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

ZHIHONG ZHANG\textsuperscript{1,2,†}, CHAO DENG\textsuperscript{1,2,†}, YANG LIU\textsuperscript{3}, XIN YUAN\textsuperscript{4,*}, JINLI SUO\textsuperscript{1,2,*}, and QIONGHAI DAI\textsuperscript{1,2,5}

\textsuperscript{1} Department of Automation, Tsinghua University, Beijing 100084, China  
\textsuperscript{2} Institute for Brain and Cognitive Science, Tsinghua University, Beijing 100084, China  
\textsuperscript{3} Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA  
\textsuperscript{4} Bell Labs, 600 Mountain Avenue, Murray Hill, New Jersey 07974, USA  
\textsuperscript{5} Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China  
\* Corresponding author: xyuan@bell-labs.com, jlsuo@tsinghua.edu.cn

1. INTRODUCTION

High resolution images are widely used in our daily life, whereas high-speed video capture is challenging due to the low frame rate of cameras working at the high resolution mode. Digging deeper, the main bottleneck lies in the low throughput of existing imaging systems. Towards 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. After this, algorithms are employed to retrieve the video frames from the coded snapshot. Recently developed Plug-and-Play (PnP) algorithms make it possible for SCI reconstruction in large-scale problems. However, the lack of high-resolution encoding systems still precludes SCI’s wide application. In this paper, we build 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 achieve a 10-mega pixel SCI system to capture high-speed scenes, leading to a high throughput of 4.6G voxels per second. Both simulation and real data experiments verify the feasibility and performance of our proposed HCA-SCI scheme. © 2021 Optical Society of America

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. INTRODUCTION

Recent advances of machine vision with applications in robotics, drones, autonomous vehicles and cellphones have driven high resolution images into our daily life. However, high-speed high-resolution videos though having wide applications in various fields such as physical phenomenon observation, biological fluorescence imaging, live broadcast of sports, on the other hand, are facing the challenge of low throughput due to the limited frame rate of existing cameras working at the high resolution mode. This is further limited by the memory, bandwidth and power. This paper aims to address this challenge by building a high-speed high-resolution imaging system using compressive sensing. Specifically, our system captures the high-speed scene in an encoded way and thus maintain the low bandwidth during capture. After this, reconstruction algorithms are employed to reconstruct the high speed high resolution scenes to achieve high throughput. Note that though the idea of video compressive sensing has been proposed before, scaling it up to 10-mega pixel in spatial resolution presents the challenges of both hardware implementation and algorithm development. Please refer to Fig. 1 for a real high-speed scene captured by our newly built camera.

While 10-mega pixel lenses and sensors are both accessible, 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, and thus makes 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 like super-resolution [1–3], deblurring [4–6], depth estimation [7–9] and so on. 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 by a two-dimensional (2D) detector. A video SCI system is typically composed of an objective lens, a temporarily 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
mask will change their directions accordingly, and 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 a novel hybrid coded aperture snapshot compressive imaging (HCA-SCI) scheme, which can provide \textit{multiplexed shifting patterns} to encode the image plane without physical movement of the mask.

- Inspired by the plug-and-play (PnP) algorithms for large-scale SCI in [17], we implement a reconstruction algorithm which 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, 30 are implemented, providing reconstructed frame rate up to 450 frames per second (fps) for a conventional camera operating at 15 fps, verifying the effectiveness of 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 frame rate of the camera 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 in recent years [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 like [23, 24], which takes advantage of the temporal shifting feature of rolling shutter cameras or streak cameras for the temporal-variant mask generation.

Paralleled to these systems, different algorithms are proposed to improve the SCI reconstruction performance as well. 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 (GMM) [27, 28] are employed, forming widely used TwiST [29], GAP-TV [25], DeSCI [26] and some other reconstruction algorithms. Generally, iterative optimization based algorithms 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,
The incident light from a scene is first collected by the primary lens 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. In the decoding, the video is reconstructed under a PnP framework incorporating deep denoising prior into a joint framework of iterative and learning-based methods (right).

**B. Encoding mask generation**

The aperture of the system (i.e., the activate 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 towards certain directions. As shown in Fig. 3, because the lithography mask is placed in front of the image plane, the light beam propagating towards certain directions. Therefore, the light beams from the same sub-aperture are focused at different virtual image planes, which can thus generate corresponding shifting masks. In our implementation, after extensive experiments, the mask is placed in front of the second image plane with a distance of 80µm, which is a good trade-off between the sharpness and the shift.

