decode the information until the intended output is obtained. The intuition is that as they are forced to compress and decompress the input, they become impervious to noise during the process (see Fig. 2). Our intent was to take advantage of this property during the image extracting process, resulting in smoother, noiseless reconstructions. Fully connected layers were chosen based on the configuration of the problem. Because the goal of the application is multiple image extraction from a single source, the architecture needs to be accommodating to an output of two images. Initially, convolutional layers were tested with a single channel input image and a dual channel output tensor. The network often returned the same image for both channels. As an alternative, the images were vectorized necessitating the use of fully connected layers. This allowed the network to produce a single vector containing both reconstructions.

2.2 Network Parameters

The network used in this work begins by reshaping the single channel $n \times n$ combined images into the vector $p = [p_1\ p_2\ p_3\ \ldots\ p_{n^2}]$ where p_i represents the pixel intensity at a given location (see Fig. 2). The input layer

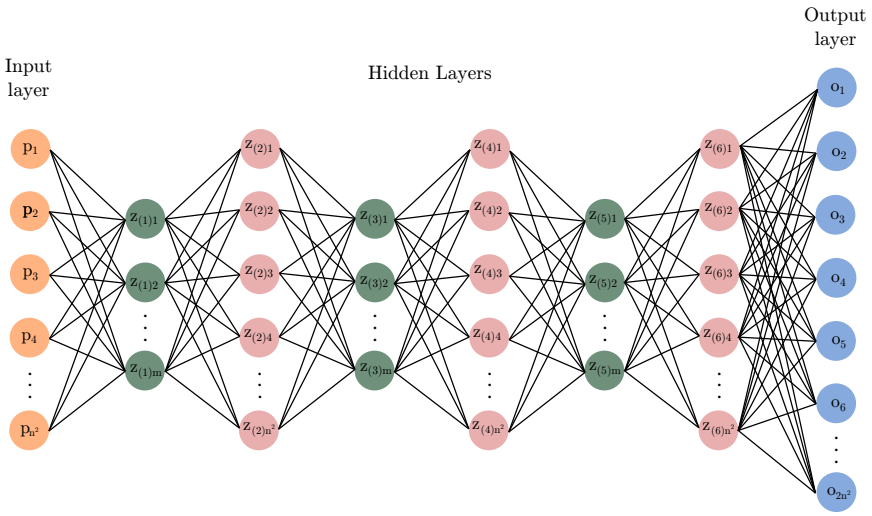


Figure 2. Deep neural network for image disambiguation. The network processes the n^2 input p through six fully connected layers $z_{(1)} - z_{(6)}$. The output layer o doubles the size of the input to allow for the extraction of the two images.

is followed by 3 stacked autoencoders where the input from the previous layer is either compressed to length m or decompressed to the size n^2. The dimension of compression (m) is a hyperparameter which requires tuning, but the dimension n is constrained to the size of the image being processed. The final layer produces the vector $o = [o_1\ o_2\ o_3\ \ldots\ o_{2n^2}]$ containing the separated images. The two images are recovered by dissociating the output vector into the components $\tilde{p}_1 = [o_1\ \ldots\ o_{n^2}]$ and $\tilde{p}_2 = [o_{n^2+1}\ \ldots\ o_{2n^2}]$ and reshaping each vector into the original $n \times n$ dimensions. After all but the final layer, the rectified linear unit (ReLU) activation function was applied to the output before being passed as the input of the next layer. After the output layer the, sigmoidal activiation function is applied to ensure that the output pixel intensities are within the range 0 to 1. The network was implemented using two different cost functions which will be discussed in a later section. In either case the network was trained using the backpropagation algorithm.

