Deep Optics for Video Snapshot Compressive Imaging

Ping Wang[1,2] Lishun Wang[2] Xin Yuan[2,*]
[1]Zhejiang University [2]School of Engineering, Westlake University
{wangping,wanglishun,xyuan}@westlake.edu.cn

Abstract

*Video snapshot compressive imaging (SCI) aims to capture a sequence of video frames with only a single shot of a 2D detector, whose backbones rest in optical modulation patterns (also known as masks) and a computational reconstruction algorithm. Advanced deep learning algorithms and mature hardware are putting video SCI into practical applications. Yet, there are two clouds in the sunshine of SCI: i) low dynamic range as a victim of high temporal multiplexing, and ii) existing deep learning algorithms' degradation on real system. To address these challenges, this paper presents a deep optics framework to jointly optimize masks and a reconstruction network. Specifically, we first propose a new type of **structural mask** to realize **motion-aware** and **full-dynamic-range** measurement. Considering the motion awareness property in measurement domain, we develop an **efficient** network for video SCI reconstruction using Transformer to capture **long-term temporal dependencies**, dubbed **Res2former**. Moreover, **sensor response** is introduced into the forward model of video SCI to guarantee end-to-end model training close to real system. Finally, we implement the learned structural masks on a digital micro-mirror device. Experimental results on synthetic and real data validate the effectiveness of the proposed framework. We believe this is a milestone for real-world video SCI. The source code and data are available at https://github.com/pwangcs/DeepOpticsSCI.*

1. Introduction

Capturing high-dynamic-range (HDR) and high-frame-rate (HFR) video is a long-term challenge in the field of computational photography. As an elegant solution of HFR, video snapshot compressive imaging (SCI) optically multiplexes a sequence of video frames, each of which is coded with a distinct modulation pattern (hereafter called mask), into a snapshot measurement of a two-dimensional (2D) detector, and computationally reconstructs a decent estimate of the original video from the measurement using an advanced algorithm. In a nutshell, video SCI is a hardware-encoder-plus-software-decoder system and its performance mainly depends on mask and reconstruction algorithm.

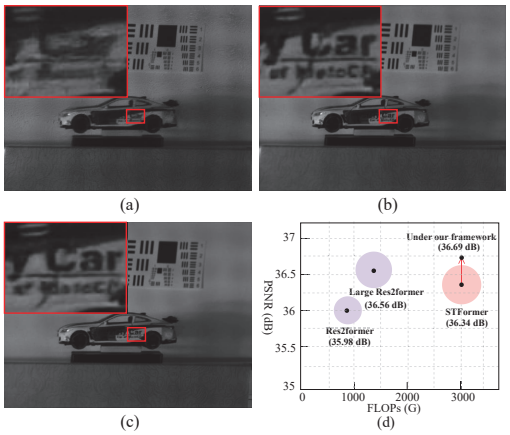


Figure 1. The proposed deep optics framework brings a significant improvement for real-world video SCI as demonstrated in real results (a), (b), and (c), got by previous SOTA STFormer [34], current STFormer (under our framework), and our Res2former, respectively. (d) summarizes the comparison between Res2former and STFormer in terms of PSNR (vertical axis), FLOPs (horizontal axis), and Parameters (circle radius). The proposed Res2former achieves competitive performance (35.98 dB) with only 28.15% FLOPs and 56.57% parameters of STFormer (36.34 dB). By increasing parameters to STFormer's level, large Res2former can lead to a better performance (36.56 dB). By the way, STFormer under our framework can increase by 0.35 dB.

For the hardware encoder, random binary mask has been widely used in both simulation and real video SCI systems, often implemented in a digital micro-mirror device (DMD) [28, 27] or liquid crystal on silicon (LCOS) [29, 9, 16]. Recently, learned binary mask was also implemented in programmable pixel sensors [21]. For the software decoder, it is an ill-posed inverse problem to retrieve high-fidelity video from the captured single measurement and various reconstruction methods [41, 17, 18, 28, 5, 35, 4, 37, 34, 23] have been developed to solve it in recent years. Conventional optimization algorithms adopt hand-crafted priors, *e.g.*, total variation [41] and non-local self-

*Corresponding author.

similarity [17], to confine the solution to the desired signal space. But optimization-based methods commonly require a long running time to get usable results. With the powerful generalization ability of deep neural networks (DNNs), deep learning methods have been increasingly developed and achieved excellent results in a little inference time, usually designed as an end-to-end (E2E) network, *e.g.*, E2E-CNN [28], BIRNAT [5], MetaSCI [35], RevSCI [4], STFormer [34], or a deep unfolding network, *e.g.*, GAP-net [23], ADMM-Net [18], SCI3D [37], ELP-Unfolding [39]. Despite these remarkable advances, particularly in deep learning reconstruction methods, there are still some practical challenges in putting video SCI into our daily life.

Due to the limited bit depth of image sensors, the higher temporal multiplexing, the lower dynamic range. For an video SCI camera using random binary masks, the measurable brightness values of video frames is approximately equal to $2^{\kappa+1}/B$, far less than the available brightness values of image sensor 2^{κ}, where B (usually $8 \leq B \leq 50$) and κ denote compressed frames and sensor bit depth, respectively. We take 8-frame video SCI camera equipped with a typical 8-bit-depth image sensor as an example, namely, 8 video frames are compressed into a single image with 256 available brightness values during measurement. If using random binary masks that take values of '1' or '0' with equal probability, at each spatial position, half of 8 frames are integrated into one pixel along temporal dimension with high probability. In this case, each frame can only be represented by 64 brightness values, which is calculated by $256/4 = 64$. Obviously, there is a significant gap between the wide range of brightness variations in natural scenes and the very limited dynamic range in previous video SCI. Such a practical problem is also widely rooted in other compressive imaging systems, *e.g.*, spectral SCI [8], compressive light field imaging [22], and single-pixel imaging [6].

Without considering sensor response, existing deep reconstruction networks have a great performance degradation when used in real system. As is well known, the performance of DNNs is closely related to the used training dataset. Without available specialized datasets, the forward model of video SCI usually need to be mathematically formulated to synthesize the training dataset from a public HFR video dataset. Accordingly, deep reconstruction networks have a high dependence on the forward model. Unfortunately, previous forward model only considers *optical transmission and modulation* but overlooks *sensor response* characterizing the used image sensor, meaning that there is a gap between previous forward model and real system. As a result, previous deep reconstruction networks show excellent performance on synthetic data but degraded performance on real data.

To address the above challenges, a deep optics framework is proposed to improve the performance of real-world video SCI. The contributions of this work are summarized as follows.

- Unlike widely-used random binary mask, a new type of *structural mask* is presented to realize *motion-aware* and *full-dynamic-range* (FDR) measurement. Motion-aware measurement contributes to video SCI reconstruction. To our best knowledge, we are the first to enable FDR video SCI.
- Considering the motion-aware property in the encoder, we tailor an *efficient* reconstruction network, dubbed Res2former, as the video SCI decoder by using Transformer to capture *long-term temporal dependencies*. Compared with the state-of-the-art (SOTA) network STFormer [34], Res2former is highly lightweight but provides competitive performance.
- We propose a deep optics framework to jointly optimize the proposed structural mask and reconstruction network, in which *sensor response* is introduced to guarantee end-to-end (E2E) training close to real system. Under this framework, Res2former and previous reconstruction networks achieve significant improvement on synthetic data and real data.

