Deep-Learning Supervised Snapshot Compressive Imaging Enabled by an End-to-End Adaptive Neural Network

Miguel Marquez, Yingming Lai, Xianglei Liu, Cheng Jiang, Shian Zhang, Henry Arguello, Senior Member, IEEE, and Jinyang Liang

Abstract—Snapshot compressive imaging (SCI) is an advanced approach for single-shot high-dimensional data visualization. Deep learning is popularly used to improve SCI’s performance. However, most existing methods are merely used as a replacement for analytical-modeling-based image reconstruction. Moreover, these models cling to the conventional random coded apertures and often presume a linear shearing operation. To overcome these limitations, we develop a new end-to-end convolutional neural network, termed deep high-dimensional adaptive net (D-HAN) that offers multifaceted supervision to SCI by optimizing the coded aperture, sensing the shearing operation, and reconstructing three-dimensional datacubes. The D-HAN is implemented in two representative SCI systems for ultrahigh-speed imaging and hyperspectral imaging. The D-HAN is envisioned to benefit SCI in system design, image reconstruction, and performance evaluation.

Index Terms—Snapshot compressive imaging, end-to-end neural networks, coded aperture design, shearing estimation, high-dimensional imaging.

I. INTRODUCTION

SNAPSHOT compressive imaging (SCI) [1], integrating the compressive sensing theory into novel optical designs, can record high-dimensional (e.g., spatiotemporal [2], spatiospectral [3], volumetric [4], and spectro-polarimetric [5]) data in a single exposure. A typical type of existing SCI systems works by first encoding a high-dimensional scene with one or multiple coded apertures, followed by a shearing operation that couples the non-spatial dimension (e.g., spectral and temporal) of the spatially encoded scene with one of its spatial dimensions, and finally integration of the encoded and sheared scene on a focal plane array [6]. Despite increasing system complexity compared to conventional imaging modalities, SCI allows using widely available two-dimensional (2D) focal plane arrays (e.g., CCD and CMOS cameras) to collect compressively recorded snapshots and then to recover the underlying datacube using computational methods. Leveraging this encoder-decoder approach, SCI sparks the development of many high-dimensional imaging systems with low bandwidth/memory requirements, low costs, and low power consumption. In the meantime, a careful design of the coded apertures and the shearing operation can transfer many advantageous properties embedded in certain mathematical models into optical systems, which exceed the conventional counterparts in imaging and sensing capabilities [7].

As an indispensable constituent in SCI, advanced image reconstruction methods provide high reconstructed image quality. To date, analytical modeling is dominantly used in SCI’s image reconstruction [13–17]. These methods solve the inverse problem by using the prior knowledge of the sensing matrix and by leveraging the sparsity of high-dimensional datacubes in specific domains (e.g., spatial, wavelet, discrete cosine transform, and gradient domains). However, they require precise prior knowledge of all operations in the SCI systems and are susceptible to...
noise. Moreover, their processing time increases exponentially in high-dimensional optimization problems. Finally, the reconstructed image quality highly depends on the empirical tuning of parameters [18].

To overcome these limitations, deep-learning approaches have been increasingly featured owing to their flexible capability of obtaining high reconstruction performance compared to the analytical-modeling approaches. Given access to rich datasets, many novel methods based on convolutional neural networks (CNNs) have been developed for SCI’s reconstruction as well as for the coded-aperture design [19]–[25]. Despite these advances, CNN approaches have ample room for improvement. First, in most cases, they are merely considered as a substitute for analytical-modeling-based reconstruction methods [22]. Their performance is hence limited by the existing arrangements in SCI systems. Moreover, most CNN-based coded-aperture designs are proposed for active codification, in which each frame in the datacube is coded by a different coded aperture [25]. In contrast, despite a few promising demonstrations [10], [26], [27], less emphasis has been placed on the optimization of a single coded aperture. Consequently, to date, most SCI systems in this category still rely on a random coded aperture, which, although universally applicable to most scenes, limits the reconstruction quality. Finally, these approaches often ignore the ideal behaviors of optical instruments. Particularly for SCI, the shearing operation, realized by various optical elements (e.g., a galvanometer scanner (GS) [28], a streak tube [29], a grating [30], or a prism [31]), are modeled with a linear function with a slope of one pixel and an offset of zero. This idealization neglects any deviation in the shearing operation induced by many possible experimental factors, including misalignment, jitter, and imperfect instrument responses [29]. Consequently, this limitation reduces CNN’s robustness in reconstruction. It also excludes the possibility of sensing information in key optical elements in SCI systems. To date, the deep-learning approaches are able to provide little feedback to evaluate SCI’s hardware performance.

To surmount these limitations, we develop a new end-to-end (E2E)-CNN—termed the deep high-dimensional adaptive net (D-HAN)—that provides multi-faceted supervision to an SCI system. In system design, the D-HAN optimizes the coded aperture and establishes SCI’s sensing geometry. In image reconstruction, the D-HAN senses the shearing operation and retrieves a three-dimensional (3D) scene. D-HAN-supervised SCI is experimentally validated using two representative systems for ultrahigh-speed imaging and hyperspectral imaging.

