PHOTONICS Research

Single-shot real-time compressed ultrahigh-speed imaging enabled by a snapshot-to-video autoencoder

Xianglei Liu,[1,†] João Monteiro,[1,†] Isabela Albuquerque,[1] Yingming Lai,[1] Cheng Jiang,[1] Shian Zhang,[2] Tiago H. Falk,[1,3] and Jinyang Liang[1,4]

[1]Centre Énergie Matériaux Télécommunications, Institut National de la Recherche Scientifique, Varennes, Québec J3X1S2, Canada
[2]State Key Laboratory of Precision Spectroscopy, East China Normal University, Shanghai 200062, China
[3]e-mail: falk@emt.inrs.ca
[4]e-mail: jinyang.liang@emt.inrs.ca

Received 11 February 2021; revised 13 October 2021; accepted 20 October 2021; posted 21 October 2021 (Doc. ID 422179); published 16 November 2021

Single-shot 2D optical imaging of transient scenes is indispensable for numerous areas of study. Among existing techniques, compressed optical-streaking ultrahigh-speed photography (COSUP) uses a cost-efficient design to endow ultrahigh frame rates with off-the-shelf CCD and CMOS cameras. Thus far, COSUP's application scope is limited by the long processing time and unstable image quality in existing analytical-modeling-based video reconstruction. To overcome these problems, we have developed a snapshot-to-video autoencoder (S2V-AE)—which is a deep neural network that maps a compressively recorded 2D image to a movie. The S2V-AE preserves spatiotemporal coherence in reconstructed videos and presents a flexible structure to tolerate changes in input data. Implemented in compressed ultrahigh-speed imaging, the S2V-AE enables the development of single-shot machine-learning assisted real-time (SMART) COSUP, which features a reconstruction time of 60 ms and a large sequence depth of 100 frames. SMART-COSUP is applied to wide-field multiple-particle tracking at 20,000 frames per second. As a universal computational framework, the S2V-AE is readily adaptable to other modalities in high-dimensional compressed sensing. SMART-COSUP is also expected to find wide applications in applied and fundamental sciences.

https://doi.org/10.1364/PRJ.422179

1. INTRODUCTION

2D optical visualization of transient phenomena in the actual time of the event's occurrence plays a vital role in the understanding of many mechanisms in biology, physics, and chemistry [1–3]. To discern spatiotemporal details in these phenomena, high-speed optical imagers are indispensable. Imaging speeds of these systems, usually determined by the frame rates of deployed CCD or CMOS cameras, can be further increased using novel sensor designs [4–6], new readout interfaces [7,8], and advanced computational imaging methods [9–11].

Among existing approaches, compressed ultrafast photography (CUP) [12–20] is an innovative coded-aperture imaging scheme [21,22] that integrates video compressed sensing [23] into streak imaging [24]. In data acquisition, a spatiotemporal (x, y, t) scene is compressively recorded by optical imaging hardware to a 2D snapshot. The ensuing reconstruction computationally recovers the datacube of the scene. Despite initially demonstrated using a streak camera, the concept of CUP was soon implemented in CCD and CMOS cameras in compressed optical-streaking ultrahigh-speed photography (COSUP) [25]. Compared to other single-shot ultrahigh-speed imaging modalities [26–29], COSUP is not bounded by the moving speed of piezo-stages [26,27] or the refresh rate of spatial light modulators [28,29]. As a cost-efficient system, COSUP has demonstrated single-shot transient imaging ability with a tunable imaging speed of up to 1.5 million frames per second (fps) based on an off-the-shelf CMOS camera with an intrinsic frame rate of tens of hertz.

Despite these hardware innovations, COSUP's video reconstruction has ample room for improvement. Existing reconstruction frameworks can be generally grouped into analytical-modeling-based methods and machine-learning-based methods [30]. Using the prior knowledge of the sensing matrix and the sparsity in the transient scene, the analytical-modeling-based methods reconstruct videos by solving an optimization problem that synthetically considers the image fidelity and the sparsity-promoted regularization. However, demonstrated methods, such as the two-step iterative shrinkage/thresholding (TwIST) algorithm [31], augmented Lagrangian algorithm [32], and an alternating direction method of a multiplier (ADMM) algorithm [29], undergo time-consuming processing that uses tens to hundreds of iterations. The excessively long reconstruction time strains these analytical-modeling-based methods from real-time (i.e., ≥16 Hz [33]) reconstruction, which excludes COSUP's application scope from tasks that need on-time feedback [34]. Moreover, the reconstructed video quality highly depends on the accuracy of prior knowledge and the empirical tuning of parameters.

To solve these problems, machine learning has become an increasingly popular choice. Instead of relying solely on prior knowledge, large amounts of training data are used for deep neural networks (DNNs) [35] to learn how to map an acquired snapshot back to a video. Upon the completion of training, DNNs can then execute non-iterative high-quality reconstruction during runtime. Thus far, DNNs that employ the architectures of the multilayer perceptrons (MLPs) [36,37] and the U-net [38–41] have shown promise for compressed video reconstruction. Nonetheless, MLPs, with fully connected structures, scale linearly with the dimensionality of input data [42]. Besides, the decomposition in the reconstruction process presumes that all information in the output video block is contained in a patch of the input image, which cannot always be satisfied [36,37]. As for the U-net, the reconstruction often starts with a pseudo-inverse operation to the input snapshot to accommodate the equality in dimensionality required by the original form of this network [43]. This initial step increases the reconstruction burden in computational time and memory. Moreover, akin to MPLs, U-net-based methods require slicing input data for reconstruction, which could cause the loss of spatial coherence [39]. Finally, inherent temporal coherence across video frames is often unconsidered in the U-net [44]. Because of these intrinsic limitations, videos reconstructed by the U-nets are often subject to spatiotemporal artifacts and a shallow sequence depth (i.e., the number of frames in the reconstructed video) [41].

Here, we propose a way to overcome these limitations using an autoencoder (AE), whose objective is to learn a mapping from high-dimensional input data to a lower-dimensional representation space, from which the original data is recovered [45]. The implementation of convolutional layers in AE's architecture provides a parameter-sharing scheme that is more efficient than MLPs. Besides, without relying on locality presumptions, deep AEs with convolutional layers can preserve the intrinsic coherence in information content. Furthermore, recent advances in combining AE with adversarial formulations [46] have allowed replacing the loss functions based on pixel-wise error calculation to settings where perceptual features are accounted for, which have enabled a more accurate capture of data distribution and increased visual fidelity [47]. In the particular case of training generative models [e.g., generative adversarial networks (GANs)] for natural scenes, recent advances have improved the reconstructed imaging quality by dividing the overall task into sub-problems, such as independent modeling of the foreground and background [48], separated learning of motion and frame content [49], and conditioning generation on optical flows [50]. Despite these advances, with popular applications in audio signal enhancement [51] and pattern recognition [52], AEs have been mainly applied to 1D and 2D reconstruction problems [53,54]. Thus, existing architectures of AEs cannot be readily implemented for video reconstruction in compressed ultrahigh-speed imaging.

To surmount these problems, we have developed a snapshot-to-video autoencoder (S2V-AE)—a new DNN that directly maps a compressively recorded 2D (x, y) snapshot to a 3D (x, y, t) video. This new architecture splits up the reconstruction process into two sub-tasks, each of which is trained individually to obtain superior quality in reconstructed videos. Implemented in compressed ultrahigh-speed imaging, such a video reconstruction framework enables the development of a single-shot machine-learning-assisted real-time (SMART) COSUP, which is applied to tracking multiple fast-moving particles in a wide field at 20,000 fps (20 kfps).

2. PRINCIPLE OF SMART-COSUP

The schematic of the SMART-COSUP system is shown in Fig. 1(a). Its operating principle contains single-shot data acquisition and real-time video reconstruction [Fig. 1(b)]. A dynamic scene, $I(x, y, t)$, is imaged by front optics onto a printed pseudo-random binary transmissive mask (Fineline Imaging, Inc., Colorado Springs, CO, USA) with encoding pixels of 25 μm × 25 μm in size. This spatial modulation operation is denoted by the operator **C**. The intensity distribution right after the encoding mask is expressed as

$$I_{\mathrm{c}}(x, y, t) = \sum_{j,k} I\left(\frac{x}{M_{\mathrm{f}}}, \frac{y}{M_{\mathrm{f}}}, t\right) C_{jk}\mathrm{rect}\left(\frac{x}{d_{\mathrm{e}}} - j, \frac{y}{d_{\mathrm{e}}} - k\right). \tag{1}$$

Here, M_{f} is the magnification of the front optics, C_{jk} denotes an element of a binary matrix representing the encoding pattern, j and k are matrix element indices, d_{e} is the size of encoding pixels on the mask, and $\mathrm{rect}(\cdot)$ represents the rectangular function.