2. Related Work

Deep optics. Deep optics takes the idea of jointly optimizing optics and algorithm to improve various computational imaging systems, *e.g.* microscopy [25], HDR imaging [21, 32, 24], depth imaging [40, 3, 36], single-pixel imaging [10], light field imaging [13], and compressing imaging [12, 40, 21, 14, 44, 33]. Mask optimization for video SCI has been increasingly studied under hardware constraints [40, 21, 14]. Based on an emerging programmable sensor SCAMP-5, a hand-held video SCI camera [21] has recently developed but its spatial-temporal resolution is very limited. These works attached great importance to the implementation of binary mask by using some heuristic sensors. This paper aims to the performance of real-world video SCI. To our best knowledge, we are the first to optimize more challenging structural mask and remove the incompatibility between temporal multiplexing and dynamic range.

Video SCI reconstruction. Video SCI reconstruction algorithms can be classified into regularization-based methods and learning-based methods. Regularization-based methods combine the idea of iterative optimization, *e.g.*, generalized alternating projection (GAP) [15] or alternating direction method of multipliers (ADMM) [2], with various prior knowledge, *e.g.*, total variation (TV) [41] and non-local low rank [17]. They provide usable results in an unsupervised manner but cannot balance fidelity and speed. In recent years, kinds of learning-based methods have been

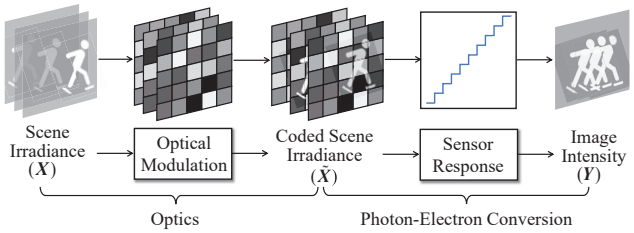


Figure 2. Illustration of video SCI encoder. High-speed scene is first optically modulated with temporally-varying masks and then integrated into a single digital image (*i.e.*, snapshot measurement) through an off-the-shelf image sensor. In the process, optical modulation and sensor response are two key ingredients.

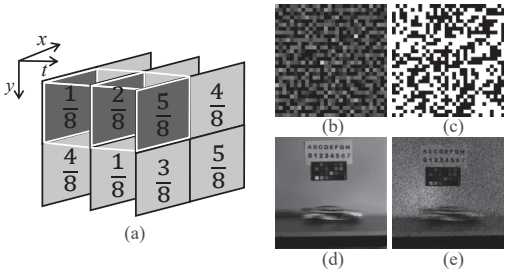


Figure 3. Proposed structural mask (b) *vs.* widely-used random binary mask (c). As demonstrated in (a), structural mask values represent the transmittance of incident light and the sum of values across temporal dimension is 1. It lead to the motion-aware measurement (d), havinging more visual information than the measurement (e) captured by random binary mask.

developed for high fidelity and low inference time. Recently, an E2E network STFormer [34] has achieved the state-of-the-art (SOTA) results using temporal and spatial Transformer, but at the cost of high parameters and complexity. In addition to E2E networks, *e.g.*, Unet [28], BIRNAT [5], MetaSCI [35], RevSCI [4], deep unfolding networks, *e.g.*, GAP-net [23], ADMM-Net [18], SCI3D [37], ELP-Unfolding [39], and plug-and-play (PnP) algorithms, *e.g.*, PnP-FFDNet [42] and PnP-FastDVDnet [43], have been developed by combining an iterative optimization framework with convolutional neural networks or a deep image denoiser. Both regularization-based methods and learning-based methods aim to solve the ill-posed inverse problem of video SCI forward model, thus their performance is susceptible to this model. Previous forward model only considers optical transmission and modulation but overlooks sensor response in practice. As a result, existing reconstruction networks lead to excellent results in synthetic data rather than real data.

3. Video SCI: from Theory to Practice

Aiming to move one step further towards real-world video SCI, we hereby make a wide appeal for modeling video SCI under hardware constraints and employing structural mask instead of random binary mask.

3.1. Mathematical Model of Practical Video SCI

As shown Fig. 2, video SCI encoder is mainly composed of optical modulation and sensor response. In the video SCI decoder, a reconstruction algorithm is employed.

Optical modulation. By implementing temporally-varying masks $\boldsymbol{M}(u,v,t)$ on B discrete time slots ($1 \le t \le B$), a dynamic scene irradiance $\boldsymbol{X}(u,v,t)$ is modulated into the coded spatial-temporal irradiance $\tilde{\boldsymbol{X}}(u,v,t)$ by

$$\tilde{\boldsymbol{X}}(u,v,t) = \boldsymbol{M}(u,v,t) \odot \boldsymbol{X}(u,v,t), \tag{1}$$

where (u,v,t) denotes the spatial-temporal coordinate and $\odot$ denotes the Hadamard (element-wise) product.

Sensor response. Given an image sensor, $\tilde{\boldsymbol{X}}(u,v,t)$ is integrated as a single digital image $\boldsymbol{Y}(u,v)$ (*i.e.*, snapshot measurement) by

$$\boldsymbol{Y}(u,v) = \mathcal{R}\left[\sum_{t=1}^{B} \tilde{\boldsymbol{X}}(u,v,t)\right] + \boldsymbol{Z}(u,v), \tag{2}$$

where $\mathcal{R}$ represents the mapping function of from scene irradiance to image pixels and $\boldsymbol{Z}(u,v)$ denotes the noise originated from measurement, read-out, *etc*. Defining the vectorization operation on a matrix as $\texttt{vec}(\cdot)$, we can reformulate Eq. (2) as the following vectorized form:

$$\boldsymbol{y} = \mathcal{R} \circ \mathcal{H}(\boldsymbol{x}), \tag{3a}$$

$$s.t. \quad \mathcal{H}(\boldsymbol{x}) = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{z}, \tag{3b}$$

where $\boldsymbol{x} = \texttt{vec}(\boldsymbol{X})$, $\boldsymbol{y} = \texttt{vec}(\boldsymbol{Y})$, $\boldsymbol{z} = \texttt{vec}(\boldsymbol{Z})$, and $\boldsymbol{\Phi} = [\texttt{diag}\left(\texttt{vec}(\boldsymbol{M}(:,:,1))\right), \cdots, \texttt{diag}\left(\texttt{vec}(\boldsymbol{M}(:,:,B))\right)]$.

