


REVIEW

Open Access



From compressive sampling to compressive tasking: retrieving semantics in compressed domain with low bandwidth

Zhihong Zhang^{1†}, Bo Zhang^{1†}, Xin Yuan^{2,3†}, Siming Zheng^{4,5}, Xiongfei Su², Jinli Suo^{1*} , David J. Brady⁶ and Qionghai Dai^{1*}

[†]Zhihong Zhang, Bo Zhang and Xin Yuan contributed equally to this work.

*Correspondence: jlsuo@tsinghua.edu.cn; qhdai@tsinghua.edu.cn

¹Department of Automation, Tsinghua University, 100084 Beijing, China

²Research Center for Industries of the Future and School of Engineering, Westlake University, 310024 Hangzhou, China

³Key Laboratory of 3D Micro/Nano Fabrication and Characterization of Zhejiang Province, Westlake University, 310024 Hangzhou, China

⁴Computer Network Information Center, Chinese Academy of Sciences, 100190 Beijing, China

⁵University of Chinese Academy of Sciences, 100049 Beijing, China

⁶Wyant College of Optical Sciences, University of Arizona, AZ 85721 Tucson, USA

Abstract

High-throughput imaging is highly desirable in intelligent analysis of computer vision tasks. In conventional design, throughput is limited by the separation between physical image capture and digital post processing. Computational imaging increases throughput by mixing analog and digital processing through the image capture pipeline. Yet, recent advances of computational imaging focus on the “compressive sampling”, this precludes the wide applications in practical tasks. This paper presents a systematic analysis of the next step for computational imaging built on snapshot compressive imaging (SCI) and semantic computer vision (SCV) tasks, which have independently emerged over the past decade as basic computational imaging platforms.

SCI is a physical layer process that maximizes information capacity per sample while minimizing system size, power and cost. SCV is an abstraction layer process that analyzes image data as objects and features, rather than simple pixel maps. In current practice, SCI and SCV are independent and sequential. This concatenated pipeline results in the following problems: *i*) a large amount of resources are spent on task-irrelevant computation and transmission, *ii*) the sampling and design efficiency of SCI is attenuated, and *iii*) the final performance of SCV is limited by the reconstruction errors of SCI. Bearing these concerns in mind, this paper takes one step further aiming to bridge the gap between SCI and SCV to take full advantage of both approaches.

After reviewing the current status of SCI, we propose a novel joint framework by conducting SCV on raw measurements captured by SCI to select the region of interest, and then perform reconstruction on these regions to speed up processing time. We use our recently built SCI prototype to verify the framework. Preliminary results are presented and the prospects for a joint SCI and SCV regime are discussed. By conducting computer vision tasks in the compressed domain, we envision that a new era of snapshot compressive imaging with limited end-to-end bandwidth is coming.

Keywords: Snapshot compressive imaging, Semantic computer vision, Computational imaging, Deep learning

Introduction

Imaging is the founding pillar of computer vision (CV), which is now a core technique in today's digital world, especially in multimedia, with a wide range of applications in consumer and military electronics as well as the emerging topics such as the metaverse. Starting from the invention of charge-coupled devices (CCD) in 1970 [1], digital imaging techniques have revolutionized CV. As life keeps moving forward, so does imaging technology. Particularly, **computational imaging** (CI) [2, 3] has come to a new stage due to recent advances in artificial intelligence, especially machine learning and deep learning. We are fortunate to witness a new era driven by CI and CV. This paper moves one step further, and aims to bridge the gap between CI and CV, in particular, between snapshot compressive imaging (SCI) [4] and semantic CV (SCV) tasks.

CI, an emerging multidisciplinary topic covering optics, sensors, image processing, signal processing and machine learning, aims to incorporate computational design into the image capturing process to improve image quality [5–8], optimize imaging systems and procedures [9–11], acquire high-dimensional visual information [12–14] or boost the performance of subsequent high-level tasks [15–17]. On the other hand, CV mainly focuses on *high-level image processing and understanding tasks* such as image classification [18, 19], semantic segmentation [20, 21], object detection and tracking [22–24], video captioning [25, 26] and so on, with little attention on the preceding image acquisition process. CI and CV are two close fields related to imaging acquisition, processing and understanding. Sharing the similar goal of intelligent imaging and driven by common engines such as machine learning, CI and CV have come to the same big stage to shed light on each other. Therefore, now it is the right time to consider CI and CV jointly. In this manner, novel frameworks that design the whole process of visual information acquisition, processing and understanding will emerge.

As two burgeoning topics in CI and CV, SCI and SCV have attracted more and more attention in their respective fields. On the one hand, with the increasing demand for ultra-high-definition video acquisition like 4K and 8K recording, the bandwidth of current imaging systems has become a bottleneck for further development. The end of Moore's Law makes it hard to break down the barrier simply by improving the hardware performance. Instead, compressive sensing (CS) based SCI may provide a feasible solution to this dilemma by conducting data compression during acquisition [4]. By investigating the intrinsic redundancy prior of natural images or videos, SCI is able to acquire high-dimensional signal such as videos and hyper-spectral images with a conventional 2D camera through well-designed coded measuring strategies. In this manner, the bandwidth can be greatly reduced with little sacrifice of useful information. Furthermore, recent advances in SCI systems and reconstruction algorithms have paved the way for SCI's applications in our daily lives.

On the other hand, in the era of big data, we are facing more and more multimedia data from social media, surveillance, self-driving, Internet-of-things, remote sensing, etc., which heavily rely on automatic information analysis. In this case, extracting high-level semantic information from massive multimedia data has become one of the most important tasks in computer vision. Equipped with the rapid development of deep learning, we have made significant progress in the field of SCV. For instance, we are constantly pushing forward the intelligent level of our automatic information processing

systems from image classification [19], semantic segmentation [27], object detection and tracking [23] to action recognition [28], facial expression recognition [29], scene understanding [30] and video captioning [31].

Although SCI and SCV are both playing increasingly important roles in CI and CV respectively, they have not been jointly considered, which hinders their mutual development. Thus, bridging the gap between SCI and SCV has become an urgent direction and prospect to draw on each other's strengths and achieve a win-win future. In this paper, we propose a **novel framework that incorporates SCI and SCV** to efficiently implement the whole process of compressive video acquisition and semantic information retrieval. To validate the feasibility of linking SCI and SCV, and meanwhile to expose the challenges to push forward the joint framework, we build an SCI prototype system which can capture coded high-speed videos with a conventional low-speed sensor. Besides, a semantic information retrieval pipeline involving *object tracking, depth estimation, and scene understanding* is designed to enable qualitative and quantitative description of the target video. We conduct an outdoor experiment and demonstrate preliminary results in this paper. As the first attempt to combine SCI and SCV for practical applications in the natural scene, as depicted in Fig. 1, we believe our work can provide new insights into the collaboration of CI and CV, and shed light on the future applications of corresponding techniques in our daily lives.

A brief review of recent advances of SCI

In this section, we first review the recent development of SCI, and then focus on the current status of joint SCI and SCV development.

Development of SCI

As an emerging and representative sub-field of computational imaging, SCI has become more and more popular in recent years. The intrinsic merits of low bandwidth and high data throughput of SCI provide great potential in various fields such as autonomous-driving, video surveillance and so on. With the maturity of SCI systems, reconstruction algorithms and related high-level tasks, it is convinced that SCI will march forward for

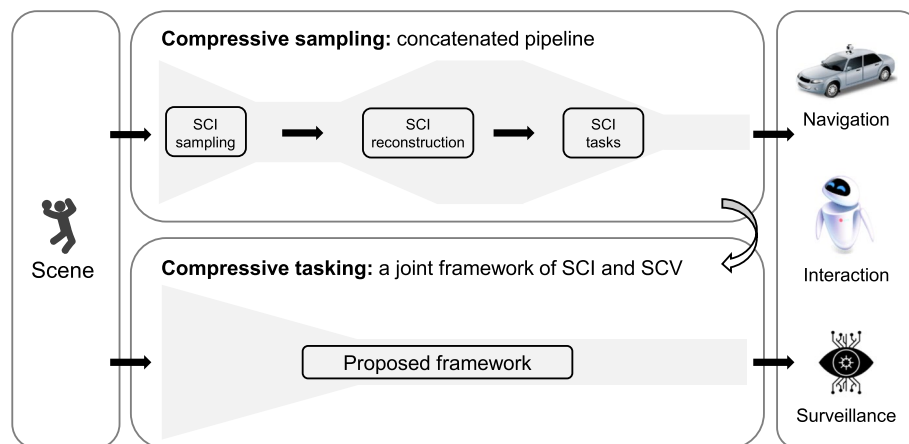


Fig. 1 Our perspective from “compressive sampling” to “compressive tasking”

practical applications in the near future. In this paper, we focus on video SCI, which compresses multiple video frames into a single shot measurement and then reconstructs the desired high-speed frames by algorithms.