3. NUMERICAL EXPERIMENTS

The architecture was developed using the open source machine learning library for Python, PyTorch. Training and testing were done using an NVIDIA Tesla P100 PCI-E GPU on the MERCED cluster. The loss functions were minimized using the PyTorch implementation of the Adam optimizer.[24]

3.1 MNIST Dataset

The dataset used to test and train the proposed architecture is a modified form of the original MNIST data set.[25] The data set consists of 70,000 28×28 images of handwritten numbers from 0-9. The data is then partitioned into 60,000 training examples and 10,000 testing examples. For the purposes of this paper, classification is not the intended objective and therefore all of the labels were disregarded. Each training and testing sample were created by first randomly selecting two images from the dataset without replacement. The images were normalized to assure that the pixel intensities ranged from 0 to 1 and then converted to a single channel tensor. The two tensors were then added together and the result was normalized and paired with the original images creating a triplet consisting of a combined image and its two sources (see Fig. 1). Because of the nature of the previously described image selection, the new training set consisted of 30,000 training instances and 5,000 testing images where both the combinations and the individual images are unique. Multiple datasets were created following a similar protocol, the difference being the intensity of one of the two target images was reduced to 75%, 50% or 25%. When describing these data sets the image with the full intensity is noted as Image A and the image with the reduced intensity will be known as Image B.

3.2 Performance

The proposed architecture was tested for a variety of configurations. In what will be referred to as the Single Image Recovery experiment, the intent was to recover only one of the two images. The image target was chosen randomly from the two potential candidates and the output layer in Fig. 2 was removed so that $z_{(6)}$ would provide an extraction of the appropriate size. The ReLU activation function implemented in $z_{(6)}$ was also substituted for the sigmoidal activation function. The Dual Image Recovery Experiment seeks to extract both images using the architecture described in Sec. 2. In both experiments, the dimension of the latent space was chosen to be $m = 256$ after hyperparameter tuning revealed this to be the optimal setting. All test images were compared to their targets using the Mean Squared Error (MSE) after the model was trained.

The normalization of the pixel intensities within a range of 0 to 1 allows us to use the Binary Cross Entropy (BCE) function as a loss function during training. Experiments were performed to determine if there was an advantage in the choice of using either the BCE or the MSE in a learning capacity. The performance of the two loss functions are comparable when using the MSE as a metric of validation on the test set (see Fig. 3). The reconstructions using the BCE had a slightly sharper appearance. This is to be expected as the MSE tends to average the pixel intensities resulting in a smoother image. For the remainder of this paper we present only the results using the BCE loss function. The Dual Image Recovery experiment indicates that image extraction of both images are comparable when they are at full intensity. As the image intensity of Image B decreases, the architecture improves on its ability to extract Image A at the cost of an accurate reconstruction of Image B (see Fig. 3). This is to be expected as the intensity of Image B is weaker. The Single Image Recovery approach underperforms in comparison to Dual Image Recovery approach when both images have the same intensity. The