In this paper, we build a novel video SCI system by 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 [17].
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 $X \in \mathbb{R}^{n_x \times n_y \times B}$, containing $B$ consecutive high speed frames is modulated by $B$ encoding masks $C \in \mathbb{R}^{n_x \times n_y \times B}$ and integrated by the sensor to generate a snapshot coded measurement $Y$. Then $Y$ can be expressed as

$$Y = \sum_{k=1}^{B} C_k \odot X_k + G,$$

where $\odot$ denotes the Hadamard (element-wise) product; $G \in \mathbb{R}^{n_x \times n_y}$ is the measurement noise; $C_k = C(\cdot; k)$ and $X_k = X(\cdot; k)$ 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

$$y = Hx + g,$$

where $y = \text{Vec}(Y) \in \mathbb{R}^n$ and $g = \text{Vec}(G) \in \mathbb{R}^n$ with $n = n_x n_y$. Corresponding video signal $x \in \mathbb{R}^{nd}$ is given by $x = [x_1^\top, \ldots, x_B^\top]^\top$, where $x_k = \text{Vec}(X_k) \in \mathbb{R}^n$. Different with traditional CS, the coding matrix $H \in \mathbb{R}^{n \times nB}$ in video SCI is sparse and has a special structure which can be written as

$$H = [D_1, \ldots, D_B],$$

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

### D. system calibration

Though 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 before 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 $I$ and a background pattern $B$ are captured with LCoS projecting white and black patterns, respectively. After that, we put on the lithography mask and capture each dynamic encoding mask $C$ 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 get the accurate encoding mask $C'$ following Eq. (5). Besides, 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.

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

### 4. RECONSTRUCTION ALGORITHM

The reconstruction of high speed videos from the snapshot coded measurement is an ill-posed problem. As mentioned before, to solve this problem, different priors and frameworks have been employed. Roughly, the algorithms can be categorized into the following 3 classes [12]: i) Regularization (or priors) based optimization algorithms with well known methods such as TwIST [29], GAP-TV [25] and DeepSCI [26]. ii) End-to-end deep learning based algorithms [15, 32, 34], like BIRNAT [30] which reaches state-of-the-art performance and recent developed MetaSCI [31] that using meta-learning to improve adaption capability for different masks in SCI reconstruction. iii) Plug-and-Play (PnP) algorithms that using deep denoising networks into the optimization framework such as ADMM and GAP.

Among these algorithms, 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 the inflexibility, i.e., re-training being required for a new system. Though recent works like MetaSCI [31] try to mitigate this problem with meta-learning or transfer learning and thus march forward to large-scale SCI problem with patch-wise reconstruction strategy, it still takes a long time for the training and adaption of 10-meta pixel scale SCI reconstruction. For example, MetaSCI takes about two weeks for the $256 \times 256 \times 10$ base model training on a single NVIDIA 2080Ti GPU, and further adaption performed on more than $570 \times 256 \times 256 \times 10$ sub-tasks (overlapped patches extracted from a 10-mega pixel image) costs about two months, which is impractical in real applications (more GPUs can be used to mitigate this challenge). By contrast, PnP has achieved a well balance of speed, flexibility and accuracy. Therefore, we employ the 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 purpose.
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

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

where $R(x)$ is a regularizer or prior being imposed on $x$, which can be a TV, sparse prior or a deep prior [39] and $\lambda$ is an auxiliary parameter.

Let $k$ index the iteration number, through a two-step iteration, the minimization in Eq. (6) can be solved as follows:

1. Solving $x$:

$$x^{(k)} = v^{(k-1)} + H^\top (HH^\top)^{-1}(y - Hv^{(k-1)}), \quad (7)$$

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

2. Solving $v$:

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

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 $v$ following Eq. (8) is equivalent to performing a denoising process on $x$. 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:** $H, y,$

1. Initialize: $v^{(0)}, \lambda_0, \xi < 1, k = 1, K_{\text{Max}}$.
2. While Not Converge and $k \leq K_{\text{Max}}$ do
   3. Update $x$ by Eq. (7).
   4. Update $v$.
   5. If $k \leq K_1$ then
      6. $v^{(k)} = TV(x^{(k)})$
     7. else
      8. $v^{(k)} = TV(x^{(k)})$
      9. $v^{(k)} = \text{FastDVDNet}(v^{(k)})$

5. RESULTS

In this section, we conduct a series of experiments on both simulation data 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 512x512 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 realized a good balance between the system’s light throughput and the encoding masks’ quality when taking the physical factors into account.

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 (1536x2048) and the video data (10-mega) in our system. Both traditional optimization-based methods and learning based methods have difficulty in dealing with data of this scale. So, the design of the optimal multiplexing pattern is still an open challenge which is worthy for future investigation.

B. Reconstruction comparison between different algorithms on simulation datasets

In order 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 256x256, 512x512, 1024x1024 and two compressive rates (Cr) of 10 and 20.

