PHOTONICS Research

Ten-mega-pixel snapshot compressive imaging with a hybrid coded aperture

Zhihong Zhang,[1,2,†] Chao Deng,[1,2,†] Yang Liu,[3] Xin Yuan,[4,6] Jinli Suo,[1,2,*] and Qionghai Dai[1,2,5]

[1]Department of Automation, Tsinghua University, Beijing 100084, China
[2]Institute for Brain and Cognitive Science, Tsinghua University, Beijing 100084, China
[3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[4]Westlake University, Hangzhou 310024, China
[5]Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
[6]e-mail: xyuan@westlake.edu.cn
*Corresponding author: jlsuo@tsinghua.edu.cn

Received 8 July 2021; revised 12 August 2021; accepted 12 August 2021; posted 19 August 2021 (Doc. ID 435256); published 26 October 2021

High-resolution images are widely used in our everyday life; however, high-speed video capture is more challenging due to the low frame rate of cameras working at the high-resolution mode. The main bottleneck lies in the low throughput of existing imaging systems. Toward this end, snapshot compressive imaging (SCI) was proposed as a promising solution to improve the throughput of imaging systems by compressive sampling and computational reconstruction. During acquisition, multiple high-speed images are encoded and collapsed to a single measurement. Then, algorithms are employed to retrieve the video frames from the coded snapshot. Recently developed plug-and-play algorithms made the SCI reconstruction possible in large-scale problems. However, the lack of high-resolution encoding systems still precludes SCI's wide application. Thus, in this paper, we build, to the best of our knowledge, a novel hybrid coded aperture snapshot compressive imaging (HCA-SCI) system by incorporating a dynamic liquid crystal on silicon and a high-resolution lithography mask. We further implement a PnP reconstruction algorithm with cascaded denoisers for high-quality reconstruction. Based on the proposed HCA-SCI system and algorithm, we obtain a 10-mega-pixel SCI system to capture high-speed scenes, leading to a high throughput of 4.6×10^9 voxels per second. Both simulation and real-data experiments verify the feasibility and performance of our proposed HCA-SCI scheme.

https://doi.org/10.1364/PRJ.435256

1. INTRODUCTION

Recent advances in machine vision with applications in robotics, drones, autonomous vehicles, and cellphones have brought high-resolution images into our daily lives. However, high-speed high-resolution videos are facing the challenge of low throughput due to the limited frame rate of existing cameras working at the high-resolution mode, although they have wide applications in various fields such as physical phenomena observation, biological fluorescence imaging, and live broadcast of sports. This is further limited by the memory, bandwidth, and power. Thus, we aim to address this challenge here by building a high-speed, high-resolution imaging system using compressive sensing. Specifically, our system captures the high-speed scene in an encoded way, thus maintaining the low bandwidth during capture. Next, reconstruction algorithms are employed to reconstruct the high-speed, high-resolution scenes to achieve high throughput. Note that although the idea of video compressive sensing has been proposed before, scaling it up to 10 mega pixels in spatial resolution presents the challenges of both hardware implementation and algorithm development. Figure 1 shows a real high-speed scene captured by our newly built camera.

While 10-mega-pixel lenses and sensors are both available, the main challenge for high-speed and high-resolution imaging lies in the deficient processing capability of current imaging systems. Massive data collected from high-speed high-resolution recording imposes dramatic pressure on the system's storage and transmission modules, thus making it impossible for long-time capturing. In recent decades, the boosting of computational photography provides researchers with creative ideas and makes breakthroughs in many imaging-related fields such as super-resolution [1–3], deblurring [4–6], and depth estimation [7–9]. Regarding the high throughput imaging, snapshot compressive imaging (SCI) has been proposed and become a widely used framework [10–12]. It aims to realize the reconstruction of high-dimensional data such as videos and hyper-spectral images from a single-coded snapshot captured

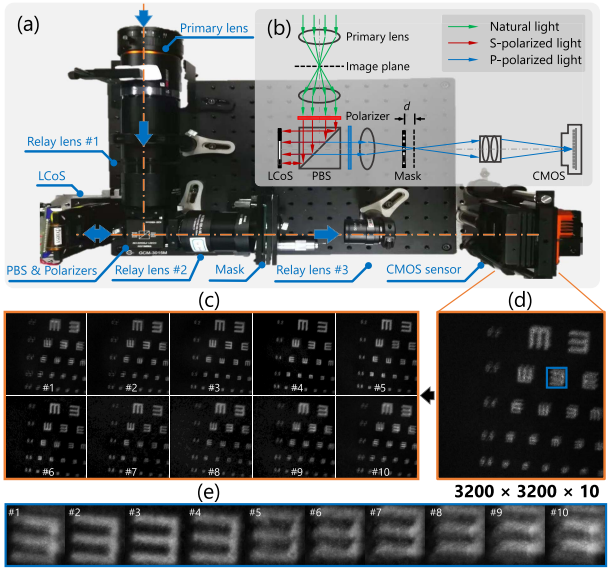

Fig. 1. Our 10-mega-pixel video SCI system (a) and the schematic (b). Ten high-speed (200 fps) high-resolution (3200 × 3200 pixels) video frames (c) reconstructed from a snapshot measurement (d), with motion detail in (e) for the small region in the blue box of (d). Different from existing solutions that only use an LCoS or a mask (thus with limited spatial resolution), our 10-mega-pixel spatio-temporal coding is generated jointly by an LCoS at the aperture plane and a static mask close to the image plane.

by a two-dimensional (2D) detector. A video SCI system is typically composed of an objective lens, a temporally varying mask, a monochrome or color sensor, and some extra relay lenses. During every single exposure, tens of temporal frames are modulated by corresponding temporal-variant masks and then integrated into a single snapshot. The high-dimensional data reconstruction in an SCI system can be formulated as an ill-posed linear model. Although different video SCI systems have been built [10,13–16], they are usually of low spatial resolution. By contrast, in this paper, we aimed to build a high-resolution video SCI system up to 10 mega pixels.

As mentioned above, the 10-mega-pixel lenses (including imaging lens and relay lens) and sensors are both commercialized products. Off-the-shelf reconstruction algorithms such as the recently developed plug-and-play (PnP) framework [17,18] can also meet our demands in most real applications. However, the 10-mega-pixel temporally varying mask is still an open challenge. Classical SCI systems usually rely on shifting masks produced by the lithography technique or dynamic patterns projected by the spatial light modulator (SLM), such as the digital micromirror device (DMD) or liquid crystal on silicon (LCoS), as temporally varying masks. The shifting mask scheme can provide high spatial resolution modulation, but it relies on the mechanical movement of the translation stage, which is inaccurate or unstable and can be hardly compact. For the masks generated by SLM or DMD, they can switch quickly with micro-mechanical controllers, but their resolution is generally limited to a mega-pixel level, which is difficult to scale up. To the best of our knowledge, there are few SCI systems that can realize 1000 × 1000 pixel-resolution imaging in real-world scenes [19]. In addition, typically, the resolution in prior works was mostly 256 × 256 [10] or 512 × 512 [15]. Therefore, it is desirable to build a high-resolution video SCI system for real applications. To bridge this research gap, in this paper, we came up with, to the best of our knowledge, a novel coded aperture imaging scheme, which leverages existing components to achieve the modulation up to 10 mega pixels. Our proposed modulation method is a hybrid approach using both a lithography mask and SLM. As depicted in Fig. 2, during the encoding capture process, two modulation modules, an LCoS and a lithography mask, are incorporated in different planes of the optical system. The LCoS with a low spatial resolution is placed at the aperture plane of the imaging system, to dynamically encode the aperture and change the directions of incident