Subsequently, the spatially modulated scene is relayed by a $4f$ imaging system that consists of a galvanometer scanner (GS, 6220 H, Cambridge Technology, Bedford, MA, USA) and two identical lenses (Lens 1 and Lens 2, AC254-075-A, Thorlabs, Inc., Newton, NJ, USA). The GS is placed at the Fourier plane of this $4f$ imaging system to conduct optical shearing in the x direction, denoted by the operator $\mathbf{S_o}$. The sheared image can be expressed as

$$I_{\mathrm{s}}(x, y, t) = I_{\mathrm{c}}(x - v_{\mathrm{s}}t, y, t), \tag{2}$$

where v_{s}, which denotes SMART-COSUP's shearing velocity, is calculated by $v_{\mathrm{s}} = \alpha V_{\mathrm{g}} f_2 / t_{\mathrm{g}}$. Here, $V_{\mathrm{g}} = 0.16$–0.64 V is the voltage added onto the GS, α is a constant that links V_{g} with GS' deflection angle with the consideration of the input waveform, $f_2 = 75$ mm is the focal length of Lens 2 in Fig. 1(a), and $t_{\mathrm{g}} = 50$ ms is the period of the sinusoidal signal added to the galvanometer scanner.

Finally, the dynamic scene is spatiotemporally integrated by a CMOS camera (GS3-U3-23S6M-C, Teledyne FLIR LLC, Wilsonville, OR, USA) to a 2D snapshot, denoted by the

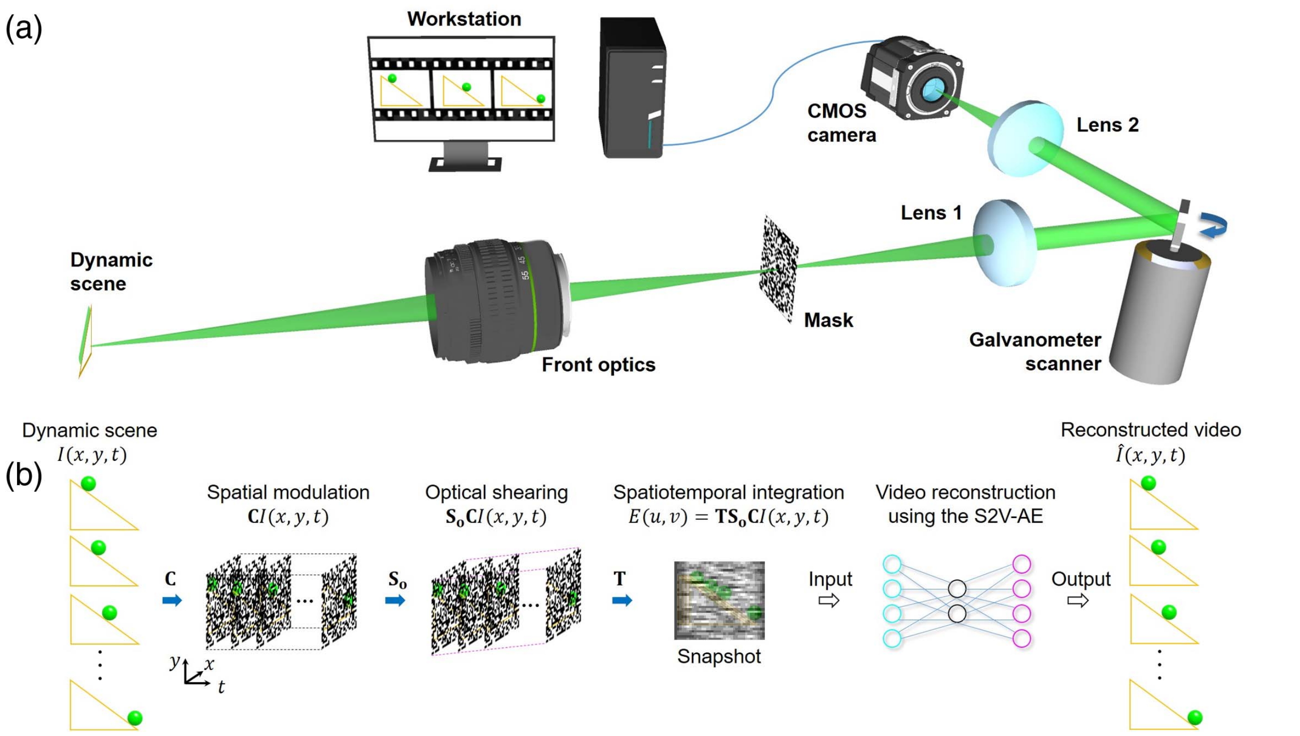

Fig. 1. Single-shot machine-learning assisted real-time (SMART) compressed optical-streaking ultrahigh-speed photography (COSUP). (a) System schematic. (b) Operating principle. S2V-AE, snapshot-to-video autoencoder.

operator **T**. The optical energy $E(m,n)$ measured at pixel (m,n), is

$$E(m,n) = \iiint I_s(x,y,t)\mathrm{rect}\left(\frac{x}{d_c} - m, \frac{y}{d_c} - n\right)\mathrm{d}x\mathrm{d}y\mathrm{d}t. \quad (3)$$

Here, m and n are the pixel indices in the x and y axes on the camera, and $d_c = 5.86\ \mu\mathrm{m}$ is the CMOS sensor's pixel size. From Eqs. (1)–(3), the forward model of SMART-COSUP is expressed by

$$E(m,n) = \mathbf{TS_o C}I(x,y,t). \quad (4)$$

In the ensuing real-time video reconstruction, the captured data is transferred to a workstation equipped with a graphic processing unit (RTX Titan, NVIDIA, Santa Clara, CA, USA). The S2V-AE retrieves the datacube of the dynamic scene in 60 ms. The frame rate of the SMART-COSUP system is derived from

$$r = \frac{v_s}{d_c}. \quad (5)$$

In this work, the reconstructed video has a frame rate of up to $r = 20$ kfps, a sequence depth of $N_t = 40$–100 frames, and a frame size of up to $N_x \times N_y = 256 \times 256$ pixels.

Compared to the previous hardware configuration [25], SMART-COSUP replaces the digital micromirror device (DMD), which functions as a 2D programmable blazed grating [55], with the transmissive mask for spatial modulation. This arrangement avoids generating a large number of unused diffraction orders, preventing a limited modulation efficiency to unblazed wavelengths, and eliminating intensity loss from the reflection from its cover glass as well as by its interpixel gap. In addition, the printed mask is illuminated at normal incidence, making it fully conjugated with both the object and the camera. Thus, the SMART-COSUP system presents a simpler, economical, and compact design with improved light throughput of the system and image quality of the captured snapshot.

3. STRUCTURE OF S2V-AE

The architecture of S2V-AE consists of an encoder and a generator [Fig. 2(a)]. The encoder (denoted as $\mathcal{E}$) converts a 2D snapshot to a series of low-dimensional latent vectors that represent particular features of the dynamic scene under study. As shown in Fig. 2(b), its architecture consists of five convolutional layers, a bidirectional long short-term memory (Bi-LSTM) recurrent layer [56], and a fully connected layer. In the convolutional layers, each convolution operation is followed by batch normalization (BN) [57] along with rectified linear unit (ReLU) activation [58]. The number of channels of feature maps, denoted by N, decreases from a preset value (512 in our experiments) to N_t. Then, the feature map, output by the last convolutional layer, is reshaped from a tensor into N_t vectors, all of which are fed into the Bi-LSTM recurrent blocks with the fully connected layer to model temporal coherence. The outputs of the encoder, referred to as latent vectors, are then input to the generator (denoted as $\mathcal{G}$). In particular, each latent vector is reshaped back to a tensor, which is fed into the generator to reconstruct one frame in the video. As shown in Fig. 2(c), the architecture of the generator consists of seven transposed convolutional layers. Akin to the encoder, BN and ReLU activation are employed after each transposed convolution, whose preset number of channels decreases from

