A Simple Low-bit Quantization Framework for Video Snapshot Compressive Imaging

Miao Cao[1,2], Lishun Wang[2], Huan Wang[2], and Xin Yuan[2](✉)

[1] Zhejiang University, Hangzhou 310058, Zhejiang, China
[2] School of Engineering, Westlake University, Hangzhou, 310030, Zhejiang, China
{caomiao,wanglishun,wanghuan,xyuan}@westlake.edu.cn

Abstract. Video Snapshot Compressive Imaging (SCI) aims to use a low-speed 2D camera to capture high-speed scene as snapshot compressed measurements, followed by a reconstruction algorithm to reconstruct the high-speed video frames. State-of-the-art (SOTA) deep learning-based algorithms have achieved impressive performance, yet with heavy computational workload. Network quantization is a promising way to reduce computational cost. However, a direct low-bit quantization will bring large performance drop. To address this challenge, in this paper, we propose a simple low-bit quantization framework (dubbed *Q-SCI*) for the end-to-end deep learning-based video SCI reconstruction methods which usually consist of a feature extraction, feature enhancement, and video reconstruction module. Specifically, we first design a high-quality feature extraction module and a precise video reconstruction module to extract and propagate high-quality features in the low-bit quantized model. In addition, to alleviate the information distortion of the Transformer branch in the quantized feature enhancement module, we introduce a shift operation on the query and key distributions to further bridge the performance gap. Comprehensive experimental results manifest that our Q-SCI framework can achieve superior performance, *e.g.*, 4-bit quantized EfficientSCI-S derived by our Q-SCI framework can theoretically accelerate the real-valued EfficientSCI-S by 7.8× with only 2.3% performance gap on the simulation testing datasets. Code is available at `https://github.com/mcao92/QuantizedSCI`.

Keywords: Computational imaging · Snapshot compressive imaging · Deep learning · Network quantization · Transformer

1 Introduction

Recently, video Snapshot Compressive Imaging (SCI) has attracted much attention because it can capture high-speed scenes using a low-speed 2D camera with low bandwidth. There are two main steps in the video SCI system: hardware encoding (shown in Fig. 1(a)) and software decoding(shown in Fig. 1(b)) [62]. In the hardware encoding process, we first modulate the dynamic scene with different masks, and then the modulated scene is compressed into a series of snapshot measurements, which are finally captured by a low-cost 2D camera. In

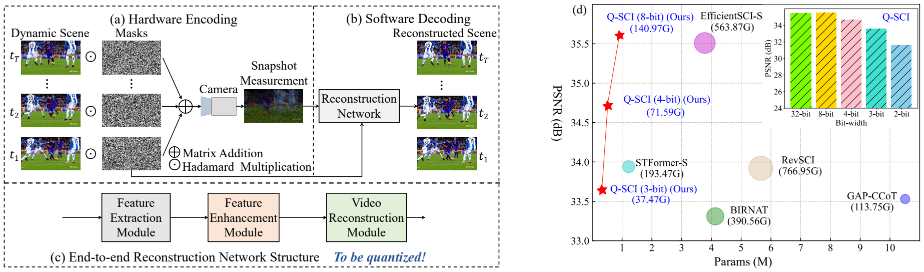


Fig. 1: (a) In the encoding process of video SCI, a dynamic scene is modulated by different masks and then the modulated scene is captured by a camera as snapshot measurements. (b) In the decoding process of video SCI, the captured measurements and the corresponding masks are fed into a reconstruction algorithm to obtain the desired video frames. (c) The pipeline of an end-to-end video SCI reconstruction network. (d) Comparison of the reconstruction quality and computational cost of different reconstruction methods.

the software decoding stage, the captured snapshot measurements and the corresponding modulation masks are fed into a reconstruction algorithm to recover the desired video frames.

On the one hand, many successful video SCI hardware encoders [9,15,37,67] have been built. On the other hand, state-of-the-art (SOTA) reconstruction algorithms [7,8,47,51,53,57], mostly based on deep neural networks, have outperformed the traditional model-based methods [27,61] by a large margin not only in the reconstruction quality but also in the running speed. Thus, it seems that the current video SCI systems can already be used in practical applications. Unfortunately, existing deep learning-based reconstruction algorithms usually need a large number of parameters and floating point operations (FLOPs) to achieve satisfactory accuracy. EfficientSCI [45] is a significant exploration to design an efficient video SCI reconstruction algorithm, which has 3.78 MB parameters and requires 563.87 GFLOPs to reconstruct the video frames with 35.51 dB on the simulation testing datasets. Our ultimate goal is to deploy video SCI reconstruction methods on AI chips, thus to forming a real end-to-end system integrating capture and reconstruction. However, it is still challenging to deploy previous reconstruction algorithms on resource-limited devices due to their high computational cost. This motivates us to move one step further to reduce the computational workload of the video SCI reconstruction algorithms while preserving the model performance as much as possible.

Substantial efforts have been made to compress and accelerate neural networks for efficient online inference. These methods can be divided into the following categories: network quantization [59,66], network pruning [12,25,43], knowledge distillation [14,41,44,68], and compact network design [6,23,24]. Among them, network quantization is suitable for deployment on resource-limited platforms because it can reduce the bit-width of network parameters and activations for efficient inference. However, directly applying low-bit quantization to video

SCI reconstruction algorithms may encounter the following issues: **i)** Directly quantizing previous reconstruction methods into low-bit will lead to a large performance drop from their full-precision counterpart. **ii)** There exists a distortion of the query and key distributions in the low-bit quantized vision Transformer.

Bearing the above concerns in mind, in this paper, we propose a simple low-bit quantization framework, dubbed *Q-SCI*, for the end-to-end deep learning-based video SCI reconstruction methods which usually consist of a feature extraction module, a feature enhancement module, and a video reconstruction module as shown in Fig. 1(c). Specifically, we first conduct extensive empirical analysis and identify that the large performance drop in low-bit quantized model mainly comes from the information loss of low-bit quantized features. Therefore, we design a high-quality feature extraction module and a precise video reconstruction module which can extract and propagate high-quality features through the low-bit quantized network. In addition, to alleviate the information distortion of the Transformer branch in the quantized feature enhancement module, we introduce a shift operation on the query and key distributions to further bridge the performance gap. Finally, following previous works [28, 32, 48] which are built and evaluated on existing backbones. We choose EfficientSCI-S as the full-precision backbone model for our proposed Q-SCI because EfficientSCI-S can achieve SOTA reconstruction quality with much cheaper computational cost. As shown in Fig. 1(d), compared with previous methods, the proposed Q-SCI framework can achieve comparable reconstruction quality with *much fewer model parameters and FLOPs*.