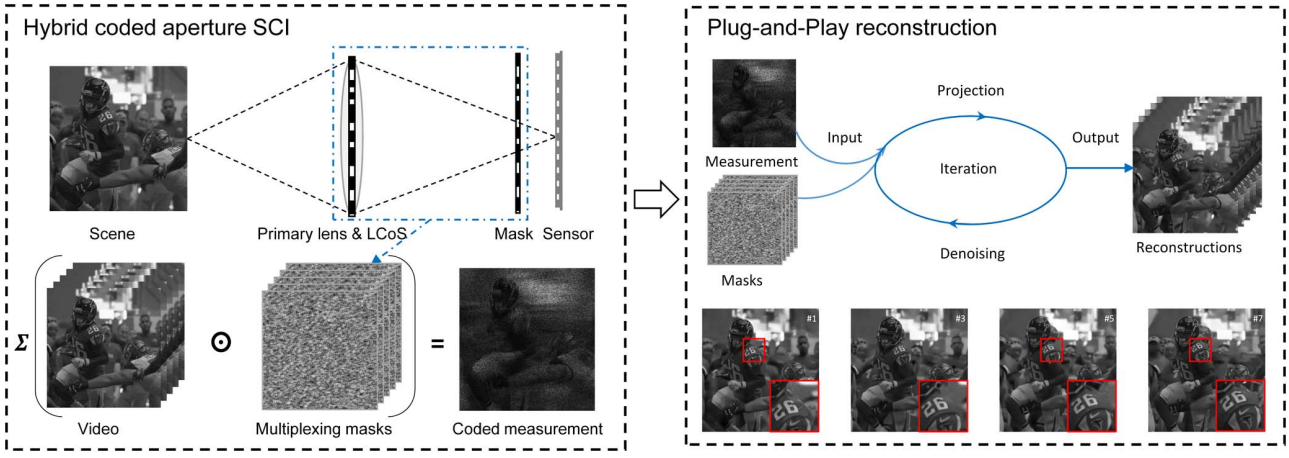

Fig. 2. Pipeline of the proposed large-scale HCA-SCI system (left) and the PnP reconstruction algorithms (right). Left: During the encoded photography stage, a dynamic low-resolution mask at the aperture plane and a static high-resolution mask close to the sensor plane work together to generate a sequence of high-resolution codes to encode the large-scale video into a snapshot. Right: In the decoding, the video is reconstructed under a PnP framework incorporating deep denoising prior and TV prior into a convex optimization (GAP), which leverages the good convergence of GAP and the high efficiency of the deep network.

lights. In addition, the static lithography mask with high spatial resolution is placed in front of the image plane of the primary lens, which can project different high-resolution patterns on the image plane. When the LCoS changes its patterns, the lights propagating toward the lithography mask will change their directions accordingly, thus leading to a different pattern. In this manner, we can implement dynamic modulation within one exposure time, up to 10 mega pixels. Specifically, this paper makes the following contributions.

- By jointly incorporating a dynamic LCoS and a high-resolution lithography mask, we proposed, to the best of our knowledge, a novel hybrid coded aperture snapshot compressive imaging (HCA-SCI) scheme, which can provide multiplexed shifting patterns to encode the image plane without physical movement of the mask.
- Inspired by the PnP algorithms for large-scale SCI in Ref. [17], we implement a reconstruction algorithm that involves cascading and series denoising processes of total variation (TV) [20] denoiser and learning-based FastDVDNet [21] denoiser. Simulation results show that the proposed algorithm can provide relative good reconstruction results in a reasonable time.
- Based on our proposed HCA-SCI scheme and the developed reconstruction algorithm, we build a 10-mega-pixel large-scale SCI system. Different compression rates of 6, 10, 20, and 30 are implemented, providing a reconstructed frame rate of up to 450 frames per second (fps) for a conventional camera operating at 15 fps, verifying the effectiveness of the proposed scheme in real scenarios.

2. RELATED WORK

SCI has been proposed to capture high-dimensional data such as videos and hyper-spectral images from a single low-dimensional coded measurement. The underlying principle is to modulate the scene at a higher frequency than the camera frame rate, and then, the modulated frames are compressively sampled by a low-speed camera. Following this, inverse algorithms are employed to reconstruct the desired high-dimensional data [12].

Various video SCI systems have been developed recently [10,13–15,22–24], and the differences among these implementations mainly lie in the coding strategies. Typically, video SCI systems contain the same components as traditional imaging systems, except for several extra relay lenses and a modulation device that generates temporal-variant masks to encode the image plane. An intuitive approach is to directly use a DMD [15,22] or an LCoS [13,14], which can project given patterns with an assigned time sequence, on the image plane for image encoding. A substitute approach in early work is to simply replace the modulation device with a physically shifting lithography mask driven by a piezo [10]. There are also some indirect modulation methods proposed in recent work [23,24], which takes advantage of the temporal shifting feature of rolling shutter cameras or streak cameras for the temporal-variant mask generation.

Parallel to these systems, different algorithms are proposed to improve the SCI reconstruction performance. Since the inverse problem is ill-posed, different prior constrains such as TV [25], sparsity [10,13,14,16], self-similarity [26], and Gaussian mixture model [27,28] are employed, forming widely used TwIST [29], GAP–TV [25], DeSCI [26], and some other reconstruction algorithms. Generally, algorithms based on iterative optimization have high computational complexity. Inspired by advances of deep learning, some learning-based reconstruction approaches are proposed and boost the reconstruction performance to a large extent [15,30–34]. Recently, a sophisticated reconstruction algorithm BIRNAT [30] based on recurrent neural network has led to state-of-the-art reconstruction performance with a significant reduction on the required time compared with DeSCI. However, despite the highest reconstruction quality achieved by learning-based methods, their main limitation is in the inflexibility resulting from inevitable training process and requirement for large-scale training data when changing encoding masks or data capture environment. Other learning-based methods such as MetaSCI [31] try to utilize meta-learning or transfer learning to realize fast mask adaption for SCI reconstruction with different masks, but the time cost is still unacceptable on the 10-mega-pixel SCI data. To solve the trilemma of reconstruction quality, time consumption, and algorithm flexibility, a joint framework of iterative and learning-based methods called PnP [17] is proposed. By integrating pre-trained deep denoisers as the prior terms into certain iterative optimization process such as generalized alternating projection (GAP) [35] and alternating direction method of multiplier (ADMM) [36], PnP-based approaches combine the advantages of both frameworks and realize the trade-off between speed, quality, and flexibility.

In this paper, we build a novel video SCI system using a hybrid coded aperture, composed of an LCoS and a physical mask shown in Fig. 1. Moreover, we modify the PnP algorithm to fit our system, leading to better results than the method proposed in Ref. [17].

3. SYSTEM

A. Hardware Implementation

The hardware setup of our HCA-SCI system is depicted in Fig. 1. It consists of a primary lens (HIKROBOT, MVL-LF5040M-F, $f = 50$ mm, $F\# = 4.0$–32), an amplitude-modulated LCoS (ForthDD, QXGA-3DM, 2048 × 1536 pixels, 4.5×10^3 refresh rate), a lithography mask (5120 × 5120 pixels, 4.5 μm × 4.5 μm pixel size), a complementary metal-oxide-semiconductor transistor camera (HIKROBOT, MV-CH250-20TC, 5120 × 5120 pixels, 4.5 μm pixel size), two achromatic doublets (Thorlabs, AC508-075-A-ML, $f =$ 75 mm), a relay lens (ZLKC, HM5018MP3, $f = 50$ mm), a polarizing beamsplitter (Thorlabs, CCM1-PBS251/M), and two film polarizers (Thorlabs, LPVISE100-A). The incident light from a scene is first collected by the primary lens and focused at the first virtual image plane. Then a 4f system consisting of two achromatic doublets transfers the image through the aperture coding module and the lithography mask, and subsequently onto the second virtual image plane. The aperture coding module positioned at the middle of the 4f system is composed of a polarizing beamsplitter, two film polarizers, and an amplitude-modulated LCoS, which are used to change the open–close states ("open" means letting the light go

through, while "close" means blocking the light) of the sub-apertures and thus modulate the light's propagation direction. Finally, the image delivered by the 4f system is relayed to the camera sensor being captured. Note that the 4f system used has a magnification of 1, and the relay lens has a magnification of 2, which on the whole provides a 1:2 mapping between the pixels of the lithography mask and the sensor. Even though the imaging model involves pixel-wise encoding, there is no need for a precise alignment for the lithography mask and the camera sensor since the actual encoding mask will be calibrated before data acquisition. During the acquisition process, the camera shutter is synchronized with the LCoS by using an output trigger signal from the LCoS driver board.