II. RELATED WORK

A. Sensing Model of SCI

In this work, the data acquisition of SCI consists of the following three operations. First, a 3D scene, denoted by \( \mathbf{F} \), is modulated by a single 2D \((x, y)\) coded aperture. This operation is termed spatial encoding. Then, information in the third dimension of the spatially encoded scene is intermixed with that in one of the spatial dimensions. This process is defined as the shearing operation. Finally, via the operation of integration, a 2D detector integrates the encoded and sheared scene as a 2D image, commonly known as a compressed measurement, denoted by \( \mathbf{G} \) [32]. The forward model is expressed as

\[
\mathbf{g} = \mathbf{G} \mathbf{f} + \mathbf{e}.
\]

Here, \( \mathbf{f} \in \mathbb{R}^{n \times 1} \) is the vector representation of a discrete version of the 3D scene \( \mathbf{F} \in \mathbb{R}^{N_x \times N_y \times N_L} \) with the size \( n = N_x N_y N_L \), where \( N_x \) and \( N_y \) represent the data lengths in the two spatial dimensions, and \( N_L \) represents the data length in the third dimension. \( \mathbf{g} \in \mathbb{R}^{m \times 1} \) is the vectorial version of the compressed measurement \( \mathbf{G} \in \mathbb{R}^{N_x \times (N_y + \gamma (N_L - 1))} \) with \( m = N_x \cdot \lceil N_y + \gamma \cdot (N_L - 1) \rceil \), \( \gamma \in \mathbb{W} \) is the shearing magnitude factor, \( \Phi \in \mathbb{R}^{m \times n} \) is the SCI’s sensing matrix, \( \varepsilon \in \mathbb{R}^{m \times 1} \) represents the noise added to the compressed measurement in data acquisition. The structure of \( \Phi \) accommodates the optical elements in the SCI system, and its entries are given by [33]

\[
\Phi_{i,j} = \begin{cases} c_v & \text{if } i = N_x \Gamma \left( \left\lfloor \frac{j}{N_y} \right\rfloor \right) \\ 0 & \text{otherwise} \end{cases}
\]

for \( v = \text{mod}(j, N_x \cdot N_y) \) with \( v \in \mathbb{W} \), \( i = \{0, \ldots, m-1\} \), and \( j = \{0, \ldots, n-1\} \). Here, \( \Gamma(v) : \mathbb{W} \rightarrow \mathbb{W} \) is the function that models the shearing operation. \( c_v \) is the value at the \( v \)th position of \( c \in \mathbb{R}^{N_x \times N_y \times 1} \), which is the vectorial version of the coded aperture \( C \in \mathbb{R}^{N_x \times N_y} \). In the scope of this work, the coded aperture is set as binary (i.e., \( c_v \in \{0, 1\} \)).

B. Integration of Deep Learning Into SCI

The increasing availability of vast amounts of image data and computational power has consolidated CNNs as one of the most potent and desired tools for SCI approaches [34]–[39]. CNN’s potential for compressive sensing was initially demonstrated in its non-iterative solution to reconstruct images from measurements generated by a random Gaussian matrix [34]. Inspired by this early development, CNN paradigms were integrated into SCI’s reconstruction by combining its forward model as an encoder, followed by convolutional layers as a fast approximation of the decoding network [35]. Despite high reconstructed image quality in simulation, these early CNN-based approaches produced a dramatically decreased reconstruction quality in experiments. This limited performance was attributed to their neglect of the operations in SCI models (e.g., direct, transpose, and inverse processes) that could work as prior information to improve the reconstruction results. In this path, ensuing works tackled this problem by taking advantage of the well-established link between analytical-modeling approaches and deep learning networks [36], resulting in the development of the unrolling approach [40], [41] and the plug-and-play (PnP) method [42], [43], both of which relied on the iterative gradient descent models for the inferences. However, the unrolling approaches were highly prone to suffer vanishing gradient issues as the number of emulated iterations increased. The PnP approaches limited the CNN’s inference potential by using them as simple denoisers. Both methods strained the CNN’s ability to design or infer the structure and behavior of related optical elements in SCI systems. These limitations, along with the well-demonstrated SCI homologous frameworks [44]–[48], call for new CNNs for SCI in design guidance, image reconstruction, and performance evaluation.
C. Deep-Learning Approaches for Spectral SCI

In these two sub-sections, we review related CNN-based approaches implemented in the spectral and temporal SCI frameworks—two representative branches in SCI for single-shot observation of hyperspectral scenes and transient phenomena, respectively. First, deep learning has assisted spectral SCI in providing higher image quality and instantaneous reconstruction after training compared to the analytical-modeling-based algorithms. Representative approaches include the HSCNN [49], the HyperReconNet [26], the spatial-spectral self-attention (TSA) net [50], and deep Gaussian scale mixture (DGSM) prior for spectral compressive imaging [51]. In particular, the HSCNN was a unified deep-learning framework for spectral SCI’s reconstruction [49]. It used a two-step iterative shrinkage/thresholding algorithm [52] to reconstruct an approximation, which was then used as input to a trained CNN to recover the missing details. To improve the correlation of these two steps presented in the HSCNN, the HyperReconNet was introduced as the first E2E-CNN that joined coded aperture design with image reconstruction [26]. However, relying on a patch-based methodology, the HyperReconNet disregarded the global spatial features, which limited its performance in both tasks. To solve the shortcomings in the HSCNN and the HyperReconNet, the TSA-net jointly modeled the spatial and spectral correlation, which reconstructed 3D datacubes with good accuracy [50]. To outperform the TSA-net, the DGSM algorithm [51] modeled the reconstruction process as a maximum a posteriori estimation problem with a learned Gaussian scale mixture model in an E2E manner. The local means of the Gaussian scale mixture models were estimated as a weighted average of the spatial-spectral neighboring pixels, which robustly reconstructed 3D datacubes.

Despite these encouraging developments in image reconstruction, existing deep-learning approaches for spectral SCI still possess several limitations. First, the CNN-based approaches relegated the coded aperture to be a non-trainable parameter generated offline with a conventional binary random distribution. Although enabling the sensing models to be incoherent with almost all sparsifying bases [53], the random features resulted in inferior reconstruction performance compared with the ones in the optimized coded apertures [26], [54]. Meanwhile, the shearing operation, as an indispensable step in SCI’s data acquisition, was often overlooked. Most works presumed it to be a linear function [50]. Others pre-determined it by calibration and theoretical calculations [55]. In this regard, how the shearing operation would affect the reconstructed image quality and hence the SCI’s overall performance is still underexplored.