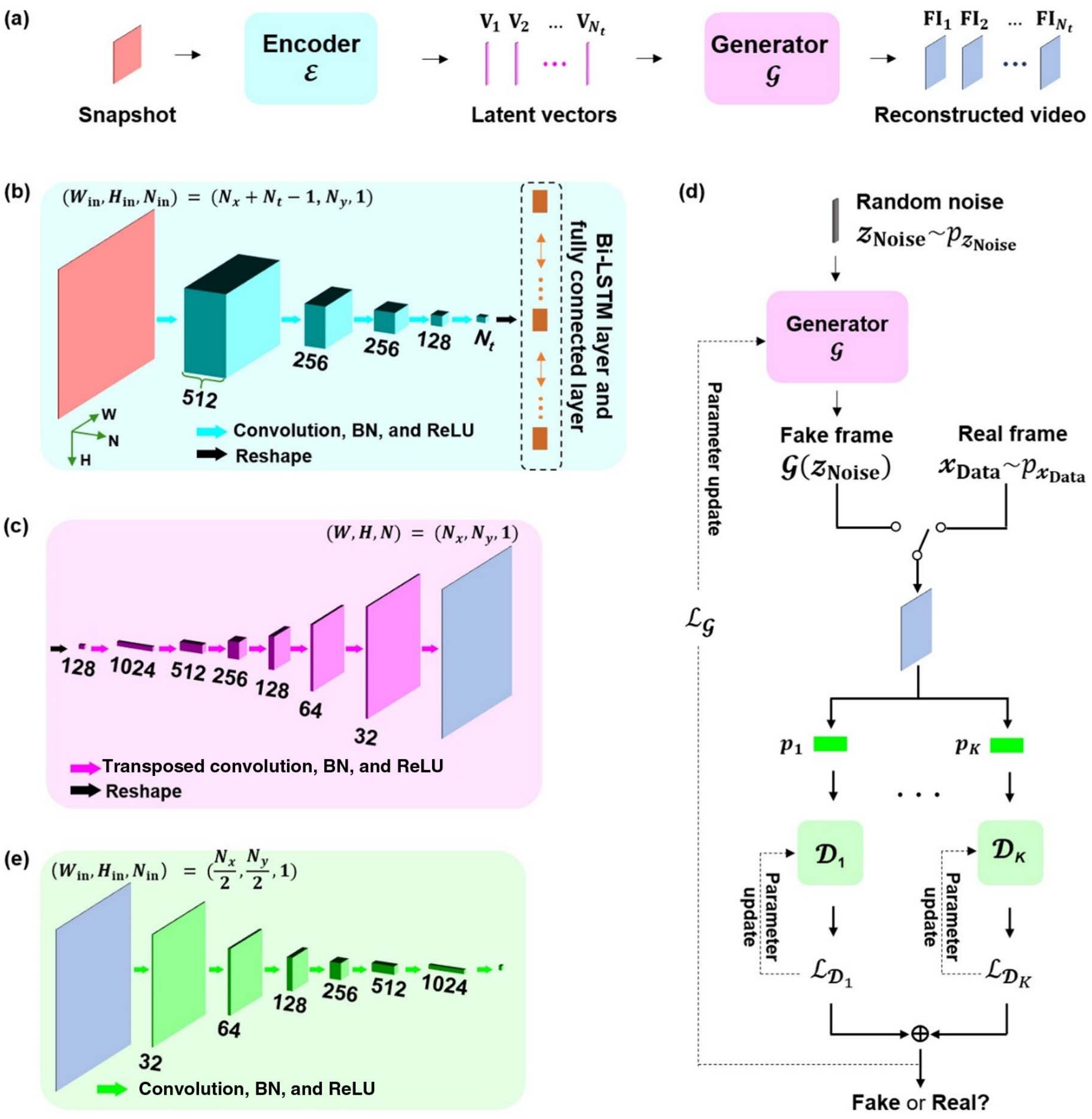
(a)
Snapshot
Encoder
$\mathcal{E}$
V_1 V_2 ... V_{N_t}
Latent vectors
Generator
$\mathcal{G}$
FI_1 FI_2 ... FI_{N_t}
Reconstructed video
(b)
$(W_{\text{in}}, H_{\text{in}}, N_{\text{in}}) = (N_x + N_t - 1, N_y, 1)$
512
256
256
128
N_t
W
N
H
Convolution, BN, and ReLU
Reshape
Bi-LSTM layer and fully connected layer
(c)
$(W, H, N) = (N_x, N_y, 1)$
128
1024
512
256
128
64
32
Transposed convolution, BN, and ReLU
Reshape
(d)
Random noise
$z_{\text{Noise}} \sim p_{z_{\text{Noise}}}$
Generator
$\mathcal{G}$
Parameter update
Fake frame
$\mathcal{G}(z_{\text{Noise}})$
Real frame
$x_{\text{Data}} \sim p_{x_{\text{Data}}}$
$\mathcal{L}_{\mathcal{G}}$
p_1
p_K
. . .
$\mathcal{D}_1$
$\mathcal{D}_K$
Parameter update
Parameter update
$\mathcal{L}_{\mathcal{D}_1}$
$\mathcal{L}_{\mathcal{D}_K}$
⊕
Fake or Real?
(e)
$(W_{\text{in}}, H_{\text{in}}, N_{\text{in}}) = (\frac{N_x}{2}, \frac{N_y}{2}, 1)$
32
64
128
256
512
1024
Convolution, BN, and ReLU

Fig. 2. Snapshot-to-video autoencoder (S2V-AE). (a) General architecture. FI, frame index. (b) Architecture of encoder showing the generation of latent vectors from a compressively recorded snapshot. Bi-LSTM, bidirectional long short-term memory; BN, batch normalization; ReLU, rectified linear unit; W, H, and N, output dimensions; W_{in}, H_{in}, and N_{in}, input dimensions. (c) Architecture of the generator showing the reconstruction of a single frame from one latent vector. (d) Generative adversarial networks (GANs) with multiple discriminators $\{\mathcal{D}_k\}$. $\mathcal{L}_{\mathcal{D}_k}$, the loss function of each discriminator; $\mathcal{L}_{\mathcal{G}}$, the loss function of the generator; and $\{p_k\}$, random projection with a kernel size of [8,8] and a stride of [2,2]. (e) Architecture of each discriminator.

1024 to 1. Each latent vector is processed by the generator to a frame of $N_x \times N_y$ in size. The composition of N_t such frames produces the reconstructed video.

The training of the encoder and the generator in the S2V-AE is executed sequentially. Training data are generated on the fly. The details of the training data collection and the training procedure are described in our open source code (see the link in Disclosure). Additional data, not included in its training phase, are used for evaluation. The generator is first trained under the setting of a GAN with multiple discriminators to ensure sufficient diversity. In brief, a random noise vector $\boldsymbol{z}_{\text{Noise}}$, sampled from a prior distribution $p_{\boldsymbol{z}_{\text{Noise}}}$ (i.e., $\boldsymbol{z}_{\text{Noise}} \sim p_{\boldsymbol{z}_{\text{Noise}}}$), is input to the generator to produce a fake frame $\mathcal{G}(\boldsymbol{z}_{\text{Noise}})$ that is expected to have visual similarity with the real frame $\boldsymbol{x}_{\text{Data}}$ with an implicit data distribution $p_{\boldsymbol{x}_{\text{Data}}}$ (i.e., $\boldsymbol{x}_{\text{Data}} \sim p_{\boldsymbol{x}_{\text{Data}}}$). The fake or real data are judged by $K = 40$ discriminators [Fig. 2(d)]. In each such discriminator, the data are first projected by a random matrix (denoted by p_k, where $k = 1, 2, \ldots, K$) to lower dimensions. Then, each discriminator (denoted as $\mathcal{D}_k$) converts the input to a number that is expected to be high for a real frame and low for a fake frame. Each discriminator, corresponding to a binary classifier as schematically shown in Fig. 2(e), consists of seven convolutional layers with the numbers of channels ranging from 1024 to 1. The loss functions of each discriminator $\{\mathcal{D}_k\}$ $(k = 1, 2, \ldots, K)$ (denoted by $\mathcal{L}_{\mathcal{D}_k}$) and the generator (denoted by $\mathcal{L}_{\mathcal{G}}$) are calculated by

$$\mathcal{L}_{\mathcal{D}_k} = -\mathbb{E}_{x_{\text{Data}}\sim p_{x_{\text{Data}}}}[\log(\mathcal{D}_k(x_{\text{Data}_k}))] \\ -\mathbb{E}_{\boldsymbol{z}_{\text{Noise}}\sim p_{z_{\text{Noise}}}}[\log(1-\mathcal{D}_k(\mathcal{G}(\boldsymbol{z}_{\text{Noise}})_k))], \tag{6}$$

$$\mathcal{L}_{\mathcal{G}} = -\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{z}_{\text{Noise}}\sim p_{z_{\text{Noise}}}}[\log(\mathcal{D}_k(\mathcal{G}(\boldsymbol{z}_{\text{Noise}})_k))]. \tag{7}$$