The main contributions of our paper can be summarized as follows:

- We propose a simple low-bit quantization framework for video SCI reconstruction. To our best knowledge, this is *the first network quantization framework* in the video SCI reconstruction task.
- We design *a high-quality feature extraction module and a precise video reconstruction module* to better extract and propagate high-quality features. Furthermore, the *generalization ability* of the proposed Q-SCI framework is verified on different end-to-end video SCI reconstruction methods.
- *A shift operation* on the query and key distributions is introduced to alleviate the information distortion in the quantized Transformer branch.
- Experimental results on the simulated and real datasets demonstrate that the low-bit quantized models derived by our proposed Q-SCI framework can achieve superior performance with *much cheaper computational cost.*

2 Related Work

2.1 Video SCI Reconstruction Algorithms

The existing video SCI reconstruction methods can be divided into two categories: *traditional model-based methods* and *deep learning-based ones.* The traditional model-based methods formulate the video SCI reconstruction process as an optimization problem with a regularization term such as total variation [61] and Gaussian mixture model [58], and solve it via iterative algorithms. The major drawback of these model-based methods is that one needs to perform time-consuming iterative optimization. In order to improve the running speed, Yuan *et al.* develop a plug-and-play (PnP) framework, in which a pre-trained denoising model is plugged into every iteration of the optimization process [63,64]. However, they still take a long time on reconstructing large-scale data.

Recently, more and more researchers begin to utilize deep neural networks in the video SCI reconstruction task. For example, BIRNAT [8] uses a bidirectional recurrent neural network to exploit the temporal correlation. To save model training memory, RevSCI [7] builds an end-to-end 3D convolutional neural network with reversible structure for video SCI reconstruction. More recently, Wang *et al.* build the first Transformer-based reconstruction network [46] with space-time factorization and local self-attention mechanism. After that, Wang *et al.* design an efficient reconstruction network [3,45] based on dense connections and space-time factorization. Combining the idea of model-based and deep learning-based methods, researchers propose the deep unfolding networks [53,57,69]. For example, Zheng *et al.* first introduce the uncertainty estimation mechanism into the deep unfolding framework [69]. Although the deep learning-based models have achieved impressive results, it is still challenging to deploy them on the resource-limited devices due to their high computational workload. In this paper, we mainly focus on developing *light-weight video SCI reconstruction algorithm* empowered by network quantization.

2.2 Network Quantization

Due to its impressive effectiveness in computational compression, network quantization has been widely applied in high-level vision [5, 21, 49, 55] and low-level vision tasks [2, 16, 19, 35, 54]. *In the high-level vision tasks*, Liu *et al.* develop a dynamic quantization scheme [30] to decide the optimal bit-widths for each image. GPUSQ-ViT [60] is a compression scheme which maximally utilize the GPU-friendly 2:4 fine-grained structured sparsity and quantization. Q-ViT [22] presents the first quantization-aware training framework for accurate and low-bit vision Transformer. ReActNet [29] learns the distribution shape and shift to improve the performance of the CNN-based binary network. Li *et al.* propose a novel post-training quantization framework specifically tailored towards the unique multi-timestep pipeline and model architecture of the diffusion models [21]. Li *et al.* develop a fully quantized network for object detection [20]. *Considering the low-level vision tasks*, PAMS [19] learns the quantization intervals of different layers to adapt to vastly distinct feature distributions in the super-resolution (SR) networks. Hong *et al.* propose CADyQ [16] to assign different bit-width for each patch and layer for image SR. Xia *et al.* design a binarized convolution unit BBCU [54] for image restoration-tasks, *e.g.*, image denoising and JPEG compression artifact reduction. FQSR [39] jointly optimize efficiency and accuracy on the image SR task. Yet, the potential of network quantization for the video SCI reconstruction task has not been explored.

3 Prerequisites of Network Quantization

In this section, we briefly introduce the network quantization architecture used in our paper. First, given the activation x and weight $\mathbf{w}$, we introduce a general asymmetric activation quantization and symmetric weight quantization scheme, expressed as follows:

$$\begin{aligned} Q_a(x) &= \lfloor \mathrm{clip}\{(x-z)/\alpha_x, -Q_n^x, Q_p^x\} \rceil, \\ \hat{x} &= Q_a(x) \times \alpha_x + z, \\ Q_{\mathbf{w}}(\mathbf{w}) &= \lfloor \mathrm{clip}\{\mathbf{w}/\alpha_{\mathbf{w}}, -Q_n^{\mathbf{w}}, Q_p^{\mathbf{w}}\} \rceil, \\ \hat{\mathbf{w}} &= Q_{\mathbf{w}}(\mathbf{w}) \times \alpha_{\mathbf{w}}, \end{aligned} \tag{1}$$

where clip$\{x, m_1, m_2\}$ constrains x into $[m_1, m_2]$ by setting x as m_1 when $x < m_1$ and m_2 when $x > m_2$. $\lfloor x \rceil$ rounds x to the nearest integer. α is a scale factor, which can divide the entire range of the input into uniform partitions. z is a zero-point, which can shift the quantized distribution into the a specific range in the asymmetric quantization. During training, α and z are initialized as 1 and 0, and then been optimized as network parameters. Following this, if we quantize the activation x to a bits and weight $\mathbf{w}$ to b bits, we have:

$$\begin{aligned} Q_n^x &= 2^{a-1}, Q_p^x = 2^{a-1} - 1, \\ Q_n^{\mathbf{w}} &= 2^{b-1}, Q_p^{\mathbf{w}} = 2^{b-1} - 1. \end{aligned} \tag{2}$$

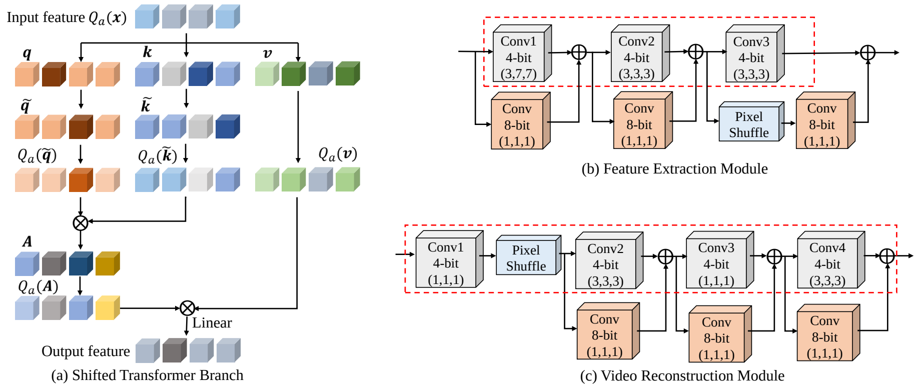
In this way, the forward and back propagation of network quantization can be formulated as Q-Linear$(x) = \hat{x} \cdot \hat{\mathbf{w}} = \alpha_x \alpha_{\mathbf{w}}((Q_a(x) + z/\alpha_x) \otimes Q_{\mathbf{w}}(\mathbf{w})$ and Eq. (3) respectively.

