Dual-Scale Transformer for Large-Scale Single-Pixel Imaging

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

Abstract

Single-pixel imaging (SPI) is a potential computational imaging technique which produces image by solving an ill-posed reconstruction problem from few measurements captured by a single-pixel detector. Deep learning has achieved impressive success on SPI reconstruction. However, previous poor reconstruction performance and impractical imaging model limit its real-world applications. In this paper, we propose a deep unfolding network with hybrid-attention Transformer on Kronecker SPI model, dubbed HATNet, to improve the imaging quality of real SPI cameras. Specifically, we unfold the computation graph of the iterative shrinkage-thresholding algorithm (ISTA) into two alternative modules: efficient tensor gradient descent and hybrid-attention multi-scale denoising. By virtue of Kronecker SPI, the gradient descent module can avoid high computational overheads rooted in previous gradient descent modules based on vectorized SPI. The denoising module is an encoder-decoder architecture powered by dual-scale spatial attention for high- and low-frequency aggregation and channel attention for global information recalibration. Moreover, we build a SPI prototype to verify the effectiveness of the proposed method. Extensive experiments on synthetic and real data demonstrate that our method achieves the state-of-the-art performance. The source code and pre-trained models are available at https://github.com/Gang-Qu/HATNet-SPI.

1. Introduction

Conventional imaging technology produces images by exploiting the light reflected or scattered by an object on a two-dimensional (2D) CCD or CMOS detector with millions of pixels. But in applications, such as the infrared or deep ultraviolet sensing, the availability of pixelated array detectors becomes expensive or impractical. As an alternative solution, single-pixel imaging (SPI) utilizes just one light-sensitive single-pixel detector (SPD) to record the total intensity of the reflected or scattered light encoded by temporally varying modulation patterns from a spatial light modulator, yielding compressed measurements, and the desired image can be estimated from the captured (compressed) measurements via iterative optimization algorithms or a deep learning model. SPI camera offers advantages over conventional cameras, such as improved detection efficiency, lower dark counts, and faster timing response. Such advantages can have significance in scenarios where the detected intensities are very weak due to scattering or absorption losses. Moreover, SPI camera is capable of sensing compressively during data acquisition, thereby reducing the data storage and communication bandwidth requirements. During the last decade, SPI has been widely used in 3D imaging [45], hyperspectral imaging [56], X-ray diffraction tomography [15], magnetic resonance imaging [16], ophthalmic imaging [27] and imaging in non-visible wavebands [18] or through scattering media [73].

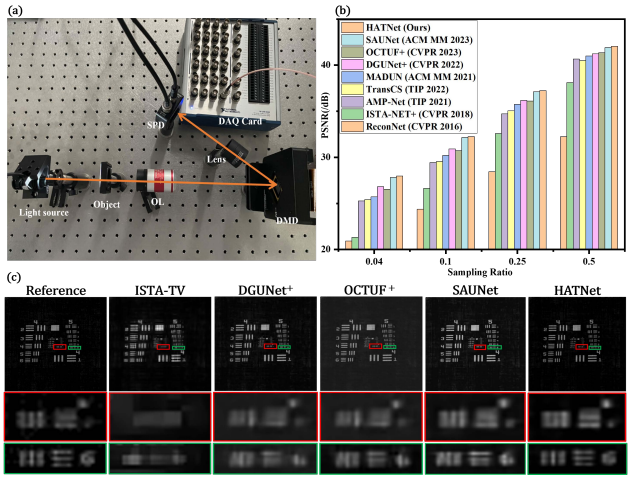

Figure 1. (a) Our built SPI prototype (OL: objective lens, DMD: digital micromirror device, DAQ card: data acquisition card). (b) Performance comparison of different methods on Set11 dataset at different sampling ratios. (c) Real experimental results of different methods at sampling ratio of 25%.

SPI is driven from the compressive sensing (CS) [5, 10, 52, 53, 63–65] theory. In CS paradigm, a 1D signal $\boldsymbol{x} \in \mathbb{R}^N$ is compressively sampled into few measurements $\boldsymbol{y} \in \mathbb{R}^M$

*Equal contribution.
†Corresponding author.

at a sub-Nyquist sampling ratio $\frac{M}{N}$ ($M \ll N$) via a linear system:

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{x}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{M \times N}$ is the measurement matrix. Due to the ill-posed nature of the inverse process of Eq. (1), an estimate $\hat{\boldsymbol{x}}$ of $\boldsymbol{x}$ could be reconstructed from $\boldsymbol{y}$ by solving the following optimization problem:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|_2^2 + \lambda\mathcal{R}(\boldsymbol{x}), \tag{2}$$

where $\frac{1}{2}\|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|_2^2$ is a data fidelity term, and $\lambda\mathcal{R}(\boldsymbol{x})$ is a regularization term. Over the past years, various optimization algorithms have been developed to solve it, among which iterative shrinkage-thresholding algorithm (ISTA) [8] is the most widely used one. ISTA is composed of two alternative operations:

$$\begin{cases} \boldsymbol{z}_k = \boldsymbol{x}_{k-1} + \rho\mathbf{A}^\top(\boldsymbol{y} - \mathbf{A}\boldsymbol{x}_{k-1}), & (3) \\ \boldsymbol{x}_k = \arg\min_{\boldsymbol{x}} \dfrac{1}{2\sigma^2}||\boldsymbol{z}_k - \boldsymbol{x}||_2^2 + \mathcal{R}(\boldsymbol{x}), & (4) \end{cases}$$

where $\rho = (1+\eta)^{-1}$ and $\sigma = \sqrt{\lambda/\eta}$ both with a penalty parameter η. Eq. (3) is a gradient descent process and Eq. (4) is equivalent to denoising image $\boldsymbol{z}_k$ with the regularization term $\mathcal{R}(\boldsymbol{x})$. Regarding $\mathcal{R}(\boldsymbol{x})$, various hand-crafted image priors have been proposed to regularize the solution in the desired signal space, such as sparsity [36, 67], total variation [62], low rank [25], and non-local self-similarity [68]. Unfortunately, hundreds and thousands of iterations lead to just passable results, thereby making it impractical for real-time and high-fidelity scenarios. Over the past few years, deep neural networks have recently gained considerable popularity in solving the inverse problem of CS, from early black-box networks [22, 28, 41, 59] to recent deep unfolding networks (DUNs) [31, 34, 39, 42, 69, 71, 72]. Early black-box networks usually learn a non-linear mapping from sampled measurements to the final reconstructed result in an end-to-end manner, with limited performance and without interpretability. By combining an optimization algorithm with a deep denoising network, DUNs enjoy the interpretability of optimization algorithms and the powerful modeling ability of deep neural networks, leading to the-state-of-art (SOTA) performance. These deep models, both black-box networks and DUNs, are developed as CS solvers.