SCI systems

The main function of SCI systems is to capture high-dimensional visual signal with low-dimensional sensors according to designed multiplexing schemes, and solve the inverse problem afterwards to recover the desired data. With the development of optics, materials and manufacturing techniques, various SCI systems have been proposed in recent years [5, 9, 32–42]. As illustrated in Fig. 2, a typical SCI system introduces an encoding module to a conventional imaging system. The encoding module is employed to generate uncorrelated encoding masks and modulate corresponding incoming images before integrating them into a single frame (coded measurement) on the sensor.

A straightforward implementation of the encoding module is to simply place a spatial light modulator such as digital mirror device (DMD) [32, 37] in the image plane. Micro-electromechanical system (MEMS) based DMDs can rapidly change the modulation patterns and realize dynamic encoding by switching their micro-mirrors' directions with articulated hinges. Liquid crystal on silicon (LCoS) collocated with a polarizing beamsplitter can achieve the same function, as the polarizing beamsplitter can transform the polarization modulation of LCoS into amplitude modulation [33, 38]. Using a shifting lithography mask translated by a piezo [34] is a simple substitute to aforementioned DMDs or LCoSs, but the mechanical translation may cause system instability and degrade the image quality.

There are other methods that use specially designed sensors to implement the encoding process. For example, sensors with pixel-wise exposure ability can perform the frame encoding directly on the sensor plane before integration, while no extra optical elements are needed compared to conventional imaging systems [40]. Most recently, a novel programmable sensor called Coded-2-Bucket camera was designed with two light-collecting buckets per pixel [43]. By switching between the two buckets during exposure, pixel-wise shutter control is realized and a pair of complementary patterns encoded measurements can be acquired during a single exposure. In addition to the above methods using programmable sensors, video frame encoding can also be achieved in indirect ways by exploiting the temporal shifting feature of streak cameras or rolling shutter cameras [5, 9, 44]. Furthermore, some sophisticated encoding strategies are also proposed to further boost the system performance or efficiency [32, 35, 42].

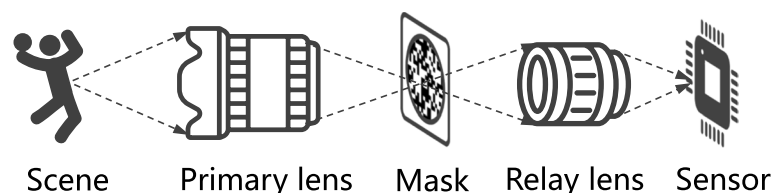


Fig. 2 Schematic of the video snapshot compressive imaging system. The video frames are encoded by uncorrelated masks in the image plane and then relayed to the sensor plane and integrated to form a coded snapshot

SCI reconstruction algorithms

Based on the underlying principle of SCI, the mathematical model of SCI can be formulated as

$$\mathbf{Y} = \sum_{k=1}^B \mathbf{M}_k \odot \mathbf{X}_k + \mathbf{G}, \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^{n_x \times n_y \times B}$ and $\mathbf{M} \in \mathbb{R}^{n_x \times n_y \times B}$ denote B high-speed video frames and corresponding encoding masks with spatial-resolution of $n_x \times n_y$, respectively; \odot denotes the Hadamard (element-wise) product. $\mathbf{Y} \in \mathbb{R}^{n_x \times n_y}$ represents the coded measurement, and $\mathbf{G} \in \mathbb{R}^{n_x \times n_y}$ is the measurement noise. We can further rewrite Eq. (1) to the following form which is coincident with the standard compressive sensing (CS) problem

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{g}, \quad (2)$$

where $\mathbf{y} = \text{Vec}(\mathbf{Y}) \in \mathbb{R}^n$ and $\mathbf{g} = \text{Vec}(\mathbf{G}) \in \mathbb{R}^n$ with $n = n_x n_y$. The original high-speed video signal $\mathbf{x} \in \mathbb{R}^{nB}$ is given by $\mathbf{x} = [\mathbf{x}_1^\top, \dots, \mathbf{x}_B^\top]^\top$, where $\mathbf{x}_k = \text{Vec}(\mathbf{X}_k) \in \mathbb{R}^n$. It is worth noting that the coding matrix $\mathbf{H} \in \mathbb{R}^{n \times nB}$ in video SCI has a special structure which can be expressed as

$$\mathbf{H} = [\mathbf{D}_1, \dots, \mathbf{D}_B], \quad (3)$$

where $\mathbf{D}_k = \text{diag}(\text{Vec}(\mathbf{M}_k)) \in \mathbb{R}^{n \times n}$. This special structure makes it possible to reduce the computational complexity in some optimization-based SCI reconstruction algorithms [45–47].

As illustrated in Eq. (2), recovering high-speed frames from the coded measurement is a highly ill-posed inverse problem. However, theoretical analysis has proved that the reconstruction error of SCI is bounded even when the compressive ratio $B > 1$ [48, 49]. To solve the inverse problem, a variety of SCI reconstruction algorithms [4, 33, 37, 45, 46, 50–56] based on different priors and frameworks have been proposed, which are summarized in Table 1.

Early methods were mainly developed under optimization frameworks and hand-crafted image/video priors. For example, the classical optimization-based method GAP-TV [46] introduces the total variation (TV) prior constrain into the forward model of SCI and iteratively solves the optimization problem with the generalized alternating projection (GAP) framework [60]. The state-of-the-art optimization-based method DeSCI [45] alternatively leverages the non-local self-similarity prior and alternating direction method of multiplier (ADMM) framework [61] to achieve higher reconstruction quality

Table 1 Summary of different frameworks and algorithms for SCI reconstruction

Category	Algorithm	Pros & Cons	Reference
optimization	TwIST, GAP-TV, DeSCI, GMM, KSVD	flexible, diverse quality, slow	[33, 45, 46, 50, 57]
deep learning	Tensor ADMM-Net, E2E-CNN, BIRNAT, MetaSCI, 3D-CNN, RevSCI	fast inference, high quality, inflexible, large GPU memory consumption, long training time, extensive training data	[35, 37, 51–55, 58]
plug-and-play	PnP-FFDNet, PnP-TV-FastDVDNet	flexible, moderate speed, moderate quality	[42, 47, 59]

but suffers from low speed at the same time. Generally, optimization-based algorithms are robust but have high computational complexity. There are also some “shallow-learning” based methods such as Gaussian mixture model (GMM) [56, 57] and dictionary learning [33] proposed for SCI reconstruction, but their performance improvement over aforementioned optimization based methods is limited.

Recent advances in deep learning have boosted the performance of classical low-level computer vision tasks such as super-resolution, denoising and deblurring to a large extent, which meanwhile inspires the emergence and rapid development of many learning-based SCI reconstruction algorithms [37, 51–55, 58]. The first deep learning architecture for SCI reconstruction was proposed in [52], which is based on a simple fully-connected neural network learning to map directly coded measurements to video frames. An unfolding network called Tensor ADMM-Net was designed by generalizing the standard tensor ADMM algorithm to a learnable deep neural network [53]. After that, an end-to-end convolutional neural network (CNN) with encoder-decoder structure and residual learning strategy was proposed [37]. As early attempts to solve SCI reconstruction problem with deep neural networks, these learning based methods have intrinsic advantage in reconstruction speed compared with iteration based methods, and achieve continuously improving reconstruction quality. However, it was not until the advent of BIRNAT [51] that learning-based methods surpassed the state-of-the-art optimization based algorithm DeSCI in terms of reconstruction quality. BIRNAT uses CNN to reconstruct the first frame, and subsequent frames are reconstructed by a bidirectional recurrent neural network (RNN) in a sequential manner. Then, RNN was further improved with optical flow and applied to dual-view video compressive sensing [35]. Most recently, a novel learning based method combining 3D CNN and deep unfolding was designed, which achieved significant improvements over previous state-of-the-art methods [55].