Computational reconstruction. Provided with the used $\boldsymbol{\Phi}$, a regularization-based or learning-based reconstruction algorithm $\mathcal{D}$ is employed to retrieve a decent estimate $\hat{\boldsymbol{x}}$ of $\boldsymbol{x}$ from $\boldsymbol{y}$ by

$$\hat{\boldsymbol{x}} = \mathcal{D}(\boldsymbol{y}) = \mathcal{D} \circ \mathcal{R} \circ \mathcal{H}(\boldsymbol{x}). \tag{4}$$

In general, $\mathcal{R}$ can be modeled as a combination of nonlinear response function f, out-of-range clipping function g, and quantization function h, *i.e.*, $\mathcal{R} = h \circ g \circ f$, leading to non-linearity, saturation error, and quantization error, respectively. These functions are generally inevitable to transform real-valued scene irradiance into digital image brightness. In most industrial cameras, the non-linearity function f can be corrected to be linear, thus $\mathcal{R} = h \circ g \circ f$ is simplified as $\mathcal{R} = h \circ g$.

Previous works [41, 17, 18, 28, 5, 35, 4, 37, 34, 23] view Eq. (3b), only considering optical modulation and sensor integration, as the forward model of video SCI. By introducing the complete sensor response, we present the forward model in Eq. (3) closer to real system.

3.2. Proposed Structural Mask

As mentioned previously, there is an incompatibility between temporal multiplexing and dynamic range in existing works due to the use of random binary mask. Here, we pro-

pose a new type of structural mask to realize full-dynamic-range (FDR) and motion-aware measurement for video SCI.

Structural mask is mathematically defined as

$$\boldsymbol{M}_\lambda(u,v,t)\in\begin{cases}\{0,1/2^\lambda,...,1-1/2^\lambda\} & \lambda\geq 2\\ \{0,1\} & \lambda=1\end{cases} \tag{5a}$$

$$s.t.\quad \sum_{t=1}^{B}\boldsymbol{M}_\lambda(:,:,t)=\mathbf{1}, \tag{5b}$$

where λ denotes the bit depth of mask. Unlike widely-used binary mask, the proposed mask has two attributes: *discretization* and *structuralization*. Discretization indicates that the mask can only take binary ($\lambda=1$) or grayscale ($\lambda\geq 2$) values. Structuralization indicates that, at all spatial coordinates, the sum across temporal dimension is fixed to 1, as demonstrated in Fig. 3 (a). Due to the undesirable performance of $\lambda=1$ (see Tab. 5), we mainly focus on the setting of $\lambda\geq 2$ in this paper. Structural mask can also be easily implemented in an off-the-shelf spatial light modulator (*e.g.*, DMD) at the cost of decreasing the pattern refresh rate. Fortunately, current DMDs' pattern refresh rate is high enough for video SCI. Taking DLP7000 DMD [1] as an example, the maximal pattern refresh rate is 32552 or 4069 for 1-bit (*i.e.*, binary) mask or 8-bit mask, respectively.

Full dynamic range (FDR). The proposed structural mask is capable of removing the incompatibility between temporal multiplexing and dynamic range, rooted in previous video SCI using random binary mask. Taking an 8-bit image sensor as an example, it can record scene irradiance under brightness range $[0,1,\ldots,255]$. For 8-frame video SCI (*i.e.*, $B=8$), the sum of random binary mask across temporal dimension is approximate to 4, equivalent to that almost 4 video frames are integrated into a single image with brightness range $[0,1,\ldots,255]$. Accordingly, the brightness range of each video frame is limited in $[0,1,\ldots,63]$, leading to a low dynamic range. Clearly, it cannot meet the wide range of brightness variations in natural scenes and worsen along with larger B. Using the proposed structural mask, each pixel of captured measurement is the weighted sum of a sequence of video pixels across temporal dimension and the total weight is 1. It means that the brightness range of measurement is equal to that of each of video frames regardless of B. Therefore, the proposed structural mask keep the dynamic range of video SCI in line with that of the used sensors, *i.e.*, full dynamic range (FDR).

Motion-aware measurement. As shown in Fig. 3 (d) using structural mask, the motionless objects, background, and motion trajectory could be greatly recorded in the captured measurement. We refer to it as motion-aware measurement. Such measurement can be viewed as a coarse estimate of original video frames. Generating the network input from a coarse estimate is essential in nearly all impressive video SCI reconstruction works [5, 35, 4, 37, 39, 34]. Unlike our direct acquisition by optics, previous works get the coarse estimate by idealizing video SCI forward model as Eq. (3b) and then computing $\sum_{t=1}^{B}\tilde{\boldsymbol{X}}(t)\Big/\sum_{t=1}^{B}\boldsymbol{M}(t)$. But their estimate becomes $\boldsymbol{Y}\Big/\sum_{t=1}^{B}\boldsymbol{M}(t)$ in practice. The gap in input initialization also makes for previous network's performance degradation in real system.

[1] https://ti.com/product/DLP7000

4. Deep Optics Framework for Video SCI

Previous video SCI reconstruction networks [28, 5, 35, 4, 34, 23, 18, 37, 39] were trained on the impractical forward model in Eq. (3b) and thus achieved impressive performance in simulation rather than real system. To bridge the performance gap, we propose an E2E deep optics framework to jointly optimize structural mask and a reconstruction network under hardware constraints.

Algorithm 1: Structural Mask Training

Input: A learnable mask $\boldsymbol{M}'\in[0,1]$ with a size of $B\times H\times W$ and the desired bit depth $\lambda\geq 2$.

Output: A λ-bit structural mask $\boldsymbol{M}_\lambda$.

Forward $\mathcal{F}(\boldsymbol{M}')$**:**

  /* Discretization */

  $L\longleftarrow 2^\lambda$

  $\boldsymbol{M}\longleftarrow\lfloor\boldsymbol{M}'\cdot L+\mathbf{0.5}\rfloor/L$

  $\boldsymbol{M}[\boldsymbol{M}==1]\longleftarrow 1-1/L$

  $\boldsymbol{M}[:,\mathtt{sum}(\boldsymbol{M},0)==0]\longleftarrow 1/L$

  /* Structuralization */

  $\boldsymbol{\Omega}\longleftarrow\mathtt{sum}(\boldsymbol{M},0)$

  $\boldsymbol{\Sigma}\longleftarrow L\cdot(\boldsymbol{\Omega}-\mathbf{1})$

  for $0\leq k<B$ **do**

    $\boldsymbol{W}\longleftarrow\boldsymbol{M}[k]\oslash\boldsymbol{\Omega}$

    $\boldsymbol{\Delta}\longleftarrow\lfloor\boldsymbol{\Sigma}\odot\boldsymbol{W}+\mathbf{0.5}\rfloor/L$

    $\boldsymbol{M}_\lambda[k]\longleftarrow\boldsymbol{M}[k]-\boldsymbol{\Delta}$

    $\boldsymbol{\Omega}\longleftarrow\boldsymbol{\Omega}-\boldsymbol{M}[k]$

    $\boldsymbol{\Sigma}\longleftarrow\boldsymbol{\Sigma}-L\times\boldsymbol{\Delta}$

  return $\boldsymbol{M}_\lambda$

Backward $\mathcal{G}(x)$**:**

  $y\longleftarrow x$

  return y

4.1. Overall Architecture

As shown in Fig. 4 (a), a real-world video SCI system is composed of a compressive camera in the physical layer and a reconstruction algorithm in the digital layer. Due to the lack of specialized video SCI dataset, the compressive camera commonly needs to be formulated as the forward model to synthesize a amount of measurement-video ($\boldsymbol{y}$, $\boldsymbol{x}$)