Here, $\mathcal{L}_{\mathcal{D}_k}$ corresponds to the cross-entropy loss [59]. After the random projection $\{p_k\}$, the input to each discriminator is either x_{Data_k} or $\mathcal{G}(\boldsymbol{z}_{\text{Noise}})_k$. Note that each discriminator is trained on two mini-batches of samples (i.e., real frames and fake frames). The notations $\mathbb{E}_{x_{\text{Data}}\sim p_{x_{\text{Data}}}}[\cdot]$ and $\mathbb{E}_{\boldsymbol{z}_{\text{Noise}}\sim p_{z_{\text{Noise}}}}[\cdot]$ indicate taking the expectations over the distribution $x_{\text{Data}} \sim p_{x_{\text{Data}}}$ and $\boldsymbol{z}_{\text{Noise}} \sim p_{z_{\text{Noise}}}$, respectively. These loss functions are estimated over mini-batches to compute the gradients of losses for each parameter update. Moreover, training iterations are such that each discriminator is first updated in the descent direction of its corresponding loss and then $\mathcal{L}_{\mathcal{G}}$'s gradients are used to update the generator's parameters. The described training game is expected to converge to equilibrium (i.e., no player can improve without changing the other player), which is not guaranteed to occur in highly non-convex cases, such as in the training of neural networks. However, the results found in practice in our setting are satisfactory. Successful training of the generator will yield parameters that enable its outputs $\mathcal{G}(\boldsymbol{z}_{\text{Noise}})$ to resemble characteristics of the real data. Leveraging this architecture, the goal of each discriminator is to distinguish the real data from the fake ones. The generator, by contrast, aims to fool all discriminators by learning how to produce frames as close as possible to real data. Parameters in the generator and discriminators are updated according to these loss functions [i.e., Eqs. (6) and (7)], which are minimized via gradient-descent-based optimization.

As the second step, the encoder is trained with the parameters of the generator fixed. The mean square error (MSE) between the recovered video $\mathcal{G}(\mathcal{E}(E))$ and the input data I is defined as the loss function denoted by $\mathcal{L}_{\mathcal{E}}$, i.e.,

$$\mathcal{L}_{\mathcal{E}} = \text{MSE}[\mathcal{G}(\mathcal{E}(E)), I]. \tag{8}$$

By minimizing $\mathcal{L}_{\mathcal{E}}$, the encoder learns how to correctly extract the latent vectors with temporal coherence from the 2D snapshot. The training of S2V-AE is finished when the reconstructed video quality stops increasing. Weight decay is employed during the training of the encoder to prevent the weights of the encoder from growing too large [60]. Hyperparameters to be trained in the encoder are defined through a search over a small grid of candidate values using cross-validation with reconstruction performance measured over a freshly generated batch of data examples.

In the training of both the generator and the encoder, the Adam optimization algorithm [61] was employed with a fixed learning rate, set to 10^{-3} for the training of the generator, and 3×10^{-4} for the training of the encoder. Adam's β_1 and β_2 parameters were set to 0.9 and 0.999 for the training of the generator and 0.5 and 0.9 for the training of the encoder, respectively. Data loading was set at training time so that both scenes and corresponding snapshots were generated on the fly, yielding a virtually infinite amount of training data. Once the completion of both the generator and the encoder training, the S2V-AE was employed to reconstruct dynamic scenes.

4. VALIDATION OF S2V-AE'S RECONSTRUCTION

To test the feasibility of S2V-AE, we simulated video reconstruction of flying handwritten digits [62]. Each dynamic scene had a size of $(N_x, N_y, N_t) = (64, 64, 40)$, which produced the snapshot of (103, 64) in size. Snapshots were generated using the forward model of SMART-COSUP [i.e., Eq. (1)]. Simulation results are summarized in Fig. 3. For the flying digits corresponding to 3, 5, and 7, six representative frames in the ground truth and the reconstructed videos are shown in Figs. 3(a)–3(c), respectively. The reconstructed videos are included in Visualization 1. To quantitatively assess the reconstructed video quality, we analyzed the peak SNR (PSNR) and the structural similarity index measure (SSIM) [63] frame by frame [Figs. 3(d) and 3(e)]. The average PSNR and SSIM of the reconstructed results are 22.9 dB and 0.93, respectively. These results demonstrate that the S2V-AE can accurately reconstruct dynamic scenes from compressively acquired snapshots.

Furthermore, to show that the S2V-AE possesses a more powerful ability in high-quality video reconstruction, we compared its performance to U-net, which is most popularly used in video compressed sensing [38]. In particular, this U-net featured a convolutional encoder–decoder architecture with residual connection and used the same loss function in Ref. [38]. To implement the optimal specifications of this U-net based technique, we used an approximate inverse operator $\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\Phi\Phi}^{\mathrm{T}})^{-1}$ to alleviate the burden in learning the forward model [38,39]. Particular to SMART-COSUP, we defined $\boldsymbol{\Phi} = \mathbf{TS_oC}$. Using the compressively recorded snapshot of the scene (i.e., E), the initialized input to the U-net is expressed as $\hat{I}_o = \boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\Phi\Phi}^{\mathrm{T}})^{-1}E$, which had the same (x, y, t) dimension as the ground truth. Both the initialized input and its ground truth were used to train the U-net to obtain a good inference ability for new training scenes generated on the fly. To compare the results between U-net and the S2V-AE, we reconstructed the flying digits of 3, 5, and 7 (see Visualization 2). Despite resembling a close trace of these moving digits to their ground truths, the U-net reconstruction failed to recognize the digits' spatial information in each frame. The limited feature extraction ability (imposed by the large number of frames in these scenes) and the requirement of high temporal coherence (broken by the fast and randomly moving traces of the digits in these scenes) are the two main reasons that attribute to the unsuccessful reconstruction using U-net. In contrast, benefiting from its two-step strategy that incorporates spatiotemporal coherence, the S2V-AE has shown superior performance, manifesting in the sharpness of reconstructed digits, the maintenance of high image quality over a large sequence depth, and the capability of handling randomly moving traces.

5. DEMONSTRATION OF SMART-COSUP

The proof-of-concept experiments of SMART-COSUP were conducted by imaging an animation of three bouncing balls, whose starting positions and moving directions were randomly

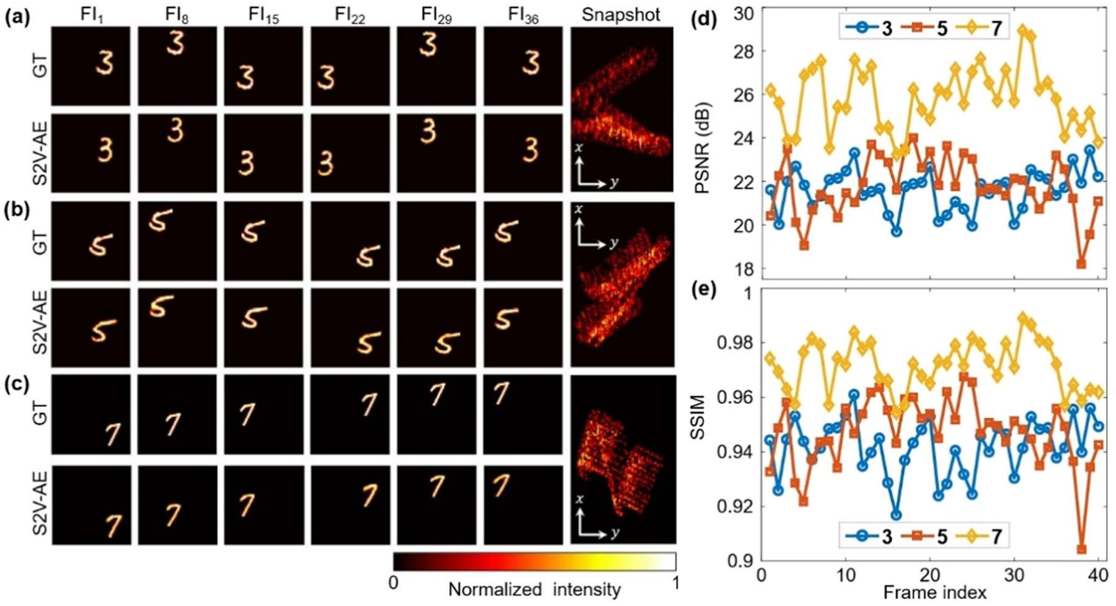

Fig. 3. Simulation of video reconstruction using the S2V-AE. (a) Six representative frames of the ground truth (GT, top row) and the reconstructed result (bottom row) of the handwritten digit "3." The snapshot is shown in the far right column. (b), (c) As (a), but showing handwritten digits 5 and 7. (d), (e) Peak SNR and the structural similarity index measure (SSIM) of each reconstructed frame for the three handwritten digits.

chosen (see Visualization 3). This scene had the size of $(N_x, N_y, N_t) = (256, 256, 100)$, which produced a snapshot with a size of (355, 256). To improve S2V-AE's reliability for experimentally captured data, we included various experimental conditions in the forward model to train the S2V-AE. In particular, an experimentally captured mask image was used for the spatial modulation operator. Moreover, with consideration of the noise level in the deployed CMOS camera, Gaussian noise with a standard deviation randomly selected from 10^{-1} to 10^{-4} was added into the training data to match the SNRs in acquired snapshots. Finally, distortion in the acquired snapshot was corrected by an established procedure [64,65].