However, there is a significant mismatch between real SPI system and CS-oriented solvers. Regular CS model in Eqs. (1) to (4) is established for vectored 1D signal but SPI cameras take aim at 2D image. Before this work, 2D image to be detected is considered as the vectorized signal when a CS-oriented solver is employed, leading a huge measurement matrix $\mathbf{A}$ and thus extremely high computational cost in Eq. (3). For example, a 512×512 image needs a measurement matrix $\mathbf{A}$ with $262,144$ columns, which is terribly large. To address this problem, most previous DUNs [34, 39, 42, 69, 71, 72] divide the whole image into several small patches to process, thus also called block-CS-oriented solvers. However, block-based sampling is impractical for mainstream SPI cameras. This is why these DUNs are outstanding on simulation metrics but are rarely deployed on real-world SPI cameras. Recently, the first practical DUN [51] has been proposed to enable full image CS reconstruction but also overlooks physical constraints in imaging.

To bridge the gap between SPI and DUNs, we propose a deep unfolding network with hybrid-attention Transformer, dubbed HATNet, on Kronecker SPI [11] by unrolling the computation graph of ISTA into two alternative modules: efficient gradient descent and HAT-based denoising. The main contributions of this work are summarized as follows.

1) To exploit global interactions of images and satisfy physical constraints of real SPI cameras, we introduce a DUN of tensor ISTA, composed of tensor gradient descent module and deep denoising module, to enable full-size sampling and reconstruction for SPI. By avoiding vectorized huge measurement matrix, the forward model of SPI and the gradient descent of DUN are significantly accelerated.
2) We propose a DUN with hybrid-attention Transformer, dubbed HATNet, powered by spatial-wise dual-scale self-attention and channel-wise self-attention. HAT is capable of aggregating high- and low-frequency information and recalibrating channel-wise global information.
3) We use HAT under an encoder-decoder architecture as the deep denoiser of deep unfolding and it achieve SOTA performance on synthetic data as reported in Fig. 1 (b). Moreover, we also verify the effectiveness of proposed method on real data as demonstrated in Fig. 1 (c), which is captured by our SPI prototype in Fig. 1 (a). To our best knowledge, we are the first to develop a SOTA deep model to improve practical SPI, particularly for large scale.

2. Related Work

2.1. Compressive Sensing Reconstruction

CS reconstruction methods could be classified into two categories: optimization-based methods [3, 4, 9, 13, 17, 21, 29, 33, 57, 70] and learning-based methods [22, 32, 34, 39–42, 44, 58, 59, 69, 71, 72]. Optimization-based methods mainly employ an iterative optimization algorithm along with hand-crafted image priors to increasingly retrieve the visual information from the sub-sampled measurement. Various iterative optimization algorithms have been proposed, including iterative shrinkage-thresholding algorithm (ISTA) [8], approximate message passing (AMP) algorithm [74], alternating direction method of multiplies (ADMM) [14], generalized alternating projection (GAP) [62] method, least absolute shrinkage and selection operator (LASSO) [47]. TVAL3 [23] utilizes the augmented Lagrangian method with total

variation (TV) prior to remove the noise and restore the details. However, optimization-based methods need hundreds and thousands of iterations and usually have a long processing time and limited reconstruction quality. In recent years, deep neural networks have been developed as powerful CS solvers and have achieved impressive success. Early networks [22, 32] usually learn a black-box mapping from the compressed measurements to the restored image. Most recently, deep unfolding networks [31, 34, 39, 69, 71, 72] use a deep denoising network to replace the proximal mapping and maintain the gradient descent of a conventional optimization algorithms, which achieve SOTA performance after few iterations. With good performance and interpretability, DUNs have become the mainstream choice for CS reconstruction. Such regime was originally applied in plug-and-play (PnP) methods, where pre-trained denoiser is employed to implicitly express the regularization term as a denoising problem [30]. Different iterative optimization algorithms foster kinds of DUNs, such as ADMM-Net [72], ISTA-Net [69], AMP-Net [72]. However, most of previous DUNs are troubled with the information loss rooted in the frequent signal-to-feature transformation. This problem indicates that early-stage high-level features cannot be efficiently used for later-stage feature refinement. Latest DUNs [42–44, 51] try to solve this problem by heuristic cross-stage information fusion designs. In addition, most of previous DUNs are developed under the block-based sampling assumption, which is impractical for real SPI.

2.2. Vision Transformer

Motivated by the power of Transformer [48] in natural language processing (NLP), ViT [2] first extends Transformer into vision tasks by conducting self-attention (SA) mechanism on non-overlapping patches. Swin Transformer [26] proposes a pioneering SA within shifted windows under a hierarchical architecture to achieve significant improvement over convolutional neural networks (CNNs) on kinds of vision tasks. Due to the remarkable performance of SA, researchers are extending Transformer into low-level vision tasks [7, 12, 35, 55, 60, 66]. PIT first introduces Transformer to image restoration and showcases its performance on several image restoration tasks [7]. Uformer [55] combines Transformer and U-Net to build multi-scale Transformer to further improve the performance. TransGAN, a mixture of generative adversarial network (GAN) and Transformer, is proposed in [20] for image generation. Restormer [66] operates self-attention along channel dimension for high-resolution image restoration. These Transformer-based methods remarkably outperform CNN-based methods and also reveal that the attention mechanism on both spatial and channel dimensions are significant for most vision tasks.