D. Deep-Learning Approaches for Temporal SCI

In recent years, deep learning has enhanced the performance of temporal SCI techniques. The resemblance of its forward model to the one of spectral SCI allows adapting many existing deep-learning approaches to reconstruct spatiotemporal datacubes. For example, an E2E-CNN with residual learning was reported for fast temporal SCI reconstruction [56]. In addition, various CNNs—such as the U-net-based DeepCUP [57] and the hybrid algorithm combining the augmented Lagrangian (AL) method with deep learning [58]—were developed to mitigate artifacts and to enhance the reconstruction speed. Moreover, a deep neural network decomposed the temporal SCI’s reconstruction into independent 2D sub-problems to improve the reconstructed image quality [59].

Besides assisting image reconstruction, deep learning has contributed to system design in temporal SCI. For example, an encoder-decoder neural network was developed to obtain a set of trained coded apertures [25]. Although resulting in superior performance compared to the traditional random counterparts, the trained coded-aperture set lost the universality to dynamic events with different compression ratios. In other works, coded apertures were jointly learned with video reconstruction using a fully connected neural network [24]. However, multiple trained coded apertures were still required in reconstruction, which limited the sequence depth (i.e., the number of frames in the reconstructed video). More importantly, the requirement of using multiple coded apertures in data acquisition disqualified both works for the single-coded-aperture SCI, which is the model of our work. Furthermore, akin to spectral SCI, existing methods assumed a linear shearing operation in the data acquisition of temporal SCI. Consequently, the image reconstruction could sense and self-adapt to any deviations in shearing. Thus far, most deep-learning approaches still focus separately on either optimizing temporal SCI’s hardware of data acquisition or improving image reconstruction. There lacks a mastermind capable of providing overall supervision of both the hardware and software for temporal SCI.

III. END-TO-END SUPERVISION BY D-HAN

The D-HAN aims to bring new capabilities to the CNN for E2E supervision of an SCI system. In brief, the D-HAN targets to couple the design and modeling of the two essential operators in the SCI sensing geometry—spatial encoding and shearing—together with the image reconstruction. Towards this goal, the D-HAN comprises three main sequential stages: four dense layers for shearing estimation, a deep-unfolding-based network to embody SCI’s sensing model, and a U-net [60] that works as a filter. The second stage is developed from the structure of an alternating direction method of multipliers (ADMM) [61]. It is used to integrate SCI’s major operations with a mathematical coherence, which are emulated as three custom layers constructed by using the coded aperture and the shearing function. These two trainable structures are modeled via a binarization and polynomial approximation. For training, these two constraints are respectively represented as a coded-aperture design layer and a set of dense layers, both of which receive the compressed measurement as input. Using this approach, the D-HAN optimizes the coded aperture based on the training data and outputs the ad-hoc shearing estimation from the dense layers.

In this section, we will provide a detailed theoretical description and simulation to show D-HAN’s multi-faceted supervision to SCI by optimizing the coded aperture, evaluating the shearing operation, and reconstructing 3D datacubes. In Sections IV and
V, we will experimentally demonstrate the D-HAN in both spectral and temporal SCI.

A. D-HAN-Supervised System Design

To optimize the coded aperture, we adopt the approach in Ref. [62]. The pixel values in the coded aperture are denoted by \( \theta_{x,y} \), where \( x = \{0, \ldots, N_x - 1\} \) and \( y = \{0, \ldots, N_y - 1\} \) are initialized with random values and trained by the D-HAN. Notice that \( \theta_{x,y} \) are not set to be binary; instead, a binarization version is estimated via a sigmoid function \( C_{x,y} = \frac{1}{1 + e^{-\theta_{x,y}}} \). This approach avoids the gradient vanishing and promotes the influence of the forward model in the coded-aperture design. Finally, a constraint function \( \mathcal{T}(\cdot) \), whose minima are obtained uniquely when \( C_{x,y} \) are 0 or 1, is used to generate the binary coded aperture by

\[
\mathcal{T}(C) = \sum_{x,y} (C_{x,y})^2 (C_{x,y} - 1)^2. \tag{3}
\]

In the training stage, the layer that designs the coded aperture is located at the beginning of the D-HAN (shown in dark green in Fig. 1).

To better sense the shearing operation, rather than a linear function (with a unity slope and a zero offset) in traditional CNNs, we model it as a polynomial function as

\[
\Gamma_a(z) = z + \sum_{k=0}^{K-1} a_k z^k, \tag{4}
\]

where \( K \in \mathbb{N} \) is the polynomial degree, \( a_k \in \mathbb{R} \) represents the polynomial weights with \( a \in \mathbb{R}^K \), and \( z = \{0, \ldots, N_L - 1\} \) represents the variable of the third dimension. In (4), the first term models the linear shearing function while the second term models the nonlinear deviation in the shearing operation. Typically, nonlinear shearing functions have only a small deviation from the linear ones. Thus, these functions, sharing similar mathematical characteristics to those of the linear shearing functions, are monotone, continuous, and smooth. Thus, a polynomial function is used to model nonlinear shearing functions. In our work, we used the fourth degree (i.e., \( K = 4 \)) polynomial, which can accurately approximate the many slow-varying functions with an error of \( \sim 1\% \). Thus, this polynomial function is mathematically convenient and permits high precision of estimation. Weight estimation is performed by a set of four dense layers with a sigmoid activation function (Fig. 1). The first layer receives the compressed measurement as an input, and the last layer returns the estimated shearing weights \( \hat{a} \), corresponding to (4). These values are relayed to the direct sensing operator layers, the inverse operator layer, and the transposed operator layers. Located at the D-HAN’s beginning, \( \mathcal{G}(\cdot) \) emulates the direct sensing process in the training step from \( \hat{F} \) and \( a \). In addition, as a fidelity regularizer, another layer of \( \mathcal{G}(\cdot) \) is added at the end of D-HAN to estimate a compressed measurement from \( \hat{F} \) and \( \hat{a} \). In this way, different from a traditional E2E-CNN with a single input/output, the D-HAN relies on a tuple of input/output parameters (i.e., the datacube and the set of weights) in the system design.