It is worth noting that the active area of the LCoS and the position of the lithography mask should be carefully adjusted. The active area of the LCoS should be as large as possible meanwhile to ensure it to serve as the aperture stop of the whole system, so that it can provide a higher light efficiency. As for the lithography mask, some fine tuning is needed to ensure that the mask's projection on the image plane can generate a shifting when different parts of the aperture are open, and meanwhile, the shifting masks can still keep sharp. In our implementation, after extensive experiments, the mask is placed in front of the second image plane with a distance of 80 μm, which is a good trade-off between the sharpness and the shift.

B. Encoding Mask Generation

The aperture of the system (i.e., the activated area of the LCoS) can be divided into several sub-apertures according to the resolution of the LCoS after pixel binning, and each sub-aperture corresponds to a light beam propagating toward certain directions. As shown in Fig. 3, because the lithography mask is placed in front of the image plane, when different sub-apertures are turned on, the light beams from the corresponding sub-apertures will project the mask onto different parts of the image plane, which can thus generate corresponding shifting encoding masks. In practice, to enhance the light throughput, multiple sub-apertures will be turned on simultaneously in one frame by assigning the LCoS with a specific multiplexing pattern to obtain a single multiplexing encoding mask. In addition, in different frames, different combinations of the sub-apertures are applied to generate different multiplexing encoding masks. Generally, we turn on 50% of the sub-apertures in one multiplexing pattern. In this multiplexing case, the final encoding mask on the image plane will be the summation of those shifting masks provided by the corresponding sub-apertures.

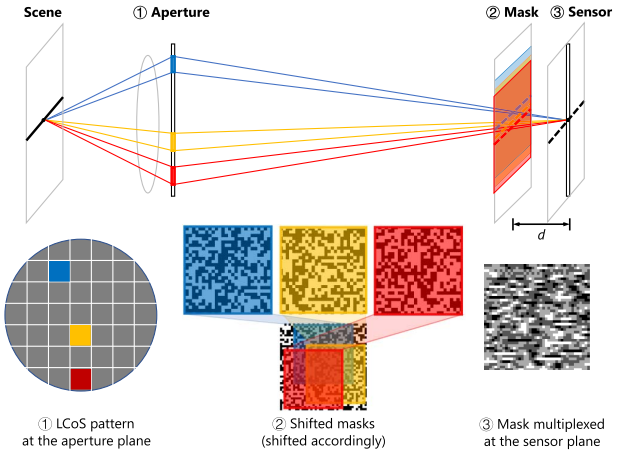

Fig. 3. Illustration of the multiplexed mask generation. For the same scene point, its images generated by different sub-apertures (marked as blue, yellow, and red, respectively) intersect the mask plane with different regions and are thus encoded with corresponding (shifted) random masks before summation at the sensor. The multiplexing would raise the light flux for high SNR recording, while doing so only with slight performance degeneration.

C. Mathematical Model

Mathematically, the encoding mask generation process can be modeled as a multiplexing of shifting masks. Let $\boldsymbol{O}$ denote the center-view mask generated by opening the central sub-aperture of the system; then each mask $\boldsymbol{C}$ generated by sub-aperture multiplexing can be formulated as

$$\boldsymbol{C} = \sum_{i=1}^{N} m_i S_i(\boldsymbol{O}), \tag{1}$$

where S_i denotes the mask-shifting operator corresponding to the i-th sub-aperture; m_i is the multiplexing indicator (a scalar) for the i-th sub-aperture, with 0 and 1 for blocking or transmitting the light, respectively; and N is the amount of sub-apertures (i.e., the number of macro-pixels in the active area of the LCoS after binning). Consider that a video $\boldsymbol{X} \in \mathbb{R}^{n_x \times n_y \times B}$, containing B consecutive high-speed frames is modulated by B encoding masks $\boldsymbol{C} \in \mathbb{R}^{n_x \times n_y \times B}$ and integrated by the sensor to generate a snapshot-coded measurement $\boldsymbol{Y}$. Then $\boldsymbol{Y}$ can be expressed as

$$\boldsymbol{Y} = \sum_{k=1}^{B} \boldsymbol{C}_k \odot \boldsymbol{X}_k + \boldsymbol{G}, \tag{2}$$

where $\odot$ denotes the Hadamard (element-wise) product; $\boldsymbol{G} \in \mathbb{R}^{n_x \times n_y}$ is the measurement noise; and $\boldsymbol{C}_k = \boldsymbol{C}(:,:,k)$ and $\boldsymbol{X}_k = \boldsymbol{X}(:,:,k)$ represent the k-th multiplexed mask and corresponding frame, respectively. Through a simple derivation, the coded measurement in Eq. (2) can be further expressed by

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}, \tag{3}$$

where $\boldsymbol{y} = \mathrm{Vec}(\boldsymbol{Y}) \in \mathbb{R}^n$ and $\boldsymbol{g} = \mathrm{Vec}(\boldsymbol{G}) \in \mathbb{R}^n$ with $n = n_x n_y$. Corresponding video signal $\boldsymbol{x} \in \mathbb{R}^{nB}$ is given by $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_B^\top]^\top$, where $\boldsymbol{x}_k = \mathrm{Vec}(\boldsymbol{X}_k) \in \mathbb{R}^n$. Different with traditional CS, the coding matrix $\boldsymbol{H} \in \mathbb{R}^{n \times nB}$ in video SCI is sparse and has a special structure, which can be written as

$$\boldsymbol{H} = [\boldsymbol{D}_1, \ldots, \boldsymbol{D}_B], \tag{4}$$

where $\boldsymbol{D}_k = \mathrm{diag}[\mathrm{Vec}(\boldsymbol{C}_b)] \in \mathbb{R}^{n \times n}$. Therefore, the compressive sampling rate in video SCI is equal to $1/B$. Theoretical analysis in Refs. [37,38] has proved that the reconstruction error of SCI is bounded even when $B > 1$.

D. System Calibration

Although pixel-wise modulation is involved in SCI systems, there is no need for pixel-to-pixel alignment during system building, as we can get the precise encoding patterns through an end-to-end calibration process prior to data capture. To be specific, we place a Lambertian white board at the objective

plane, and provisionally take away the lithography mask. Then, an illumination pattern $\boldsymbol{I}$ and a background pattern $\boldsymbol{B}$ are captured with LCoS projecting white and black patterns, respectively. After that, we put on the lithography mask and capture each dynamic encoding mask $\boldsymbol{C}$ with LCoS projecting corresponding multiplexing patterns directly. To eliminate the influence of background light and nonuniform illumination caused by light source or system vignetting on the actual encoding masks, we conduct calibration to obtain the accurate encoding mask $\boldsymbol{C}'$ following Eq. (5). In addition, after encoding acquisition, the encoded measurements will also subtract the background pattern to accord with the providing mathematical model, and illumination will be regarded as a part of the scene itself:

$$\boldsymbol{C}' = \frac{\boldsymbol{C} - \boldsymbol{B}}{\boldsymbol{I} - \boldsymbol{B}}. \tag{5}$$

4. RECONSTRUCTION ALGORITHM

The reconstruction of high-speed videos from the snapshot-coded measurement is an ill-posed problem. As mentioned before, to solve it, different priors and frameworks have been employed. Roughly, the algorithms can be categorized into the following three classes [12]: i) regularization (or priors)-based optimization algorithms with well-known methods such as TwIST [29], GAP–TV [25], and DeSCI [26]; ii) end-to-end deep-learning-based algorithms [15,32,34], such as BIRNAT [30], which reaches state-of-the-art performance, and recently developed MetaSCI [31] that uses meta-learning to improve adaption capability for different masks in SCI reconstruction; and iii) PnP algorithms that use deep-denoising networks in the optimization framework such as ADMM and GAP.