3. Proposed Method

3.1. Tensor ISTA Unfolding Framework

In SPI paradigm, assume $\mathbf{X} \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}}$ is a 2D image, and its 2D compressed measurements $\mathbf{Y} \in \mathbb{R}^{\sqrt{M}\times\sqrt{M}}$ can be obtained by a linear measurement system:

$$\mathbf{Y} = \mathbf{\Phi}\mathbf{X}\mathbf{\Psi}^{\top}. \tag{5}$$

where $\mathbf{\Phi} \in \mathbb{R}^{\sqrt{M}\times\sqrt{N}}$, $\mathbf{\Psi} \in \mathbb{R}^{\sqrt{M}\times\sqrt{N}}$ are two independent measurement matrices, simultaneously compressing image along horizontal and vertical dimensions. In real imaging systems, such dual modulation is impossible to implement. Eq. (5) is equivalent to the vectorized CS form in Eq. (1) through the Kronecker product [11]:

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{x}, \quad s.t. \quad \begin{cases} \boldsymbol{x} = \mathrm{vec}\left(\mathbf{X}\right), \\ \boldsymbol{y} = \mathrm{vec}\left(\mathbf{Y}\right), \\ \mathbf{A} = \mathbf{\Psi} \otimes \mathbf{\Phi}, \end{cases} \tag{6}$$

where $\mathrm{vec}(\cdot)$ denotes the vectorization operation and $\otimes$ represents the Kronecker product. The ill-posed inverse process of Eq. (5) can be conducted by solving the following optimization problem:

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \frac{1}{2}\left\|\mathbf{Y} - \mathbf{\Phi}\mathbf{X}\mathbf{\Psi}^{\top}\right\|_F^2 + \lambda\mathcal{R}(\mathbf{X}), \tag{7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The above optimization problem can be solved by the tensor version of ISTA [51], namely

$$\begin{cases} \mathbf{Z}_k = \mathbf{X}_{k-1} + \rho\mathbf{\Phi}^{\top}(\mathbf{Y} - \mathbf{\Phi}\mathbf{X}_{k-1}\mathbf{\Psi}^{\top})\mathbf{\Psi}, & (8) \\ \mathbf{X}_k = \arg\min_{\mathbf{X}} \frac{1}{2\sigma^2}||\mathbf{Z}_k - \mathbf{X}||_F^2 + \mathcal{R}(\mathbf{X}), & (9) \end{cases}$$

Eq. (8) is a tensor gradient descent (TGD) with a step size ρ. Eq. (9) is a proximal mapping, usually can be seen as a denoising problem [6] with the noise level σ from the perspective of Bayesian probability. By alternatively repeating the above two steps enough times, a decent estimate $\hat{\mathbf{X}}$ would well approach to the ground truth $\mathbf{X}$. In this manner, image reconstruction task in Eq. (7) is converted to a multi-stage image denoising task, which has extensively studied in low-level vision [24, 26]. Recently, learning-based denosiers have shown great performance gains over conventional optimization-based denosiers.

By unfolding the computation graph of tensor ISTA into deep neural network, we propose a deep unfolding network with hybrid-attention Transformer (HATNet), which can be formulated as

$$\begin{cases} \mathbf{Z}_k = \mathbf{X}_{k-1} + \rho_{k-1}\mathbf{\Phi}^{\top}(\mathbf{Y} - \mathbf{\Phi}\mathbf{X}_{k-1}\mathbf{\Psi}^{\top})\mathbf{\Psi}, & (10) \\ \mathbf{X}_k = \mathcal{D}_{(\theta,k)}(\mathbf{Z}_k), & (11) \end{cases}$$

where ρ_{k-1} is a learnable step size controlling the intensity of k-th gradient descent and $\mathcal{D}_{(\theta,k)}$ is a stage-specific deep