B. D-HAN-Supervised Image Reconstruction

In the D-HAN, the image reconstruction relies on the joint effort of three sequential stages: the dense layers for shearing estimation, a set of parallel layers emulating the closed-form solution of SCI’s inverse problem modeled by the ADMM, and the U-net structure [60]. The dense layers output the estimated shearing weights (denoted by \( \hat{a} \)). To avoid the gradient vanishing in the dense layers, a direct connection is set between these dense layers and the sensing operator layer located at the end of the D-HAN. By regularizing the compressed measurement, this connection increases the fidelity of the estimated shearing function. Meanwhile, the parallel layers and the U-net manifest the “splitting-and-optimization” approach embedded in the ADMM with mathematical coherence.

The ADMM-based inverse problem is incorporated by using the Eckstein and Bertsekas lemma [63], which is expressed as

\[
\{\mathbf{f, w}\} = \arg\min_{\mathbf{f,w}} \left\|\Phi_{\mathbf{f}} - \mathbf{d}\right\|_2^2 + P_1 (\mathbf{f}) + P_2 (\mathbf{w}). \tag{5}
\]

Here, \( P_1 (\cdot) : \mathbb{R}^n \to \mathbb{R}, P_2 (\cdot) : \mathbb{R}^q \to \mathbb{R} \) with \( q \in \mathbb{N} \) as an arbitrary constant. \( \mathbf{w} \in \mathbb{R}^{q \times 1} \) is a splitting variable that serves as the argument of \( P_2 (\cdot) \) with \( \mathbf{w} = \mathbf{I} \mathbf{f} \), where \( \mathbf{I} \) is the identity matrix with an arbitrary size. \( \mu \geq 0 \) is called the AL penalty parameter. \( \mathbf{d} \in \mathbb{R}^{n \times 1} \) is the Lagrange multiplier vector. To integrate this ADMM structure into the D-HAN, \( P_1 (\mathbf{f}) = ||\mathbf{f} - \mathbf{g}||^2_2 \) is established with \( \mathbf{g} \) generated via (2) and (4), and \( P_2 (\mathbf{w}) \) is relegated to an arbitrary function where its structure is irrelevant in our approach. Then, (5) is decoupled into two analytical inverse models

\[
\mathbf{f} = \arg\min_{\mathbf{f}} ||\Phi_{\mathbf{f}} - \mathbf{g}||^2_2 + \frac{\mu}{2} ||\mathbf{f} - \mathbf{d}||^2_2, \text{ and}
\]

Authorized licensed use limited to: Inst Natl de la Recherche Scientific EMT. Downloaded on July 17,2022 at 13:45:58 UTC from IEEE Xplore. Restrictions apply.
\[
    w = \text{argmin}_w P_2(w) + \frac{\mu}{2}||w - f - d||^2. \tag{6}
\]

Note that in (6), the analytical inverse model of \( f \) refers to a quadratic problem, while the analytical model of \( w \) is a denoising problem. To solve the first inverse model in (6), we exploit the Sherman Woodbury Morrison (SWM) matrix inversion lemma [64] and the full-column rank properties to obtain the closed-form solution that simultaneously involves the sensing process, the sensing matrix conditionality, and the gradient solution:

\[
    f = \tilde{\mu}^{-1} \left[ I - \Phi^T [\tilde{\mu} I + \Phi \Phi^T]^{-1} \Phi \right] \left[ \Phi^T g + \tilde{\mu}^{-1} (w - d) \right],
\]

where \( \Phi \Phi^T \in \mathbb{R}^{m \times m} \) represents a matrix product resulting in a diagonal matrix and \( \tilde{\mu} = \mu/2 \).

To construct the D-HAN using (7), the first step is to define the operators of SCI’s data acquisition using the 3D scene, the coded aperture, and shearing coefficients. First, for the direct sensing process (whose layer is shown in magenta in Fig. 1), the operator is expressed by

\[
    G(F, C, a) = \sum_{i=0}^{N_L-1} R(\Gamma_a(i)), F_{l} \circ \Phi \Phi^T C. \tag{8}
\]

Here, \( G(\cdot) : \mathbb{R}^{N_s \times N_y \times N_L} \to \mathbb{R}^{N_s \times [N_y + \gamma (N_L - 1)]} \) is the direct sensing operator. The operator \( R(\cdot) : \mathbb{R}^{N_s \times N_y} \to \mathbb{R}^{N_s \times [N_y + \gamma (N_L - 1)]} \) introduces a right-horizontal circular shifting of \( \Gamma_a(i) \) pixels to a right-zero-padded version of the compressed measurement (i.e., \( [F_{l}, i, 0] \) with \( 0 \in \mathbb{R}^{N_s \times [\gamma (N_L - 1)]} \)).

Then, the transpose sensing operator (whose layer is shown in red in Fig. 1) is defined as

\[
    F(G, C, a) = S(\Gamma_a(i), G \circ R(\Gamma_a(i)), C). \tag{9}
\]

Here, \( F(\cdot) : \mathbb{R}^{N_s \times [N_y + \gamma (N_L - 1)]} \to \mathbb{R}^{N_s \times N_y \times N_L} \) is an operator that returns a datacube from a 2D compressed measurement. \( S(\cdot) \) is an operator of left-horizontal circular shifting of \( \Gamma_a(i) \) pixels and then removes the last \( \gamma \cdot (N_L - 1) \) columns in the resulting shifted matrix.

Finally, the inverse operator (whose layer is shown in brown in Fig. 1) is defined as

\[
    I(G, C, a) = G \circ \left[ \sum_{i=0}^{N_L-1} R(\Gamma_a(i)), C^{\circ 2} + \tilde{\mu} I \right]^{-\circ 1}. \tag{10}
\]

Here \( I(\cdot) : \mathbb{R}^{N_s \times [N_y + \gamma (N_L - 1)]} \to \mathbb{R}^{N_s \times N_y \times N_L} \), \( (\cdot)^{\circ 2} \) and \( (\cdot)^{-\circ 1} \) represent the Hadamard quadratic power and the Hadamard inverse operation, respectively.