Among these, regularization-based algorithms are usually too slow, and end-to-end deep learning networks need a large amount of data and also a long time to train the network, in addition to showing inflexibility (i.e., re-training being required for a new system). Although recent works such as MetaSCI [31] try to mitigate this problem with meta-learning or transfer learning and thus march forward to a large-scale SCI problem with a patch-wise reconstruction strategy, it still takes a long time for the training and adaption of 10-mega-pixel-scale SCI reconstruction. For example, MetaSCI takes about 2 weeks for the $256 \times 256 \times 10$ base model training on a single NVIDIA 2080Ti GPU, and further adaption performed on more than 570 $256 \times 256 \times 10$ sub-tasks (overlapped patches extracted from a 10-mega-pixel image) takes about two months, which is impractical in real applications (more GPUs can be used to mitigate this challenge). By contrast, PnP has achieved a good balance of speed, flexibility, and accuracy. Therefore, we employ a PnP framework in our work and further develop the PnP–TV–FastDVDNet to achieve promising results for our high-resolution HCA-SCI scheme. Meanwhile, we use GAP–TV and BIRNAT as baselines for comparison.

In the following, we review the main steps of PnP–GAP [17] and then present our PnP–TV–FastDVDNet algorithm for HCA-SCI in Algorithm 1.

A. PnP–GAP

In GAP, the SCI inversion problem is modeled as

$$\arg\min_{\boldsymbol{x},\boldsymbol{v}} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}\|_2^2 + \lambda R(\boldsymbol{x}), \quad \text{s.t. } \boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}, \tag{6}$$

where $R(\boldsymbol{x})$ is a regularizer or prior being imposed on $\boldsymbol{x}$, which can be a TV, sparse prior, or a deep prior [39], and $\boldsymbol{v}$ is an auxiliary parameter. Let k index be the iteration number; through a two-step iteration, the minimization in Eq. (6) can be solved as follows.

- Solving $\boldsymbol{x}$:

$$\boldsymbol{x}^{(k)} = \boldsymbol{v}^{(k-1)} + \boldsymbol{H}^{\top}(\boldsymbol{H}\boldsymbol{H}^{\top})^{-1}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{v}^{(k-1)}). \tag{7}$$

By utilizing the special structure of $\boldsymbol{H}$ shown in Eq. (4), this subproblem can be solved efficiently via element-wise operation rather than calculating the inversion of a huge matrix.

- Solving $\boldsymbol{v}$:

$$\boldsymbol{v}^{(k)} = \mathcal{D}_{\sigma}(\boldsymbol{x}^{(k)}), \tag{8}$$

where $\mathcal{D}_{\sigma}(\cdot)$ represents a denoising process with $\sigma = \sqrt{\lambda}$. Here, different denoising algorithms can be used, such as TV (thus GAP–TV), WNNM [40] (thus DeSCI [26]), and FFDnet [39] (thus PnP–FFDnet [17]).

B. PnP–TV–FastDVDNet

Recall that solving $\boldsymbol{v}$ following Eq. (8) is equivalent to performing a denoising process on $\boldsymbol{x}$. By plugging various denoisers into the GAP iteration steps, we can make a trade-off between different aspects of reconstruction performance. In fact, more than one denoiser can be employed in a series manner (i.e., one after another in each iteration), or in a cascading manner (i.e., the first several iterations using one denoiser, while the next several iterations using another). In this way, we can further balance the strengths and drawbacks of different denoisers.

Algorithm 1. PnP–TV–FastDVDNet for HCA-SCI

Require $\boldsymbol{H}$, $\boldsymbol{y}$.
1: Initialize: $\boldsymbol{v}^{(0)}, \lambda_0, \xi < 1, k = 1, K_1, K_{\text{Max}}$.
2: **while** Not Converge **and** $k \le K_{\text{Max}}$ **do**
3: Update $\boldsymbol{x}$ by Eq. (7).
4: Update $\boldsymbol{v}$:
5: **if** $k \le K_1$ **then**
6: $\boldsymbol{v}^{(k)} = \mathcal{D}_{\text{TV}}(\boldsymbol{x}^{(k)})$
7: **else**
8: $\boldsymbol{v}' = \mathcal{D}_{\text{TV}}(\boldsymbol{x}^{(k)})$
9: $\boldsymbol{v}^{(k)} = \mathcal{D}_{\text{FastDVDNet}}(\boldsymbol{v}')$

In this paper, considering the high video denoising performance of recently proposed FastDVDNet [21], we jointly employ the TV denoiser and the FastDVDNet denoiser in a PnP–GAP framework, implementing a reconstruction algorithm of PnP–TV–FastDVDNet, which involves cascading and series denoising processes. The algorithm pipeline is shown in Algorithm 1. In each iteration, the updating of $\boldsymbol{x}$ follows Eq. (7), and the updating of $\boldsymbol{v}$ (i.e., the denoising process), differs in different periods. To be specific, in the first period, a single TV denoiser is employed, while in the second period, joint TV and FastDVDNet denoisers are involved one after another.

5. RESULTS

In this section, we conduct a series of experiments on both simulation and real data to validate the feasibility and performance of our proposed HCA-SCI system. Four reconstruction algorithms, including iterative-optimization-based algorithm GAP–TV [25], plug-and-play-based algorithm PnP–FFDNet [17], our proposed PnP–TV–FastDVDNet, and the state-of-the-art learning-based algorithm BIRNAT [30] are employed.

A. Multiplexing Pattern Design

Multiplexing pattern design plays an important role in our HCA-SCI system. In simulation, we used the random squares multiplexing scheme (shown in the first row of Fig. 4), which contains 12 sub-apertures with each sub-aperture containing 512 × 512 binning pixels. In each multiplexing pattern, there are 50% randomly selected sub-apertures being open. In real experiments, the optical aberration and diffraction that are not considered in simulation experiments will cause blur to the encoding mask, and thus introduce more correlation among the encoding masks, which degrades the encoding effect. Considering the fact that the calibration process is conducted in an end-to-end manner in real experiments, we thus have enough degree of freedom to design any possible multiplexing patterns without being limited by the binning mode or the sub-aperture amount. We empirically tried many different schemes and finally found that the "rotating-circle" scheme (shown in the second row of Fig. 4) can realize a good balance between the system's light throughput and the encoding masks' quality when taking the physical factors into account.

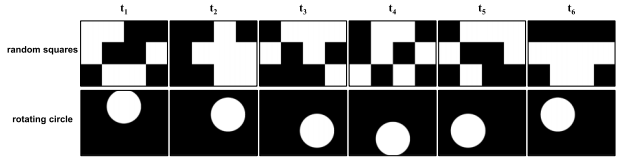

Fig. 4. Multiplexing pattern schemes used in our experiments (taking Cr = 6 for an example). Top row: multiplexing patterns for simulation experiments. Each pattern contains 50% open sub-apertures, and each sub-aperture is a 512 × 512 binning macro pixel on the LCoS. Bottom row: multiplexing patterns for real experiments. Each pattern contains an open circle with a radius of about 400 pixels, and the circles in adjacent patterns have a rotation of 360/Cr degrees.

Currently, the design of the multiplexing patterns is heuristic, and the challenge of the algorithm-based multiplexing pattern design mainly lies in the large size of the multiplexing pattern (1536 × 2048) and the coded measurement (3200 × 3200) in our system. Both traditional optimization-based methods and learning based methods have difficulty in dealing with data of this scale. Thus, the design of the optimal multiplexing pattern is still an open challenge that is worthy of future investigation.