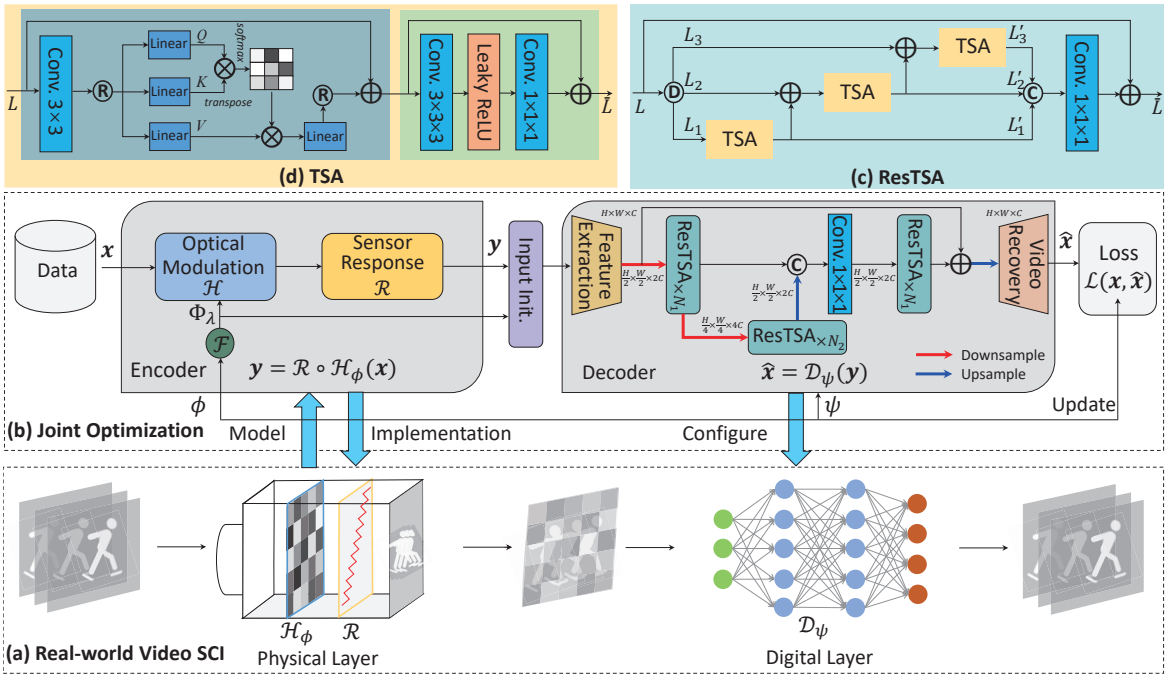


Figure 4. Deep optics framework for the joint optimization of structural mask and a deep reconstruction network. $\oplus$, ⓒ/ⓓ, and $\otimes$ denote element-wise addition, channel concentration/division, and matrix multiplication, respectively. By default, $(N_1, N_2) = (3, 3)$.

pairs as training dataset. During this process, we introduce sensor response to be close to real compressive camera.

As shown in Fig. 4 (b), the joint optimization framework includes the modeled encoder and the designed deep decoder, in which learnable mask weights ϕ and reconstruction network weights ψ are trained in an E2E manner. The proposed structural mask $\boldsymbol{M}_\lambda$ is updated from ϕ by a differentiable transformer $\mathcal{F}$, which is detailedly introduced in Sec. 4.2. Given the input 3D video $\boldsymbol{X}$, the encoder compresses it into a 2D measurement $\boldsymbol{Y}$ by

$$\boldsymbol{Y} = \mathcal{R}\left[\sum_{t=1}^{B} \boldsymbol{M}_\lambda \odot \boldsymbol{X}\right]. \tag{6}$$

Before fed into the decoder, the captured 2D measurement $\boldsymbol{Y}$ is initialized into a 3D datacube $\boldsymbol{X}'$ by

$$\boldsymbol{X}' = \boldsymbol{Y} + \boldsymbol{Y} \odot \boldsymbol{M}. \tag{7}$$

The structural mask $\boldsymbol{\Phi}_\lambda$ in the encoder and a deep reconstruction network $\mathcal{D}$ as the decoder are jointly optimized by the following vectorized loss function:

$$\underset{\{\phi,\psi\}}{\arg\min} \sum_{k=1}^{K} \|\mathcal{D}_\psi \circ \mathcal{R} \circ \mathcal{H}_\phi(\boldsymbol{x}_k) - \boldsymbol{x}_k\|_2^2, \tag{8}$$

where K denotes the number of training samples, ϕ and ψ represents the parameters of learnable mask $\boldsymbol{M}_\lambda$ and deep decoder $\mathcal{D}$ (described in Sec. 4.3), respectively. Since sensor's non-linearity is easily calibrated, we model the sensor response of an 8-bit image sensor as $\mathcal{R}(x) = \lfloor 255 \cdot x + 0.5 \rfloor / 255$, a composition of out-of-range clipping and quantization. As a hard thresholding function, $\mathcal{R}$ doesn't yield useful gradients and it follows the training strategy of mask optimization.

4.2. Structural Mask Optimization

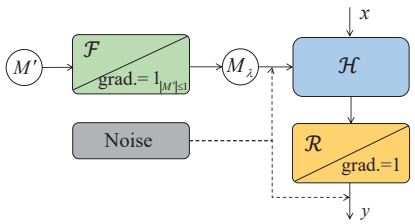


Figure 5. Illustration of mask optimization with non-differentiable hardware encoder. During forward propagation, $\boldsymbol{y} = \mathcal{R}[\mathcal{F}(\boldsymbol{\Phi}') \cdot \boldsymbol{x}]$. During back propagation, the derivative of $\mathcal{R}$ and $\mathcal{F}$ (see Alg. 1) are set to 1. Noise should be considered into the encoder when error caused by measurement noise and physical mask miscalibration is non-negligible.

By considering mask as learnable weights, jointly optimizing mask with a deep reconstruction network could contribute to video SCI as demonstrated in previous binary mask optimization works [12, 40, 21], which is generally faced with difficulties in forward-propagation *binarization* and back-propagation *differentiability*. Compared with them, optimizing the proposed structural mask is more challenging due to the difficulties in forward-propagation *discretization* and *structuralization*, and back-propagation *differentiability*.

A pioneering work [11] indicated that back-propagation gradients through discretization can be considered to be invariant as long as the forward-propagation input is lim-

ited in $[-1, 1]$. Inspired by this work, we propose a differentiable structural mask transformer $\mathcal{F}$ to generate λ-bit structural mask $\boldsymbol{M}_\lambda$ from a learnable loating-point mask $\boldsymbol{M}' \in [0, 1]$. As illustrated in Alg. 1, the input $\boldsymbol{M}$ is first discretized and then structuralized in the desired discrete domain during forward propagation, and following the training strategy in [11], the gradient of $\mathcal{F}$ is set to 1 during backward propagation. In the discretization process, the discretized mask $\boldsymbol{M}$ is fine-tuned to meet the structure of the temporal sum being 1 at all spatial position. The same back propagation strategy is also used for the non-differentiable sensor response $\mathcal{R}$. $\boldsymbol{\Phi}'$ and $\boldsymbol{\Phi}_\lambda$ denote the measurement matrix form of $\boldsymbol{M}'$ and $\boldsymbol{M}_\lambda$, respectively. The whole training framework is depicted in Fig. 5.