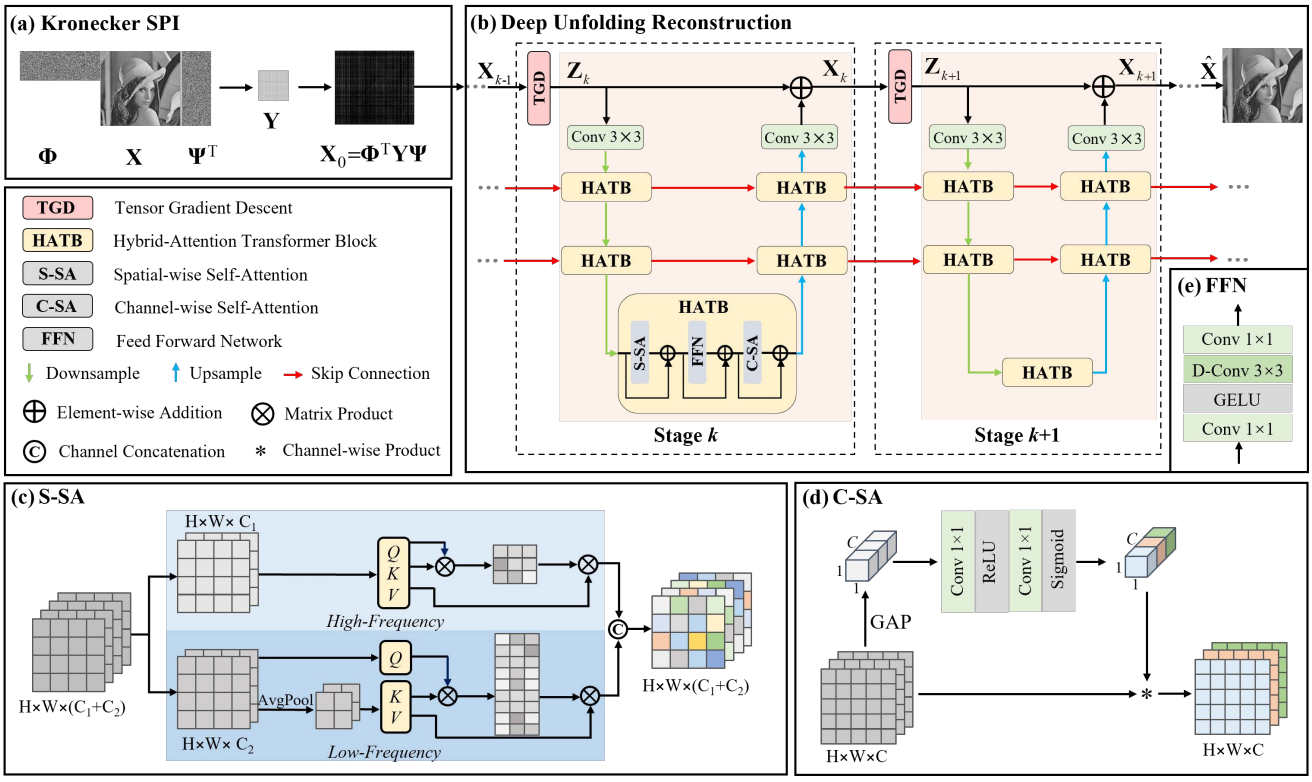

Figure 2. Illustration of the proposed method. (a) demonstrates the Kronecker SPI model. As shown in (b), our DUN aims to reconstruct a high-fidelity image $\hat{\mathbf{X}}$ from the initialization input $\mathbf{X}_0$, which is composed of multiple stages with skip connections and each stage involves a tensor gradient descent (TGD) operator in Eq. (10) and a U-shaped deep denoiser as Eq. (11). The deep denoiser is powered by the proposed HATB, each of which consists of residual dual-scale spatial-wise self-attention (S-SA), feed-forward network (FFN), and channel-wise self-attention (C-SA). The structure of S-SA and C-SA are shown in (c) and (d), respectively.

denoiser with learnable parameters θ. The initialization input is $\mathbf{X}_0 = \mathbf{\Phi}^\top \mathbf{Y} \mathbf{\Psi}$. Regarding the design of $\mathcal{D}_{(\theta,k)}$, we propose a hybrid-attention Transformer (HAT) as building block, where spatial-wise dual-scale attention for long-range high- and low-frequency aggregation and channel-wise attention for global information recalibration are established, which will be introduced in detail in Sec. 3.2.

3.2. Deep Denoiser with HAT

In this sub-section, we give the details of deep denoiser $\mathcal{D}_{(\theta,k)}$ used in Eq. (11). Different-stage denoisers have the same network structure with independent learnable parameters and thus we introduce just one of it.

Overall Architecture. As shown in Fig. 2 (b), k-th stage denoiser is a symmetric encoder-decoder architecture built by multiple hybrid-attention Transformer blocks (HATBs) to generate the residual image with degraded input. Each HATB is powered by a residual spatial-wise self-attention (S-SA), feed-forward network (FFN), and channel-wise self-attention (C-SA). As illustrated in Fig. 2 (e), FFN is composed of two 1×1 convolutions which increases or decreases the channel dimensions, one 3×3 depth-wise convolution (D-Conv), and one non-linear activation GELU between them. The downsampling layer uses a 2×2 convolution with a stride of 2. The upsampling layer uses a point-wise convolution (1×1 Conv) to double the channel dimensions and then is followed by a pixel shuffle operation. At each stage, the encoder features are concatenated with the decoder features via skip connections and then a point-wise convolution is used to reduce channel dimensions by half for efficient feature fusion and refinement. At two adjacent stages, previous-stage decoder features are also fused with current-stage encoder features to avoid the potential information loss caused by the signal-feature transformation of deep unfolding. Such dense skip connections between both intra-stage and inter-stage HATBs enhance the performance of proposed method clearly as demonstrated in the ablation experiments shown in Tab. 2. As two core components of the proposed HAT, S-SA and C-SA can realize spatial high- and low-frequency aggregation and channel-wise recalibration respectively. Next, we will describe the details.

Spatial-wise Self-Attention (S-SA). S-SA conducts multi-head self-attention mechanism on dual scales within shifted windows. Specifically, given the input feature $\mathbf{F} \in$

$\mathbb{R}^{H\times W\times C}$, S-SA generates two groups of *query*, *key*, and *value* through the following linear projection:

$$\begin{cases} \mathbf{Q}^h = \mathbf{F}\mathbf{W}_q^h, & \mathbf{Q}^l = \mathbf{F}\mathbf{W}_q^l, \quad (12) \\ \mathbf{K}^h = \mathbf{F}\mathbf{W}_k^h, & \mathbf{K}^l = \texttt{AvgPool}(\mathbf{F})\mathbf{W}_k^l, \quad (13) \\ \mathbf{V}^h = \mathbf{F}\mathbf{W}_v^h, & \mathbf{V}^l = \texttt{AvgPool}(\mathbf{F})\mathbf{W}_v^l, \quad (14) \end{cases}$$