Following the definition of these three SCI operators, the next step is to model the SWM matrix approach in (7). Toward this goal, (7) is split into two main equations \( \Phi \Phi^T g + \tilde{\mu}^{-1} (w - d) \) and \( \tilde{\mu}^{-1} - \tilde{\mu}^{-1} \Phi^T [\tilde{\mu} I + \Phi \Phi^T]^{-1} \Phi \). In the D-HAN, the first equation is reflected as a transpose operator coupled to two 2D convolutional layers (shown in cyan in Fig. 1), each of which with a ReLU activation function (referred to hereafter as a conventional+ReLU layer). Subsequently, the second equation is represented by two parallel arms. The upper arm, corresponding to \( \tilde{\mu}^{-1} \Phi^T [\tilde{\mu} I + \Phi \Phi^T]^{-1} \Phi \), is composed of the direct sensing operator as a first layer followed by an inverse and transpose sensing operator along with four 2D convolutional+ReLU layers. The bottom arm, which corresponds to \( \tilde{\mu}^{-1} I \), has three 2D convolutional+ReLU layers. The outputs of both arms are subtracted and given as the input to the U-net in the D-HAN that reflects the second equation in (6). The output from this section is the reconstructed datacube (denoted by \( \tilde{F} \)).

The loss function \( L(\cdot) \), used to learn D-HAN’s weights, is established as

\[
    L(\tilde{C}, F, a) = l_1(F, \tilde{F}) + l_1(G, \tilde{G}) + l_1(a, \bar{a}) + \ell_{SSIM}(F, \tilde{F}) + \ell_{SSIM}(C, \tilde{C}),
\]

where \( \tilde{F}, \tilde{G}, \tilde{C}, \bar{a} \) are the D-HAN outputs, \( l_1 \) is the \( l_1 \)-norm, and \( \ell_{SSIM}(\cdot) \) represents the structural similarity index measure metric. At inference in simulation, the D-HAN is run with \( F \) and a sample extracted from the test set. For experiments, the parameters from learning are directly translated from simulation to hardware.

IV. SIMULATION OF D-HAN IN SCI

We used five popular databases for simulations: “SumMe” [65], “Need for Speed” [66], “Sports Videos in the Wild” [67], “Sparse Recoveries of Hyperspectral Signal from Natural RGB Images” [68], and “Challenge on Spectral Reconstruction from RGB Images” [69]. All simulations were implemented in Tensorflow and trained on Google Colabory with a Tesla P100-PCI-E GPU using the ADAM optimizer [70]. To train the coded aperture and the shearing weights, we randomly selected and cropped 1000 datacubes with \( N_x \times N_y = 256 \times 256 \) and \( N_L = 25 \). Based on (4), the nonlinear deviation in the shearing operator was modeled using a fourth-degree polynomial function (i.e., \( K = 4 \)), where the weights were initiated randomly. After the training, the same coded aperture was used to reconstruct the underlying datacube and the shearing weights.

We conducted comprehensive ablation studies on how the terms \( l_1(F, \tilde{F}), \ell_{SSIM}(F, \tilde{F}) \), and \( l_1(F, \tilde{F}) + \ell_{SSIM}(F, \tilde{F}) \) in the loss function [i.e., (11)] would affect the performance of D-HAN in reconstruction (Supplementary Note 1 and Fig. S1). The investigation of \( l_1(F, \tilde{F}) + \ell_{SSIM}(F, \tilde{F}) \) was inspired by Ref. [71]. Results show that implementing the \( l_1(F, \tilde{F}) + \ell_{SSIM}(F, \tilde{F}) \), which integrates the robust training capability of \( l_1(F, \tilde{F}) \) and the ability to preserve the detailed features using \( \ell_{SSIM}(F, \tilde{F}) \), allows the D-HAN to exceed reconstruction performance when these two terms are considered individually by 1.15 dB and 0.63 dB, respectively (Table S1). For this ablation analysis, the terms concerning the coded aperture optimization and shearing estimation were deliberately set aside because removing any of them led to D-HAN’s divergence.

We evaluated D-HAN’s performance by the impact of the designed coded aperture on the sensing model, the accuracy of the estimated shearing function, and the quality of reconstruction. For the first analysis, Fig. 2(a) illustrates a visual comparison between a conventional random coded aperture and the one designed by the D-HAN. We calculated 2D cross-correlation between different compressed measurements using these two
coded apertures in 100,000 executions [referred to as “Optimized” and “Random” in Fig. 2(b)], where the compressed measurements were obtained by developing a data augmentation. A unit-slope, zero-offset linear shearing function was used to compute these compressed measurements. We also included two well-known and ill-posed scenarios: no coded aperture but with linear shearing [referred to as “No CA” in Fig. 2(b)] and no coded aperture and no shearing [referred to as “No CA and no shearing” in Fig. 2(b)]. The result shows high cross-correlation values of 0.95 and 0.90 for the scenarios without using a coded aperture. Conventional random coded aperture reduces the value of 2D cross-correlation to 0.80. The optimized coded aperture further decreases the value to 0.65, which verifies its superior performance in mitigating the issue of ambiguity in compressed measurements generated from closely related datacubes [72]–[74]. Furthermore, we analyzed these two coded apertures in the spatial frequency domain (Supplementary Note 2). As shown in Fig. S2, the spatial frequency spectra show that the optimized coded aperture prioritizes the high-frequency features and suppresses low-frequency components. To better comprehend the influence of this characteristic, we calculated the 2D cross-correlation average value for the optimized coded aperture and the random one, resulting in 0.4903 and 0.5030, respectively. These results echo the claims presented in Fig. 2(b) as well as literature [75]–[77]. Thus, the optimized coded aperture leads to a more incoherent encoding operation, which contributes to better image quality in reconstruction.