This animation was displayed on a DMD (AJD 4500, Ajile Light Industries, Gloucester, ON, Canada) at a pattern refreshing rate of 5 kHz. The trajectories of all three balls were blind to the S2V-AE. A collimated laser beam from a 640 nm continuous-wave laser [MRL-III-640-50 mW, Changchun New Industries Optoelectronics Tech. Co., Ltd. (CNI), Changchun, China] illuminated this DMD at an incident angle of ∼24°, as shown in Fig. 4(a). A camera lens (Fujinon HF75SA1, Fujifilm Holdings Corp, Tokyo, Japan) was used as the front optics. The SMART-COSUP system imaged this event at 5 kfps. A captured 2D snapshot is shown as the inset in Fig. 4(a). Video reconstruction using the S2V-AE was compared to those using TwIST and plug-and-play (PnP)-ADMM with the BM3D denoiser [13]. In terms of the reconstruction speed, the execution of algorithms of S2V-AE, TwIST, and PnP-ADMM took 0.06, 5, and 220 s, respectively. Thus, the S2V-AE offers speed enhancements of ∼80× and ∼3700× compared to TwIST and PnP-ADMM, respectively. The S2V-AE also provides superior quality in the real-time reconstructed images. Figure 4(b) shows five representative frames of ground truth and their corresponding reconstructed results using the three methods (see the full reconstructed videos in Visualization 3). For both TwIST and PnP-ADMM, the reconstructed balls appear blurry and part of the balls are lost in certain frames. In contrast, the S2V-AE provides the best results, in which each ball is fully recovered with a clean background. To quantitatively compare these results, we plotted the PSNRs and SSIMs for all frames, as shown in Figs. 4(c) and 4(d). The reconstructed frames of S2V-AE have an average PSNR of 25.62 dB, superior to 15.09 dB of TwIST and 16.30 dB of PnP-ADMM. The results from the S2V-AE have an average SSIM of 0.94, considerably better than 0.76 of TwIST and 0.85 of PnP-ADMM. Moreover, we traced the centroids of each ball over time. To further evaluate the reconstruction's accuracy, we calculated the standard deviations of reconstructed centroids, which are shown in Table 1. On average, the S2V-AE improves the accuracy by ∼3× compared to the TwIST reconstruction and by ∼2× to the PnP-ADMM reconstruction.

Furthermore, the three centroids in each frame were used as vertices to build a triangle. Figures 4(e) and 4(f) show the time histories of the geometric center of this triangle generated from the results of the three reconstruction methods. The standard deviations in the x and y directions averaged over time were calculated as (25.4 μm, 17.0 μm), (14.8 μm, 14.5 μm), and (8.3 μm, 6.7 μm) for TwIST, PnP-ADMM, and S2V-AE, respectively. These results show that the S2V- AE has delivered superior performance in image quality and measurement accuracy.

6. APPLICATION OF SMART-COSUP TO MULTIPLE-PARTICLE TRACKING

To show the broad utility of SMART-COSUP, we applied it to tracking multiple fast-moving particles. In the setup, white

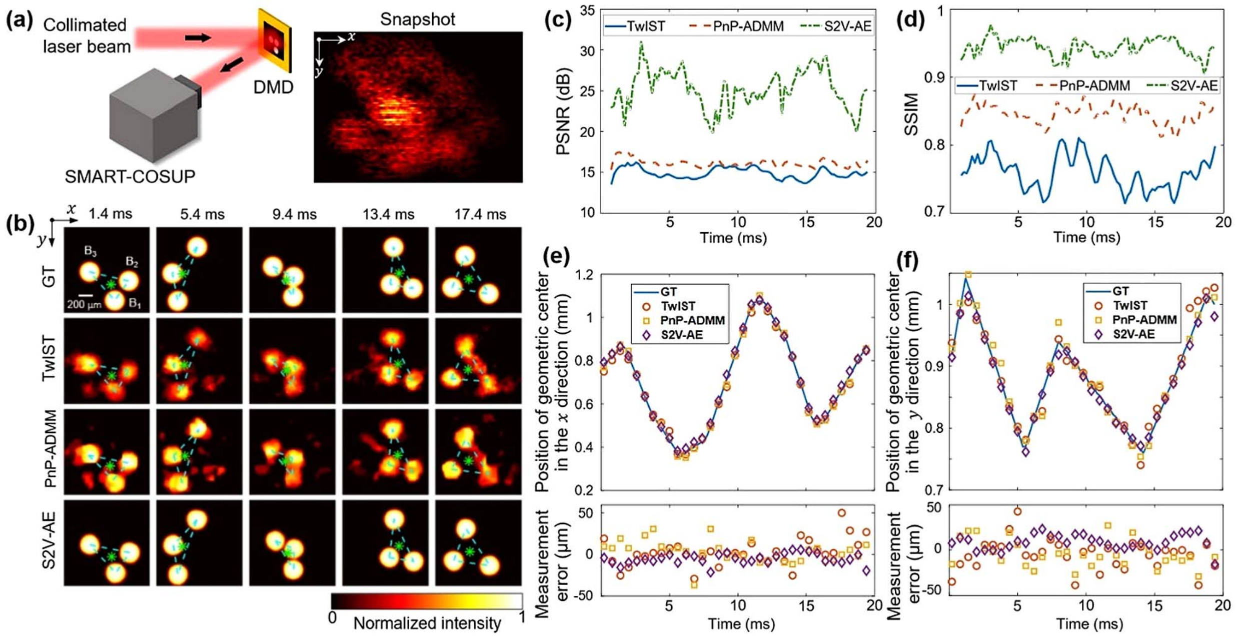

Fig. 4. SMART-COSUP of animation of bouncing balls at 5 kfps. (a) Experimental setup. DMD, digital micromirror device. Inset: an experimentally acquired snapshot. (b) Five representative frames with 4 ms intervals in the ground truth (GT) and the videos reconstructed by TwIST, PnP-ADMM, and S2V-AE, respectively. Centroids of the three balls are used as vertices to build a triangle (delineated by cyan dashed lines), whose geometric center is marked with a green asterisk. (c), (d) PSNR and SSIM at each reconstructed frame. (e) Comparison of the positions of the geometric center between the GT and the reconstructed results in the *x* direction. (f) As (e), but showing the results in the *y* direction.

Table 1. Standard Deviations of Reconstructed Centroids of Each Ball Averaged over Time (Unit: μm)

	1		2		3		
Algorithm	*x*	*y*	*x*	*y*	*x*	*y*	**Mean**
TwIST	37.5	36.3	39.4	35.7	43.2	34.9	37.8
PnP-ADMM	27.6	26.2	25.6	25.3	28.6	30.5	27.3
S2V-AE	15.0	12.3	11.0	12.6	15.3	16.0	13.7

microspheres were scattered on a surface that rotated at 6800 revolutions per minute [Fig. 5(a)]. The 640 nm continuous-wave laser was used to illuminate the rotating microspheres at an incident angle of ~50°. To visualize the beads' continuous motion while capturing a sufficiently long trace, the scattered light was captured by the SMART-COSUP system at 20 kfps. An objective lens (CF Achro 4×, Nikon Corp., Tokyo, Japan) was used as the front optics. Figure 5(b) shows a static image of three microspheres (marked as M_1–M_3) around the rotation center. Figure 5(c) shows a time-integrated image of this dynamic event acquired using the CMOS camera in the SMART-COSUP system at its intrinsic frame rate of 20 fps. Due to the low imaging speed, this time-integrated image cannot discern any spatiotemporal details. In contrast, imaging at 20 kfps, SMART-COSUP captures the trajectory of each microsphere, as shown in Visualization 4. The top image in Fig. 5(d) provides a color-coded overlay of five reconstructed frames (from 0.55 ms to 4.55 ms with a 1 ms interval), which are shown individually in the bottom row of Fig. 5(d). The rotation of M_1 and M_3 at two different radii [i.e., r_{M_1} and r_{M_3} labeled in Fig. 5(b)] is evident.