where $\mathbf{W}_q^h, \mathbf{W}_k^h, \mathbf{W}_v^h \in \mathbb{R}^{C\times C_1}$ and $\mathbf{W}_q^l, \mathbf{W}_k^l, \mathbf{W}_v^l \in \mathbb{R}^{C\times C_2}$ are learnable projection matrices with biases omitted, and `AvgPool` represents an average pooling operator with the window resolution p. The resulting $\mathbf{Q}^h, \mathbf{K}^h, \mathbf{V}^h \in \mathbb{R}^{H\times W\times C_1}$ belong to the high-frequency group, similar to that of regular Transformer [26]. The resulting $\mathbf{Q}^l \in \mathbb{R}^{H\times W\times C_2}$ and $\mathbf{K}^l, \mathbf{V}^l \in \mathbb{R}^{h\times w\times C_2}$ belong to low-frequency group, similar to that of PVT [54], where $h = \frac{H}{\sqrt{p}}$ and $w = \frac{W}{\sqrt{p}}$. The total channel number of high-frequency and low-frequency branches is the same with that of the input feature, namely $C_1 + C_2 = C$. Then, $\mathbf{Q}^h, \mathbf{K}^h, \mathbf{V}^h$ and $\mathbf{Q}^l, \mathbf{K}^l, \mathbf{V}^l$ are partitioned into non-overlapping windows and then flatted into token sequences. For high-frequency *query*, *key*, and *value*, the window resolution is N and the reshaped results can be represented as $\bar{\mathbf{Q}}^h, \bar{\mathbf{K}}^h, \bar{\mathbf{V}}^h \in \mathbb{R}^{\frac{HW}{N}\times N\times C_1}$. For low-frequency *query*, *key*, and *value*, the window resolution is pN for $\mathbf{Q}^l$ and N for $\mathbf{K}^l, \mathbf{V}^l$, and the reshaped results can be represented as $\bar{\mathbf{Q}}^l \in \mathbb{R}^{\frac{hw}{N}\times pN\times C_1}$ and $\bar{\mathbf{K}}^l, \bar{\mathbf{V}}^l \in \mathbb{R}^{\frac{hw}{N}\times N\times C_1}$. Next, they are split into m heads, namely $\{\bar{\mathbf{Q}}_i^h\}_{i=1}^m$, $\{\bar{\mathbf{K}}_i^h\}_{i=1}^m$, $\{\bar{\mathbf{V}}_i^h\}_{i=1}^m$, $\{\bar{\mathbf{Q}}_i^l\}_{i=1}^m$, $\{\bar{\mathbf{K}}_i^l\}_{i=1}^m$, $\{\bar{\mathbf{V}}_i^l\}_{i=1}^m$. The channel dimension of each head is $d = \frac{C_1}{m}$ for high-frequency group and $d = \frac{C_2}{m}$ for low-frequency group. The illustration of Fig. 2 (c) is the case with $m = 1$. For i-th head, high-frequency output $\bar{\mathbf{E}}_i^h$ and low-frequency output $\bar{\mathbf{E}}_i^l$ are calculated by

$$\begin{aligned} \bar{\mathbf{E}}_i^h &= \texttt{softmax}(\frac{\bar{\mathbf{Q}}_i^h \bar{\mathbf{K}}_i^{h\top}}{\sqrt{d}})\bar{\mathbf{V}}_i^h, \\ \bar{\mathbf{E}}_i^l &= \texttt{softmax}(\frac{\bar{\mathbf{Q}}_i^l \bar{\mathbf{K}}_i^{l\top}}{\sqrt{d}})\bar{\mathbf{V}}_i^l. \end{aligned} \quad (15)$$

As a result, high-frequency feature $\mathbf{E}^h \in \mathbb{R}^{H\times W\times C_1}$ and low-frequency feature $\mathbf{E}^l \in \mathbb{R}^{H\times W\times C_2}$ can be got by reshaping and concatenating $\{\bar{\mathbf{E}}_i^h\}_{i=1}^m$ and $\{\bar{\mathbf{E}}_i^h\}_{i=1}^m$ separately. The final output is obtained by fusing $\mathbf{E}^h \in \mathbb{R}^{H\times W\times C_1}$ and $\mathbf{E}^l \in \mathbb{R}^{H\times W\times C_2}$. This process is formulated as

$$\text{S–SA}(\mathbf{F}) = \texttt{Concat}(\mathbf{E}^h\mathbf{W}_h, \mathbf{E}^l\mathbf{W}_l), \quad (16)$$

where $\mathbf{W}^h \in \mathbb{R}^{C_1\times C_1}, \mathbf{W}^l \in \mathbb{R}^{C_2\times C_2}$ are two learnable projection matrices and `Concat` denotes the channel concatenation.

As illustrated in Fig. 2 (c), the high-frequency attention performs regular attention within N-pixel windows, and the low-frequency attention performs cross-scale attention between *query* and average-pooled *key*, *value* within pN-pixel windows. As average pooling can act as a low-pass filter [35], such dual-scale attention has two sizes of receptive fields on the input and the average-pooled input, enabling high- and low-frequency aggregation.

Channel-wise Self-Attention (C-SA). Since that S-SA focuses on capturing spatial information within local windows, we incorporate a channel-wise self-attention (C-SA) to capture channel-wise global information [19]. As illustrated in Fig. 2 (d), C-SA squeezes the spatial information into channels first and then a multilayer perceptron applies to it to calculate the channel attention, which will be used to weight the feature map. Given an input $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$, the output of C-SA is formulated as

$$\text{C–SA}(\mathbf{F}) = \mathbf{F} * \texttt{Sigmoid}(\texttt{ReLU}(\texttt{GAP}(\mathbf{F})\mathbf{W}_1)\mathbf{W}_2), \quad (17)$$

where $*$ denotes channel-wise multiplication, $\mathbf{W}_1 \in \mathbb{R}^{C\times\frac{C}{\beta}}$, $\mathbf{W}_2 \in \mathbb{R}^{\frac{C}{\beta}\times C}$ are two fully-connected layers with a non-linear activation `ReLU` inside, `GAP` indicates the global average pooling operation, and `Sigmoid` limits the channel attention map in $(0, 1)$. β is a channel shrinking factor. C-SA plays two important roles, that is, global information aggregation and channel-wise recalibration.

4. Experiment

4.1. Implementation Details

In the proposed HATNet, each-stage denoiser is a three-level symmetric U-shaped structure, powered by proposed S-SA and C-SA. From level-1 to level-3, the number of HATB are $[1, 1, 1]$ and the dimensions of head is 16. Toward S-SA, the size of shifted windows is 4×16 or 16×4 and the kernel size of average pooling operator is 2×2, namely $N = 64$ and $p = 4$. Toward C-SA, the channel shrinking factor is $\beta = 16$. Following previous works [34, 39–42, 44, 51, 72], we adopt 400 images from BSD500 [1] as the training dataset. Data augmentation operations, including random horizontal flipping, random scaling, and random cropping, are performed to generate 20,000 images as the training dataset. Proposed method is implemented by PyTorch on NVIDIA A100 GPUs. All models are trained through 100 epochs with learning rate 1×10^{-3} and then fine-tuned through 20 epochs with learning rate 1×10^{-4} using Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$). Similar to previous works [34, 44, 51], two measurement matrices of Kronecker SPI are set to be learnable for fair comparison on simulation. In real SPI, they are set to be cake-cutting Hadamard matrices [37, 49, 61]. For testing on synthetic data, we evaluate the proposed method with different sampling ratios (SRs) $\{4\%, 10\%, 25\%, 50\%\}$ on a commonly-used Set11 dataset. For testing on real data, we build a SPI prototype to verify the effectiveness of our method. Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) are used to estimate the performance in our experiments.