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial x} &= \frac{\partial \mathcal{J}}{\partial \hat{x}} \frac{\partial \hat{x}}{\partial x} = \begin{cases} \frac{\partial \mathcal{J}}{\partial \hat{x}} & \text{if } x \in [-Q_n^x, Q_p^x] \\ 0 & \text{otherwise} \end{cases}, \\ \frac{\partial \mathcal{J}}{\partial \mathbf{w}} &= \frac{\partial \mathcal{J}}{\partial x} \frac{\partial x}{\partial \hat{\mathbf{w}}} \frac{\partial \hat{\mathbf{w}}}{\partial \mathbf{w}} = \begin{cases} \frac{\partial \mathcal{J}}{\partial x} \frac{\partial x}{\partial \hat{\mathbf{w}}} & \text{if } \mathbf{w} \in [-Q_n^{\mathbf{w}}, Q_p^{\mathbf{w}}] \\ 0 & \text{otherwise} \end{cases}, \end{aligned} \tag{3}$$

where $\mathcal{J}$ denotes the loss function, $Q(\cdot)$ is used in the forward process while the straight-through estimator [1] is adopted to retain the derivation of gradient in the backward propagation process, and $\otimes$ denotes the efficient bit-wise matrix multiplication operation.

Recent methods [45,46] begin to explore the use of Transformer in the end-to-end video SCI reconstruction methods. Thus, we define the quantization process of the self-attention module in the following way. First, we denote the quantized computation on the query $\boldsymbol{q}$, key $\boldsymbol{k}$ and value $\boldsymbol{v}$ as $\boldsymbol{q} = Q$-Linear$_q(x)$, $\boldsymbol{k} = Q$-Linear$_k(x)$, $\boldsymbol{v} = Q$-Linear$_v(x)$, where Q-Linear$_q$, Q-Linear$_k$ and Q-Linear$_v$ represent the quantized linear layers for $\boldsymbol{q}$, $\boldsymbol{k}$ and $\boldsymbol{v}$, respectively. Then, the attention weight can be expressed as:

$$\begin{aligned} \mathbf{A} &= \tfrac{1}{\sqrt{d}}(Q_a(\boldsymbol{q}) \otimes Q_a(\boldsymbol{k})^{\top}), \\ Q_{\mathbf{A}} &= Q_a(\mathrm{softmax}(\mathbf{A})). \end{aligned} \tag{4}$$



(a) Shifted Transformer Branch

(b) Feature Extraction Module

(c) Video Reconstruction Module

Fig. 2: Illustration of the proposed Q-SCI framework: In (a), we introduce a *shift operation* on the query and key to alleviate their distribution distortion in the low-bit Transformer branch. In (b) and (c), convolutional layers in orange boxes are plugged to extract and propagate high-quality features through the low-bit quantized network.

4 Proposed Methods

In this section, we propose an accurate and fully quantized framework for the end-to-end video SCI reconstruction algorithms. Firstly, let us give a brief introduction of the previous SOTA EfficientSCI which is used as the full-precision backbone model for our proposed Q-SCI framework. The ResDNet module with N residual style ResDNet blocks is used for feature enhancement in EfficientSCI. In each ResDNet block, we put several CFormer blocks with *temporal Transformer branch*. Please refer to [45] and the supplementary material for more details about EfficientSCI and the video SCI mathematical model. Then, we conduct extensive empirical analysis on the severe performance drop of the low-bit quantized model. Following this, we design a high-quality feature extraction module, a precise video reconstruction module and a shifted Transformer branch to improve the performance of the low-bit quantized model. Afterwards, we customize four different quantized network variants based on our Q-SCI framework. Finally, the loss function is defined for the training process.

4.1 Performance Analysis on Low-bit Quantization

In the experiments, compared with the full-precision backbone model EfficientSCI-S, we observe a large performance drop (4.11 dB) in the 4-bit quantized baseline model which directly quantize each network layer of EfficientSCI-S into 4-bit. Thus, we try to detect the source of this severe performance drop in the experiments. *First*, based on the 8-bit quantized baseline model which directly quantize each network layer into 8-bit, we set the bit-width of one module (the feature extraction module, ResDNet module or video reconstruction module) to 4-bit while maintaining that of the other two modules. As we can see in Tab. 1,

Table 1: Reconstruction quality of different models which directly quantizing the feature extraction module, ResDNet module and video reconstruction module of EfficientSCI-S into 4-bit.

Feature Extraction Module	ResDNet Module	Video Reconstruction Module	PSNR	SSIM
			35.23	0.968
✓			33.01	0.941
	✓		34.71	0.963
		✓	34.74	0.962

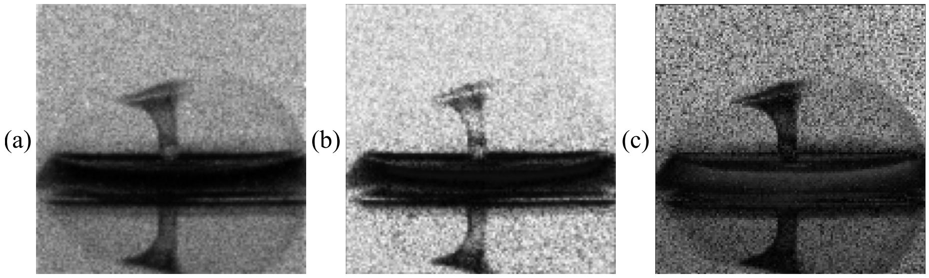


Fig. 3: The output feature map of the feature extraction module under different models: (a) Full-precision model, (b) 4-bit quantized baseline model, and (c) Using the proposed high-quality feature extraction module in 4-bit quantized baseline model.

there exists a 2.22 dB performance drop when we quantize the feature extraction module into 4-bit. However, the performance drop is relatively small with 0.52 dB and 0.49 dB for the 4-bit quantized ResDNet module and video reconstruction module. Therefore, we infer that the severe performance drop mainly comes from the low-bit quantized feature extraction module. *Stepping forward*, we visualize the output feature map of the feature extraction module to further verify our assumption. As shown in Fig. 3, compared with full-precision model (Fig. 3(a)), the output feature map of the feature extraction module in 4-bit quantized baseline model (Fig. 3(b)) degrades severely. Therefore, we can conclude that the large performance drop of the low-bit quantized model mainly comes from the low-bit quantized feature extraction module.

4.2 Proposed Quantization Framework

Quantized Feature Extraction Module We have noticed from the empirical analysis (detailed in Sec. 4.1) that high-quality features play an important role in the performance improvement of the low-bit quantized model. Therefore, in this section, we present a simple and high-quality feature extraction module to improve the performance of the low-bit quantized reconstruction methods. First, shortcut connections are widely used to propagate high-quality information. However, there exists a dimension mismatch problem during feature re-

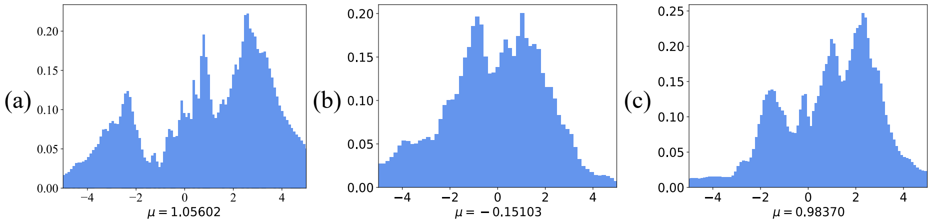


Fig. 4: The histogram of the query distribution in the first CFormer block of the first ResDNet block with different quantized models: (a) Full-precision model, (b) 8-bit quantized baseline model, and (c) Using the proposed shifted Transformer branch in 8-bit quantized baseline model.

shaping in the previous feature extraction module (shown in the red box of Fig. 2(b)). Thus, we propose to adopt several $1 \times 1 \times 1$ convolution layers as shortcut connections. Note that a pixel shuffle operation is performed before the last shortcut connected convolution layer to calibrate the spatial size mismatch. Then, we set the bit-width of the shortcut connected convolution layers to 8-bit to precisely propagate high-quality features. In this way, the proposed high-quality feature extraction module is capable of extracting and propagating high-quality features. Finally, when we compare Fig. 3(b) and Fig. 3(c), obvious quality improvement of the output feature map can be observed, which verifies the effectiveness of our proposed feature extraction module.