To quantitatively analyze these images, we calculated the time histories of *x* and *y* positions and the corresponding velocities of these microspheres. M_2, sitting at the rotation center, barely changes its position. The time histories of the positions and velocities of M_1 and M_3 follow sinusoidal functions expressed as

$$v_{i(x\,\mathrm{or}\,y)}(t) = \boldsymbol{\omega}_{\mathrm{F}} r_{\mathrm{M}_i} \sin(\boldsymbol{\omega}_{\mathrm{F}} t + \alpha_{i(x\,\mathrm{or}\,y)}). \tag{9}$$

Here, $i = 1$ or 3, $\boldsymbol{\omega}_{\mathrm{F}}$ denotes the angular velocity, whose value was preset at 0.71 rad/ms (i.e., 6800 rounds per minute), and r_{M_i} denotes the radius of each microsphere's rotation trajectory. In this experiment, $r_{\mathrm{M}_1} = 0.44$ mm and $r_{\mathrm{M}_3} = 0.64$ mm. $\alpha_{i(x\,\mathrm{or}\,y)}$ is the initial phase of the *i*th microsphere in either the *x* direction or the *y* direction. Thus, the theoretical linear speeds of M_1 and M_3 are 0.31 m/s and 0.45 m/s, respectively.

Based on the above analysis, we used single sinusoidal functions to fit the measured velocities. The fitted maximum velocities in the *x* direction and the *y* direction are 0.30 m/s and 0.32 m/s for M_1, and 0.46 m/s and 0.45 m/s for M_3. The fitted angular speeds in the *x* direction and the *y* direction are 0.71 rad/ms and 0.70 rad/ms for M_1 and 0.71 rad/ms and 0.72 rad/ms for M_3. The experimentally measured values have a good agreement with the preset angular speed of the rotating surface.

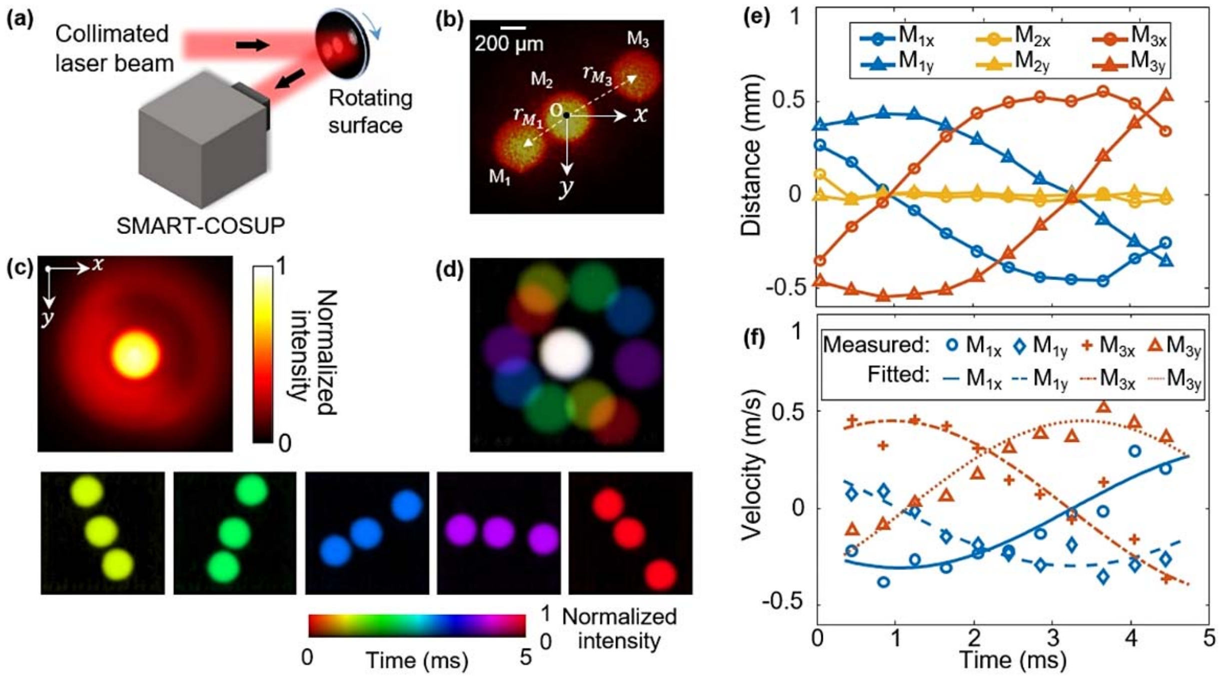

Fig. 5. SMART-COSUP of multiple-particle tracking at 20 kfps. (a) Experimental setup. (b) Static image of three microspheres (labeled as M_1–M_3) and the radii (labeled as r_{M_1} and r_{M_3}). (c) Time-integrated image of the rotating microspheres imaged at the intrinsic frame rate of the CMOS camera (20 fps). (d) Color-coded overlay (top image) of five reconstructed frames (bottom row) with a 1 ms interval. (e) Time histories of the microspheres' centroids. (f) Measured velocities of microspheres with fitting.

7. DISCUSSION AND CONCLUSIONS

The S2V-AE offers a new real-time reconstruction paradigm to compressed ultrahigh-speed imaging, as shown in Fig. 2(a). The new architecture of the encoder allows mapping a compressively recorded snapshot into a set of low-dimensional latent vectors. After that, the GAN-trained generator maps such latent vectors into frames of the reconstructed video. Using this scheme, the training procedure is divided into two distinct phases: to train a generative model of static frames and to train an encoding model aiming to sample from the generator. By doing so, unlike direct reconstruction approaches, high quality in frame-wise reconstruction can be ensured by the initially trained generator, while the encoding model needs to learn only how to query coherently across time. This scheme brings in benefits to the reconstructed videos in terms of both quality and flexibility. The encoder in S2V-AE preserves coherence in both space and time. Different from previous works [36,37,39], no artificial segmentation is conducted in the S2V-AE, which avoids generating artifacts due to the loss of spatial coherence. The S2V-AE also explicitly models temporal coherence across frames with the Bi-LSTM. Both innovations ensure artifact-free and high-contrast video reconstruction of sophisticated moving trajectories. Meanwhile, the S2V-AE presents a flexible structure with a higher tolerance for input data. In particular, the generator, used in a PnP setting [66], is independent of the system's data acquisition, which is important for adaptive compressed sensing applications.

The multiple-discriminator framework implemented in the S2V-AE improves training diversity. While able to generate high-quality, natural-looking samples, generators trained under the framework of the GAN have known drawbacks that have to be accounted for at training time. Namely, mode collapse refers to cases where trained generators can generate only a small fraction of the data support [67]. Standard GAN settings do not account for the diversity of the generated data, but instead, the generator is usually rewarded if its outputs are individually close to the real data instances. As such, a large body of recent literature has tackled the mode collapse using different approaches to improve the diversity of the GAN generators [67,68]. Mode collapse is especially critical in the application we consider here. The generator in the S2V-AE must be able to generate any possible frame, which means being able to output images containing any objects (e.g., balls or digits) in any position. To ensure that the generator is sufficiently diverse, the S2V-AE implements the multiple-discriminator framework [69,70]. Moreover, each such discriminator is augmented with a random projection layer at its input. More random views of the data distribution aid the generator in producing results that are approximate to the real data distribution.

The S2V-AE enables the development of SMART-COSUP. This new technique has demonstrated the largest sequence depth (i.e., 100 frames) in existing DNNs-based compressed ultrahigh-speed imaging methods [36–41]. The sequence depth, as a tunable parameter, could certainly exceed 100 frames. In this aspect, the performance of the S2V-AE mainly depends on the encoder [Fig. 2(b)] since it needs to extract the same number of latent vectors as the sequence depth. Although a large sequence depth may bring in training instabilities due to vanishing/exploding gradients, our choice of the Bi-LSTM architecture in the S2V-AE could alleviate gradient-conditioning issues relative to standard recurrent neural networks [71]. Thus, we expect the limit of sequence depth to be up to 1000 frames in the current setup. Moreover, although we only experimentally demonstrated the 20 kfps imaging speed in this work, the S2V-AE could be extended to reconstruct videos with much higher imaging speeds. As shown in Eq. (5), SMART-COSUP's imaging speed is determined completely by the hardware. Regardless of the imaging speed, the operation of the S2V-AE—reconstruction of a 3D datacube from a 2D snapshot—remains the same. Moreover, considering the link between imaging speeds and SNRs, the successful reconstruction of snapshots with different SNRs during the training procedure, as discussed in Section 5, indicates S2V-AE's applicability to reconstruct videos with a wide range of imaging speeds. Furthermore, SMART-COSUP replaces the DMD with a printed transmissive mask. Despite being inflexible, the implemented pseudo-random binary pattern has better compatibility with diverse dynamic scenes, improves light throughput and image quality, as well as offers a simpler, more compact system arrangement. Along with its real-time image reconstruction, the SMART-COSUP system is advancing toward real-world applications.