Apart from reconstruction quality and speed, flexibility is also an important factor that needs to be considered in practical applications. Although BIRNAT leads state-of-the-art reconstruction quality and has near real-time inference speed, its flexibility is greatly limited by the time-consuming training process and requirement of large GPU memory. This problem becomes even severe in real scenarios, where encoding masks will inevitably change over time requiring retraining of the network consequently. To mitigate this problem, MetaSCI leverages meta-learning to achieve fast mask adaption to new encoding masks in SCI reconstruction, which significantly reduces retraining or adaptation time. Another way to balance the trade-off among quality, speed and flexibility is using Plug-and-Play (PnP) based methods like PnP-FFDNet [47] and PnP-FastDVDNet [59], which plug a pre-trained network as prior into the optimization frameworks to speed up the convergence, avoid tedious training process and improve the reconstruction quality.

Joint development of SCI and SCV

Different from SCI that focuses on the capture side, the processing framework of semantic computer vision in dynamic scenes usually includes the following stages: scene understanding, object detection and tracking, behavior classification and description, human or vehicles identification, and quantitative and qualitative data fusion.

As two sequential modules in the visual information acquisition and processing pipeline, hardware-oriented CI and task-driven CV are naturally complementary. Thus, to improve the overall performance has become a new trend to combine SCI and SCV [15, 16, 62–64].

Kwan et al. came up with a novel object tracking and classification scheme using compressive measurements captured by pixel-wise coded exposure cameras, and demonstrated the efficacy of the proposed approach with extensive experiments using visible light, short-wave infrared (SWIR), mid-wave infrared (MWIR) and long-wave infrared (LWIR) videos, respectively [63, 65, 66]. They retrained YOLO with synthetic compressive measurements to extract target locations and linked them to create trajectories. Then, the classification was separately conducted using a ResNet classifier trained on the compressive measurements. It's worth noting that this work directly regarded the coded measurement as a single image and performed subsequent computer vision tasks, but neglected the video property of the measurements, which attenuated the advantages of SCI in high-throughput high-speed imaging. Hu et al. further extended this work by proposing VODS, i.e., video object detection from one single coded image through optoelectronic neural network [15]. In this work, they modelled the optical encoding process with an optical neural network with the pixel values of the encoding masks as trainable parameters. Then a CNN decoder and an object detection network were cascaded sequentially to perform the following object detection and classification tasks. During training, all trainable parameters in the optical encoder, CNN decoder and object detection network were jointly updated to achieve overall-optimized performance. By taking advantage of SCI's high-speed imaging feature and deep learning's superiority in object detection, VODS made it possible for high-speed object detection with low-speed sensors. Similarly, Okawara et al. implemented reconstruction-free action recognition from a single coded image through combining SCI and deep neural networks [16]. Apart from the co-optimization strategy, they also designed a shiftvariant convolution to adapt for spatial unsmoothness caused by spatial-temporal encoding of SCI. Finally, they achieved a relatively high recognition accuracy with a single coded image, which was on par with traditional 3D convolutional networks with original high-speed videos.

Our perspective: A novel joint framework of SCI and SCV

To bridge the gap between emerging SCI techniques and SCV tasks, we propose a novel joint framework (Fig. 3) to take advantage of both parties and efficiently implement the entire process of coded high-speed video acquisition, qualitative description generation and quantitative description generation. We believe that the proposed framework can provide a possible route to blur the boundary between high-throughput video capture and corresponding high-level semantic information retrieval. In this section, we will describe the detailed design of the proposed framework from the following three aspects, i.e., SCI for coded high-speed video acquisition, measurement domain SCV for efficient *semantic information retrieval* and video domain SCV for the trade-off among accuracy, speed and field of view (FOV). The flowchart of the proposed framework is shown in Fig. 3.

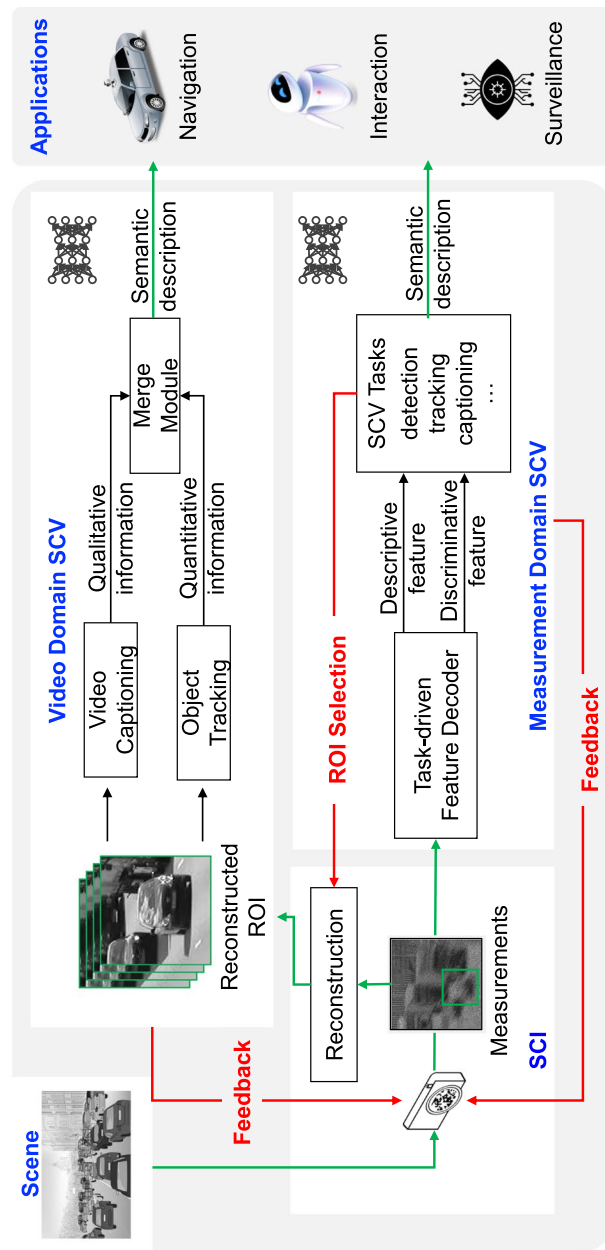


Fig. 3 Flowchart of the proposed joint SCI and SCV framework. SCI records the high-speed dynamic scenes and sends the coded measurements to the measurement domain SCV for fast semantic information retrieval. When fine semantic information is required for some regions of interest (ROI), the ROI selection signal will be generated and sent to the video domain SCV. Then, the video domain SCV will be awakened and conduct detailed semantic information retrieval after reconstructing the ROI videos. While providing semantic information to the terminal application, the SCV modules also send feedback to the SCI system to improve its self-adaptability and robustness

SCI system

At the beginning of the flowchart in Fig. 3, an SCI system is employed to achieve low-bandwidth high-throughput video capture from natural scenes with a conventional low-speed sensor. To improve flexibility and robustness in outdoor environment, the SCI system is designed to use a spatial light modulator as an encoding module, which can project arbitrary patterns without any mechanical translation [33, 38, 42]. The output of the SCI module is a series of coded measurements that store the raw visual information in a compressive manner through designed multiplexing scheme.

The inherent high-throughput and low-bandwidth feature of SCI can facilitate subsequent SCV tasks significantly. On the one hand, SCI has the ability to record high-speed dynamic scenes with a low-speed sensor and avoid information aliasing (i.e., “blurring” in video domain) through mask encoding. Thus, it can provide more detailed features for subsequent SCV tasks and boost their performance. On the other hand, compared with conventional video capture, the data format of coded measurements can reduce the data size by dozens of times and relieve the pressure on CPU/GPU memory during SCV processing.

Measurement domain SCV

Regarding the SCV part, there are two branches based on measurement domain processing and video domain processing, respectively. For large-scale SCI imaging, the reconstruction process is time-consuming, and the reconstructed video will occupy a large amount of memory, making the subsequent semantic information retrieval steps challenging. Considering that the coded measurements from SCI contain nearly all the useful information of the original scene, performing semantic information retrieval directly on coded measurements is theoretically feasible and has become a new trend [15, 16, 62–64].

Meanwhile, it is worth noting that the features of measurement domain and video domain are quite different. Therefore, novel task-driven feature decoders are required to extract descriptive features and discriminative features from the coded measurements and then feed them into corresponding modules of SCV tasks. Afterwards, with the extracted descriptive features, we can further retrieve the qualitative semantic information through scene understanding [30] or video captioning [31] methods. In addition, the discriminative features can be used to retrieve quantitative semantic information by object objection/tracking [23] and distance/velocity estimation approaches [67, 68]. Note that, the deep neural networks for these SCV tasks should be retrained on the extracted features to adapt to the measurement domain. Further adaption of network structures or training strategies could be employed to improve the final performance, too.