Quantized Transformer Branch As shown in Q-ViT [22], there exists a distortion of the query and key distributions in the low-bit quantized vision Transformer branch, which brings performance drop. Therefore, we plot the distribution of the query activation in Fig. 4 to better measure the distribution distortion of the low-bit quantized Transformer branch. We can see from Fig. 4(a) and Fig. 4(b) that the mean of the query distribution in the first CFormer block of the first ResDNet block between 8-bit quantized baseline model and its full-precision counterpart is 1.207 (1.056 *v.s.* -0.151). Unfortunately, we cannot directly adapt Q-ViT into the reconstruction methods due to the absence of bell-shaped distributions as shown in Fig. 4(b). Therefore, we adapt a shift operation into the Transformer branch to rectify the query and key distributions. Specifically, given the query $\boldsymbol{q}$ and key $\boldsymbol{k}$, we define a shift operation as $\widetilde{\boldsymbol{q}} = \boldsymbol{q} + \beta_{\boldsymbol{q}}$ and $\widetilde{\boldsymbol{k}} = \boldsymbol{k} + \beta_{\boldsymbol{k}}$, where $\beta_{\boldsymbol{q}}$ and $\beta_{\boldsymbol{k}}$ denote the learnable shift bias of the distributions. Finally, comparing Fig. 4(a) with Fig. 4(c), we observe that the mean of the query distribution in the first CFormer block of the first ResDNet block of 8-bit quantized baseline model equipped with our proposed shifted Transformer branch stays close to that of its full-precision counterpart (1.056 *v.s.* 0.984), which verifies the effectiveness of the proposed approach. Due to the space limitation, please refer to the supplementary material for more distribution histograms of different CFormer blocks.

Quantized Video Reconstruction Module In this section, to better propagate high-quality features through the network, we design a precise video reconstruction module to further bridge the performance gap with the following considerations: **i)** As demonstrated in Sec. 4.2, the proposed high-quality feature extraction module can extract and then propagate high-quality features to the ResDNet module. **ii)** As illustrated in Fig. 1(c), there are already some skip connections in the ResDNet module of EfficientSCI, which are able to propagate the high-quality features to the video reconstruction module. Therefore, we want to further propagate the high-quality features to the end of the network. As the video reconstruction module is also stacked with some convolution layers, we present a similar design with Sec. 4.2. Specifically, as shown in Fig. 2(c), several $1 \times 1 \times 1$ shortcut connected convolution layers are added above the previous video reconstruction module (shown in the red box of Fig. 2(c)). Finally, we also set the bit-width of the shortcut connected convolution layers as 8-bit to ensure precise feature propagation.

4.3 Architecture and Variants

In this section, we choose EfficientSCI-S as a full-precision backbone model to customize quantized networks due to its superior efficiency performance. However, this does not mean that our proposed Q-SCI framework is equal to a low-bit quantized EfficientSCI-S, because our Q-SCI framework can also be applied on the other end-to-end video SCI reconstruction methods (See Tab. 3).

After that, we customize four quantized network variants including Q-SCI (8-bit), Q-SCI (4-bit), Q-SCI (3-bit), and Q-SCI (2-bit). The network quantization settings are as follows: **i)** In the Q-SCI (8-bit) network, we directly set the bit-width of all the network layers to 8-bit and then adopt the proposed shifted Transformer branch. **ii)** In the Q-SCI (4-bit), Q-SCI (3-bit), and Q-SCI (2-bit) network, we first set the bit-width of all the network layers to 4-bit, 3-bit, and 2-bit, respectively. Then, to ensure high-quality reconstruction results, we use our proposed high-quality feature extraction module, precise video reconstruction module and shifted Transformer branch. Note that previous work [11] also sets a part of the network layers as 8-bit to ensure high performance. Additionally, since most layers of the low-bit quantized reconstruction networks are set to be (2, 3, or 4)-bit, it can still be considered as a (2, 3, or 4)-bit quantized model.

4.4 Loss Function

Our proposed low-bit quantized network takes the measurement ($\mathbf{Y}$) and the corresponding masks ($\{\mathbf{M}_t\}_{t=1}^{T}$) as inputs, and then generates the reconstructed video frames ($\{\hat{\mathbf{X}}_t\}_{t=1}^{T} \in \mathbb{R}^{n_x \times n_y}$). To train the network, we choose the mean squared error (MSE) as our loss function,

$$\mathcal{L}_{MSE} = \frac{1}{T n_x n_y} \sum_{t=1}^{T} \|\hat{\mathbf{X}}_t - \mathbf{X}_t\|_2^2, \tag{5}$$

where $\{\mathbf{X}_t\}_{t=1}^{T}$ denotes the ground truth.

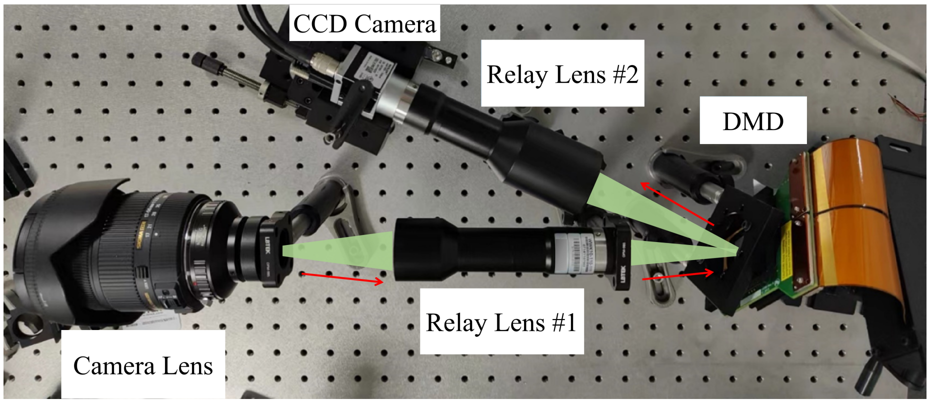


Fig. 5: Illustration of our real built video SCI system.

5 Experimental Results

Video SCI Hardware: The optical setup of the real video SCI system is shown in Fig. 5. The hardware encoding process of the video SCI system is as follows: First, the reflected light from the dynamic scene is imaged onto the surface of the digital micromirror device (DMD) via a camera lens (Sigma, 17-50/2.8, EX DC OS HSM) and the first relay lens (Coolens, WWK10-110-111). Following this, the projected high-speed scene on the DMD (TI, 2560×1600 pixels with 7.6 μm pixel pitch) is modulated by the pre-stored random binary masks. Finally, the encoded compressed measurement is projected onto the surface of a low-cost CCD camera (Basler acA1920, 1920×1200 pixels with 4.8 μm pitch pixel) with the second relay lens (Coolens, WWK066-110-110), which is captured in a single exposure time. Note that the CCD camera works at 50 FPS when capturing real data and the compression ratio of the video SCI system is Cr, then the equivalent sampling rate of our real video SCI system is 50×Cr FPS.