4.3. Res2former as Deep Decoder

Building spatial-temporal interactions is the key of video SCI reconstruction. Temporal features are considered to be as important as spatial features in previous reconstruction networks [28, 35, 4, 37, 39, 34]. Previous SOTA STFormer [34] has a long-term spatial-temporal feature extraction ability but the computational complexity and memory occupation is too high to enable real-world large-scale video SCI. Considering the motion-aware property caused by the proposed structural mask, we tailor an highly efficient reconstruction network, dubbed Res2former, as the deep decoder. Res2former is the first to put most of computations into capturing long-term temporal dependencies using Transformers.

As demonstrated in the decoder of Fig. 4 (b), Res2former is composed of a feature extraction module, a two-level U-shaped network built by multiple ResTSA modules, and video recovery module. Feature extraction module is to extract low-level features from measurement domain, composed of two 3D convolutional layers with kernel sizes of $3 \times 3 \times 3$ and $1 \times 3 \times 3$ respectively. From the perspective of U-Net [30], N_1 ResTSA modules work with two downsampling/upsampling operations as encoder/decoder and N_2 ResTSA modules as the bottleneck. Such an architecture can enable Res2former to learn high-level feature residuals from the low-level feature embeddings computationally efficiently. The video reconstruction module is composed of pixelshuffle [31] and two 3D convolution layers with kernel sizes of $1 \times 1 \times 1$ and $3 \times 3 \times 3$ respectively. The main novelties of Res2former is ResTSA module and its temporal self-attention (TSA) mechanism. Next, we introduce them in detail.

ResTSA Module. Previous works [38, 19] have indicated that channel grouping calculations can effectively reduce model complexity and layered interactions between groups can effectively improve the multiple-scale representation ability [7]. As shown in Fig. 4 (c), ResTSA module is also a hierarchical and residual-like structure built by multiple TSA branches. Given an input $\boldsymbol{X}_r$, a P-level ResTSA module can be formulated as

$$\begin{aligned}
&\boldsymbol{L}_1, \boldsymbol{L}_2, ..., \boldsymbol{L}_P = \texttt{Div}(\boldsymbol{L}),\\
&\boldsymbol{L}'_1 = \text{TSA}(\boldsymbol{L}_1),\\
&\boldsymbol{L}'_2 = \text{TSA}(\boldsymbol{L}_2 + \boldsymbol{L}'_1),\\
&\quad\vdots\\
&\boldsymbol{L}'_P = \text{TSA}(\boldsymbol{L}_P + \boldsymbol{L}'_{P-1}),\\
&\bar{\boldsymbol{L}} = \texttt{Conv}_{1\times1\times1}(\texttt{Concat}(\boldsymbol{L}'_1, \boldsymbol{L}'_2, ..., \boldsymbol{L}'_P)) + \boldsymbol{L}.
\end{aligned} \tag{9}$$

where `Div` and `Concat` denote the channel division and concatenate respectively.

TSA Branch. With a global perception ability, Transformer can mitigate the shortcomings caused by CNNs' limited receptive field and has achieved SOTA performance for video SCI reconstruction [34]. However, self-attention computation along spatial-temporal (3D) dimensions leads to a computational bottleneck for real-world large-scale video SCI applications. Inspired by [1, 34], in each ResTSA module, self-attention computation is limited in the temporal dimension. Given an input $\boldsymbol{L} \in \mathbb{R}^{B\times H\times W\times C}$, we first use a 2D convolution to establish local interrelations and then reshape the output into $\boldsymbol{L}_t \in \mathbb{R}^{HW\times B\times C}$, *i.e.*, $\boldsymbol{L}_t = \texttt{Reshape}(\texttt{Conv}_{3\times3}(\boldsymbol{L}))$. Next, we can obtain *query* $\boldsymbol{Q} \in \mathbb{R}^{HW\times B\times \frac{C}{2}}$, *key* $\boldsymbol{K} \in \mathbb{R}^{HW\times B\times \frac{C}{2}}$, and *value* $\boldsymbol{V} \in \mathbb{R}^{HW\times B\times \frac{C}{2}}$ by the following linear projection:

$$\boldsymbol{Q} = \boldsymbol{L}_t\boldsymbol{W}^Q, \boldsymbol{K} = \boldsymbol{L}_t\boldsymbol{W}^K, \boldsymbol{V} = \boldsymbol{L}_t\boldsymbol{W}^V, \tag{10}$$

where $\{\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V\} \in \mathbb{R}^{C\times\frac{C}{2}}$ denote the linear projection matrices. Note that the output dimension is reduced to half of the input dimension, further decreasing the computational complexity. Then, $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are divided into N heads along the feature channel: $\boldsymbol{Q} = \{\boldsymbol{Q}_k\}_1^N$, $\boldsymbol{K} = \{\boldsymbol{K}_k\}_1^N$, $\boldsymbol{V} = \{\boldsymbol{V}_k\}_1^N \in \mathbb{R}^{HW\times B\times\frac{C}{2N}}$. For k-th *head*, the attention can be calculated by

$$head_k = \boldsymbol{A}_k * \boldsymbol{V}_k, \tag{11}$$

where $\boldsymbol{A}_k = \texttt{softmax}(\boldsymbol{Q}_j\boldsymbol{K}_j^T/\sqrt{d}) \in \mathbb{R}^{HW\times B\times B}$ represents an attention map with a scaling parameter $d = \frac{C}{2N}$. Finally, we concatenate the outputs of N *heads* along the channel dimension and perform a linear mapping to obtain the final output $\boldsymbol{L}' \in \mathbb{R}^{B\times H\times W\times C}$:

$$\boldsymbol{L}' = \boldsymbol{L} + \texttt{Reshape}(\boldsymbol{W}(\texttt{Concat}[head_1, ..., head_N])), \tag{12}$$

where $\boldsymbol{W} \in \mathbb{R}^{\frac{C}{2}\times C}$ is the linear projection matrix. After temporal self-attention calculations, long-term correlation have been established. Next, we use the feed-forwad network, composed of two 3D convolutions with kernel sizes of $3 \times 3 \times 3$ and $1 \times 1 \times 1$, respectively, to further improve the model capacity and the local detail refinement ability, which can be formulated as

$$\bar{\boldsymbol{L}} = \boldsymbol{L}' + \texttt{Conv}_{1\times1\times1}(\texttt{LeakyReLU}(\texttt{Conv}_{3\times3\times3}(\boldsymbol{L}'))). \tag{13}$$

Table 1. Definition of different encoders.