Video domain SCV

Although measurement domain processing provides a promising way to efficiently connect SCI and SCV, its drawback currently lies in inferior performance compared to video domain processing [15, 16, 64]. This conclusion is intuitive, as coded measurements have much less redundancy and suffer from “broken” natural image prior, which hinders the subsequent SCV tasks empowered by data-driven deep neural networks. Optimized

measurement feature decoders or powerful neural network structures with stronger representation ability may mitigate this problem in the future. But for now, we can circumvent it by employing video domain SCV for regions of interest (ROI).

To be specific, in normal circumstances, the measurement domain SCV is employed to get rough results directly from the output of SCI efficiently. When we want to focus on some regions containing objects of interest, an ROI selection signal can be generated from the object detection module of measurement domain SCV and passed to the reconstruction module of video domain SCV. Then, the video domain SCV will be awakened and retrieve detailed semantic information based on the reconstructed videos of selected measurement ROIs. As video domain SCV is coincident with conventional CV approaches, we can directly use existing pretrained neural network models or finetune them to finish the semantic information retrieval tasks. In a nutshell, the ROI selection signal acts as a bridge connecting the measurement domain SCV and video domain SCV. In this manner, we can achieve a good balance between FOV and semantic information retrieval accuracy, which helps the framework to flexibly adapt to different scenarios.

Closing the loop

Digging deeper into the collaboration of SCI and SCV, we can further introduce the feedback from SCV to SCI to adjust SCI system's compressive ratio [69] or optimize its encoding masks according to the semantic information retrieval results. In this manner, the framework's self-adaptability, robustness, and overall performance can be further improved [16].

To sum up, the proposed framework has the capacity to efficiently extract visual features from natural scenes and convert them into semantic descriptions, which facilitates many terminal applications such as navigation, interaction, and surveillance. To validate the feasibility of the framework and accumulate experience for further development, we conduct an outdoor experiment and demonstrate preliminary results in the next section.

Preliminary results and analysis

Hardware implementation and calibration

The optical diagram and real image of our prototype setup are illustrated in Fig. 4. Our system consists of a commercial primary lens (KOWA LM50HC, $f=50\text{ mm}$), two high-quality relay lenses (Chiopt, LS1610A), a polarizing beamsplitter (PBS) (Thorlabs, CCM1-PBS251/M), an LCoS (ForthDD, QXGA-3DM, 2048×1536 pixels, 4.5k refresh rate), and a CMOS sensor (JAI, GO-5000M-USB, 2560×2048 pixels). The primary lens first captures the scene and focuses it onto the virtual image plane. Then, the image of the scene is relayed to the LCoS plane which is conjugate to the first image plane. The LCoS then encodes the image with fast-changing random binary masks. Finally, multiple encoded images are integrated into a single coded snapshot measurement by the sensor.

Due to inevitable system aberrations and instabilities, the actual encoding masks in the sensor plane will slightly differ from the corresponding patterns projected by LCoS and will experience jitter over time. Therefore, a calibration step is required before each acquisition to guarantee the accuracy of the masks used for subsequent reconstruction. To be specific, we will place a Lambertian whiteboard on the objective plane. Then, each encoding mask M will be recorded sequentially with LCoS projecting corresponding

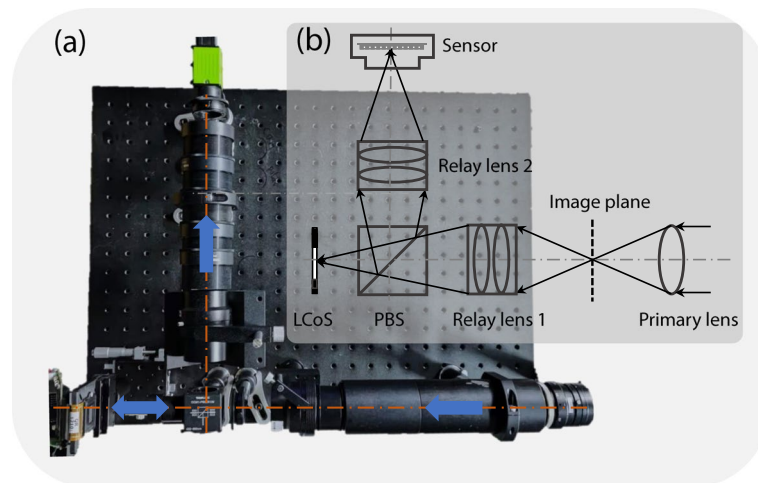


Fig. 4 The real image (a) and optical diagram (b) of our SCI system. Incident light from the scene is first collected by the primary lens and focused on the image plane. Then the image is relayed to the LCoS after passing through the polarizing beamsplitter (PBS). LCoS encodes the image with a random binary mask and reflects it back through the PBS. Finally, the encoded image is relayed to the sensor and integrated into a snapshot measurement

patterns. Besides, to eliminate the influence of background light and nonuniform illumination caused by ambient light or system vignetting, a background image R and an illumination image I will also be captured with LCoS projecting black and white patterns, respectively. Finally, the actual encoding mask M' can be calculated with Eq. (4). Note that, in order to improve the signal to noise ratio (SNR), all the images mentioned above will be recorded 50 times and averaged to a single image for calibration.

$$M' = (M - R) \oslash (I - R), \quad (4)$$

where \oslash denotes the element-wise division.

During acquisition, the camera and LCoS are synchronized by a signal generator. We set the capture frame rate of the camera to 20 frames per second (FPS), and 8 consecutive frames are encoded and collapsed to a single coded measurement. In this way, we can achieve a final frame rate of 160 FPS in the reconstructed video.

SCI reconstruction

Existing algorithms for SCI can be roughly divided into two categories, i.e., learning-based algorithms and optimization-based algorithms, each of which has pros and cons. For learning-based algorithms, although having higher inference speed and better reconstructed results, they lack the flexibility to masks and can hardly scale to different spatial sizes, especially large scale. On the contrary, optimization-based algorithms enjoy flexibility to different masks and scales, but suffer from low speed due to required numerous iterations.

In practical applications [70] with a large field-of-view, a model for large-scale reconstruction and high inference speed is urgently needed. To this end, we employ a physics-driven two-stage model [71] for the reconstruction. It combines the merits of

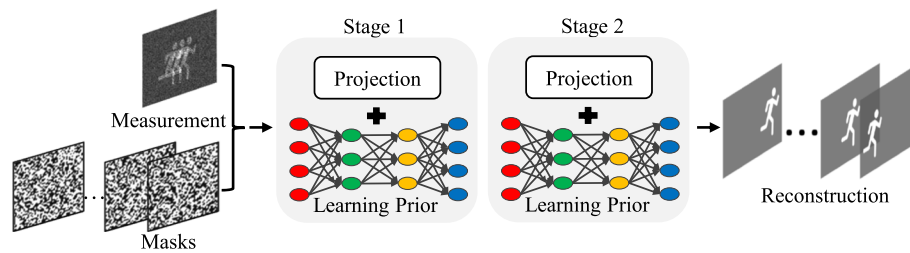


Fig. 5 The architecture of the physics-driven two-stage unrolling model for SCI reconstruction. Each stage in the model consists of a projection step and a prior learning step. By taking advantage of the unrolling design, this model can be trained with small patches and then finetuned to large-scale data efficiently

learning-based and optimization-based algorithms by utilizing an unfolding/unrolling framework. Models for large-scale data reconstruction can be trained by data with smaller spatial sizes, even with different masks. This makes network training no longer limited by GPU memory as before, thus breaking the bottleneck of existing algorithms.

The architecture of our model is shown in Fig. 5. It unfolds the iterations used in conventional optimization algorithms into 2 stages, and each stage is composed of a generalized alternating projection step [60] and a prior learning step. Note that we employ an invertible network [72] as the backbone in prior learning step to save memory during training.

SCV description

In the proposed framework, we classify the semantic information into two categories, i.e., qualitative information and quantitative information. To be specific, qualitative information refers to the objects, events and background contained in a scene, which is a sketch of the surrounding environment. Quantitative information represents the measured values including distance, speed and so on, which give us an accurate description about the objects of interest. In this preliminary experiment, we incorporate existing SCV approaches into the proposed framework and facilitate the retrieval of qualitative and quantitative information in both measurement domain and video domain.

To generate the qualitative description, we first use a commonly-used scene recognition network called Places365-CNN [73] to obtain the category of the scene. Places365-CNN is trained on the Places dataset that consists of 10 million images with over 400 unique scene categories, so it is able to distinguish common scenes in our daily life.