For the second analysis, we tested four shearing functions with weights of $a_0 = 0$, $a_1 \in \{0.8, 0.85, 0.9, 0.95\}$, $a_2 = -0.0072$, and $a_3 = 0.0013$ using 400 different scenes. The relative errors between the ground truth (GT) and the corresponding estimations were calculated to quantify accuracy, which yielded an average error and a standard deviation of 0.16% and 0.21%, respectively [Fig. 2(c)]. To further validate D-HAN’s accurate sensing of the shearing operation, we compared peak signal-to-noise ratios (PSNRs) of the reconstructions of these scenes by using these four shearing functions with and without shearing estimation [Fig. 2(d)]. The D-HAN gains up to 8 dB and reduces the standard deviation by up to 1 dB on average with shearing estimation [refer to as “On” in Fig. 2(d)] with respect to the reconstruction without shearing estimation [refer to as “Off” in Fig. 2(d)] trained using a linear shearing function (i.e., $a_k = 0$) and a conventional random coded aperture, which is the configuration currently used in many state-of-the-art SCI reconstruction algorithms.

To further investigate the effectiveness of D-HAN, we carried out an ablation analysis of the coded aperture and the shearing function (Supplementary Note 3 and Fig. S3). The results, echoing Fig. 2(c)–(e), show the necessity of implementing shearing estimation status and the optimized coded aperture for high-quality reconstruction. Moreover, D-HAN’s reconstruction shows comparable performance with the GT where the shearing function is made known to the reconstruction algorithm (Table S2). This analysis demonstrates D-HAN’s high-quality reconstruction without pre-calibration. This robustness also proves D-HAN’s shearing estimation to be a viable strategy for SCI.

To further compare the image quality, a representative reconstructed result of the “Bird” scene (with $a_1 = 0.8$) is shown in Fig. 2(e). In particular, three selected frames of the GT
implementation of D-HAN in SCI systems

A. Experimental Demonstration in Temporal SCI

To demonstrate the D-HAN’s capability of multi-faceted supervision in SCI, we conducted experimental validations using two representative modalities. First, for temporal SCI, we implemented the D-HAN in a compressed optical-streaking ultrahigh-speed photography (COSUP) system [28] [Fig. 3(a)]. To generate an ultrahigh-speed scene that matched the COSUP’s speed, binary-pattern sequences, multiplied with the optimized coded aperture [i.e., Fig. 2(a)], were displayed by a digital micromirror device [DMD, (Digital Light Innovations, D4100 7’’ VIS XGA)] at 20 kfps. Each encoding pixel occupied 4×4 DMD pixels. A collimated 532-nm laser beam from a continuous-wave laser (M-squared, EQUINOX) shone onto the DMD at an incident angle of ~24°. The spatially encoded scene was imaged by Lens 1 (Thorlabs, AC254-075-A) to the intermediate image plane with a 1× magnification ratio. An iris (Thorlabs, ID20) was placed on the back focal plane of Lens 1 for spatial filtering. Then, the intermediate image was relayed onto a CMOS camera (FLIR, GS3-U3-23S6M-C) by a 4× and a GS (Thorlabs, GVS-001) that was placed at its Fourier plane for temporal shearing. Finally, the spatially encoded and temporally sheared scene was spatiotemporally integrated by the CMOS camera to form a compressed measurement with a size of up to 208×307 binned CMOS pixels (5.86 µm × 5.86 µm pixel size, 4×4 binning). After data acquisition, the D-HAN reconstructed a datacube with the size of $N_x \times N_y = 208 \times 208$ pixels and a sequence depth of $N_z = \{50, 100\}$ frames. The imaging speed, controlled by the CMOS camera (5 ms exposure) and the sinusoidal signal (60 Hz frequency, 1.12 V peak-to-peak voltage) applied to the GS (with a deflection rate of 4 °/V), was 20 kfps.

To demonstrate D-HAN’s ability to estimate the shearing function, we imaged an animation of three moving balls. The radii of these balls were also slightly varied in time. The sinusoidal signal applied to the GS was intentionally set to a nonlinear zone. The shearing weights, verified by calibration (detailed in Supplementary Note 5 and Fig. S5), were $a_0 = 0$, $a_1 = 0.175$, $a_2 = -0.007$, and $a_3 = 3.6 \times 10^{-5}$. The D-HAN reconstructed a $N_x \times N_y \times N_z = 208 \times 208 \times 100$ datacube and estimated the shearing operation. The full evolution is included in Video S2. Fig. 3(b) shows a comparison of the GT with reconstructions of the D-HAN without and with shearing estimation. Without shearing estimation, the reconstruction fails to recover the balls when the shearing function deviates from the linear function at the end of the sequence. In contrast, the D-HAN accurately recovers the position, size, and shape of each ball in the entire sequence. To quantitatively analyze D-HAN’s reconstruction, the centroids of all three balls [labeled in the first frames of Fig. 3(b)] were traced. The time courses of relative errors of each centroid in the reconstruction of the D-HAN without and with shearing estimation are presented in Fig. 3(c). The results show that the D-HAN produces a higher reconstructed image accuracy than the D-HAN without shearing estimation, especially in the duration of 4–5 ms. Finally, Fig. 3(d) presents the precise matching between the shearing function output from

**Fig. 3.** Single-shot imaging of fast-moving ball patterns using D-HAN-supervised temporal SCI. (a) Schematic of the experimental setup. (b) Comparison of the GT with the reconstructions by the D-HAN without and with shearing estimation. (c) Time courses of relative errors of each ball’s centroid traced by using the D-HAN without and with shearing estimation. (d) Comparison of the estimated shearing function with the GT and the estimation output from the D-HAN. A linear function is also plotted as a reference.

are compared with their corresponding frames reconstructed by the D-HAN without and with shearing estimation. An RGB composited version is also included to highlight the changes in the third dimension. The full movie is shown in Video S1. The D-HAN produces a higher reconstructed image quality than the result without shearing estimation, showing fewer artifacts in the background and better preservation of the bird’s eye. These results reveal that the incorporation of modeling the nonlinear deviation in the shearing operation results in a more robust CNN with the potential to sense hardware’s behavior in the SCI system.