Table 1. Average PSNR/SSIM of different methods on Set11 datasets with different SRs. The best and second best results are highlighted in **bold** and underlined, respectively.

Dataset	Method	Sampling Ratio (SR)			
		4%	10%	25%	50%
Set11	ReconNet [22]	20.93/0.5897	24.38/0.7301	28.44/0.8531	32.25/0.9177
	ISTA-Net+ [69]	21.32/0.6037	26.64/0.8087	32.59/0.9254	38.11/0.9707
	CSNet+ [40]	24.83/0.7480	28.34/0.8580	33.34/0.9387	38.47/0.9796
	SCSNet [41]	24.29/0.7589	28.52/0.8616	33.43/0.9373	39.01/0.9769
	OPINE-Net+ [71]	25.69/0.7920	29.81/0.8884	34.86/0.9509	40.17/0.9797
	AMP-Net [72]	25.27/0.7821	29.43/0.8880	34.71/0.9532	40.66/0.9827
	TransCS [39]	25.41/0.7883	29.54/0.8877	35.06/0.9548	40.49/0.9815
	MADUN [42]	25.71/0.8042	30.20/0.9016	35.76/0.9601	41.00/0.9837
	DGUNet+ [34]	26.82/0.8230	30.93/0.9088	36.18/0.9616	41.24/0.9837
	OCTUF+ [44]	26.54/0.8150	30.73/0.9036	36.10/0.9607	41.35/**0.9838**
	SAUNet [51]	27.80/0.8353	32.15/0.9147	37.11/0.9628	41.91/**0.9838**
	HATNet (ours)	**27.98/0.8382**	**32.26/0.9182**	**37.24/0.9634**	**42.05/0.9838**

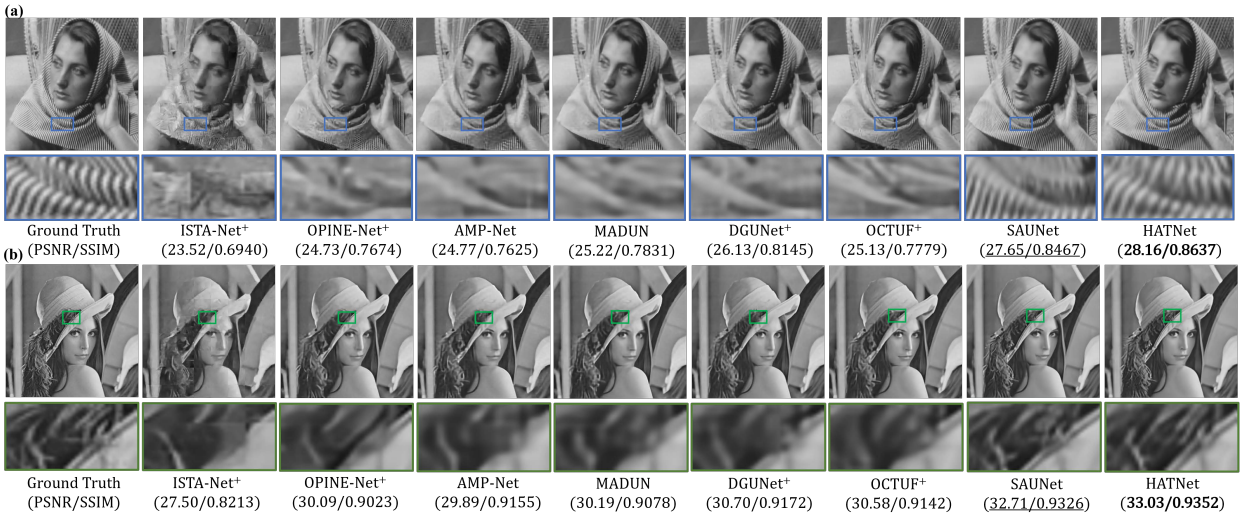

Figure 3. Visualization of different methods on (a) `Barbara` and (b) `Lena` at SR = 10%.

4.2. Results on Synthetic Data

To evaluate the performance of proposed HATNet, we compare it with previous methods, including ReconNet [22], ISTA-Net+ [69], CSNet+ [40], SCSNet [41], OPINENet+ [71], AMP-Net [72], TransCS [39], MADUN [42], DGUNet+ [34], OCTUF+ [44], and SAUNet [51]. Tab. 1 reports the average PSNR/SSIM of different methods. Our method outperforms previous methods at all SRs. Fig. 3 visualizes the reconstruction results of our HATNet and previous competitive methods. Obviously, our HATNet has a significant improvement in image details and textures, as highlighted in the zoom-in regions.

4.3. Results on Real Data

SPI Prototype Details. To evaluate the real performance of our proposed method, we build a SPI Prototype to capture real data as illustrated in Fig. 1 (a), which mainly consists of a digital micro-mirror device (DMD), and a single-pixel detector (SPD). A DMD is used to spatially filter light by selectively redirecting parts of an incident light beam. A DMD is used to measure the total filtered intensity. An object is illuminated and imaged onto the DMD, where a sequence of binary patterns displayed on the DMD are used to mask the image, and then integrated into one pixel detected by SPD. In view of practicality, we use cake-cutting Hadamard matrix (CCH) [37, 49, 61], a variant of Hadamard matrix, as the measurement matrices, whose each row is a binary pattern to be displayed on the DMD.