B. Reconstruction Comparison between Different Algorithms on Simulation Datasets

To investigate the reconstruction performance of different algorithms on the proposed HCA-SCI system, we first perform experiments on simulated datasets, which involve three different scales of 256 × 256, 512 × 512, and 1024 × 1024, and two compression rates (Cr) of 10 and 20. Two datasets named Football and Hummingbird used in Ref. [17] and three datasets named ReadySteadyGo, Jockey, and YachtRide provided in Ref. [41] are employed in our simulation. According to the forward model depicted in Section 5.C, to simulate the encoding masks, we first calculate the mask-shifting distance with respect to the center-view mask for each sub-aperture in the current multiplexing pattern (shown in the first row of Fig. 4) based on geometry optics. Then we shift the center-view mask accordingly to get each sub-aperture's shifting mask. Finally, the shifting masks are added together to obtain the final multiplexing encoding mask on the image plane. After that, we can generate six groups of simulation datasets by modulating different scales of videos (containing 10 or 20 frames) with multiplexed shifting masks generated from the HCA-SCI system and then collapsing the coded frames to a single (coded) measurement.

The reconstruction peak signal to noise ratio (PSNR) and structural similarity index measure (SSIM) for each algorithm are summarized in Tables 1 and 2 for Cr = 10 and 20, respectively. It is worth noting that due to the limited memory of our GPU (GeForce RTX 3090 with 24 GB memory), we only test BIRNAT on 256 × 256 scale datasets with Cr equal to 10 and 20, and the adversarial training was not involved. For GAP–TV and PnP–TV–FastDVDNet, the reconstruction is conducted on a platform equipped with an Intel Core i7-9700K CPU (3.60 GHz, 32 G memory) and a GeForce RTX 2080 GPU with 8 G memory; 160 and 250 iterations are taken for Cr = 10 and

Table 1. Average Results of PSNR in dB (left entry in each cell) and SSIM (right entry in each cell) by Different Algorithms (Cr = 10)[a]

Scales	Algorithms	Football	Hummingbird	ReadySteadyGo	Jockey	YachtRide	Average
256 × 256	GAP–TV	27.82, 0.8280	29.24, 0.7918	23.73, 0.7499	31.63, 0.8712	26.65, 0.8056	27.81, 0.8093
	PnP–FFDNet	27.06, 0.8264	25.52, 0.6912	21.68, 0.6859	31.14, 0.8493	23.69, 0.7035	25.82, 0.7513
	PnP–TV–FastDVDNet	**31.31, 0.9123**	**31.19, 0.8264**	**26.18, 0.8276**	**31.36, 0.8817**	**28.90, 0.8841**	**29.79, 0.8664**
	BIRNAT	34.67, 0.9719	34.33, 0.9546	29.50, 0.9389	36.24, 0.9711	31.02, 0.9431	33.15, 0.9559
512 × 512	GAP–TV	29.19, 0.8854	28.32, 0.7887	25.94, 0.7918	31.30, 0.8718	26.59, 0.7939	28.27, 0.8263
	PnP–FFDNet	28.57, 0.8952	28.02, 0.8363	24.32, 0.7457	29.81, 0.8248	23.45, 0.6793	26.83, 0.7963
	PnP–TV–FastDVDNet	**30.92, 0.9333**	**32.24, 0.8834**	**27.04, 0.8246**	**32.11, 0.8839**	**27.87, 0.8487**	**30.04, 0.8748**
1024 × 1024	GAP–TV	30.63, 0.9022	29.16, 0.8459	28.92, 0.8698	31.59, 0.8953	29.03, 0.8470	29.87, 0.8720
	PnP–FFDNet	29.87, 0.9023	27.70, 0.7869	27.70, 0.8483	29.88, 0.8412	25.55, 0.7211	28.14, 0.8200
	PnP–TV–FastDVDNet	**30.35, 0.9265**	**31.71, 0.8909**	**29.42, 0.8913**	**31.59, 0.9014**	**30.44, 0.8713**	**30.70, 0.8963**

[a]BIRNAT fails at large-scale due to limited GPU memory.

Table 2. Average Results of PSNR in dB (left entry in each cell) and SSIM (right entry in each cell) by Different Algorithms (Cr = 20)[a]

Scales	Algorithms	Football	Hummingbird	ReadySteadyGo	Jockey	YachtRide	Average
256 × 256	GAP–TV	25.01, 0.7544	26.33, 0.6893	20.48, 0.6326	28.13, 0.8318	23.56, 0.7129	24.70, 0.7242
	PnP–FFDNet	21.67, 0.6657	22.13, 0.5835	17.27, 0.5340	27.78, 0.7994	20.39, 0.6024	21.85, 0.6370
	PnP–TV–FastDVDNet	**27.83, 0.8459**	**28.65, 0.7520**	**23.28, 0.7381**	**29.51, 0.8597**	**26.34, 0.8235**	**27.12, 0.8038**
512 × 512	BIRNAT	27.91, 0.9021	28.58, 0.8800	23.79, 0.8279	31.35, 0.9467	26.14, 0.8585	27.55, 0.8830
	GAP–TV	23.97, 0.8179	24.50, 0.6719	22.12, 0.6975	26.99, 0.8297	23.13, 0.6930	24.14, 0.7420
	PnP–FFDNet	22.00, 0.7661	23.62, 0.7245	19.35, 0.6133	25.32, 0.7924	19.48, 0.5418	21.95, 0.6876
	PnP–TV–FastDVDNet	**25.63, 0.8852**	**28.36, 0.7778**	**23.80, 0.7499**	**28.79, 0.8553**	**25.36, 0.7784**	**26.39, 0.8093**
1024 × 1024	GAP–TV	24.82, 0.8353	25.53, 0.7296	24.98, 0.8128	26.63, 0.8388	25.80, 0.7759	25.55, 0.7985
	PnP–FFDNet	23.55, 0.8098	23.02 0.6039	22.48, 0.7702	24.48, 0.7968	21.67, 0.6414	23.04, 0.7244
	PnP–TV–FastDVDNet	**26.26, 0.8729**	**28.68, 0.8076**	**26.31, 0.8399**	**29.18, 0.8773**	**28.07, 0.8194**	**27.70, 0.8434**

[a]BIRNAT fails at large-scale due to limited GPU memory.

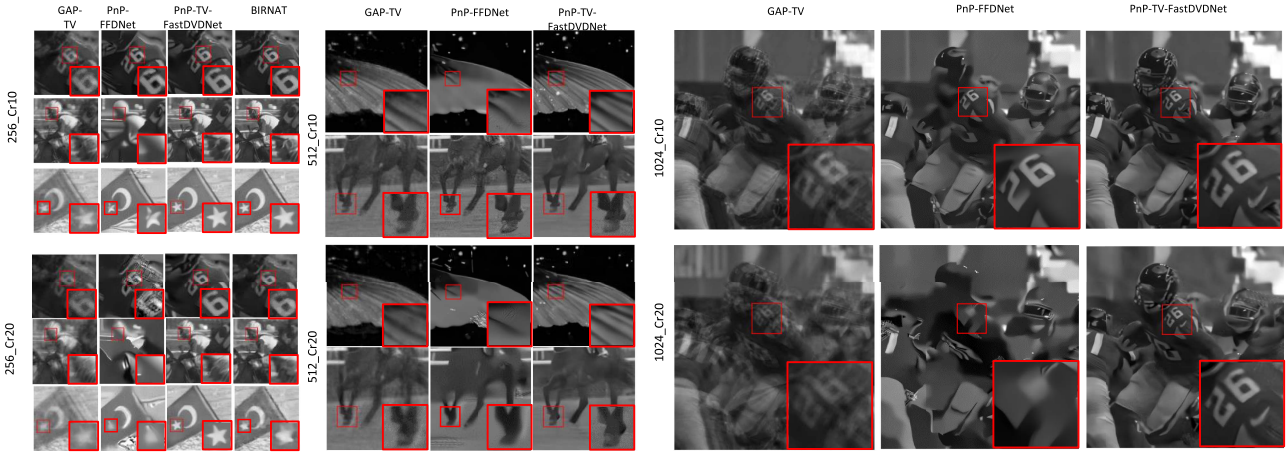

Fig. 5. Reconstruction results and comparison with state-of-the-art algorithms on simulated data at different resolutions (left: 256 × 256, middle: 512 × 512, right: 1024 × 1024) and with different compression ratios (top: Cr = 10, bottom: Cr = 20). The BIRNAT results are not available for 512 × 512 and 1024 × 1024 since the model training will be out of memory. See Visualization 1, Visualization 2, Visualization 3, Visualization 4, Visualization 5, and Visualization 6 for the reconstructed videos.