Encoder	Configuration
RBw/oSR	Random Binary Mask without Sensor Response
RBwSR	Random Binary Mask with Sensor Response
LSwSR	Learned Structural Mask with Sensor Response

Table 2. Average PSNR (left), SSIM (center) and Q-Score (right) of different networks on six grayscale benchmark datasets.

Network	Train: *RBw/oSR* Test: *RBw/oSR*	Train: *RBw/oSR* Test: *RBwSR*
U-net [28]	29.45, 0.882, 47.31	27.02, 0.878, 46.82
BIRNAT [5]	33.31, 0.951, 50.30	29.72, 0.935, 48.80
MetaSCI [35]	31.72, 0.926, 48.34	28.84, 0.921, 47.92
RevSCI [4]	33.92, 0.956, 51.21	29.71, 0.939, 49.43
SCI3D [37]	35.26, 0.968, 52.70	30.97, 0.952, 50.94
ELP-Unfolding [39]	35.41, 0.969, 53.02	30.77, 0.955, 51.53
STFormer [34]	36.34, 0.974, 54.00	31.78, 0.962, 52.15

Table 3. Average PSNR, SSIM and Q-Score of the re-trained U-net, RevSCI, SCI3D, and STFormer.

Network	Train & Test: *RBwSR*
U-net[28]	24.67, 0.878, 46.25
RevSCI [4]	26.46, 0.897, 46.57
SCI3D [37]	27.54, 0.939, 49.51
STFormer [34]	27.66, 0.941, 49.69

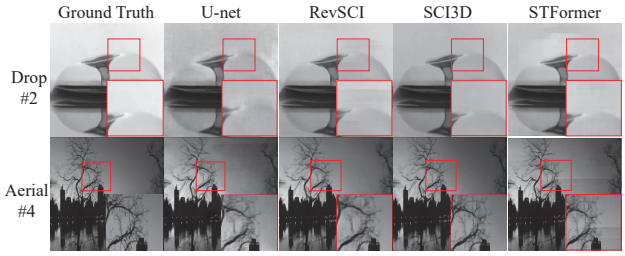


Figure 6. Visual results of the re-trained Unet, RevSCI, SCI3D, and STFormer.

4.4. Compared with Previous Framework

Previous reconstruction networks [28, 5, 35, 4, 34, 23, 18, 37, 39] were trained using random binary mask without considering sensor response and thus have achieved impressive performance on simulation rather than real system. The proposed deep optics framework aims to this problem. As mentioned previously, the modeled encoder is essential for training and simulated testing. As shown in Tab. 1, different encoders are distinguished into

- Previous framework: i) training a deep decoder with the *RBw/oSR* encoder; ii) deploying the well-trained deep decoder into real video SCI systems.
- Our framework: i) training a deep decoder with the *LSwSR* encoder in an E2E fashion; ii) deploying the learned structural mask and the well-trained deep decoder into real video SCI systems.

5. Experiments

In this section, we validate the effectiveness of the proposed deep optics framework, including learned structural mask and reconstruction network Res2former. We evaluate the accuracy of different networks' reconstruction by peak signal-to-noise-ratio (PSNR), structured similarity index metrics (SSIM), and Q-Score of HDR-VDP-2 [20] (dynamic range metric) and by our built **real system**.

5.1. Datasets and Implementation Details

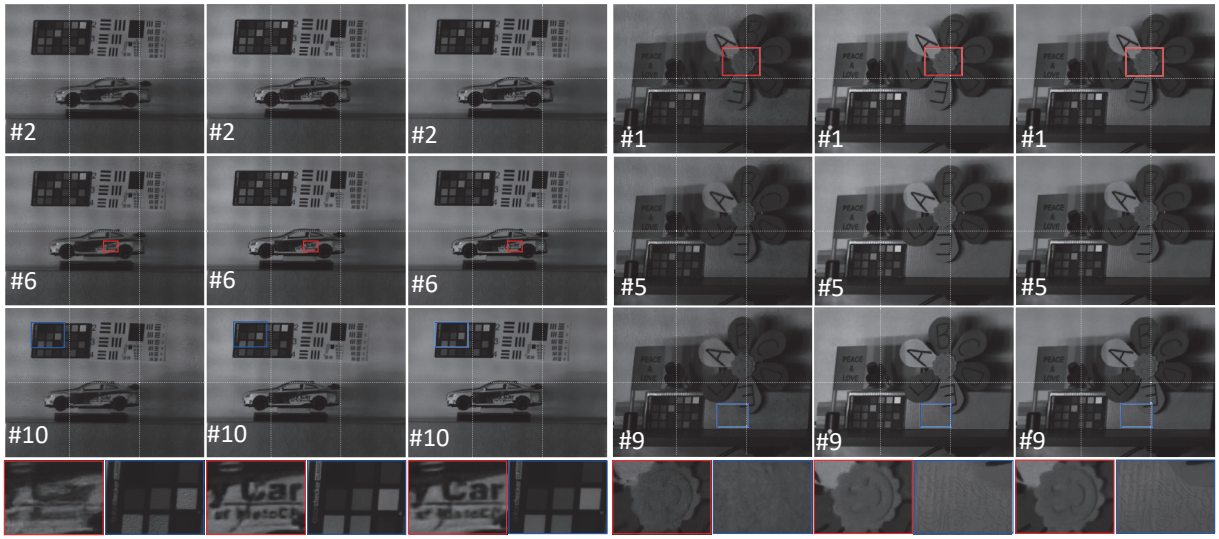
Following previous works [28, 5, 35, 4, 34, 23, 18, 37, 39], we employ DAVIS2017 [26] as the training dataset. For the simulation test, 6 benchmark datasets including `Kobe`, `Runner`, `Drop`, `Traffic`, `Aerial`, and `Vehicle` with a size of $256 \times 256 \times 8$ are used. For the real data, we built a video SCI prototype using a DLP7000 DMD, whose details are in supplementary materials (SM). The real data with a size of $768 \times 1024 \times 10$ is captured from `Car` and `Windmill` scenes by our prototype. The proposed E2E network is trained on Pytorch with 8 A40 GPUs. Adam optimizer is used to minimize the loss function with the learning rate of 10^{-4}.

5.2. Results on Synthetic Data

Before evaluating the proposed method, we give a clear insight into the performance degradation of previous video SCI reconstruction networks in real system, including Unet [28], BIRNAT [5], MetaSCI [35], RevSCI [4], SCI3D [37], ELP-Unfolding [39], and the SOTA STFormer [34]. These well-trained networks (based on the *RBw/oSR* encoder) in their works are now tested with the *RBwSR* encoder (close to a real system). To avoid overexposure caused by binary mask, automatic aperture is simulated by value scaling before sensor response. As shown in Tab. 2, there is a serious degradation in both structural information (PSNR, SSIM) and dynamic range (Q-Score). Obviously, it is caused by the gap between training without sensor response and testing with sensor response. Unfortunately, the *RBw/oSR*-training-plus-*RBwSR*-using framework is prevailing even though sensor response is inevitable in hardware encoder. It could be a natural explanation for why the performance of previous networks [28, 5, 35, 4, 34, 23, 18, 37, 39] on real data is far from that on simulation. Moreover, we have re-trained four representative networks, including Unet [28], RevSCI [4], SCI3D [37], and STFormer [34] with the *RBwSR* encoder. The results are shown in Tab. 3 and Fig. 6. All these re-trained networks lead to worse results and their reconstructed video frames have an visually clear degradation in the dynamic range. It is because these networks are incapable of simultaneously resolving compressive video reconstruction and dynamic range reconstruction caused by random binary mask. With unavoidable sensor response in real