Two datasets named Football and Hummingbird used in [17] and three datasets named ReadySteadyGo, Jockey and YachtRide provided in [41] are employed in our simulation. According to the forward model depicted in Section C, to simulate the encoding masks, we first

![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 512x512 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.](image-url)
Table 1. The average results of PSNR in dB (left entry in each cell) and SSIM (right entry in each cell) by different algorithms (Cr = 10). BIRNAT fails at large-scale due to limited GPU memory.

| 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 |
|          | BIRNAT              | 30.63, 0.9022 | 29.16, 0.8459 | 28.92, 0.8698 | 31.59, 0.8953 | 29.03, 0.8470 | 29.87, 0.8720 |
| 1024 × 1024| 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 |

Table 2. The average results of PSNR in dB (left entry in each cell) and SSIM (right entry in each cell) by different algorithms (Cr = 20). BIRNAT fails at large-scale due to limited GPU memory.

| 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 |
|          | 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 |
| 512 × 512| 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 |

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 1-6 for the reconstructed videos.
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. And 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 PSNR and SSIM for each algorithm are summarized in Table 1 and Table 2 for Cr = 10 and 20, respectively. It is worth nothing that due to the limited memory of our GPU (GeForce RTX 3090 with 24GB memory), we only test BIRNAT on 256 × 256 scale datasets with Cr equaling to 10 and 20, and the adversarial training is not involved. For GAP-TV and PnP-TV-FastDVDNet, the reconstruction is conducted on a platform equipped with a Intel(R) Core(TM) i7-9700K CPU (3.60GHz, 32G memory) and a GeForce RTX 2080 GPU with 8G memory; 160 and 250 iterations are taken for Cr = 10 and 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 1.5dB and 4dB on average, respectively. ii) On the 256 × Cr10 and 256 × Cr20 datasets, BIRNAT leads the best performance (with sufficient training) exceeding PnP-TV-FastDVDNet for about 3.4dB and 0.4dB, respectively. iii) BIRNAT is the fastest algorithm during inference period with hundreds times shorter time than GAP-TV and PnP-TV-FastDVDNet. However, the training process of BIRNAT is quite time-consuming, and it takes about one 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 PnP-FFDNet leads to some unpleasant artifacts.

![Fig. 6. Noise robustness comparison between multiplexed masks and non-multiplexed masks.](image)

C. Validation of multiplexed shifting mask’s noise robustness on simulation datasets

As mentioned in Section 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 of experiments on simulation datasets containing different level 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 are utilized here as the clean original frames. And we normalize the masks with a light throughput factor calculating 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, 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 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 the noise get larger, it will also impose significant impact on the reconstruction process until totally disrupt the reconstruction. So 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 build 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 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 4.6G 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 which, 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 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. But when Cr gets larger, the reconstructed frames of GAP-TV tend to be blurry and have more background noise. 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.
Figure 7. Reconstruction results of 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 of sizes around 400×400 to demonstrate the high-speed motion.
Fig. 8. Reconstruction comparison between 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 of size of 512×512 to demonstrate the high-speed motion. See Visualization 7 for the reconstructed videos.
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 4.6G 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 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 450Hz since the hybrid coding scheme further decreases the light throughput to some extend compared with conventional coding strategies. In the future, a worthwhile extension is to introduce new photon-efficient aperture coding devices to raise the signal-to-noise ratio 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 performances of existing physical components, designing novel algorithms which are capable of SCI reconstruction with non-uniform masks may be a economical and feasible way. Meanwhile, time-efficient reconstruction algorithms and feasible multiplexing pattern design methods for large-scale (like 10-mega pixel or even Giga pixel) SCI reconstruction is still an open challenge in the foreseeable future. Besides, 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.

Acknowledgments. Zhihong Zhang and Chao Deng (marked with † in the author list) contributed equally to this paper.

FINANCIAL DISCLOSURES. X. Y., Nokia (E).

Data Availability Statement. 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.