Finally, we compared D-HAN’s reconstruction performance with two state-of-the-art SCI-CNNs (i.e., the TSA-net and DGSM method [50], [51]) and one analytical-modeling-based method (i.e., PnP-ADMM [2]) (Supplementary Note 4 and Fig. S4). For compressive measurements generated using a linear shearing function, the D-HAN is comparable to the best performance of these methods. In the case of a nonlinear shearing function, the D-HAN slightly decreases its performance while the competing methods manifest considerable degradations of averaged PSNR of 3–10 dB (Table S3). Finally, the D-HAN exhibits higher computational efficiency and less computational complexity, compared to the three competing algorithms. This result proves the D-HAN’s competitiveness with recent reconstruction methods in SCI.
the D-HAN to the GT obtained in the calibration. The average error in shearing function estimation was calculated to be 1%. The results show D-HAN’s robustness in accurately sensing the GS’s shearing behavior and recovering SCI measurements of nonlinear shearing.

To further demonstrate D-HAN with a real-world scene, we imaged a grayscale animation of a rotating spinner using the same experimental setup shown in Fig. 3(a). An error diffusion algorithm was implemented to convert each grayscale frame in the datacube to a binary DMD pattern [78], which was then multiplied with the optimized coded aperture. The diameter of the iris was limited to $\sim 1$ mm. In this way, a spatially encoded high-speed grayscale movie was generated using this DMD-based image projector [79]. Here, the sinusoidal signal applied to the GS was the same as the one used in imaging the moving-ball patterns. However, the camera’s exposure time was reduced from 5 ms to 2.5 ms to generate a slight deviation from linear shearing. The imaging speed was kept at 20 kfps while the sequence depth was reduced to $N_L = 50$ frames. Fig. 4(a) presents six selected frames of the GT and the corresponding reconstruction by the D-HAN (the full movie is included in Video S3). Time courses of the traced centroids on the three rings on the spinner [Fig. 4(b)–(c)] show average relative errors of $\sim 2\%$. Besides the accurate reconstruction, the D-HAN could also precisely output the shearing function [Fig. 4(d)] with an average error of $\sim 1\%$.

Finally, to show D-HAN’s ability to image complex scenes with moving subjects, we conducted simulations using ten videos captured with the commercially available cameras [67]. For this analysis, we followed the ablation structure presented in Supplemental Note 3. The full video of these results is shown in Video S4. These results were obtained from compressed measurements generated using an optimized coded aperture and a random one, when they are used with and without shearing estimation. This analysis also included the baseline result (i.e., with a known shearing function and the optimized coded aperture). D-HAN’s result has a comparable reconstruction quality to the baseline result, which demonstrates its high potential to reconstruct real-world dynamic scenes.

**B. Experimental Demonstration in Spectral SCI**

For spectral SCI, we implemented the D-HAN in a coded aperture snapshot spectral imaging (CASSI) system [Fig. 5(a)]. This system used a matched achromatic doublet pair (Thorlabs, MAP10100100-A) to image a hyperspectral scene onto a DMD (Texas Instruments, D4120) for spatial encoding. Each encoding pixel occupied $2 \times 2$ DMD pixels. Then, a standard relay lens (Thorlabs, AC254-100-A-ML) and a customized double Amici prism coupled to a rotation mount (Thorlabs, CRM1P) were used to image the spatially encoded and spectrally sheared scene to a monochrome CCD camera (AVT, Stingray F-145B). By calibrating the prism’s shearing effect (see Supplementary Note 6 and Fig. S6), we determined the GT of shearing weights be $a_0 = 0$, $a_1 = 1.22$, $a_2 = -0.0072$, and $a_3 = 0.0011$.

The experiment evaluated the “Tayrona” scene. The compressed measurement exhibited $256 \times 280$ binned CCD pixels (6.45 µm × 6.45 µm pixel size, $2 \times 2$ binning) in size. The reconstruction recovered a hyperspectral datacube of $N_x \times N_y \times N_L = 256 \times 256 \times 25$ in size. Fig. 5(b) depicts representative frames and the RGB composite of the reconstruction. Moreover, Fig. 5(c) shows the comparison of reconstructed spectra of two selected points [marked as $p_1$ and $p_2$ in the top-right panel in Fig. 5(b)] with respect to the GT measured with a spectrometer (Ocean Optics, Flame S-VIS-NIR-ES). The high resemblance of the two spectra is reflected by the small spectral angle mapper.
values of 2.9° for $p_1$ and 2.6° for $p_2$. Finally, Fig. 5(d) shows the estimated shearing function output from the D-HAN. The average error to the GT was calculated to be $\sim 1\%$.

Finally, to demonstrate the importance of the optimized coded aperture to D-HAN’s reconstruction, four additional scenes were acquired in the CASSI system using the random coded aperture and optimized one. As shown in the first row in Fig. S7, compared to the results from the random coded aperture, the compressed measurements generated by the optimized coded aperture show a more uniform pixel intensity distribution and a higher dynamic range, both of which promote an optimal distribution of the information on the sensor. Leveraging this advantage, the compressed measurements acquired using the optimized coded aperture are more robust to additive noise effects, which transfer to superior reconstructed image quality (Fig. S7 and Video S5).

VI. CONCLUSION

We have developed the D-HAN that provides multi-faceted supervision to an SCI system. This new E2E-CNN integrates optimization of the coded aperture, estimation of the shearing function, and reconstruction of the 3D datacube in one framework. Numerical simulations verify the feasibility of the proposed paradigm (Section III) with an optimized coded aperture applicable to diverse scenes, accurate estimation of nonlinear shearing functions, and enhanced image reconstruction capability. Finally, the D-HAN is successfully implemented in two representative SCI systems (i.e., COSUP and CASSI) for ultrahigh-speed imaging and hyperspectral imaging (Sections IV and V).