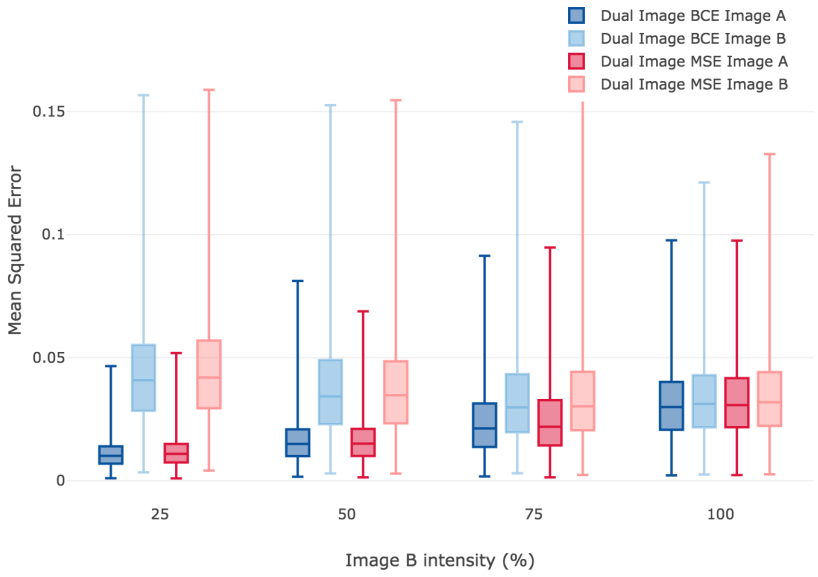


Figure 3. Boxplots of the Mean Squared Error (MSE) for 5,000 MNIST test images using the Dual Image Recovery method with the Binary Cross Entropy (BCE) [in blue] and the MSE [in red] loss functions. The MSEs are reported for both Images A and B under varying intensities of Image B in the measurements. As the intensity of Image B weakens, the accuracy of its recovery naturally worsens while the accuracy of Image A's recovery improves.

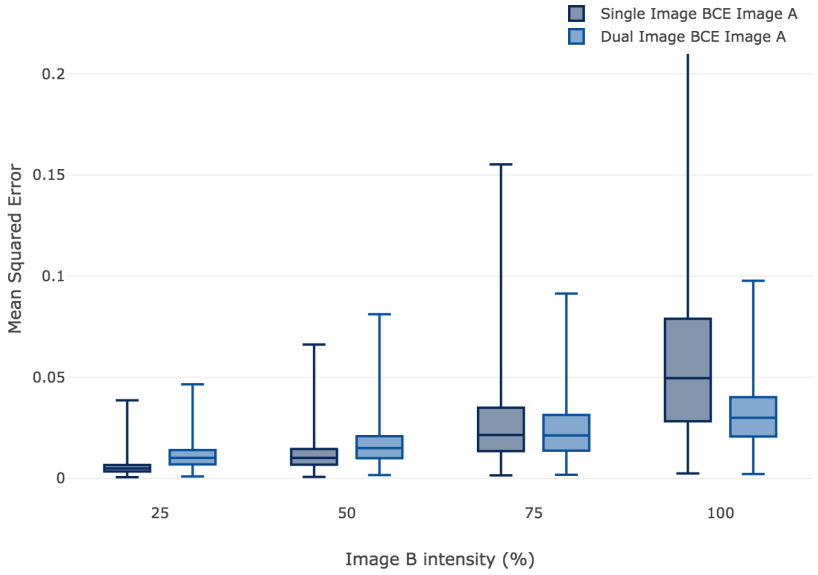


Figure 4. Boxplots comparing the Dual and Single Image Recovery methods to recover Image A using the Binary Cross Entropy (BCE) loss function. The MSE is reported for the recovered Image A from measurements with varying intensities of Image B.

Single Image Recovery approach is prone to recovering more elements from Image B. As the intensity of Image B decreases, the extraction of Image A improves, ultimately outperforming the Dual Image Recovery method (see Fig. 4). It is under the previously stated conditions that the Single Image Recovery approach views Image B as noise and performs its intended task as that of denoising.