Datasets: Three types of dataset are used in this paper: training dataset, simulated testing dataset, and real testing dataset. **i)** First, following BIRNAT [8], we choose the `DAVIS2017` dataset [34] with spatial resolution 480×894 (480p) as the training dataset. **ii)** We then test our Q-SCI framework on six simulated testing datasets, including `Kobe`, `Traffic`, `Runner`, `Drop`, `Crash`, and `Aerial` (with spatial resolution 256×256 and compression ratio is 8). For the simulated testing datasets, average Peak Signal-to-Noise-Ratio (PSNR) and average Structured Similarity Index Metrics (SSIM) [52] are used as the evaluation metrics of reconstruction quality. Besides, following previous works [2, 32, 35], model size (Params) and number of operations (OPs) are used as the evaluation metrics of algorithm efficiency. **iii)** Finally, we test the proposed Q-SCI framework on two real testing datasets, including `Domino` and `Water Balloon` (with spatial resolution 512×512 and compression ratio is 10) captured by our real built video SCI system described in the "Video SCI Hardware" section. Moreover, please refer to the supplementary material for more details about the real data capture process.

Table 2: The average PSNR in dB (left entry), SSIM (right entry), model size and computational cost of different reconstruction algorithms on six simulated testing datasets. The best results are shown in bold and the second-best results are underlined.

Method	Kobe	Traffic	Runner	Drop	Crash	Aerial	Average	Params (M)	OPs (G)
MetaSCI [51]	30.12, 0.907	26.95, 0.888	37.02, 0.967	40.61, 0.985	27.33, 0.906	28.31, 0.904	31.72, 0.926	2.07	39.85
BIRNAT [8]	32.71, 0.950	29.33, 0.942	38.70, 0.976	42.28, 0.992	27.84, 0.927	28.99, 0.917	33.31, 0.951	4.13	390.56
RevSCI [7]	33.72, 0.957	30.02, 0.949	39.40, 0.977	42.93, 0.992	28.12, 0.937	29.35, 0.924	33.92, 0.956	5.66	766.95
STFormer-S [46]	33.19, 0.955	29.19, 0.941	39.00, 0.979	42.84, 0.992	29.26, 0.950	30.13, 0.934	33.94, 0.958	1.22	193.47
Dense3D-Unfolding [53]	35.00, 0.969	31.76, 0.096	40.03, 0.980	44.96, 0.995	29.33, 0.956	30.46, 0.943	35.26, 0.968	61.91	3975.83
ELP-Unfolding [57]	34.41, 0.966	31.58, 0.962	41,16, 0.986	44.99, 0.995	29.65, 0.959	30.68, 0.944	35.41, 0.969	565.73	4634.94
EfficientSCI-S [45]	34.79, 0.968	31.21, 0.961	41.34, 0.986	44.61, 0.994	30.34, 0.965	30.78, 0.945	35.51, **0.970**	3.78	563.87
Q-ViT (8-bit) [22]	34.60, 0.966	30.65, 0.957	40.94, 0.984	43.91, 0.993	30.14, 0.962	30.75, 0.944	35.17, 0.967	0.95	141.04
Q-SCI (2-bit) (Ours)	30.88, 0.920	26.78, 0.899	36.18, 0.957	39.04, 0.975	28.02, 0.917	28.84, 0.901	31.62, 0.928	**0.25**	**19.85**
Q-SCI (3-bit) (Ours)	32.93, 0.949	29.08, 0.938	38.62, 0.972	41.70, 0.985	29.25, 0.945	30.10, 0.929	33.62, 0.953	0.37	37.47
Q-SCI (4-bit) (Ours)	34.15, 0.962	30.44, 0.954	40.09, 0.980	43.18, 0.990	29.80, 0.956	30.49, 0.938	34.69, 0.963	0.48	72.69
Q-SCI (8-bit) (Ours)	34.95, 0.968	31.24, 0.961	41.60, 0.985	44.27, 0.993	30.34, 0.963	31.03, 0.945	**35.57**, 0.969	0.95	140.95

Training Process: We build our models with PyTorch, and conduct training on 4 NVIDIA RTX 3090 GPUs. First, regular data augmentation operations such as random cropping, random flipping and random scaling are performed on the training dataset. Then, we initialize our Q-SCI with the full-precision EfficientSCI-S and adopt the Adam [17] optimizer to optimize the model with an initial learning rate of 0.0001. After iterating for 100 epochs on the training data with a 128×128 spatial size and 20 epochs on the training data with a 256×256 spatial size, we adjust the learning rate to 0.00001 and continue to iterate for 20 epochs on the training data with a 256×256 spatial size.

5.1 Results on Simulated Testing Datasets

In this section, we compare our Q-SCI networks with seven deep learning-based video SCI reconstruction methods (including MetaSCI, BIRNAT, RevSCI, Dense3D-Unfolding, ELP-Unfolding, STFormer-S, and EfficientSCI-S) on six simulated testing datasets. As shown in Tab. 2, the proposed Q-SCI networks can achieve comparable reconstruction quality with *much fewer parameters and cheaper computational cost*. Specifically, **i)** The proposed Q-SCI (2-bit) model achieves comparable reconstruction quality with MetaSCI, while the OPs is reduced by about 2×. **ii)** Our proposed Q-SCI (3-bit) model achieves 0.31 dB in PSNR higher than BIRNAT, and the OPs is reduced by about 10.4×. **iii)** The proposed Q-SCI (4-bit) model achieves 1.38 dB and 0.77 dB in PSNR higher than BIRNAT and RevSCI, and the OPs is reduced by about 5.4× and 10.6×. **iv)** Our Q-SCI (8-bit) model can achieve 0.31 dB and 0.16 dB in PSNR higher than Dense3D-Unfolding and ELP-Unfolding, and the OPs is reduced by about 28.2× and 32.9×. **v)** Our proposed Q-SCI (8-bit) model achieves comparable performance with EfficientSCI-S, while the OPs is reduced by about 4×.