Then, a state-of-the-art object tracking network called CenterTrack [74] is employed to extract information about the objects. CenterTrack abandons the widely used tracking-by-detection strategy and performs object detection and tracking simultaneously to achieve real-time inference with a high accuracy. As an online algorithm, CenterTrack takes only the current frame, previous frame, and prior detection result as inputs, and links detection results in these two frames with a simple greedy matching algorithm based on predicted 2D displacement. Apart from 2D tracking, CenterTrack can also be applied to 3D cases by training on 3D object tracking datasets like nuScenes [75], a multi-model dataset containing visual information from 6 cameras, 5 radars and 1 lidar. In this experiment, we use the pretrained model of CenterTrack on nuScenes dataset to track objects. The tracking result contains abundant qualitative and quantitative

information including the category, identity, 3D location, and 3D size of the objects. In this way, we can easily generate quantitative semantic description containing distance and velocity information of each object through simple calculation.

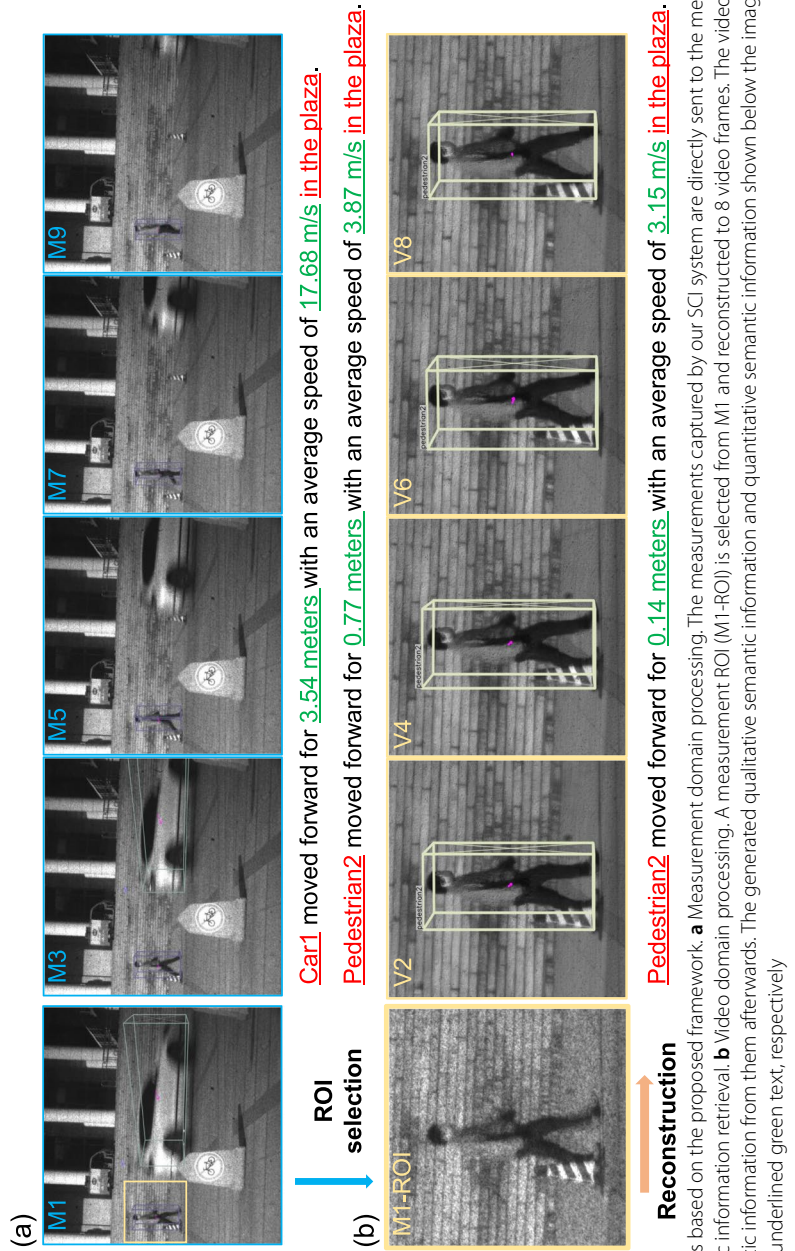
Results and discussion

Equipped with the aforementioned tools, we conducted an outdoor experiment and demonstrate the preliminary results in Fig. 6. To the best of our knowledge, this is the first time that SCI has been used in the outdoor environment for real natural scenario capture. The end-to-end framework combining SCI and SCV for high-efficiency semantic information retrieval is also a novel attempt.

As shown in Fig. 6, sub-figure (a) illustrates processing results in the measurement domain. Benefiting from Place365-CNN and CenterTrack's excellent generalization ability, we can directly use the pretrained model to process coded measurements and obtain coarse semantic information. Note that, due to the large domain gap, existing pretrained models suffer from poor performance in some measurements and may misjudge the scene category or lose the occluded tracking targets (such as the car in measurement M5, M7, and M9). Consequently, the accuracy of distance and speed estimation will also be decreased accordingly. The performance of measurement domain SCV can be improved by incorporating an efficient task-driven feature decoder and optimizing subsequent information retrieval networks, as shown in the proposed framework. But the specific design is beyond the scope of this paper, and we leave it for future work.

As mentioned in the previous section, to obtain a fine result, we can switch to video domain for semantic information retrieval of selected ROIs. Here, we select the region with a person in measurement M1 as the ROI, and show the video domain processing results in sub-figure (b). SCI reconstruction is first conducted to restore the original video frames of the ROI, and video frame V2, V4, V6 and V8 from the reconstructed 8 frames are shown in sub-figure (b). As can be seen from the figure, the details of the person and background stairs have been clearly restored by the physics driven two-stage SCI reconstruction network. Thanks to the higher temporal resolution of the reconstructed video and the advantage of video domain SCV, we can get a fine semantic description result shown in sub-figure (b) with the pretrained models of Place365-CNN and CenterTrack. Note that, in terms of the ROI reconstruction, different SCI reconstruction algorithms can be employed to balance the requirement trade-off among speed, quality, and flexibility.

Benefiting from the proposed joint framework of SCI and SCV, the overall bandwidth of the video acquisition and understanding pipeline has been greatly reduced. Specifically, on the front-end of the pipeline, conventional cameras will generate about 1 GB data per second to capture a high-speed video under the frame rate of 160 FPS. On the contrary, only 1/8 bandwidth is required for our system to achieve the same frame rate by compressively collapsing 8 frames into a coded measurement during acquisition. Besides, this promotion can be further extended by leveraging a higher compressive ratio. On the back-end of the framework, traditional SCV algorithms that perform on video domain requires the reconstruction of the compressive measurements, thus imposing great pressure on the machine memory due to the large data volume of the



reconstructed videos. Instead, the proposed framework can efficiently save the computation bandwidth by compressive domain processing.

During the experiment, we also found that the accuracy of scene classification and the estimation of the objects distance and speed rely heavily on SCI acquisition settings. For instance, using a higher compressive ratio improves the frame rate of the reconstructed video, but also increases the information capacity of the measurement (making it appear blurrier) and introduces more artifacts during reconstruction. For measurement domain processing, measurements multiplexing of massive visual information will increase the difficulty of feature extraction and subsequent semantic information retrieval process. And in terms of video domain processing, the degradation in reconstructed video will result in declined semantic retrieval performance as well. Therefore, a good balance between different modules of the framework is the guarantee of the final high performance.

Outlook and discussion

We have briefed an overview of incubation of SCI and SCV, ready to extend it to more general CI and CV. An end-to-end framework has been demonstrated with a prototype and some preliminary results. This has led to in-depth thinking and analysis of the joint framework from various perspectives, some of which are listed below.

SCI system and reconstruction algorithms

Field of view (FOV)

FOV is a key characteristic of an imaging system. A broad FOV can facilitate scene understanding and other high-level computer vision tasks. For conventional cameras, FOV is largely defined by the lens and sensor. Differently, an SCI system consists of a cascade of optical components, each of which might reduce the FOV to some extent. In other words, the final FOV is determined by the minimum angle of view across the light path. In our preliminary prototype, we employed a pair of doublets as relay optics, which suffers from severe aberrations such as spherical aberration and vignetting effect especially in peripheral areas, thus limiting the FOV. Therefore, we employed a high-quality relay lens to reduce aberrations. Currently, The FOV of our SCI system is about 30 degrees. However, in some extreme cases such as wide-angle lens or even fisheye lens, current designs might be inapplicable to such large FOV. Besides, while in conventional photography it is easy to swap between lenses of different focal lengths, in SCI we need to adjust imaging setup to take full advantage of the FOV of the primary lens. One possible solution is using relay lenses with a sufficiently large FOV, which would result in a bulky design and limit its portability. Designing assembly targeted at different FOVs via integrating the primary lens and other optics for compressive sensing is another option. Overall, an SCI with flexible FOVs is highly desirable but leaves a series of unexplored engineering problems for future research.