20, respectively. It can be observed that i) on all six groups of simulated datasets, our proposed PnP–TV–FastDVDNet outperforms GAP–TV and PnP–FFDNet for about 1.5 and 4 dB on average, respectively. ii) On the 256_Cr10 and 256_Cr20 datasets, BIRNAT shows the best performance (with sufficient training), exceeding the PnP–TV–FastDVDNet for about 3.4 and 0.4 dB, respectively. iii) BIRNAT is the fastest algorithm during the inference period that is hundreds times shorter than that of GAP–TV and PnP–TV–FastDVDNet. However, the training process of BIRNAT is quite time-consuming, and it takes about a week to train the network without adversarial training for 25 epochs. iv) From the selected reconstruction video frames in Fig. 5, we can see that our proposed PnP–TV–FastDVDNet provides higher visualization quality with sharp edges and less artifacts. By contrast, GAP–TV produces noisy results, while the PnP–FFDNet leads to some unpleasant artifacts.

C. Validation of Multiplexed Shifting Mask's Noise Robustness on Simulation Datasets

As mentioned in Section 5.B, our proposed HCA-SCI system leverages multiplexing strategy for the improvement of light throughput, which can thus gain a higher signal-to-noise ratio (SNR) and enhance the system's robustness to noise in real-world applications. In this subsection, we conduct a series

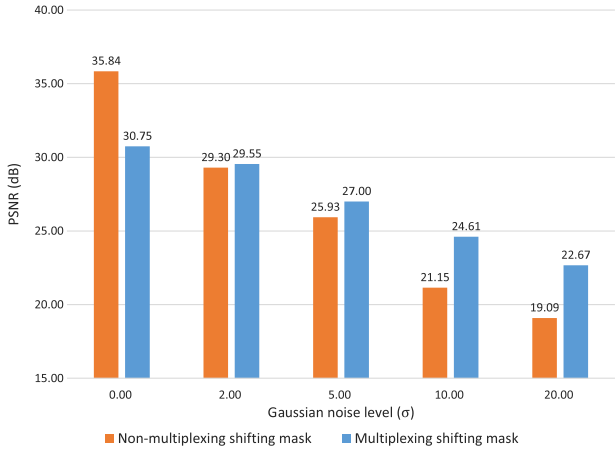

Fig. 6. Noise robustness comparison between multiplexed and non-multiplexed masks.

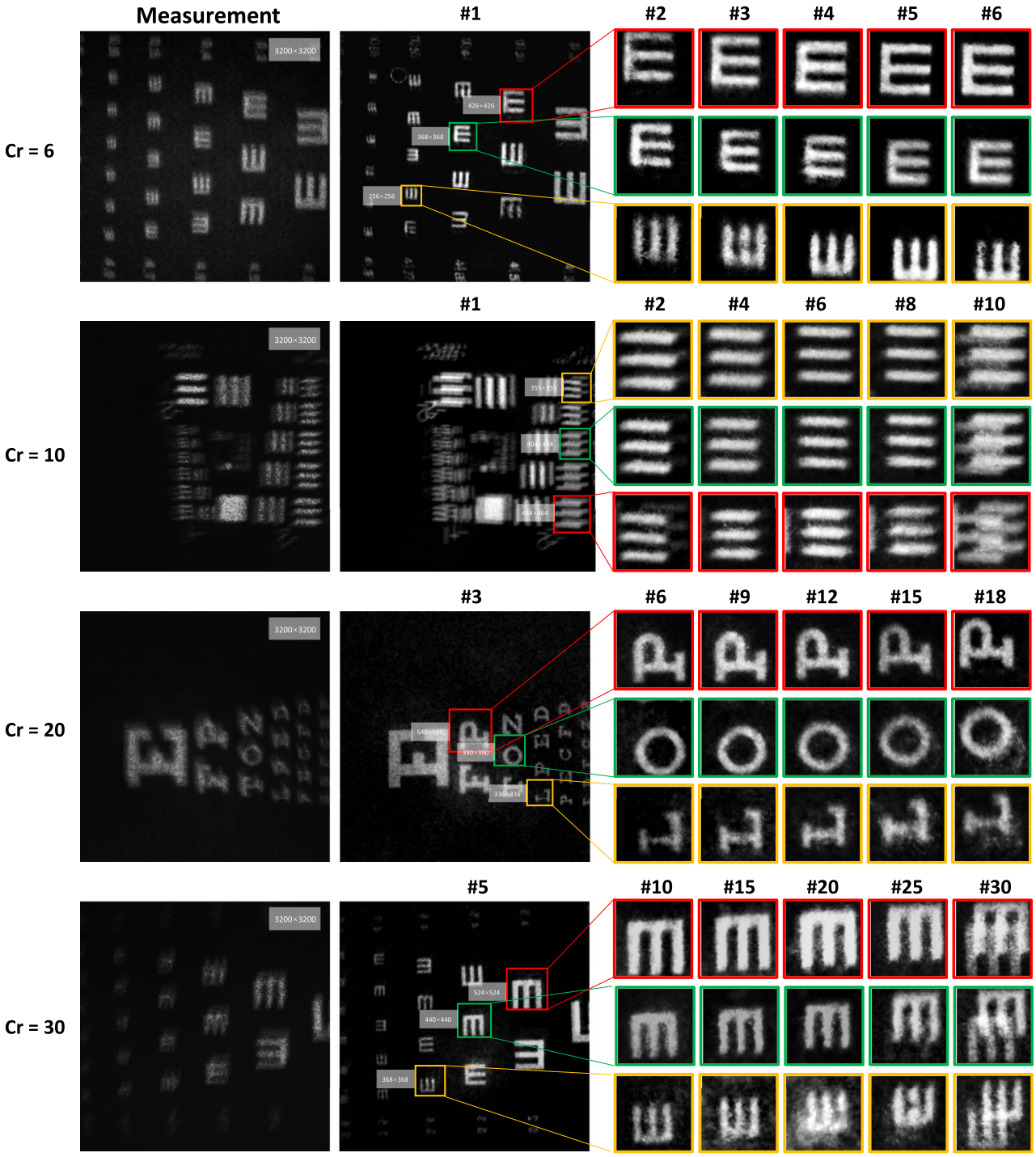

Fig. 7. Reconstruction results of the PnP–TV–FastDVDNet on real data captured by our HCA-SCI system (Cr = 6, 10, 20, and 30). Note the full frames are of 3200 × 3200, and we plot small regions about 400 × 400 in size to demonstrate the high-speed motion.

of experiments on simulation datasets containing different levels of noise to compare the reconstruction performance of multiplexed shifting masks used in HCA-SCI and non-multiplexed ones. The 256 × 256 size dataset mentioned above is utilized here as the clean original frames. In addition, we normalize the masks with a light throughput factor calculated from the proportion of the masks' active area with respect to the whole aperture. Zero-mean Gaussian noise with standard deviation ranging from 0 to 20 is added to the measurements (with values in [0,255]) to simulate the real-world acquisition process.