For visualization purposes, we also visualize some reconstructed video frames in Fig. 6, where we can see from the zooming areas in each selected video frame that our proposed Q-SCI (4-bit) model can provide much clearer reconstructed images than BIRNAT, RevSCI, and STFormer-S. Further, the proposed Q-SCI (8-bit) model can provide comparable high-quality reconstructed images with

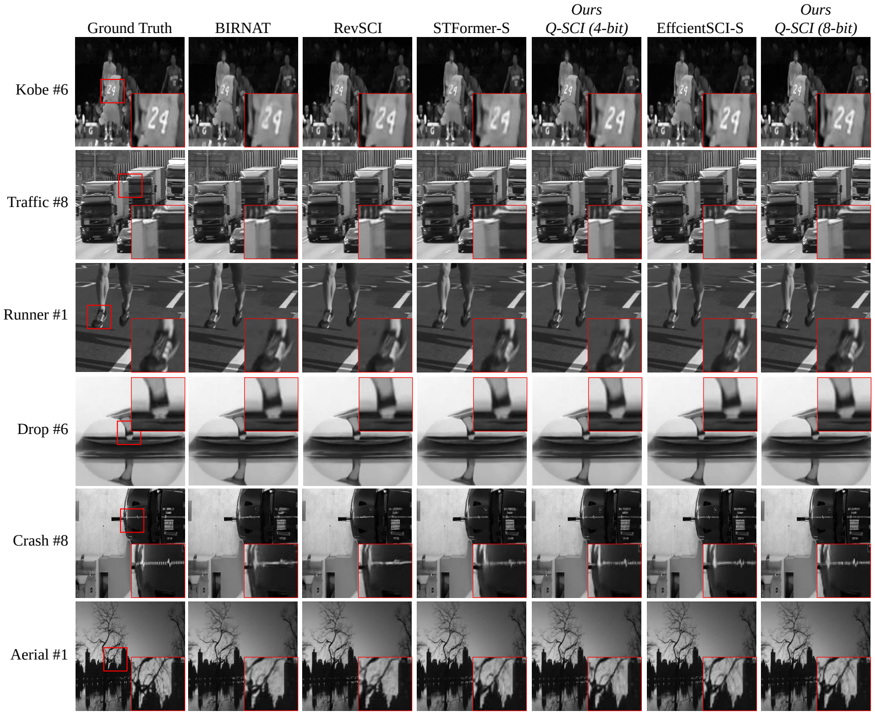


Fig. 6: Reconstructed video frames of the simulated testing datasets. For a better view, we zoom in on a local area as shown in the small red boxes of each ground truth image, and do not show the small red boxes again for simplicity.

EfficientSCI-S. Especially for the `Traffic`, `Crash`, and `Aerial` scenes, we can observe sharp edges and more details in the reconstructed frames of our proposed models, which verify the superior performance of the proposed Q-SCI quantization framework.

Moreover, comparison with previous SOTA quantization method Q-ViT is also conducted. Note that, we adapt both Q-ViT and Q-SCI into EfficientSCI-S for a fair comparison. The computational complexity and reconstruction quality are given in Tab. 2. We can see from Tab. 2 that Q-SCI (8-bit) outperforms Q-ViT (8-bit) by about 0.4 dB with less computational cost.

Finally, we adapt the proposed Q-SCI framework into previous SOTA end-to-end reconstruction method STFormer-S. As shown in Tab. 3, compared with 4-bit quantized baseline model, adapting our proposed high-quality feature extraction module will bring a 3.23 dB reconstruction quality improvement for STFormer-S. Then, further adapting the proposed precise video reconstruction module can lead to a 0.25 dB reconstruction quality improvement for STFormer-S. Therefore, the proposed Q-SCI framework can *well generalize* to other end-to-end video SCI reconstruction methods.

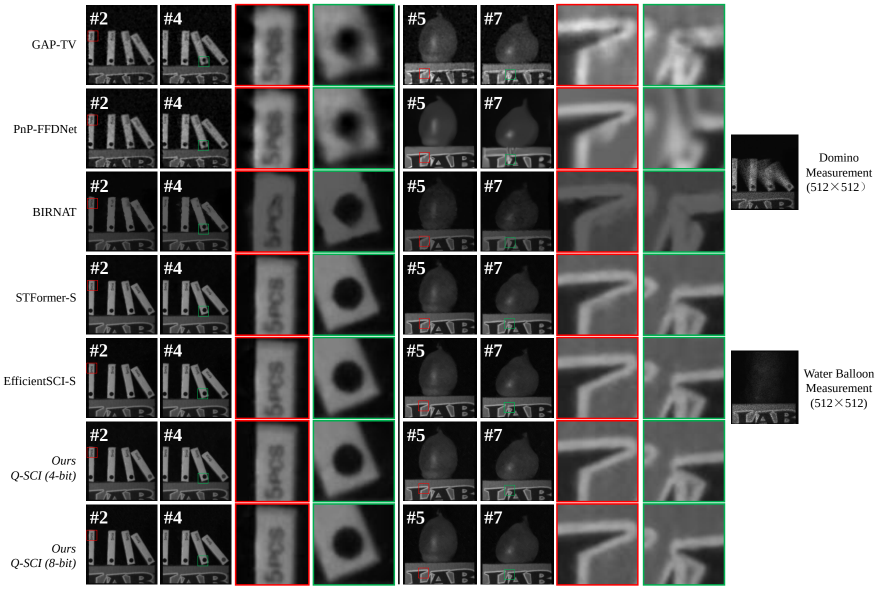


Fig. 7: Reconstructed video frames of the real data. See Supplementary Movie 1 for the complete video.

5.2 Results on Real Testing Datasets

In this section, we test the proposed Q-SCI (8-bit) and Q-SCI (4-bit) models on the real testing datasets. Due to the uncertain noises of the real video SCI system, it is more difficult to reconstruct real measurements. As shown in Fig. 7, our proposed Q-SCI (4-bit) model can reconstruct clearer borders than previous real-valued models including GAP-TV, PnP-FFDNet, and BIRNAT. Moreover, we can clearly recognize the letters on the `Domino` on the reconstructed frames of the proposed Q-SCI (4-bit) and Q-SCI (8-bit). Finally, our proposed Q-SCI (8-bit) model can present sharper boarders and easier-recognized letters on the `Domino` when compared with the real-valued STFormer-S and EfficientSCI-S.

5.3 Ablation Study

Here, we conduct a break-down ablation towards higher performance. We first build a baseline method by directly quantizing each layer of EfficientSCI-S into 4-bit. As shown in Tab. 4, the baseline model yields 31.40 dB in PSNR and 0.931 in SSIM. When we adopt the shifted Transformer branch, the model achieves a 0.53 dB improvement (which is a moderately large improvement in the video SCI reconstruction task) with neglect computational burden. After that, we apply the high-quality feature extraction module and the precise video reconstruction module successively. The performance improvement are 2.35 dB and 0.43 dB, respectively. Finally, we obtain the Q-SCI (4-bit) model with 34.71 dB which is

Table 3: Verification on the **generalization ability** of our proposed Q-SCI framework, where "FEM" and "VRM" stand for adapting the proposed high-quality feature extraction and precise video reconstruction modules into 4-bit baseline model of STFormer-S.

Method	PSNR	SSIM
4-bit Baseline	30.03	0.903
4-bit Baseline+FEM	33.26	0.949
4-bit Baseline+FEM+VRM	33.51	0.952

Table 4: Break-down ablation study towards higher performance, where "RDM", "FEM", and "VRM" represent using our shifted Transformer branch, high-quality feature extraction module, and precise video reconstruction module into 4-bit baseline model of EfficientSCI-S.

Method	PSNR	SSIM	Params(M)	OPs(G)
4-bit Baseline	31.40	0.931	0.473	70.477
+RDM	31.93	0.929	0.473	70.477
+RDM+FEM	34.28	0.959	0.482	71.857
+RDM+FEM+VRM	34.71	0.963	0.483	72.692

a totally 3.31 dB improvement against the 4-bit quantized baseline model, while the computational cost only grows by 3.14%. These experimental results verify the effectiveness of the proposed module designs of our Q-SCI framework.