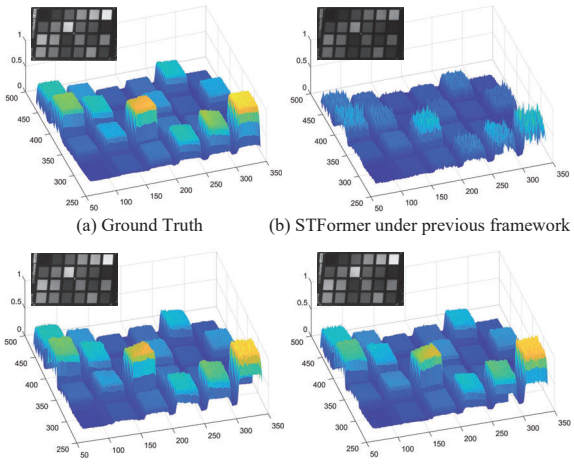
Table 4. Average PSNR, SSIM and Q-Score of different networks on six grayscale benchmark datasets.

Network	Under Previous Framework (with impractical *RBwoSR* encoder)	Under Our Framework (with practical *LSwSR* encoder)	Gain ↑	Parameters (M)	FLOPs (G)	Runing Time (s)
U-net [28]	29.45, 0.882, 47.31	32.42, 0.940, 49.72	2.97, 0.058, 2.41	0.82	53.49	0.01
RevSCI [4]	33.92, 0.956, 51.21	34.81, 0.965, 52.74	0.89, 0.009, 3.31	5.66	766.95	0.19
STFormer [34]	36.34, 0.974, 54.00	36.69, 0.976, 55.08	0.35, 0.002, 1.08	19.48	3060.75	0.49
Res2former	NA	35.98, 0.972, 54.31	NA	11.02	861.76	0.19
Res2former-L	NA	36.56, 0.975, 54.93	NA	17.70	1362.51	0.42



(1a) (1b) (1c) (2a) (2b) (2c)

Figure 7. Results of real data. (1a)-(2a), (1b)-(2b), and (1c)-(2c) are reconstructed by STFormer under previous framework, STFormer under our framework, and Res2former under our framework, respectively. Our deep optics framework brings a significant improvement compared with previous framework. In real data, the proposed Res2former is as good as STFormer.



(a) Ground Truth (b) STFormer under previous framework

(c) STFormer under our framework (d) Res2former under our framework

Figure 8. Dynamic range comparison through the 3D heat map of a standard ColorChecker placed in scenes, which is an average effect of the recovered 10 video frames. Obviously, Res2former and STFormer under our framework can retrieve wider dynamic range (close to ground truth) than STFormer under previous framework.

system, random binary mask is therefore not the best choice unless using expensive high bit depth sensor (defying the main motivation of video SCI).

Next, we valuate the generalization of the proposed deep optics framework and Res2former's effectiveness on balancing reconstruction performance and computational load. Unet [28], RevSCI [4], and STFormer [34] are re-trained under our deep optics framework as competitors, in which only Res2former is replaced and 4-bit structural mask is jointly optimized. As shown in Tab. 4, three other networks have archived improvement with various degrees. The jointly-optimized U-net has achieved a significant improvement in PSNR and SSIM and the improvement of the jointly-optimized RevSCI is the best in terms of Q-Score, resulting from their promotion space is greater than STFormer. The Res2former can achieve the result close to the jointly-optimized STFormer ($< 1dB$) with only 28.15% FLOPs and 56.57% parameters of STFormer and its running time is far less than STFormer. We have also tried to increase the parameters of Res2former to close to that of STFormer. The large Res2former, dubbed Res2former-L, is generated by increasing channels from 96 to 128 and the depth of ResTSA module from N_1=3 to N_1=5. Res2former-L can achieve the same level as the jointly-optimized STFormer and is better than the original STformer.

5.3. Results on Real Data

Table 5. Ablation study with different kinds of structural mask.

Mask	Random	Learned
1-bit	34.25, 0.964, 50.06	34.62, 0.964, 51.45
2-bit	35.03, 0.967, 53.17	35.87, 0.971, 53.91
3-bit	34.91, 0.965, 52.54	35.90, 0.971, 54.17
4-bit	34.95, 0.965, 52.85	35.98, 0.972, 54.31

We validate the effectiveness of the proposed deep optics framework and Res2former in our prototype whose details can be got in SM. Previous SOTA STFormer is regarded as the benchmark reconstruction network. We conduct real system test in the following three settings: i) STFormer under previous framework; ii) STFormer under our framework; ii) Res2former under our framework. Two kinds of high-speed scenes (`Car` and `Windmill`) are modulated by random binary masks (previous framework) or the learned structural masks (our framework) and then captured into single-shot 768×1024 measurement frames by an off-the-shelf camera with 50 fps. The compressive sampling ratio $1/10$. To ensure motion uniformity `Car` and `Windmill` are driven by an electric linear gateway and a rotating motor, respectively. As shown in Fig. 7, the reconstructed results of STFormer under our framework is far better than that of the original version of STFormer in dynamic and static region. STFormer and Res2former are too close to call in real data. Moreover, we analyze the dynamic range of these recovered results. The proposed framework can eliminate the dynamic range degradation (rooted in previous works) completely and achieve FDR video SCI, as demonstrated in Fig. 8. More results are in SM.

5.4. Ablation Study

To verify the proposed deep optics framework, we conduct two ablation experiments: i) learning for different bit of structural mask; ii) training with learnable structural mask or fixed structural mask (generated through Alg. 1 with an random input). All experiments are conducted on Res2former and tested on the 6 grayscale benchmark datasets. As shown in Tab. 5, mask optimization can contribute to reconstruction regardless of the bit depth (space). The larger the learnable mask space is, the better the results are, which follows the conclusion about mask conditioning in [33]. With randomly generated structural mask, Res2former cannot achieve its full potential.

6. Conclusion

Aiming to move one step further of video SCI towards practical applications, we have proposed a deep optics framework to jointly optimize the proposed structural mask and reconstruction network Res2former. As validated in simulation and real system, our framework can bring a significant improvement for other networks. Besides, our Res2former can provide competitive performance in a computationally efficient manner.