Middle-Scale Results. We use our SPI prototype to capture real measurements of different scenes with 256×256 pixels, and then they are reconstructed by ISTA-TV [62], DGUNet+ [34], OCTUF+ [44], and SAUNet [51]. ISTA-TV is a representative optimization algorithm and SAUNet is

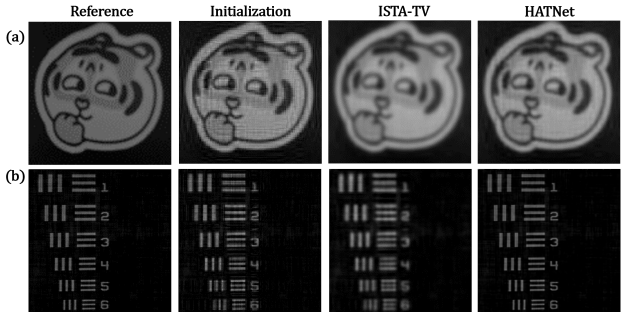

Figure 4. Experimental results of (a) `cartoon tiger` and (b) `resolution target` reconstructed by different methods at SR = 25%.

the first practical deep unfolding network. Note that original DGUNet$^+$ and OCTUF$^+$ are impractical for real SPI cameras due to their block-based sampling, thus we re-trained them under Kronecker SPI for full-size sampling. Middle-scale reconstructed results are visualized in Fig. 1 (c) and Fig. 4. The reference images in the first column are captured through full-sampling (*i.e.*, uncompressed) SPI. Full-sampling SPI can be formulated as $\boldsymbol{x} = \mathbf{A}^\top \boldsymbol{y}$ s.t. $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$, where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is a orthogonal Hadamard matrix. In theory, full-sampled image is lossless due to the orthogonality of Hadamard matrix. The initialization images are simply computed through $\boldsymbol{x} = \mathbf{A}^\top \boldsymbol{y}$, where $\mathbf{A} \in \mathbb{R}^{M\times N}$ $(M \ll N)$ is a sub-sampling matrix. However, noise is not avoidable in real optical system and thus they serve as the references. Clearly, our method leads to the best visual results.

Large-Scale Results. As mentioned previously, previous SPI methods [22, 38, 50], vectorizing 2D image into 1D signal to process, leads to a huge measurement matrix and thus high computational costs in the forward model and the gradient descent projection, which makes it difficult to train on large-scale images. Our method utilizes Kronecker SPI to replace a huge measurement matrix with two small measurement matrices. The maximal resolution of our DMD is 768×1024. We try to capture image with 768×1024 pixels at the sampling ratio 12.5%, namely compress 768,432 pixels into 78,304 measurements. The size of two measurement matrices are 256×768 and 384×1024. We use the two matrices to train our HATNet on 20,000 images with 768×1024 pixels, which are cropped from DIV2K dataset. To relieve memory and computational overheads, we properly reduce the number of stage and channel. Due to the high training costs, we do not re-train previous methods for comparison. Fig. 5 reports the large-scale reconstructed results of our HATNet and ISTA-TV. Our HATNet outperforms ISTA-TV by a large margin in the case of large-scale SPI reconstruction.

Illumination-Varying Results. We also conduct experiments with varying light intensity, from 100 lux to 1,000 lux, to evaluate the generalization ability of our HATNet. In general, stronger the illumination intensity is, the higher signal-to-noise-ratio (SNR) is. Fig. 6 reports the reconstructed

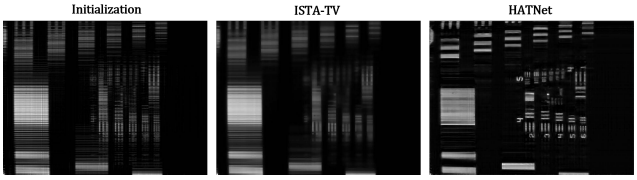

Figure 5. Large-scale experimental results with 768×1024 pixels at SR = 12.5%.

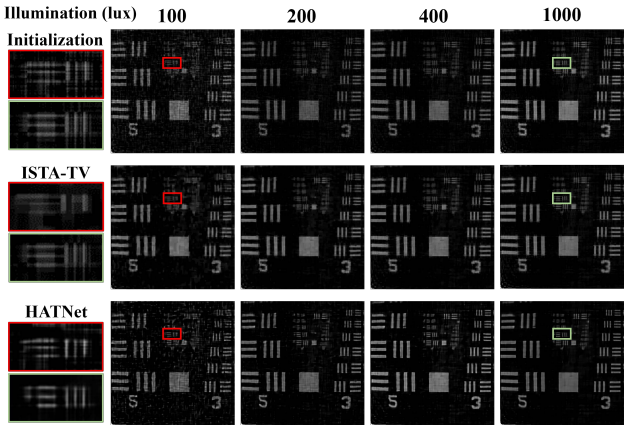

Figure 6. Experimental results of different illumination intensity at SR = 25%.

results of ISTA-TV and our HATNet in different illumination intensities. Clearly, the reconstructed results become better as the illumination intensity increases. Our HATNet shows a great generalization ability in both low- and high-light conditions.

Optical Resolution. Under the same data throughput, we are curious whether our sub-sampling method has truly improved the optical resolution compared to full-sampling method. To this end, we apply full-sampling SPI to capture $4,096$ measurements using a orthogonal Hadamard matrix and then form a theoretically lossless 64×64 image. We apply sub-sampling SPI to capture $4,096$ measurements at SR = 6.25% and then the proposed HATNet to reconstruct a 256×256 image. The full-sampling and sub-sampling images are visualized in Fig. 7. The results demonstrate our method, a pipeline of sub-sampling plus deep reconstruction, can improve the optical resolution significantly.