6 Conclusion and Future Work

In this paper, we propose the first low-bit quantization framework for the end-to-end video SCI reconstruction methods, which is *simple and effective*. Specifically, we design a high-quality feature extraction module and a precise video reconstruction module to extract and propagate high-quality features in the low-bit quantized model. Additionally, we introduce a shift operation on the query and key distributions of the low-bit quantized Transformer branch to further bridge the performance gap. Finally, we verify that our proposed low-bit quantization framework can generalize well to different end-to-end reconstruction methods (including EfficientSCI and STFormer).

While our proposed Q-SCI can largely reduce computational cost with small performance drop, there is still room for further improvement, which we discuss as follows. First of all, deploying Q-SCI on the resource-limited devices (*i.e.*, smart phone and autonomous vehicle) requires extensive engineering and hardware support, making it a challenging task. Secondly, although Q-SCI is a well-designed network quantization framework for video SCI, a novel mobile-friendly reconstruction network will be more suitable than previous reconstruction methods for the quantized video SCI reconstruction task. Finally, algorithm-hardware codesign should be considered to better fit the resource-limited devices and the video SCI hardware encoding system.

Regarding the future work, we can further explore other model compression methods including network pruning [25, 40] and knowledge distillation [18, 42] to further optimize the efficiency performance. Moreover, we expect to extend the proposed Q-SCI framework to various imaging systems such as high-speed imaging [10, 31, 70], hyperspectral imaging [13, 26, 33], depth of field imaging [4, 38, 65], and single-pixel imaging [36, 50, 56].

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant number 62271414), Zhejiang Provincial Distinguished Young Scientist Foundation (grant number LR23F010001), Zhejiang "Pioneer" and "Leading Goose" R&D Program (grant number 2024SDXHDX0006, 2024C03182), the Key Project of Westlake Institute for Optoelectronics (grant number 2023GD007), and the 2023 International Sci-tech Cooperation Projects under the purview of the "Innovation Yongjiang 2035" Key R&D Program (grant number 2024Z126).

References

1. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013)
2. Cai, Y., Zheng, Y., Lin, J., Yuan, X., Zhang, Y., Wang, H.: Binarized spectral compressive imaging. In: NeurIPS (2023)
3. Cao, M., Wang, L., Zhu, M., Yuan, X.: Hybrid cnn-transformer architecture for efficient large-scale video snapshot compressive imaging. IJCV pp. 1–20 (2024)
4. Chen, M.K., Liu, X., Wu, Y., Zhang, J., Yuan, J., Zhang, Z., Tsai, D.P.: A metadevice for intelligent depth perception. Advanced Materials **35**(34), 2107465 (2023)
5. Chen, S., Wang, W., Pan, S.J.: Deep neural network quantization via layer-wise optimization using limited training data. In: AAAI (2019)
6. Chen, Y., Dai, X., Chen, D., Liu, M., Dong, X., Yuan, L., Liu, Z.: Mobile-former: Bridging mobilenet and transformer. In: CVPR (2022)
7. Cheng, Z., Chen, B., Liu, G., Zhang, H., Lu, R., Wang, Z., Yuan, X.: Memory-efficient network for large-scale video compressive sensing. In: CVPR (2021)
8. Cheng, Z., Chen, B., Lu, R., Wang, Z., Zhang, H., Meng, Z., Yuan, X.: Recurrent neural networks for snapshot compressive imaging. TPAMI **45**(2), 2264–2281 (2022)
9. Deng, C., Zhang, Y., Mao, Y., Fan, J., Suo, J., Zhang, Z., Dai, Q.: Sinusoidal sampling enhanced compressive camera for high speed imaging. TPAMI **43**(4), 1380–1393 (2019)
10. Dou, Y., Cao, M., Wang, X., Liu, X., Yuan, X.: Coded aperture temporal compressive digital holographic microscopy. Optics Letters **48**(20), 5427–5430 (2023)
11. Esser, S.K., McKinstry, J.L., Bablani, D., Appuswamy, R., Modha, D.S.: Learned step size quantization. arXiv preprint arXiv:1902.08153 (2019)
12. Fang, G., Ma, X., Song, M., Mi, M.B., Wang, X.: Depgraph: Towards any structural pruning. In: CVPR (2023)
13. Fang, J., Huang, K., Qin, R., Liang, Y., Wu, E., Yan, M., Zeng, H.: Wide-field mid-infrared hyperspectral imaging beyond video rate. Nature Communications **15**(1), 1811 (2024)
14. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. IJCV **129**(6), 1789–1819 (2021)
15. Hitomi, Y., Gu, J., Gupta, M., Mitsunaga, T., Nayar, S.K.: Video from a single coded exposure photograph using a learned over-complete dictionary. In: ICCV (2011)
16. Hong, C., Baik, S., Kim, H., Nah, S., Lee, K.M.: Cadyq: Content-aware dynamic quantization for image super-resolution. In: ECCV (2022)

17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
18. Kordopatis-Zilos, G., Tzelepis, C., Papadopoulos, S., Kompatsiaris, I., Patras, I.: Dns: Distill-and-select for efficient and accurate video indexing and retrieval. IJCV **130**(10), 2385–2407 (2022)
19. Li, H., Yan, C., Lin, S., Zheng, X., Zhang, B., Yang, F., Ji, R.: Pams: Quantized super-resolution via parameterized max scale. In: ECCV (2020)
20. Li, R., Wang, Y., Liang, F., Qin, H., Yan, J., Fan, R.: Fully quantized network for object detection. In: CVPR (2019)
21. Li, X., Liu, Y., Lian, L., Yang, H., Dong, Z., Kang, D., Zhang, S., Keutzer, K.: Q-diffusion: Quantizing diffusion models. In: ICCV (2023)
22. Li, Y., Xu, S., Zhang, B., Cao, X., Gao, P., Guo, G.: Q-vit: Accurate and fully quantized low-bit vision transformer. In: NeurIPS (2022)
23. Li, Y., Hu, J., Wen, Y., Evangelidis, G., Salahi, K., Wang, Y., Tulyakov, S., Ren, J.: Rethinking vision transformers for mobilenet size and speed. In: ICCV (2023)
24. Li, Y., Yuan, G., Wen, Y., Hu, J., Evangelidis, G., Tulyakov, S., Wang, Y., Ren, J.: Efficientformer: Vision transformers at mobilenet speed. In: NeurIPS (2022)
25. Li, Y., Adamczewski, K., Li, W., Gu, S., Timofte, R., Van Gool, L.: Revisiting random channel pruning for neural network compression. In: CVPR (2022)
26. Lin, C.H., Huang, S.H., Lin, T.H., Wu, P.C.: Metasurface-empowered snapshot hyperspectral imaging with convex/deep (code) small-data learning theory. Nature Communications **14**(1), 6979 (2023)
27. Liu, Y., Yuan, X., Suo, J., Brady, D.J., Dai, Q.: Rank minimization for snapshot compressive imaging. TPAMI **41**(12), 2990–3006 (2018)
28. Liu, Z., Cheng, K.T., Huang, D., Xing, E.P., Shen, Z.: Nonuniform-to-uniform quantization: Towards accurate quantization via generalized straight-through estimation. In: CVPR (2022)
29. Liu, Z., Shen, Z., Savvides, M., Cheng, K.T.: Reactnet: Towards precise binary neural network with generalized activation functions. In: ECCV (2020)
30. Liu, Z., Wang, Y., Han, K., Ma, S., Gao, W.: Instance-aware dynamic neural network quantization. In: CVPR (2022)
31. Luo, R., Cao, M., Liu, X., Yuan, X.: Snapshot compressive structured illumination microscopy. Optics Letters **49**(2), 186–189 (2024)
32. Ma, Y., Jin, T., Zheng, X., Wang, Y., Li, H., Wu, Y., Jiang, G., Zhang, W., Ji, R.: Ompq: Orthogonal mixed precision quantization. In: AAAI (2023)
33. Meng, Z., Ma, J., Yuan, X.: End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In: ECCV (2020)
34. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
35. Qin, H., Zhang, Y., Ding, Y., Liu, X., Danelljan, M., Yu, F., et al.: Quantsr: Accurate low-bit quantization for efficient image super-resolution. In: NeurIPS (2024)
36. Qu, G., Wang, P., Yuan, X.: Dual-scale transformer for large-scale single-pixel imaging. In: CVPR (2024)
37. Reddy, D., Veeraraghavan, A., Chellappa, R.: P2c2: Programmable pixel compressive camera for high speed imaging. In: CVPR (2011)
38. Shen, Z., Zhao, F., Jin, C., Wang, S., Cao, L., Yang, Y.: Monocular metasurface camera for passive single-shot 4d imaging. Nature Communications **14**(1), 1035 (2023)
39. Wang, H., Chen, P., Zhuang, B., Shen, C.: Fully quantized image super-resolution networks. In: ACM MM (2021)