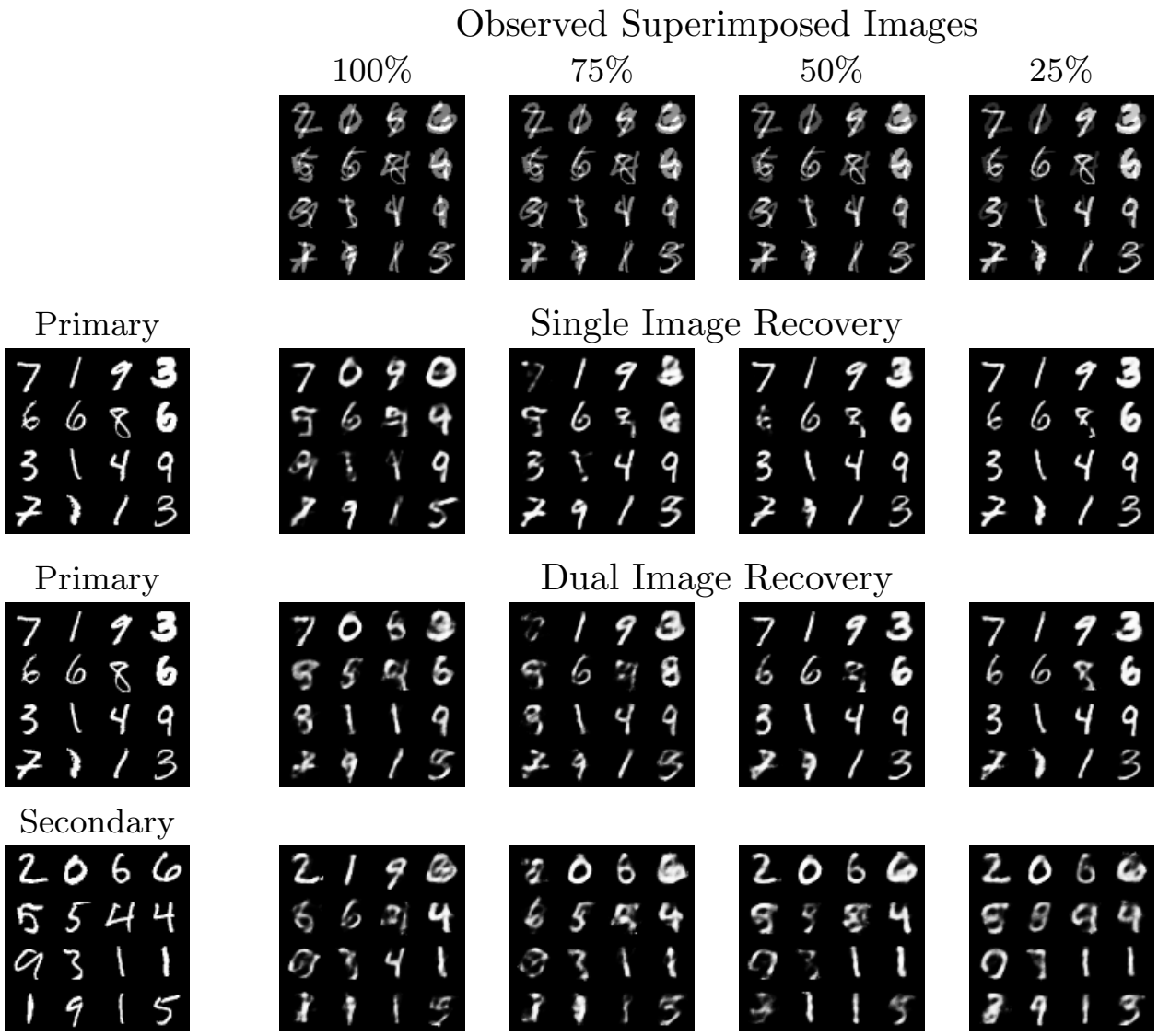


Figure 5. The first row contains the superimposed images which are composed of the primary and secondary images. They vary in that the intensity of the secondary image was altered to the percentage noted above. Given these input images, we present the reconstructions for the dual image recovery experiments and the single image recovery experiments.

4. CONCLUSION

In this paper we implemented a deep neural network in order to solve the ill-posed image separation problem. By relying on the denoising properties of the stacked autoencoder structure, we have shown that these types of networks can be effective in the recovery of two superimposed images. While the choice of the loss function between the mean squared error and the binary cross entropy function is negligible in terms of MSE validation, the binary cross entropy function provides a slightly sharper advantage. The results show that when the objective is the recovery of both images, the reconstruction of the secondary image with a weaker intensity (Image B) slightly suffers while the recovery of the primary image (Image A) improves. By shifting the focus to a single image extraction the quality of the reconstruction of the primary image improves on the dual image extraction method. For future work we look forward to extending these methods to more complex images by improving the structure of the architecture. Furthermore, we look forward to extending this work to other modalities.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation Grant IIS-1741490. The authors gratefully acknowledge computing time on the Multi-Environment Computer for Exploration and Discovery (MERCED) cluster at UC Merced, which was funded by National Science Foundation Grant No. ACI-1429783.