In summary, we have developed the S2V-AE for fast, high-quality video reconstruction from a single compressively acquired snapshot. This new DNN has facilitated the development of the SMART-COSUP system, which has demonstrated single-shot ultrahigh-speed imaging of transient events in both macroscopic and microscopic imaging at up to 20 kfps with a real-time reconstructed video size of $(N_x, N_y, N_t) = (256, 256, 100)$. This system has been applied to multiple-particle tracking. Despite demonstrated only with the SMART-COSUP system, the S2V-AE could be easily extended to other modalities in compressed temporal imaging [19] and single-shot hyperspectral imaging [72,73]. Moreover, by implementing the variational AE [74], the dependence of the encoder on the sensing matrix could be further reduced. SMART-COSUP's ability to track multiple fast-moving particles in a wide field may enable new applications on particle imaging velocimetry [75] and flow cytometry [76]. All of these topics are promising research directions in the future.

Funding. Natural Sciences and Engineering Research Council of Canada (CRDPJ-532304-18, I2IPJ-555593-20, RGPAS-507845-2017, RGPIN-2017-05959); Canada Foundation for Innovation and Ministère de l'Économie et de l'Innovation du Québec (37146); Fonds de recherche du Québec–Nature et technologies (2019-NC-252960); Fonds de Recherche du Québec–Santé (267406, 280229); Ministère des Relations internationales et de la Francophonie du Québec; Compute Canada; Calcul Québec.

Acknowledgment. The authors thank Patrick Kilcullen for experimental assistance.

Disclosures. The authors declare no conflicts of interest. The software of the S2V-AE and representative data of SMART-COSUP can be downloaded at https://github.com/joaomonteirof/SMART_COUSP_Reconstruction.

[†]These authors contributed equally to this work.

REFERENCES

1. M. Kannan, G. Vasan, C. Huang, S. Haziza, J. Z. Li, H. Inan, M. J. Schnitzer, and V. A. Pieribone, "Fast, *in vivo* voltage imaging using a red fluorescent indicator," Nat. Methods **15**, 1108–1116 (2018).
2. M. Sasaki, A. Matsunaka, T. Inoue, K. Nishio, and Y. Awatsuji, "Motion-picture recording of ultrafast behavior of polarized light incident at Brewster's angle," Sci. Rep. **10**, 7638 (2020).
3. P. R. Poulin and K. A. Nelson, "Irreversible organic crystalline chemistry monitored in real time," Science **313**, 1756–1760 (2006).
4. K. Toru, T. Yoshiaki, K. Kenji, T. Mitsuhiro, T. Naohiro, K. Hideki, S. Shunsuke, A. Jun, S. Haruhisa, G. Yuichi, M. Seisuke, and T. Yoshitaka, "A 3D stacked CMOS image sensor with 16 Mpixel global-shutter mode and 2 Mpixel 10000 fps mode using 4 million interconnections," in *IEEE Symposium on VLSI Circuits* (2015), pp. C90–C91.
5. T. Etoh, V. Dao, K. Shimonomura, E. Charbon, C. Zhang, Y. Kamakura, and T. Matsuoka, "Toward 1Gfps: evolution of ultra-high-speed image sensors-ISIS, BSI, multi-collection gates, and 3D-stacking," in *IEEE IEDM* (2014), pp. 11–14.
6. T. York, S. B. Powell, S. Gao, L. Kahan, T. Charanya, D. Saha, N. Roberts, T. Cronin, N. Marshall, S. Achilefu, S. Lake, B. Raman, and V. Gruev, "Bioinspired polarization imaging sensors: from circuits and optics to signal processing algorithms and biomedical applications," Proc. IEEE **102**, 1450–1469 (2014).
7. D. Calvet, "A new interface technique for the acquisition of multiple multi-channel high speed ADCs," IEEE Trans. Nucl. Sci. **55**, 2592–2597 (2008).
8. M. Hejtmánek, G. Neue, and P. Voleš, "Software interface for high-speed readout of particle detectors based on the CoaXPress communication standard," J. Instrum. **10**, C06011 (2015).
9. G. Barbastathis, A. Ozcan, and G. Situ, "On the use of deep learning for computational imaging," Optica **6**, 921–943 (2019).
10. A. Ehn, J. Bood, Z. Li, E. Berrocal, M. Aldén, and E. Kristensson, "FRAME: femtosecond videography for atomic and molecular dynamics," Light Sci. Appl. **6**, e17045 (2017).
11. Z. Li, R. Zgadzaj, X. Wang, Y.-Y. Chang, and M. C. Downer, "Single-shot tomographic movies of evolving light-velocity objects," Nat. Commun. **5**, 3085 (2014).
12. D. Qi, S. Zhang, C. Yang, Y. He, F. Cao, J. Yao, P. Ding, L. Gao, T. Jia, J. Liang, Z. Sun, and L. V. Wang, "Single-shot compressed ultrafast photography: a review," Adv. Photon. **2**, 014003 (2020).
13. P. Wang, J. Liang, and L. V. Wang, "Single-shot ultrafast imaging attaining 70 trillion frames per second," Nat. Commun. **11**, 2091 (2020).
14. J. Liang, L. Zhu, and L. V. Wang, "Single-shot real-time femtosecond imaging of temporal focusing," Light Sci. Appl. **7**, 42 (2018).
15. Y. Lai, Y. Xue, C. Y. Côté, X. Liu, A. Laramée, N. Jaouen, F. Légaré, L. Tian, and J. Liang, "Single-shot ultraviolet compressed ultrafast photography," Laser Photon. Rev. **14**, 2000122 (2020).
16. J. Liang, P. Wang, L. Zhu, and L. V. Wang, "Single-shot stereo-polarimetric compressed ultrafast photography for light-speed observation of high-dimensional optical transients with picosecond resolution," Nat. Commun. **11**, 5252 (2020).
17. C. Yang, F. Cao, D. Qi, Y. He, P. Ding, J. Yao, T. Jia, Z. Sun, and S. Zhang, "Hyperspectrally compressed ultrafast photography," Phys. Rev. Lett. **124**, 023902 (2020).
18. J. Liang, C. Ma, L. Zhu, Y. Chen, L. Gao, and L. V. Wang, "Single-shot real-time video recording of a photonic Mach cone induced by a scattered light pulse," Sci. Adv. **3**, e1601814 (2017).
19. X. Liu, S. Zhang, A. Yurtsever, and J. Liang, "Single-shot real-time sub-nanosecond electron imaging aided by compressed sensing: analytical modeling and simulation," Micron **117**, 47–54 (2019).