40. Wang, H., Fu, Y.: Trainability preserving neural pruning. In: ICLR (2022)
41. Wang, H., Li, Y., Wang, Y., Hu, H., Yang, M.H.: Collaborative distillation for ultra-resolution universal style transfer. In: CVPR (2020)
42. Wang, H., Lohit, S., Jones, M.N., Fu, Y.: What makes a" good" data augmentation in knowledge distillation-a statistical perspective. In: NeurIPS (2022)
43. Wang, H., Qin, C., Zhang, Y., Fu, Y.: Neural pruning via growing regularization. In: ICLR (2021)
44. Wang, H., Ren, J., Huang, Z., Olszewski, K., Chai, M., Fu, Y., Tulyakov, S.: R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In: ECCV (2022)
45. Wang, L., Cao, M., Yuan, X.: Efficientsci: Densely connected network with space-time factorization for large-scale video snapshot compressive imaging. In: CVPR (2023)
46. Wang, L., Cao, M., Zhong, Y., Yuan, X.: Spatial-temporal transformer for video snapshot compressive imaging. TPAMI **45**(7), 9072–9089 (2022)
47. Wang, L., Wu, Z., Zhong, Y., Yuan, X.: Snapshot spectral compressive imaging reconstruction using convolution and contextual transformer. Photonics Research **10**(8), 1848–1858 (2022)
48. Wang, L., Dong, X., Wang, Y., Liu, L., An, W., Guo, Y.: Learnable lookup table for neural network quantization. In: CVPR (2022)
49. Wang, P., Chen, Q., He, X., Cheng, J.: Towards accurate post-training network quantization via bit-split and stitching. In: ICML (2020)
50. Wang, Y., Huang, K., Fang, J., Yan, M., Wu, E., Zeng, H.: Mid-infrared single-pixel imaging at the single-photon level. Nature Communications **14**(1), 1073 (2023)
51. Wang, Z., Zhang, H., Cheng, Z., Chen, B., Yuan, X.: Metasci: Scalable and adaptive reconstruction for video compressive sensing. In: CVPR (2021)
52. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP **13**(4), 600–612 (2004)
53. Wu, Z., Zhang, J., Mou, C.: Dense deep unfolding network with 3d-cnn prior for snapshot compressive imaging. In: ICCV (2021)
54. Xia, B., Zhang, Y., Wang, Y., Tian, Y., Yang, W., Timofte, R., Van Gool, L.: Basic binary convolution unit for binarized image restoration network. In: ICLR (2022)
55. Xu, S., Li, H., Zhuang, B., Liu, J., Cao, J., Liang, C., Tan, M.: Generative low-bitwidth data free quantization. In: ECCV (2020)
56. Xu, Y., Lu, L., Saragadam, V., Kelly, K.F.: A compressive hyperspectral video imaging system using a single-pixel detector. Nature Communications **15**(1), 1456 (2024)
57. Yang, C., Zhang, S., Yuan, X.: Ensemble learning priors unfolding for scalable snapshot compressive sensing. In: ECCV (2022)
58. Yang, J., Liao, X., Yuan, X., Llull, P., Brady, D.J., Sapiro, G., Carin, L.: Compressive sensing by learning a gaussian mixture model from measurements. TIP **24**(1), 106–119 (2014)
59. Yao, Z., Yazdani Aminabadi, R., Zhang, M., Wu, X., Li, C., He, Y.: Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. In: NeurIPS (2022)
60. Yu, C., Chen, T., Gan, Z., Fan, J.: Boost vision transformer with gpu-friendly sparsity and quantization. In: CVPR (2023)
61. Yuan, X.: Generalized alternating projection based total variation minimization for compressive sensing. In: ICIP (2016)

62. Yuan, X., Brady, D.J., Katsaggelos, A.K.: Snapshot compressive imaging: Theory, algorithms, and applications. IEEE Signal Processing Magazine **38**(2), 65–88 (2021)
63. Yuan, X., Liu, Y., Suo, J., Dai, Q.: Plug-and-play algorithms for large-scale snapshot compressive imaging. In: CVPR (2020)
64. Yuan, X., Liu, Y., Suo, J., Durand, F., Dai, Q.: Plug-and-play algorithms for video snapshot compressive imaging. TPAMI **44**(10), 7093–7111 (2021)
65. Yuan, X., Llull, P., Liao, X., Yang, J., Brady, D.J., Sapiro, G., Carin, L.: Low-cost compressive sensing for color video and depth. In: CVPR (2014)
66. Yuan, Z., Xue, C., Chen, Y., Wu, Q., Sun, G.: Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In: ECCV (2022)
67. Zhang, Z., Deng, C., Liu, Y., Yuan, X., Suo, J., Dai, Q.: Ten-mega-pixel snapshot compressive imaging with a hybrid coded aperture. Photonics Research **9**(11), 2277–2287 (2021)
68. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: CVPR (2022)
69. Zheng, S., Yuan, X.: Unfolding framework with prior of convolution-transformer mixture and uncertainty estimation for video snapshot compressive imaging. In: ICCV (2023)
70. Zheng, Y., Zheng, L., Yu, Z., Huang, T., Wang, S.: Capture the moment: High-speed imaging with spiking cameras through short-term plasticity. TPAMI **45**(7), 8127–8142 (2023)