Light efficiency

The SCI imaging setup multiplexes a set of sequential frames into a snapshot and theoretically has a higher signal-to-noise ratio. However, in our prototype, this advantage is largely attenuated by implementation details and fabrication errors. Firstly, when LCoS

is adopted as the spatial light modulator, a PBS is required to generate polarized light, which absorbs around half of photons. During acquisition, the 50% duty ratio of random binary encoding masks blocks around half of the remaining incident photons as well. Secondly, the reflectance of the liquid crystal and transmissivity/reflectance of other components (e.g., lenses and PBS) are imperfect, so resulting in loss of photons to some extent. For our prototype, we need to use an additionally strong illumination light source for indoor photography due to the dim lighting condition.

To enhance light efficiency of our system, we envision that a compact and miniaturized optical design with fewer levels of relay lens can be elaborated to decrease photon loss. Besides, we can substitute the LCoS with other modulation devices such as DMD, or directly employ a pixel-wise programmable sensor like Coded-2-Bucket camera [43] for simultaneous modulation and capture.

System calibration

Theoretically, the forward imaging process of SCI can be precisely described by the patterns displayed on the LCoS. However, due to inevitable system aberrations and misalignment, the actual encoding masks on the sensor plane will be slightly different from the displayed ones, and will experience drift over time. Therefore, a calibration step is required to attain the accurate encoding masks before each acquisition. Although the calibration can be performed automatically with a control script using the software development toolkit (SDK) of the LCoS and sensor, this tedious calibration procedure and the requirements for a large Lambertian whiteboard and a relatively even illumination is still troublesome, especially for practical applications in outdoor environment. Moreover, varying calibrated masks will also impose great pressure on subsequent learning-based reconstruction or SCV algorithms, which may require a long time to adapt to new masks.

To pave the way of SCI in practical applications, the current prototype system is expected to be improved in aspects of stability against external disturbance and robustness in different environment conditions such as temperature, humidity, etc. We believe that, with aid of recent advances in optical and industrial engineering, these problems can be properly addressed in commercial production in the future. On the other hand, these problems can also be mitigated from software side. Possible solutions include but are not limited to designing fault-tolerant/self-calibrated algorithms, or introducing feedback control from the task side to optimize the system in real time as illustrated in our proposed framework.

Reconstruction algorithms

Quality, speed, and flexibility are three major aspects in the design of SCI reconstruction algorithms, especially in the joint framework of SCI and SCV. The reconstruction quality of coded measurements has a direct impact on the performance of subsequent SCV tasks, as mentioned in the preliminary experiments. However, it is worth noting that in the proposed framework, the ultimate goal is to retrieve semantic information from coded measurements, rather than restore a high-quality video. Therefore, a task-oriented SCI reconstruction algorithm employing a joint optimization strategy may have a greater potential to boost the overall performance of the framework. Besides,

scene-guided coding pattern design is also a promising direction to improve the encoding efficiency and reconstruction quality.

The speed and flexibility of the reconstruction algorithms play a significant role in practical applications. Generally, real-time inference is an indispensable feature in many practical applications, especially in auto-driving, human-machine interaction, and so on. Compared to measurement domain processing, video domain processing spends extra time on the reconstruction step, making it more difficult to achieve real-time inference. But by using a light-weight network design or some model pruning strategies, there is still plenty of room for improving the reconstruction efficiency.

Moreover, we are supposed to take the flexibility of SCI reconstruction algorithms into consideration in real scenarios as well. Most existing learning-based reconstruction algorithms are designed for a certain set of encoding masks, and time-consuming retraining or fine-tune is needed when the masks change. Recently, some algorithms based on meta-learning or PnP framework have been proposed and mitigate this problem to some extent [47, 54]. In the future, a promising approach is to dig deeper into the analysis of mask uncertainty, and design algorithms that can tolerate calibration error and mask drift in real scenarios.

SCV tasks and approaches

Video retrieval and captioning

The task of content-based retrieval of visual information is to retrieve video clips from video databases according to video contents, based on image and video understanding. At present, research in video retrieval focuses on low-level perceptive representations of raw data (such as color, texture, shape, etc.) and simple motion information [76]. These retrieval techniques cannot accurately and effectively search the videos for sequences related to specified behaviors. Semantic-based video retrieval (SBVR) aims to bridge the gap between low-level features and high-level semantic meanings. Based on automatic interpretation of video content, SBVR may classify and further access the corresponding video clips related to specific behaviors, providing a more advanced, more intuitive, and more humanistic retrieval mode. However, semantic-based video retrieval brings the following challenges: automatic extraction of semantic features, combination of low-level visual features, and behavior description and hierarchical organization of video features. Inspired by the SCI framework with measurement domain and reconstructed video domain jointly understanding, low-level features and high-level semantic meanings will be tightly integrated to improve the accuracy of video retrieval and captioning tasks.

Understanding algorithms

One of the objectives of semantic computer vision is to analyze and interpret individual behaviors and interactions between objects to recognize. For example, whether people are carrying, depositing or exchanging objects, whether a person is getting on or off a vehicle, or whether a vehicle is overtaking another one, etc. Equipped with the ability of behavior understanding, our proposed framework can provide more accurate semantic descriptions for subsequent applications. Recently, related research still focuses on some basic problems like gesture and simple behavior recognition. Some progress has been

made in building statistical models of human behaviors using machine learning algorithms. Behavior recognition is complicated, as the same behavior may have several different meanings, depending on the scene and task context in which it is performed. This ambiguity is exacerbated when there are multiple objects in a scene [77]. The following problems within behavior understanding are challenging: statistical learning for modeling behaviors, context-sensitive learning from example images, real-time performance required by behavior interpretation, classification and labeling of motion trajectories of tracked objects, automatic learning of prior knowledge [78] implied in object behaviors, visually mediated interaction, and attention mechanisms.

Action description and prediction

In order to recognize, describe, and even predict the actions of objects in the video, not only interpretation of individual actions but also interactions between different objects should be analyzed. Then the sequence of actions in the temporal domain is challenging to be considered optimally. For example, sitting and falling down are difficult to distinguish for learning models without motion velocity, joint variables of human beings and the situation. Current tasks focus on some certain action recognition such as gesture recognition, which limits the general use of these methods.

Enhanced by SCI with qualitative and quantitative information merging modules, novel approaches can be designed to deal with these semantic description difficulties. Since SCI is an imaging system with effective acquisition of low-level visual information in both compressive domain and reconstruction domain, it is key to analyze the actions of moving objects with different levels. We can use the correspondence between low-level descriptive features and semantic discriminative features to explore this problem.

Organizing recognized concepts and further predicting the possible object actions has been a hot topic, especially in the visual self-driving program over long periods of time. In addition, the synchronous description, where descriptions are given before an action is complete (while the action is still in progress), is also a challenge [76]. In this case, the combination of SCI and SCV takes advantage of hierarchical features due to the representative features in the measurement domain, which concludes scores of frames in one measurement clip. Thus, it is easier to conduct real-time understanding and prediction algorithms such as graph neural networks by linking corresponding objects in one measurement.

Conclusion

SCI and SCV, two closely related fields related to image acquisition, processing, and understanding, have a great potential to develop together towards a win-win future.

While the previous research of SCI focuses on compressive sampling, we extend SCI to compressive tasking, aiming to bring SCI closer to practical applications. Specifically, to bridge the gap between SCI and SCV, we review the current status of both fields and propose a novel joint framework that implements an end-to-end pipeline of visual information capture and semantic information extraction. The framework takes advantage of SCI in low-bandwidth high-throughput imaging to achieve high-speed video acquisition using a conventional camera, and comes up with an adaptive measurement/video domain information processing strategy to improve the efficiency of

semantic information retrieval. To validate the effectiveness of the framework and discover potential problems in practical applications, we have performed outdoor experiments and presented preliminary results with detailed analysis. Finally, we give a prospective outlook and provide possible directions for the future development of the framework. We believe that, with the rapid booming of SCI and SCV, the proposed framework will soon become available in practical applications and blaze a new trail for massive visual information acquisition and corresponding intelligent analysis.