REFERENCES

[1] Comon, P. and Jutten, C., [*Handbook of Blind Source Separation: Independent component analysis and applications*], Academic Press (2010).

[2] Treeaporn, V., Ashok, A., and Neifeld, M. A., "Increased field of view through optical multiplexing," *Optics Express* **18**(21), 22432–22445 (2010).

[3] Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., and Moulines, E., "A blind source separation technique using second-order statistics," *IEEE Transactions on Signal Processing* **45**(2), 434–444 (1997).

[4] Cardoso, J.-F., "Blind signal separation: statistical principles," *Proceedings of the IEEE* **86**(10), 2009–2025 (1998).

[5] Zibulevsky, M. and Pearlmutter, B. A., "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computation* **13**(4), 863–882 (2001).

[6] Amari, S., Cichocki, A., and Yang, H. H., "A new learning algorithm for blind signal separation," in [*Advances in Neural Information Processing Systems*], 757–763 (1996).

[7] Chen, C.-Y., Yang, T.-T., and Sun, W.-S., "Optics system design applying a micro-prism array of a single lens stereo image pair," *Optics Express* **16**(20), 15495–15505 (2008).

[8] Uttam, S., Goodman, N. A., Neifeld, M. A., Kim, C., John, R., Kim, J., and Brady, D., "Optically multiplexed imaging with superposition space tracking," *Optics Express* **17**(3), 1691–1713 (2009).

[9] Horisaki, R. and Tanida, J., "Multi-channel data acquisition using multiplexed imaging with spatial encoding," *Optics Express* **18**(22), 23041–23053 (2010).

[10] Shepard, R. H., Rachlin, Y., Shah, V., and Shih, T., "Design architectures for optically multiplexed imaging," *Optics Express* **23**(24), 31419–31435 (2015).

[11] Cichocki, A., Zdunek, R., Phan, A. H., and Amari, S., [*Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*], John Wiley & Sons (2009).

[12] Bofill, P. and Zibulevsky, M., "Underdetermined blind source separation using sparse representations," *Signal Processing* **81**(11), 2353–2362 (2001).

[13] Marcia, R. F., Kim, C., Eldeniz, C., Kim, J., Brady, D. J., and Willett, R. M., "Superimposed video disambiguation for increased field of view," *Optics Express* **16**(21), 16352–16363 (2008).

[14] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A., "Extracting and composing robust features with denoising autoencoders," in [*Proceedings of the 25th International Conference on Machine Learning*], 1096–1103, ACM (2008).

[15] Goodfellow, I., Bengio, Y., and Courville, A., [*Deep learning*], vol. 1, MIT Press Cambridge (2016).

[16] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A., "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research* **11**(Dec), 3371–3408 (2010).

[17] Kingma, D. P. and Welling, M., "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114* (2013).

[18] Masci, J., Meier, U., Cireşan, D., and Schmidhuber, J., "Stacked convolutional auto-encoders for hierarchical feature extraction," in [*International Conference on Artificial Neural Networks*], 52–59, Springer (2011).

[19] Xie, J., Xu, L., and Chen, E., "Image denoising and inpainting with deep neural networks," in [*Advances in Neural Information Processing Systems*], 341–349 (2012).

[20] Lore, K. G., Akintayo, A., and Sarkar, S., "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition* **61**, 650–662 (2017).

[21] Mousavi, A., Patel, A. B., and Baraniuk, R. G., "A deep learning approach to structured signal recovery," in [*Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*], 1336–1343, IEEE (2015).

[22] Wang, N. and Yeung, D.-Y., "Learning a deep compact image representation for visual tracking," in [*Advances in Neural Information Processing Systems*], 809–817 (2013).

[23] Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., and Madabhushi, A., "Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images," *IEEE Transactions on Medical Imaging* **35**(1), 119–130 (2016).

[24] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2014).

[25] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P., "Gradient-based learning applied to document recognition," *Proceedings of the IEEE* **86**(11), 2278–2324 (1998).