Acknowledgements: This work was supported by National Natural Science Foundation of China (62271414), Zhejiang Provincial Natural Science Foundation of China (LR23F010001) and Research Center for Industries of the Future (RCIF) at Westlake University.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 6

[2] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, January 2011. 2

[3] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10193–10202, 2019. 2

[4] Ziheng Cheng, Bo Chen, Guanliang Liu, Hao Zhang, Ruiying Lu, Zhengjue Wang, and Xin Yuan. Memory-efficient network for large-scale video compressive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16246–16255, June 2021. 1, 2, 3, 4, 6, 7, 8

[5] Ziheng Cheng, Ruiying Lu, Zhengjue Wang, Hao Zhang, Bo Chen, Ziyi Meng, and Xin Yuan. BIRNAT: Bidirectional recurrent neural networks with adversarial training for video snapshot compressive imaging. In *Proceedings of the European conference on computer vision (ECCV)*, August 2020. 1, 2, 3, 4, 7

[6] Marco F Duarte, Mark A Davenport, Dharmpal Takhar, Jason N Laska, Ting Sun, Kevin F Kelly, and Richard G Baraniuk. Single-pixel imaging via compressive sampling. *IEEE signal processing magazine*, 25(2):83–91, 2008. 2

[7] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019. 6

[8] M. E. Gehm, R. John, D. J. Brady, R. M. Willett, and T. J. Schulz. Single-shot compressive spectral imaging with a dual-disperser architecture. *Opt. Express*, 15(21):14013–14027, Oct 2007. 2

[9] 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*, pages 287–294, Nov 2011. 1

[10] Ryoichi Horisaki, Yuka Okamoto, and Jun Tanida. Deeply coded aperture for lensless imaging. *Optics Letters*, 45(11):3131–3134, 2020. 2

[11] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. *Advances in neural information processing systems*, 29, 2016. 5, 6

[12] Michael Iliadis, Leonidas Spinoulas, and Aggelos K. Katsaggelos. Deepbinarymask: Learning a binary mask for video compressive sensing. *Digital Signal Processing*, 96:102591, 2020. 2, 5

[13] Yasutaka Inagaki, Yuto Kobayashi, Keita Takahashi, Toshiaki Fujii, and Hajime Nagahara. Learning to capture light fields through a coded aperture camera. In *Proceedings of the*

European Conference on Computer Vision (ECCV), pages 418–434, 2018. 2

[14] Y. Li, M. Qi, R. Gulve, M. Wei, R. Genov, K. N. Kutulakos, and W. Heidrich. End-to-end video compressive sensing using anderson-accelerated unrolled networks. In *2020 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12, 2020. 2

[15] Xuejun Liao, Hui Li, and Lawrence Carin. Generalized alternating projection for weighted-2,1 minimization with applications to model-based compressive sensing. *SIAM Journal on Imaging Sciences*, 7(2):797–823, 2014. 2

[16] Dengyu Liu, Jinwei Gu, Yasunobu Hitomi, Mohit Gupta, Tomoo Mitsunaga, and Shree K Nayar. Efficient space-time sampling with pixel-wise coded exposure for high-speed imaging. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):248–260, 2013. 1

[17] Y. Liu, X. Yuan, J. Suo, D. J. Brady, and Q. Dai. Rank minimization for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):2990–3006, Dec 2019. 1, 2, 3

[18] Jiawei Ma, Xiaoyang Liu, Zheng Shou, and Xin Yuan. Deep tensor admm-net for snapshot compressive imaging. In *IEEE/CVF Conference on Computer Vision (ICCV)*, 2019. 1, 2, 3, 4, 7

[19] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018. 6

[20] Rafał Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4), jul 2011. 7

[21] Julien NP Martel, Lorenz K Mueller, Stephen J Carey, Piotr Dudek, and Gordon Wetzstein. Neural sensors: Learning pixel exposures for hdr imaging and video compressive sensing with programmable sensors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1642–1653, 2020. 1, 2, 5

[22] Kshitij Marwah, Gordon Wetzstein, Yosuke Bando, and Ramesh Raskar. Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Transactions on Graphics (TOG)*, 32(4):1–12, 2013. 2

[23] Ziyi Meng, Xin Yuan, and Shirin Jalali. Deep unfolding for snapshot compressive imaging. *International Journal of Computer Vision*, pages 1–26, 2023. 1, 2, 3, 4, 7

[24] Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1375–1385, 2020. 2

[25] Elias Nehme, Daniel Freedman, Racheli Gordon, Boris Ferdman, Lucien E Weiss, Onit Alalouf, Tal Naor, Reut Orange, Tomer Michaeli, and Yoav Shechtman. Deepstorm3d: dense 3d localization microscopy and psf design by deep learning. *Nature methods*, 17(7):734–740, 2020. 2

[26] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 7

[27] Mu Qiao, Xuan Liu, and Xin Yuan. Snapshot spatial–temporal compressive imaging. *Opt. Lett.*, 45(7):1659–1662, Apr 2020. 1

[28] M. Qiao, Z. Meng, J. Ma, and X. Yuan. Deep learning for video compressive sensing. *APL Photonics*, 5(3):030801, 2020. 1, 2, 3, 4, 6, 7, 8

[29] D. Reddy, A. Veeraraghavan, and R. Chellappa. P2c2: Programmable pixel compressive camera for high speed imaging. In *CVPR 2011*, pages 329–336, June 2011. 1

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 6

[31] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 6

[32] Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1386–1396, 2020. 2

[33] Edwin Vargas, Julien NP Martel, Gordon Wetzstein, and Henry Arguello. Time-multiplexed coded aperture imaging: Learned coded aperture and pixel exposures for compressive imaging systems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2692–2702, 2021. 2, 9

[34] Lishun Wang, Miao Cao, Yong Zhong, and Xin Yuan. Spatial-temporal transformer for video snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2022. 1, 2, 3, 4, 6, 7, 8

[35] Zhengjue Wang, Hao Zhang, Ziheng Cheng, Bo Chen, and Xin Yuan. Metasci: Scalable and adaptive reconstruction for video compressive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2083–2092, June 2021. 1, 2, 3, 4, 6, 7

[36] Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. Phasecam3d—learning phase masks for passive single view depth estimation. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2019. 2

[37] Zhuoyuan Wu, Jian Zhang, and Chong Mou. Dense deep unfolding network with 3d-cnn prior for snapshot compressive imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4892–4901, 2021. 1, 2, 3, 4, 6, 7

[38] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 6

[39] Chengshuai Yang, Shiyu Zhang, and Xin Yuan. Ensemble learning priors unfolding for scalable snapshot compressive sensing. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 2, 3, 4, 6, 7

[40] Michitaka Yoshida, Akihiko Torii, Masatoshi Okutomi, Kenta Endo, Yukinobu Sugiyama, Rin-ichiro Taniguchi, and Hajime Nagahara. Joint optimization for compressive video sensing and reconstruction under hardware constraints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 634–649, 2018. 2, 5

[41] X. Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2539–2543, Sept 2016. 1, 2, 3

[42] Xin Yuan, Yang Liu, Jinli Suo, and Qionghai Dai. Plug-and-play algorithms for large-scale snapshot compressive imaging. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1444–1454, 2020. 3

[43] Xin Yuan, Yang Liu, Jinli Suo, Frédo Durand, and Qionghai Dai. Plug-and-play algorithms for video snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7093–7111, 2022. 3

[44] Bo Zhang, Xin Yuan, Chao Deng, Zhihong Zhang, Jinli Suo, and Qionghai Dai. End-to-end snapshot compressed super-resolution imaging with deep optics. *Optica*, 9(4):451–454, Apr 2022. 2