20. L. Gao, J. Liang, C. Li, and L. V. Wang, "Single-shot compressed ultrafast photography at one hundred billion frames per second," Nature **516**, 74–77 (2014).
21. J. Liang and L. V. Wang, "Single-shot ultrafast optical imaging," Optica **5**, 1113–1127 (2018).
22. J. Liang, "Punching holes in light: recent progress in single-shot coded-aperture optical imaging," Rep. Prog. Phys. **83**, 116101 (2020).
23. J. Yang, X. Yuan, X. Liao, P. Llull, D. J. Brady, G. Sapiro, and L. Carin, "Video compressive sensing using Gaussian mixture models," IEEE Trans. Image Process. **23**, 4863–4878 (2014).
24. C. Wang, Z. Cheng, W. Gan, and M. Cui, "Line scanning mechanical streak camera for phosphorescence lifetime imaging," Opt. Express **28**, 26717–26723 (2020).
25. X. Liu, J. Liu, C. Jiang, F. Vetrone, and J. Liang, "Single-shot compressed optical-streaking ultra-high-speed photography," Opt. Lett. **44**, 1387–1390 (2019).
26. P. Llull, X. Liao, X. Yuan, J. Yang, D. Kittle, L. Carin, G. Sapiro, and D. J. Brady, "Coded aperture compressive temporal imaging," Opt. Express **21**, 10526–10545 (2013).
27. R. Koller, L. Schmid, N. Matsuda, T. Niederberger, L. Spinoulas, O. Cossairt, G. Schuster, and A. K. Katsaggelos, "High spatio-temporal resolution video with compressed sensing," Opt. Express **23**, 15992–16007 (2015).
28. D. Reddy, A. Veeraraghavan, and R. Chellappa, "P2C2: programmable pixel compressive camera for high speed imaging," in *IEEE CVPR* (2011), pp. 329–336.
29. Y. Liu, X. Yuan, J. Suo, D. J. Brady, and Q. Dai, "Rank minimization for snapshot compressive imaging," IEEE Trans. Pattern Anal. Mach. Intell. **41**, 2990–3006 (2018).
30. A. Lucas, M. Iliadis, R. Molina, and A. K. Katsaggelos, "Using deep neural networks for inverse problems in imaging beyond analytical methods," IEEE Signal Process. Mag. **35**, 20–36 (2018).
31. J. M. Bioucas-Dias and M. A. Figueiredo, "A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration," IEEE Trans. Image Process. **16**, 2992–3004 (2007).
32. C. Yang, D. Qi, F. Cao, Y. He, X. Wang, W. Wen, J. Tian, T. Jia, Z. Sun, and S. Zhang, "Improving the image reconstruction quality of compressed ultrafast photography via an augmented Lagrangian algorithm," J. Opt. **21**, 035703 (2019).
33. J. Hui, Y. Cao, Y. Zhang, A. Kole, P. Wang, G. Yu, G. Eakins, M. Sturek, W. Chen, and J.-X. Cheng, "Real-time intravascular photoacoustic-ultrasound imaging of lipid-laden plaque in human coronary artery at 16 frames per second," Sci. Rep. **7**, 1417 (2017).
34. M. Kreizer, D. Ratner, and A. Liberzon, "Real-time image processing for particle tracking velocimetry," Exp. Fluids **48**, 105–110 (2010).
35. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436–444 (2015).
36. M. Iliadis, L. Spinoulas, and A. K. Katsaggelos, "Deep fully-connected networks for video compressive sensing," Digit. Signal Process. **72**, 9–18 (2018).
37. M. Yoshida, A. Torii, M. Okutomi, K. Endo, Y. Sugiyama, R.-I. Taniguchi, and H. Nagahara, "Joint optimization for compressive video sensing and reconstruction under hardware constraints," in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 634–649.
38. M. Qiao, Z. Meng, J. Ma, and X. Yuan, "Deep learning for video compressive sensing," APL Photon. **5**, 030801 (2020).
39. Y. Ma, X. Feng, and L. Gao, "Deep-learning-based image reconstruction for compressed ultrafast photography," Opt. Lett. **45**, 4400–4403 (2020).
40. C. Yang, Y. Yao, C. Jin, D. Qi, F. Cao, Y. He, J. Yao, P. Ding, L. Gao, and T. Jia, "High-fidelity image reconstruction for compressed ultrafast photography via an augmented-Lagrangian and deep-learning hybrid algorithm," Photon. Res. **9**, B30–B37 (2021).
41. A. Zhang, J. Wu, J. Suo, L. Fang, H. Qiao, D. D.-U. Li, S. Zhang, J. Fan, D. Qi, and Q. Dai, "Single-shot compressed ultrafast photography based on U-net network," Opt. Express **28**, 39299–39310 (2020).
42. M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," Atmos. Environ. **32**, 2627–2636 (1998).
43. O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional Networks for Biomedical Image Segmentation* (Springer, 2015), pp. 234–241.
44. Z. Cheng, R. Lu, Z. Wang, H. Zhang, B. Chen, Z. Meng, and X. Yuan, "BIRNAT: bidirectional recurrent neural networks with adversarial training for video snapshot compressive imaging," in *ECCV* (2020), pp. 258–275.
45. M. Tschannen, O. Bachem, and M. Lucic, "Recent advances in autoencoder-based representation learning," arXiv:1812.05069 (2018).
46. A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug & play generative networks: conditional iterative generation of images in latent space," in *IEEE CVPR* (2017), pp. 4467–4477.
47. A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *PMLR International Conference on Machine Learning* (2016), pp. 1558–1566.
48. C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," Adv. Neural Inf. Process Syst. **29**, 613–621 (2016).
49. S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: decomposing motion and content for video generation," in *IEEE CVPR* (2018), pp. 1526–1535.
50. K. Ohnishi, S. Yamamoto, Y. Ushiku, and T. Harada, "Hierarchical video generation from orthogonal information: optical flow and texture," arXiv:1711.09618 (2017).
51. O. Plchot, L. Burget, H. Aronowitz, and P. Matejka, "Audio enhancing with DNN autoencoder for speaker recognition," in *IEEE ICASSP* (2016), pp. 5090–5094.
52. J. Yu, X. Zheng, and S. Wang, "A deep autoencoder feature learning method for process pattern recognition," J. Process Control **79**, 1–15 (2019).
53. M. A. Ranzato, C. Poultney, S. Chopra, and Y. L. Cun, "Efficient learning of sparse representations with an energy-based model," in *Advances in Neural Information Processing Systems* (2007), pp. 1137–1144.
54. P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion," J. Mach. Learn. Res. **11**, 3371–3408 (2010).
55. J. Liang, M. F. Becker, R. N. Kohn, and D. J. Heinzen, "Homogeneous one-dimensional optical lattice generation using a digital micromirror device-based high-precision beam shaper," J. Micro/Nanolithogr. MEMS MOEMS **11**, 023002 (2012).
56. X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF," arXiv:1603.01354 (2016).
57. S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning* (2015), pp. 448–456.
58. V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning* (2010), pp. 807–814.
59. Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in Neural Information Processing Systems* (2018), pp. 8778–8788.
60. A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Advances in Neural Information Processing Systems* (1992), pp. 950–957.
61. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980 (2014).
62. L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," IEEE Signal Process. Mag. **29**, 141–142 (2012).
63. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. Image Process. **13**, 600–612 (2004).

64. MathWorks, “Register images using registration estimator app,” https://www.mathworks.com/help/images/register-images-using-the-registration-estimator-app.html.
65. C. Jiang, P. Kilcullen, Y. Lai, T. Ozaki, and J. Liang, “High-speed dual-view band-limited illumination profilometry using temporally interlaced acquisition,” Photon. Res. **8**, 1808–1817 (2020).
66. X. Yuan, Y. Liu, J. Suo, and Q. Dai, “Plug-and-play algorithms for large-scale snapshot compressive imaging,” in *CVPR* (2020), pp. 1447–1457.
67. Z. Lin, A. Khetan, G. Fanti, and S. Oh, “PACGAN: the power of two samples in generative adversarial networks,” in *Advances in Neural Information Processing Systems* (2018), pp. 1498–1507.
68. A. Jolicoeur-Martineau, “The relativistic discriminator: a key element missing from standard GAN,” arXiv:1807.00734 (2018).
69. B. Neyshabur, S. Bhojanapalli, and A. Chakrabarti, “Stabilizing GAN training with multiple random projections,” arXiv:1705.07831 (2017).
70. I. Albuquerque, J. Monteiro, T. Doan, B. Considine, T. Falk, and I. Mitliagkas, “Multi-objective training of generative adversarial networks with multiple discriminators,” in *Proceedings of the 36th International Conference on Machine Learning* (2019), pp. 202–211.
71. P. Razvan, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in *Proceedings of the 30th International Conference on Machine Learning* (2013), pp. 1310–1318.
72. P. Ding, Y. Yao, D. Qi, C. Yang, F. Cao, Y. He, J. Yao, C. Jin, Z. Huang, L. Deng, L. Deng, T. Jia, J. Liang, Z. Sun, and S. Zhang, “Single-shot spectral-volumetric compressed ultrafast photography,” Adv. Photon. **3**, 045001 (2021).
73. Z. Meng and X. Yuan, “Perception inspired deep neural networks for spectral snapshot compressive imaging,” in *ICIP* (2021), pp. 2813–2817.
74. Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, “Variational autoencoder for deep learning of images, labels and captions,” in *Advances in Neural Information Processing Systems* (2016), pp. 2352–2360.
75. A. Ten Cate, C. H. Nieuwstad, J. J. Derksen, and H. E. A. Van den Akker, “Particle imaging velocimetry experiments and lattice-Boltzmann simulations on a single sphere settling under gravity,” Phys. Fluids **14**, 4012–4025 (2002).
76. N. Nitta, T. Sugimura, A. Isozaki, H. Mikami, K. Hiraki, S. Sakuma, T. Iino, F. Arai, T. Endo, Y. Fujiwaki, H. Fukuzawa, M. Hase, T. Hayakawa, K. Hiramatsu, Y. Hoshino, M. Inaba, T. Ito, H. Karakawa, Y. Kasai, K. Koizumi, S. Lee, C. Lei, M. Li, T. Maeno, S. Matsusaka, D. Murakami, A. Nakagawa, Y. Oguchi, M. Oikawa, T. Ota, K. Shiba, H. Shintaku, Y. Shirasaki, K. Suga, Y. Suzuki, N. Suzuki, Y. Tanaka, H. Tezuka, C. Toyokawa, Y. Yalikun, M. Yamada, M. Yamagishi, T. Yamano, A. Yasumoto, Y. Yatomi, M. Yazawa, D. Di Carlo, Y. Hosokawa, S. Uemura, Y. Ozeki, and K. Goda, “Intelligent image-activated cell sorting,” Cell **175**, 266–276 (2018).