Abbreviations

ADMM	Alternating direction method of multiplier
CCD	Charge-coupled device
CI	Computational imaging
CMOS	Complementary-metal-oxide-semiconductor
CNN	Convolutional neural network
CS	Compressive sensing
CV	Computer vision
DMD	Digital mirror device
FOV	Field of view
FPS	Frames per second
GAP	Generalized alternating projection
GMM	Gaussian mixture model
LCoS	Liquid crystal on silicon
LWIR	Long-wave infrared
MEMS	Micro-electromechanical system
MWIR	Mid-wave infrared
PBS	Polarizing beamsplitter
PnP	Plug-and-Play
RNN	Recurrent neural network
ROI	Region of interest
SBVR	Semantic-based video retrieval
SCI	Snapshot compressive imaging
SCV	Semantic computer vision
SDK	Software development toolkit
SNR	Signal to noise ratio
SWIR	Short-wave infrared
TV	Total variation

Acknowledgements

X. Yuan would like to thank the Research Center for Industries of the Future (RCIF) at Westlake University for supporting this work and the funding from Lochn Optics.

Authors' contributions

XY and JS proposed the idea; BZ and ZZ built the prototype system and conducted the outdoor experiments; ZZ, SZ, and XS processed the data and analyzed the results. ZZ, XY, XS, BZ, and SZ prepared the manuscript; JS, DJB, and QD reviewed and polished the manuscript; JS and QD supervised the project. All authors read and approved the final manuscript.

Funding

This work was supported by the Ministry of Science and Technology of the People's Republic of China [grant number 2020AAA0108202] and the National Natural Science Foundation of China [grant numbers 61931012, 62088102].

Availability of data and materials

The datasets generated and analysed during the current study are available from figshare under link <https://doi.org/10.6084/m9.figshare.20431911>.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 13 June 2022 Accepted: 20 August 2022

Published online: 06 September 2022

References

1. Boyle WS, Smith GE. Charge coupled semiconductor devices. *Bell Syst Tech J.* 1970;49(4):587–93.
2. Altmann Y, McLaughlin S, Padgett MJ, Goyal VK, Hero AO, Faccio D. Quantum-inspired computational imaging. *Science.* 2018;361(6403):eaat2298. <https://doi.org/10.1126/science.aat2298>.
3. Mait JN, Euliss GW, Athale RA. Computational imaging. *Adv Opt Photonics.* 2018;10(2):409–83.
4. Yuan X, Brady DJ, Katsaggelos AK. Snapshot compressive imaging: theory, algorithms, and applications. *IEEE Signal Process Mag.* 2021;38(2):65–88.
5. Gao L, Liang J, Li C, Wang LV. Single-shot compressed ultrafast photography at one hundred billion frames per second. *Nature.* 2014;516(7529):74–7.
6. Raskar R, Agrawal A, Tumblin J. Coded exposure photography: motion deblurring using fluttered shutter. *ACM Trans Graphics.* 2006;25(3):795–804.
7. Sitzmann V, Diamond S, Peng Y, Dun X, Boyd S, Heidrich W, et al. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Trans Graphics.* 2018;37(4):1–13.
8. Sun Q, Zhang J, Dun X, Ghanem B, Peng Y, Heidrich W. End-to-end learned, optically coded super-resolution SPAD camera. *ACM Trans Graph.* 2020;39(2):1–14.
9. Antipa N, Oare P, Bostan E, Ng R, Waller L. Video from stills: lensless imaging with rolling shutter. In: 2019 IEEE International Conference on Computational Photography (ICCP). IEEE; 2019. p. 1–8.
10. Asif MS, Ayremlou A, Sankaranarayanan A, Veeraraghavan A, Baraniuk RG. FlatCam: thin, lensless cameras using coded aperture and computation. *IEEE Trans Comput Imaging.* 2017;3(3):384–97.
11. Cai Z, Chen J, Pedrini G, Osten W, Liu X, Peng X. Lensless light-field imaging through diffuser encoding. *Light Sci Appl.* 2020;9(1):143.
12. Hu C, Huang H, Chen M, Yang S, Chen H. FourierCam: a camera for video spectrum acquisition in a single shot. *Photon Res.* 2021;9(5):701.
13. Liang CK, Lin TH, Wong BY, Liu C, Chen HH. Programmable aperture photography: multiplexed light field acquisition. *ACM Trans Graph.* 2008;27(3):391–400.
14. Lv X, Li Y, Zhu S, Guo X, Zhang J, Lin J, et al. Snapshot spectral polarimetric light field imaging using a single detector. *Opt Lett.* 2020;45(23):6522.
15. Hu C, Huang H, Chen M, Yang S, Chen H. Video object detection from one single image through opto-electronic neural network. *APL Photon.* 2021;6(4):046104.
16. Okawara T, Yoshida M, Nagahara H, Yagi Y. Action recognition from a single coded image. In: 2020 IEEE International Conference on Computational Photography (ICCP). IEEE; 2020. p. 1–11.
17. Wu Y, Boominathan V, Chen H, Sankaranarayanan A, Veeraraghavan A. PhaseCam3D — learning phase masks for passive single view depth estimation. In: 2019 IEEE International Conference on Computational Photography (ICCP). IEEE; 2019. p. 1–12.
18. Audebert N, Le Saux B, Lefevre S. Deep learning for classification of hyperspectral data: a comparative review. *IEEE Geosci Remote Sens Mag.* 2019;7(2):159–73.
19. Rawat W, Wang Z. Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput.* 2017;29(9):2352–449.
20. Asgari Taghanaki S, Abhishek K, Cohen JP, Cohen-Adad J, Hamarneh G. Deep semantic segmentation of natural and medical images: a review. *Artif Intell Rev.* 2021;54(1):137–78.
21. Yu H, Yang Z, Tan L, Wang Y, Sun W, Sun M, et al. Methods and datasets on semantic segmentation: a review. *Neurocomputing.* 2018;304:82–103.
22. Jiao L, Wang D, Bai Y, Chen P, Liu F. Deep learning in visual tracking: a review. *IEEE Trans Neural Netw Learn Syst.* 2021;1(1):1–20.
23. Pal SK, Pramanik A, Maiti J, Mitra P. Deep learning in multi-object detection and tracking: state of the art. *Appl Intell.* 2021;51(9):6400–29.
24. Zhu H, Wei H, Li B, Yuan X, Kehtarnavaz N. A review of video object detection: datasets, metrics and methods. *Appl Sci.* 2020;10(21):7834.
25. Aafaq N, Mian A, Liu W, Gilani SZ, Shah M. Video description: a survey of methods, datasets, and evaluation metrics. *ACM Comput Surv.* 2020;52(6):1–37.
26. Hossain MZ, Sohel F, Shiratuddin MF, Laga H. A comprehensive survey of deep learning for image captioning. *ACM Comput Surv.* 2019;51(6):1–36.
27. Guo Y, Liu Y, Georgiou T, Lew MS. A review of semantic segmentation using deep neural networks. *Int J Multimed Info Retr.* 2018;7(2):87–93.
28. Herath S, Harandi M, Porikli F. Going deeper into action recognition: a survey. *Image Vision Comput.* 2017;60:4–21.
29. Li S, Deng W. Deep facial expression recognition: a survey. *IEEE Trans Affect Comput.* 2020;1(1):1–10.
30. Pawar PG, Devendran V. Scene understanding: a survey to see the world at a single glance. In: 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT). IEEE; 2019. p. 182–6.
31. Chen S, Yao T, Jiang YG. Deep learning for video captioning: a review. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI). International Joint Conferences on Artificial Intelligence Organization; 2019. p. 6283–90.
32. Deng C, Zhang Y, Mao Y, Fan J, Suo J, Zhang Z, et al. Sinusoidal sampling enhanced compressive camera for high speed imaging. *IEEE Trans Pattern Anal Mach Intell.* 2021;43(4):1380–93.
33. Hitomi Y, Gu J, Gupta M, Mitsunaga T, Nayar SK. Video from a single coded exposure photograph using a learned over-complete dictionary. In: 2011 International Conference on Computer Vision (ICCV). IEEE; 2011. p. 287–94.
34. Lull P, Liao X, Yuan X, Yang J, Kittle D, Carin L, et al. Coded aperture compressive temporal imaging. *Opt Express.* 2013;21(9):10526.
35. Lu R, Chen B, Liu G, Cheng Z, Qiao M, Yuan X. Dual-view snapshot compressive imaging via optical flow aided recurrent neural network. *Int J Comput Vision.* 2021;129(12):3279–98.
36. Qiao M, Liu X, Yuan X. Snapshot spatial-temporal compressive imaging. *Opt Lett.* 2020;45(7):1659–62.
37. Qiao M, Meng Z, Ma J, Yuan X. Deep learning for video compressive sensing. *APL Photonics.* 2020;5(3):030801.