The change of reconstruction PSNR over increasing noise levels is shown in Fig. 6. As can be seen from the figure, shifting masks with no multiplexing outperform the multiplexed ones when there is no noise, which is probably because the

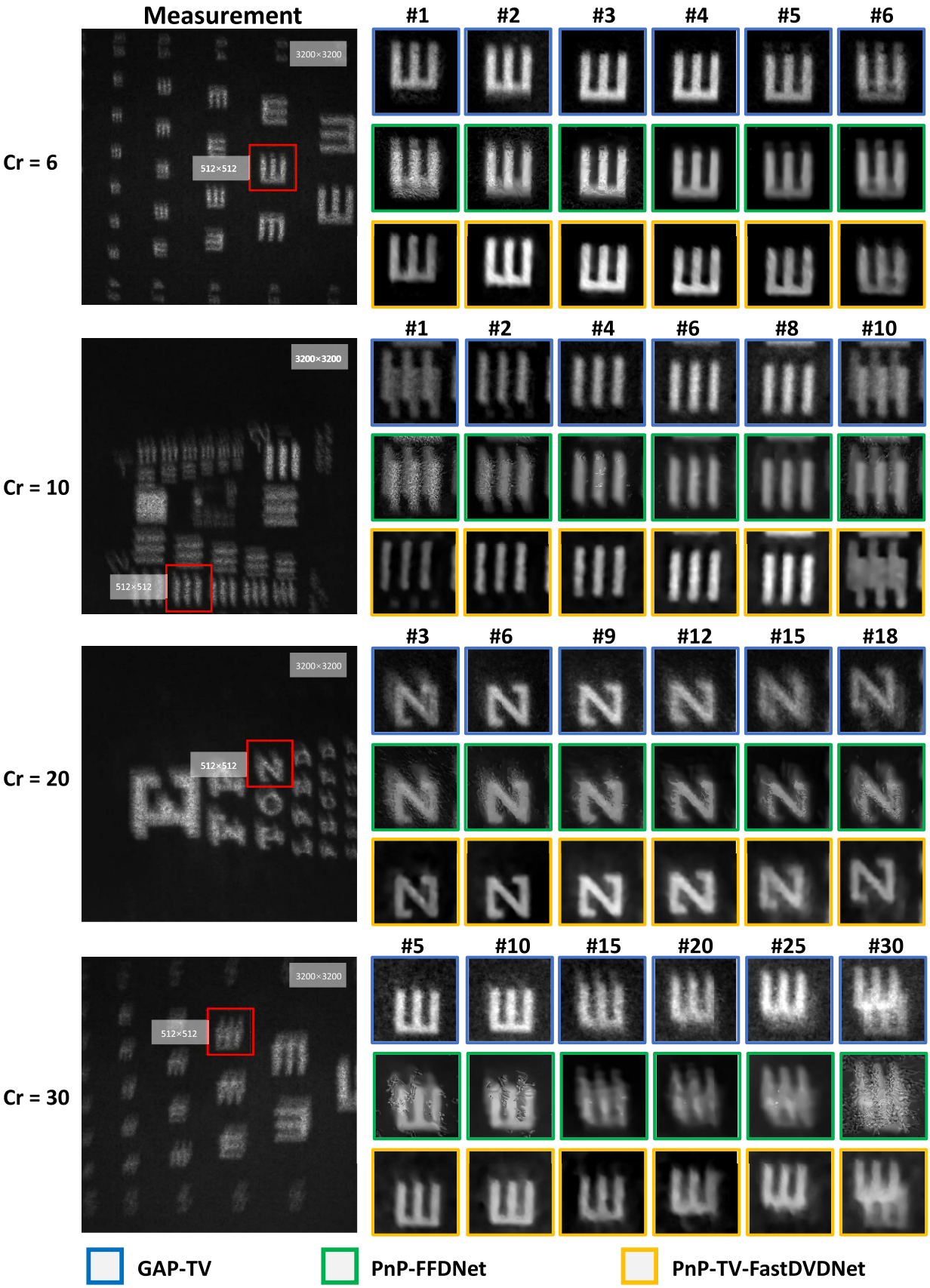

Fig. 8. Reconstruction comparison between the GAP–TV, PnP–FFDNet, and PnP–TV–FastDVDNet on real data captured by our HCA-SCI system (Cr = 6, 10, 20, and 30). Note the full frames are of 3200 × 3200, and we plot small regions 512 × 512 in size to demonstrate the high-speed motion. See Visualization 7 for the reconstructed videos.

multiplexing operation brings in some correlation among the masks that is not desired for the encoding and reconstruction. However, as the noise increases, the superiority in SNR of the multiplexing scheme shows up, and the reconstruction PSNR of non-multiplexed shifting masks drops rapidly, while that of the multiplexed masks decreases more slowly and exceeds the PSNR of the non-multiplexed ones. In real scenes, noise is inevitable, and as it increases, it will also have a significant impact on the reconstruction process until totally disrupting the reconstruction. Thus, the multiplexing strategy leveraged in HCA-SCI can equip the physical system with more robustness in real applications, especially those at relatively large noise levels.

D. Real-Data Results

We built a 10-mega-pixel snapshot compressive camera prototype illustrated in Fig. 1 for dynamic video recording. Empirically, the multiplexing patterns (shown in the second row of Fig. 4) projected by the LCoS are designed to be rotationally symmetric, which ensures the consistency in the final coding patterns and provides an adequate incoherence among these masks. Before acquisition, we first calibrate the groups of coding masks with a Lambertian whiteboard placed at the objective plane, and each calibrated pattern is averaged over 50 repetitive snapshots to suppress the sensor noise. Then, during data capture, the camera operates at a fixed frame rate of 20 fps when $Cr = 6$, 10, and 20, providing reconstruction video frame rate of 120, 200, and 400 fps, respectively. For $Cr = 30$, the camera operates at 15 fps to extend the exposure time and provide a higher light throughput, which can reach a reconstruction frame rate up to 450 fps. We determine the spatial resolution of our system as 3200×3200 pixels. We thus have achieved the throughput of 4.6×10^9 voxels per second in the reconstructed video.

Three moving test charts printed on A4 papers are chosen as the dynamic objects. In Fig. 7, we show the coded measurements and final reconstruction of the test charts. From that, one can see that the proposed HCA-SCI system and PnP–GAP–FastDVDNet reconstruction algorithm can effectively capture and restore the moving details of dynamic objects, which will be blurry when captured directly with conventional cameras. For the reconstruction of real data, we also find that, in some cases, the start and end frames tend to be blurry (refer to real-data results of $Cr = 10$ and $Cr = 30$ in Fig. 7), which might be caused by the synchronization imperfection and initialization delay of the LCoS when switching between the projection sequences.

We further compare the reconstruction of GAP–TV, PnP–FFDNet, and PnP–TV–FastDVDNet on real datasets captured by our HCA-SCI system. From the reconstruction results shown in Fig. 8, we can find that when $Cr = 6$, all these three algorithms can produce clear reconstruction frames with sharp details, especially in the intermediate frames. However when Cr gets larger, the reconstructed frames of GAP–TV tend to be blurry and have more background noise. The PnP–FFDNet will generate severe artifacts in the reconstructed frames and make the motion invisible. But our proposed PnP–TV–FastDVDNet can still reconstruct the motion details with a little increasing of noise in the background.

6. CONCLUSION

We have proposed a new computational imaging scheme capable of capturing 10-mega-pixel videos with low bandwidth and developed corresponding algorithms for computational reconstruction, providing a high-throughput (up to 4.6×10^9 voxels per second) solution for high-speed, high-resolution videos.