REFERENCES

1. G. Carles, J. Downing, and A. R. Harvey, “Super-resolution imaging using a camera array,” Opt. letters 39, 1889–1892 (2014).
2. S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, “Fast and robust multiframe super-resolution,” IEEE Transactions on Image Process. 13, 1327–1344 (2004).
3. Q. Sai, N. Xia, and G. Sapiro, “Coded aperture super-resolution image reconstruction,” in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (IEEE, Las Vegas, NV, USA, 2008), pp. 833–836.
4. T. S. Cho, A. Levin, F. Durand, and W. T. Freeman, “Motion blur removal with orthogonal parabolic exposures,” in 2010 IEEE International Conference on Computational Photography (ICCP), (IEEE, Cambridge, MA, USA, 2010), pp. 1–8.
5. R. Raskar, A. Agrawal, and J. Tumblin, “Coded exposure photography: Motion deblurring using fluttered shutter,” ACM Transactions on Graph. 25, 795–804 (2006).
6. C. Zhou and S. Nayar, “What are good apertures for defocus deblurring?” in 2009 IEEE International Conference on Computational Photography (ICCP), (IEEE, San Francisco, CA, USA, 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 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (IEEE, Boston, MA, USA, 2015), pp. 3497–3506.
10. P. Llul, 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 (2013).
11. A. Wagadarikar, R. John, R. Willett, and D. Brady, “Single disperser design for coded aperture snapshot spectral imaging,” Appl. Opt. 47, B44 (2008).
12. X. Yuan, D. J. Brady, and A. K. Matsaggelos, “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 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (IEEE, Colorado Springs, CO, USA, 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 2011 International Conference on Computer Vision (ICCV), (IEEE, Barcelona, Spain, 2011), pp. 287–294.
15. M. Qiao, Z. Meng, J. Ma, and X. Yuan, “Deep learning for video compressive sensing,” APL Photonics 5, 030801 (2020).
16. X. Yuan, P. Llul, X. Liao, J. Yang, D. J. Brady, G. Sapiro, and L. Carin, “Low-Cost Compressive Sensing for Color Video and Depth,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (IEEE, Columbus, OH, USA, 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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (IEEE, Seattle, WA, USA, 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,” IEEE Transactions on Pattern Analysis Mach. Intell. pp. 1–1 (2021).
19. Y. Sun, X. Yuan, and S. Pang, “Compressive high-speed stereo imaging,” Opt. Express 25, 18182 (2017).
20. L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” Phys. D: Nonlinear Phenom. 60, 259–268 (1992).
21. M. Tassano, J. Delon, and T. Veit, “FastDVDnet: Towards Real-Time Deep Video Denoising Without Flow Estimation,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (IEEE, Seattle, WA, USA, 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 Transactions on Pattern Analysis 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 2019 IEEE International Conference on Computational Photography (ICCP), (IEEE, Tokyo, Japan, 2019), pp. 1–8.
25. X. Yuan, “Generalized alternating projection based total variation minimization for compressive sensing,” in 2016 IEEE International Conference on Image Processing (ICIP), (IEEE, Phoenix, AZ, USA, 2016), pp. 2539–2543.
26. Y. Liu, X. Yuan, J. Suo, D. Brady, and Q. Dai, “Rank minimization for snapshot compressive imaging,” IEEE Transactions on Pattern Analysis Mach. Intell. 41, 2990–3006 (2019).
27. J. Yang, X. Yuan, X. Liao, P. Llul, D. J. Brady, G. Sapiro, and L. Carin, “Video Compressive Sensing Using Gaussian Mixture Models,” IEEE Transactions on Image Process. 23, 4863–4878 (2014).
28. Jianbo Yang, Xuejun Liao, Xin Yuan, P. Llul, D. J. Brady, G. Sapiro, and L. Carin, “Compressive Sensing by Learning a Gaussian Mixture Model,” IEEE Transactions on Image Process. 23, 2015–2027 (2014).
29. J. Bioucas-Dias and M. Figueiredo, “A New TwIST: Two-Step Iterative Shrinkage/Thresholding Algorithms for Image Restoration,” IEEE Transactions on Image Process. 24, 106–119 (2015).

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 Computer Vision – ECCV 2020, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds. (Springer International Publishing, Cham, 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 2021 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 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (IEEE, Las Vegas, NV, USA, 2016), pp. 449–458.

34. J. Ma, X.-Y. Liu, Z. Shou, and X. Yuan, “Deep Tensor ADMM-Net for Snapshot Compressive Imaging,” in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), (IEEE, Seoul, Korea (South), 2019), pp. 10222–10231.

35. X. Liao, H. Li, and L. Carin, “Generalized Alternating Projection for Weighted-$\ell_{2,1}$ Minimization with Applications to Model-Based Compressive Sensing,” SIAM J. on Imaging 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 2018 IEEE International Symposium on Information Theory, (IEEE, 2018).

38. S. Jalali and X. Yuan, “Snapshot Compressed Sensing: Performance Bounds and Algorithms,” IEEE Transactions on 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 Transactions on 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 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (IEEE, Columbus, OH, USA, 2014), pp. 2862–2869.

41. A. Mercat, M. Viitanen, and J. Vanne, “UVG dataset: 50/120fps 4K sequences for video codec analysis and development,” in Proceedings of the 11th ACM Multimedia Systems Conference, (ACM, Istanbul Turkey, 2020), pp. 297–302.