38. Reddy D, Veeraraghavan A, Chellappa R. P2C2: programmable pixel compressive camera for high speed imaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2011. p. 329–36.
39. Shedligeri P, S A, Mitra K. A unified framework for compressive video recovery from coded exposure techniques. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE; 2021. p. 1600–9.
40. Yoshida M, Sonoda T, Nagahara H, Endo K, Sugiyama Y, Taniguchi R. High-speed imaging using CMOS image sensor with quasi pixel-wise exposure. *IEEE Trans Comput Imaging*. 2020;6:463–76.
41. Yuan X, Llull P, Liao X, Yang J, Brady DJ, Sapiro G, et al. Low-cost compressive sensing for color video and depth. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2014. p. 3318–25.
42. Zhang Z, Deng C, Liu Y, Yuan X, Suo J, Dai Q. Ten-mega-pixel snapshot compressive imaging with a hybrid coded aperture. *Photonics Res*. 2021;9(11):2277.
43. Wei M, Sarhangnejad N, Xia Z, Gusev N, Katic N, Genov R, et al. Coded two-bucket cameras for computer vision. In: European Conference on Computer Vision (ECCV). Springer; 2018. p. 54–71.
44. Wang P, Liang J, Wang LV. Single-shot ultrafast imaging attaining 70 trillion frames per second. *Nat Commun*. 2020;11(1):2091.
45. Liu Y, Yuan X, Suo J, Brady DJ, Dai Q. Rank minimization for snapshot compressive imaging. *IEEE Trans Pattern Anal Mach Intell*. 2019;41(12):2990–3006.
46. Yuan X. Generalized alternating projection based total variation minimization for compressive sensing. In: 2016 IEEE International Conference on Image Processing (ICIP). IEEE; 2016. p. 2539–43.
47. Yuan X, Liu Y, Suo J, Dai Q. Plug-and-play algorithms for large-scale snapshot compressive imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2020. p. 1444–54.
48. Jalali S, Yuan X. Snapshot compressed sensing: performance bounds and algorithms. *IEEE Trans Inf Theory*. 2019;65(12):8005–24.
49. Jalali S, Yuan X. Compressive imaging via one-shot measurements. In: 2018 IEEE International Symposium on Information Theory (ISIT). IEEE; 2018. p. 416–20.
50. Bioucas-Dias JM, Figueiredo MAT. A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Trans Image Process*. 2007;16(12):2992–3004.
51. Cheng Z, Lu R, Wang Z, Zhang H, Chen B, Meng Z, et al. BIRNAT: bidirectional recurrent neural networks with adversarial training for video snapshot compressive imaging. In: European Conference on Computer Vision (ECCV). Springer; 2020. p. 258–75.
52. Iliadis M, Spinoulas L, Katsaggelos AK. Deep fully-connected networks for video compressive sensing. *Digit Signal Process*. 2018;72:9–18.
53. Ma J, Liu XY, Shou Z, Yuan X. Deep tensor ADMM-Net for snapshot compressive imaging. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). IEEE; 2019. p. 10222–31.
54. Wang Z, Zhang H, Cheng Z, Chen B, Yuan X. MetaSCI: scalable and adaptive reconstruction for video compressive sensing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2021. p. 2083–92.
55. Wu Z, Zhang J, Mou C. Dense deep unfolding network with 3D-CNN prior for snapshot compressive imaging. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). IEEE; 2021. p. 4892–901.
56. Yang J, Liao X, Yuan X, Llull P, Brady DJ, Sapiro G, et al. Compressive sensing by learning a Gaussian mixture model from measurements. *IEEE Trans Image Process*. 2015;24(1):106–19.
57. Yang J, Yuan X, Liao X, Llull P, Brady DJ, Sapiro G, et al. Video compressive sensing using Gaussian mixture models. *IEEE Trans Image Process*. 2014;23(11):4863–78.
58. Cheng Z, Chen B, Liu G, Zhang H, Lu R, Wang Z, et al. Memory-efficient network for large-scale video compressive sensing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2021. p. 16246–55.
59. Yuan X, Liu Y, Suo J, Durand F, Dai Q. Plug-and-play algorithms for video snapshot compressive imaging. *IEEE Trans Pattern Anal Mach Intell*. 2021;1(1):1–18.
60. Liao X, Li H, Carin L. Generalized alternating projection for weighted- $\ell_{2,1}$ minimization with applications to model-based compressive sensing. *SIAM J Imaging Sci*. 2014;7(2):797–823.
61. Boyd S. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found Trends[®] Mach Learn*. 2010;3(1):1–122.
62. Bethi YRT, Narayanan S, Rangan V, Chakraborty A, Thakur CS. Real-time object detection and localization in compressive sensed video. In: 2021 IEEE International Conference on Image Processing (ICIP). IEEE; 2021. p. 1489–93.
63. Kwan C, Chou B, Yang J, Rangamani A, Tran T, Zhang J, et al. Target tracking and classification using compressive measurements of MWIR and LWIR coded aperture cameras. *J Signal Inf Process*. 2019;10(03):73–95.
64. Lu S, Yuan X, Shi W. Edge compression: an integrated framework for compressive imaging processing on CAVs. In: 2020 IEEE/ACM Symposium on Edge Computing (SEC). IEEE; 2020. p. 125–38.
65. Kwan C, Chou B, Yang J, Rangamani A, Tran T, Zhang J, et al. Deep learning-based target tracking and classification for low quality videos using coded aperture cameras. *Ah S Sens*. 2019;19(17):3702.
66. Kwan C, Chou B, Yang J, Rangamani A, Tran T, Zhang J, et al. Target tracking and classification using compressive sensing camera for SWIR videos. *Signal Image Video Process*. 2019;13(8):1629–37.
67. Rezaei M, Terauchi M, Klette R. Robust vehicle detection and distance estimation under challenging lighting conditions. *IEEE Trans Intell Transp Syst*. 2015;16(5):2723–43.
68. Zhe T, Huang L, Wu Q, Zhang J, Pei C, Li L. Inter-vehicle distance estimation method based on monocular vision using 3D detection. *IEEE Trans Veh Technol*. 2020;69(5):4907–19.
69. Yuan X, Yang J, Llull P, Liao X, Sapiro G, Brady DJ, et al. Adaptive temporal compressive sensing for video. In: 2013 IEEE International Conference on Image Processing (ICIP). IEEE; 2013. p. 14–8.
70. Zheng S, Wang C, Yuan X, Xin HL. Super-compression of large electron microscopy time series by deep compressive sensing learning. *Patterns*. 2021;2(7):100292.
71. Zheng S, Yang X, Yuan X. Two-stage is enough: a concise deep unfolding reconstruction network for flexible video compressive sensing. *arXiv preprint arXiv:2201.05810*. 2022;1(1):1–10.

72. Gomez AN, Ren M, Urtasun R, Grosse RB. The reversible residual network: backpropagation without storing activations. In: Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS). vol. 30. Curran Associates, Inc.; 2017. p. 1-10.
73. Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: a 10 million image database for scene recognition. *IEEE Trans Pattern Anal Mach Intell.* 2018;40(6):1452–64.
74. Zhou X, Koltun V, Krähenbühl P. Tracking objects as points. In: European Conference on Computer Vision (ECCV). Springer; 2020. p. 474-90.
75. Caesar H, Bankiti V, Lang AH, Vora S, Liong VE, Xu Q, et al. nuScenes: a multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2020. p. 11618-28.
76. Hu W, Tan T, Wang L, Maybank S. A survey on visual surveillance of object motion and behaviors. *IEEE Syst Man Cybern Mag.* 2004;34(3):334–52.
77. Zhao ZQ, Zheng P, Xu St, Wu X. Object detection with deep learning: a review. *IEEE Trans Neural Netw Learn Syst.* 2019;30(11):3212-32.
78. Feng D, Haase-Schütz C, Rosenbaum L, Hertlein H, Glaeser C, Timm F, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges. *IEEE Trans Intell Transp Syst.* 2020;22(3):1341–60.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)