The hardware design bypasses the pixel count limitation of available spatial light modulators via joint coding at aperture and close to image plane. The results demonstrate the feasibility of high-throughput imaging under a snapshot compressive sensing scheme and hold great potential for future applications in industrial visual inspection or multi-scale surveillance.

So far, the final reconstruction is limited to 450 Hz since the hybrid coding scheme further decreases the light throughput to some extent, compared with conventional coding strategies. In the future, a worthwhile extension would be to introduce new photon-efficient aperture-coding devices to raise the SNR for coded measurements. Another limitation of the current system lies in the non-uniformity of the encoding masks along the radial direction, which is a common problem for large-scale imaging systems due to non-negligible off-axis aberration. Considering the challenge for improving optical performance of existing physical components, designing novel algorithms capable of SCI reconstruction with non-uniform masks may be economical and feasible. Meanwhile, time-efficient reconstruction algorithms and feasible multiplexing pattern design methods for large-scale (like 10-mega-pixel or even giga-pixel) SCI reconstruction are still an open challenge in the foreseeable future. Moreover, extending proposed imaging scheme for higher throughput (e.g., giga pixel) or other dimensions (e.g., light field, hyperspectral imaging) may also be a promising direction.

Funding. Ministry of Science and Technology of the People's Republic of China (2020AAA0108202); National Natural Science Foundation of China (62088102, 61931012).

Disclosures. The authors declare no conflicts of interest.

Data Availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

[†]These authors contributed equally to this paper.

REFERENCES

1. G. Carles, J. Downing, and A. R. Harvey, "Super-resolution imaging using a camera array," Opt. Lett. **39**, 1889–1892 (2014).
2. S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," IEEE Trans. Image Process. **13**, 1327–1344 (2004).
3. R. F. Marcia and R. M. Willett, "Compressive coded aperture superresolution image reconstruction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2008), pp. 833–836.
4. T. S. Cho, A. Levin, F. Durand, and W. T. Freeman, "Motion blur removal with orthogonal parabolic exposures," in *IEEE International Conference on Computational Photography (ICCP)* (2010), pp. 1–8.

5. R. Raskar, A. Agrawal, and J. Tumblin, "Coded exposure photography: motion deblurring using fluttered shutter," ACM Trans. Graph. **25**, 795–804 (2006).
6. C. Zhou and S. Nayar, "What are good apertures for defocus deblurring?" in *IEEE International Conference on Computational Photography (ICCP)* (2009), pp. 1–8.
7. C. Zhou, S. Lin, and S. K. Nayar, "Coded aperture pairs for depth from defocus and defocus deblurring," Int. J. Comput. Vis. **93**, 53–72 (2011).
8. A. Sellent and P. Favaro, "Optimized aperture shapes for depth estimation," Pattern Recognit. Lett. **40**, 96–103 (2014).
9. S. Suwajanakorn, C. Hernandez, and S. M. Seitz, "Depth from focus with your mobile phone," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3497–3506.
10. 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).
11. A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," Appl. Opt. **47**, B44–B51 (2008).
12. X. Yuan, D. J. Brady, and A. K. Katsaggelos, "Snapshot compressive imaging: theory, algorithms, and applications," IEEE Signal Process Mag. **38**, 65–88 (2021).
13. D. Reddy, A. Veeraraghavan, and R. Chellappa, "P2C2: programmable pixel compressive camera for high speed imaging," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2011), pp. 329–336.
14. Y. Hitomi, J. Gu, M. Gupta, T. Mitsunaga, and S. K. Nayar, "Video from a single coded exposure photograph using a learned over-complete dictionary," in *International Conference on Computer Vision (ICCV)* (2011), pp. 287–294.
15. M. Qiao, Z. Meng, J. Ma, and X. Yuan, "Deep learning for video compressive sensing," APL Photon. **5**, 030801 (2020).
16. X. Yuan, P. Llull, X. Liao, J. Yang, D. J. Brady, G. Sapiro, and L. Carin, "Low-cost compressive sensing for color video and depth," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 3318–3325.
17. X. Yuan, Y. Liu, J. Suo, and Q. Dai, "Plug-and-play algorithms for large-scale snapshot compressive imaging," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 1444–1454.
18. X. Yuan, Y. Liu, J. Suo, F. Durand, and Q. Dai, "Plug-and-play algorithms for video snapshot compressive imaging," arXiv:2101.04822 (2021).
19. Y. Sun, X. Yuan, and S. Pang, "Compressive high-speed stereo imaging," Opt. Express **25**, 18182–18190 (2017).
20. L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," Physica D **60**, 259–268 (1992).
21. M. Tassano, J. Delon, and T. Veit, "FastDVDnet: towards real-time deep video denoising without flow estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 1351–1360.
22. C. Deng, Y. Zhang, Y. Mao, J. Fan, J. Suo, Z. Zhang, and Q. Dai, "Sinusoidal sampling enhanced compressive camera for high speed imaging," IEEE Trans. Pattern Anal. Mach. Intell. **43**, 1380–1393 (2021).
23. 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).
24. N. Antipa, P. Oare, E. Bostan, R. Ng, and L. Waller, "Video from stills: lensless imaging with rolling shutter," in *IEEE International Conference on Computational Photography (ICCP)* (2019), pp. 1–8.
25. X. Yuan, "Generalized alternating projection based total variation minimization for compressive sensing," in *IEEE International Conference on Image Processing (ICIP)* (2016), pp. 2539–2543.
26. Y. Liu, X. Yuan, J. Suo, D. Brady, and Q. Dai, "Rank minimization for snapshot compressive imaging," IEEE Trans. Pattern Anal. Mach. Intell. **41**, 2990–3006 (2019).
27. 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).
28. J. Yang, X. Liao, X. Yuan, P. Llull, D. J. Brady, G. Sapiro, and L. Carin, "Compressive sensing by learning a Gaussian mixture model from measurements," IEEE Trans. Image Process. **24**, 106–119 (2015).
29. J. Bioucas-Dias and M. Figueiredo, "A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration," IEEE Trans. Image Process. **16**, 2992–3004 (2007).
30. 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 *European Conference on Computer Vision*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds. (2020), pp. 258–275.
31. Z. Wang, H. Zhang, Z. Cheng, B. Chen, and X. Yuan, "MetaSCI: scalable and adaptive reconstruction for video compressive sensing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 2083–2092.
32. M. Iliadis, L. Spinoulas, and A. K. Katsaggelos, "Deep fully-connected networks for video compressive sensing," Digit. Signal Process. **72**, 9–18 (2018).
33. K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, and A. Ashok, "ReconNet: non-iterative reconstruction of images from compressively sensed measurements," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 449–458.
34. J. Ma, X.-Y. Liu, Z. Shou, and X. Yuan, "Deep tensor ADMM-net for snapshot compressive imaging," in *IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 10222–10231.
35. X. Liao, H. Li, and L. Carin, "Generalized alternating projection for weighted-$l_{2,1}$ minimization with applications to model-based compressive sensing," SIAM J. Imag. Sci. **7**, 797–823 (2014).
36. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," Foundations Trends Mach. Learn. **3**, 1–122 (2011).
37. S. Jalali and X. Yuan, "Compressive imaging via one-shot measurements," in *IEEE International Symposium on Information Theory* (2018), pp. 416–420.
38. S. Jalali and X. Yuan, "Snapshot compressed sensing: performance bounds and algorithms," IEEE Trans. Inf. Theory **65**, 8005–8024 (2019).
39. K. Zhang, W. Zuo, and L. Zhang, "FFDNet: toward a fast and flexible solution for CNN based image denoising," IEEE Trans. Image Process. **27**, 4608–4622 (2018).
40. S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 2862–2869.
41. A. Mercat, M. Viitanen, and J. Vanne, "UVG dataset: 50/120 fps 4K sequences for video codec analysis and development," in *11th ACM Multimedia Systems Conference* (2020), pp. 297–302.