D-HAN’s architecture is highly suitable for solving the inverse problem in SCI. First, the ADMM-based inverse problem is implemented into the first three blocks in the D-HAN (Fig. 1) to promote a higher resonance between the D-HAN’s weights and the coded aperture design. In particular, a set of two parallel layers are built to model the direct sensing, transpose process, and inverse operations involved in SCI’s inverse problems [see (8)–(10)]. Moreover, matrix inversion is solved by taking advantage of the SWM formula and the SCI’s diagonal matrix property [see (8)]. The four dense layers, which estimate the shearing function directly from the compressed measurement, are connected to the sensing operator layer located at the end of the D-HAN. Finally, the SCI’s direct sensing operator is added at the end of D-HAN to regularize the estimated shearing weights. From this perspective, distinct from previous works, the D-HAN-supervised SCI can adapt the actual shearing function in retrieving the 3D datacube. This attractive feature, along with its optimized coded aperture, results in a more robust CNN with higher reconstructed image quality.

The D-HAN is expected to improve the reliability and stability of a variety of SCI modalities. Besides the enhanced performance of coded aperture and augmented capability of image reconstruction, its ability to sense the shearing function will provide on-time feedback on the necessity of system re-alignment, protocol re-calibration, and device replacement. More importantly, D-HAN will allow taking into consideration many experimental imperfections that are difficult to control, such as the flicker noise of a spatial light modulator in a CASSI system [74] and the electronic jitter of a GS in a COSUP system [26], which will boost its capability of capturing real-world scenes. Moreover, as a CNN based on deep unfolding methods, the D-HAN could be adapted to the shearing models of other SCI sensing modalities, such as coded-aperture compressive temporal imaging and dual-dispersive CASSI, by rewriting the sensing operators. Besides spectral and temporal SCI modalities [80]–[84], many other SCI methods in three-dimensional profilometry [85] and light field photography [86] could be benefited. Finally, the D-HAN’s multi-faceted supervision paradigm could be extended to other deep unfolding frameworks, such as the generalized alternating projection method [37] or the iterative shrinkage/thresholding algorithm [87]. All these topics are promising directions for future research.

REFERENCES


Miguel Marquez received the B.Sc. degree in computer science in 2015, and the M.Sc. degree in applied mathematics in 2018 from the Universidad Industrial de Santander, Bucaramanga, Colombia, where he is currently working toward the Ph.D. degree in physics. In 2021, he was an Intern with the Institut National de la Recherche Scientifique (INRS)–Université du Québec, Quebec City, QC, Canada. His main research interests include optical and computational imaging, compressive sensing, high dimensional signal processing, and optimization algorithms.

Yingming Lai received the B.Sc. degree in optoelectronics from the Southern University of Science and Technology, Shenzhen, China, in 2019, and the M.Sc. degree in energy and materials science from the Institut National de la Recherche Scientifique (INRS)–Université du Québec, Quebec City, QC, Canada in 2021. He is currently working toward the Ph.D. degree with the Laboratory of Applied Computational Imaging, INRS. His main research interests include computational imaging, compressive sensing, and ultrafast optical imaging.
Xianglei Liu received the B.Sc. degree in optical information science and technology from Tiangong University, Tianjin, China, in 2014 and the M.Sc. degree in optical engineering from the University of Chinese Academy of Sciences, Beijing, China, in 2017. He is currently working toward the Ph.D. degree in energy and materials science with the Institut National de la Recherche Scientifique (INRS)–Université du Québec, Quebec City, QC, Canada. His research interests include ultrafast imaging, compressive sensing, and optical system design.

Cheng Jiang received the B.E. degree in optoelectronics information science and engineering from Zhejiang University, Hangzhou, China, in 2018. He is currently working toward the Ph.D. degree in energy and materials science with the Institut National de la Recherche Scientifique (INRS)–Université du Québec, Quebec City, QC, Canada. His research interests include laser beam shaping, machine vision, and biomedical imaging.

Shian Zhang received the B.S. degree from Fujian Normal University, Fuzhou, China, in 2001, and the Ph.D. degree from East China Normal University, Shanghai, China, in 2006. Then he was with Spectra-Physics as a Senior Engineer for one year. In 2007, he joined East China Normal University as an Associate Professor and became a Full Professor in 2012. In 2008—2009, he was with Arizona State University, Tempe, AZ, USA, as a Postdoctoral. In 2015—2016, he was with Washington University, Saint Louis, MO, USA, as a Visiting Scholar. His research interests include ultrafast optical imaging, including compressed ultrafast photography and nonlinear optical microscopic imaging.

Henry Arguello (Senior Member, IEEE) received the B.Sc. Eng. degree in electrical engineering, the M.Sc. degree in electrical power from the Universidad Industrial de Santander, Bucaramanga, Colombia, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from the University of Delaware, Newark, DE, USA, in 2013. He is currently an Associate Professor with the Department of Systems Engineering, Universidad Industrial de Santander. In first semester 2020, he was a Visiting Professor with Stanford University, Stanford, CA, USA, funded by Fulbright. His research interests include high-dimensional signal processing, optical imaging, compressed sensing, hyperspectral imaging, and CI.

Jinyang Liang received the B.E. degree in optoelectronic engineering from the Beijing Institute of Technology, Beijing, China, in 2007, and the M.S. and Ph.D. degrees in electrical engineering from the University of Texas at Austin, Austin, TX, USA, in 2009 and 2012, respectively. From 2012 to 2017, he was a Postdoctoral Trainee with Washington University in St. Louis, MO, USA, and the California Institute of Technology, Pasadena, CA, USA. He is currently an Assistant Professor with the Institut National de la Recherche Scientifique (INRS)–Université du Québec, Quebec City, QC, Canada. His research interests include ultrafast imaging, computational optics, optical physics, and biophotonics.