4.4. Ablation Study

Different Components of HATNet. Our HATNet is mainly powered by the following designs: cross-stage skip connections (CSSC), high-frequency (HF) and low-frequency (LF) aggregation of spatial-wise self-attention (S-SA), and channel-wise self-attention (C-SA). To demystify the influence of different components, we conduct thorough ablation experiments on Set11 dataset at SR = 10%. The average PSNR and SSIM are reported in Tab. 2. Baseline model (a) involves complete components and yields the best re-

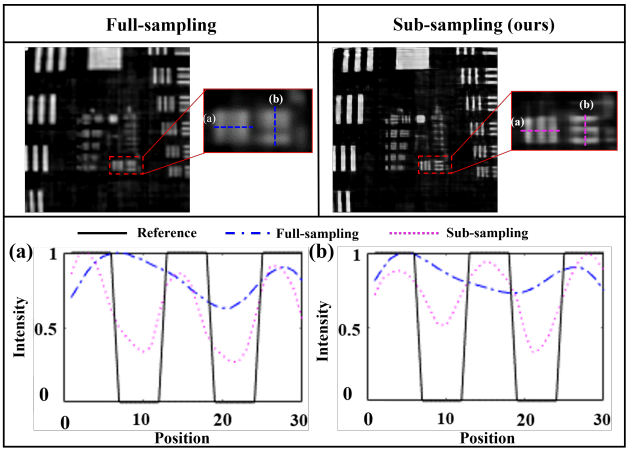

Figure 7. Optical resolution comparison under the same measurements. Full-sampling result (left) versus our sub-sampling result (right). The bottom tables visualize the intensity curve of highlighted lines.

sult of 32.26 dB/0.9182. Towards model (b) without CSSC, there is an average 0.78 dB /0.0129 performance degradation, revealing that regular deep unfolding has an inherent loss information issue due to its frequent signal-to-feature transformation. Towards model (c) with only LF aggregation in S-SA, there is an average 0.70 dB/0.0108 performance degradation. Towards model (d) with only HF aggregation in S-SA, there is an average 0.48 dB/0.0070 performance degradation. It reveals that the proposed S-SA has a good modeling ability for high- and low-frequency information. Towards model (c) without C-SA, there is an average 0.11 dB/0.0042 performance degradation, meaning that our used channel-wise global information recalibration is effective for SPI reconstruction.

Kronecker SPI. By virtue of Kronecker SPI, our HATNet has advantages over previous vectorized methods [22, 38, 50] and block-based methods [34, 44]. Vectorized methods, modeling 2D image into 1D signal, can be deployed in real SPI cameras but their performance is greatly limited by high computational complexity. For example, ReconNet [22] has an average 24.38 dB/0.7301 result at SR = 10%, far lower than 32.26 dB/0.9182 of our HATNet. Block-based methods [34, 44], divide image into small-size patches to process and their block-based sampling is impractical in mainstream SPI cameras. Our HATNet has both practicality and SOTA performance. We conduct experiments to reveal the superiority of Kronecker SPI. The comparison between Kronecker SPI and vectorized SPI is shown in Tab. 3. By shifting HATNet from Kronecker SPI to vectorized SPI, GPU memory occupation and inference time increase from 3.02 G and 0.38 s to 10.74 G and 0.55 s, revealing the efficiency of our HATNet. Full-size sampling of Kronecker SPI is compared with previous block-based sampling as shown in Tab. 4, where we re-train DGUNet$^+$ [34] and OCTUF$^+$ [44] on Kronecker SPI and re-train HATNet under block-based pipeline. Clearly, the re-train DGUNet$^+$ and OCTUF$^+$ achieve a clear improvement and the re-train HATNet have a drop on performance, revealing the effectiveness of Kronecker SPI.

Table 2. Ablation study for different components in HATNet.

Model	CSSC	HF	LF	C-SA	PSNR (dB)	SSIM
(a)	✓	✓	✓	✓	**32.26**	**0.9182**
(b)		✓	✓	✓	31.48	0.9053
(c)	✓		✓	✓	31.56	0.9074
(d)	✓	✓		✓	31.78	0.9112
(e)	✓	✓	✓		32.15	0.9140

Table 3. Comparison between Kronecker SPI and Vectorized SPI at SR$=25\%$.

Method	Kronecker SPI	Vectorized SPI
	$\mathbf{X}\in\mathbb{R}^{N\times N}$, $\mathbf{Y}\in\mathbb{R}^{M\times M}$ $\mathbf{\Phi}\in\mathbb{R}^{M\times N}$, $\mathbf{\Psi}\in\mathbb{R}^{M\times N}$, $\alpha=M^2/N^2$	$\boldsymbol{x}=\text{vec}(\mathbf{X})$, $\boldsymbol{y}=\text{vec}(\mathbf{Y})$ $\mathbf{A}\in\mathbb{R}^{M^2\times N^2}$, $\alpha=M^2/N^2$
Measurement	$\mathbf{Y}=\mathbf{\Phi X\Psi}^\top \Rightarrow \boldsymbol{y}=\mathbf{A}\boldsymbol{x}, \mathbf{A}=\mathbf{\Psi}\otimes\mathbf{\Phi}$	$\boldsymbol{y}=\mathbf{A}\boldsymbol{x}$
Gradient descent	$\mathbf{Z}_k=\mathbf{X}_{k-1}+\rho\mathbf{\Phi}^\top(\mathbf{Y}-\mathbf{\Phi X}_{k-1}\mathbf{\Psi}^\top)\mathbf{\Psi}$	$\boldsymbol{z}_k=\boldsymbol{x}_{k-1}+\rho\mathbf{A}^\top(\boldsymbol{y}-\mathbf{A}\boldsymbol{x}_{k-1})$
Complexity	$\mathcal{O}((\sqrt{\alpha}+\alpha)N^3)$	$\mathcal{O}(\alpha N^4)$
GPU memory (G)	3.02	10.74
Inference time (s)	0.38	0.55

Table 4. Comparison between block-based sampling and full-size sampling of Kronecker SPI at SR$=10\%$.

Method	DGUNet$^+$ [34]	OCTFU$^+$ [44]	HATNet (ours)
Block-based	30.92/0.9088	30.73/0.9037	31.62/0.9115
Full-size	31.65/0.9110	31.51/0.9102	32.26/0.9154

5. Conclusion

Towards real-world SPI cameras, previous vectorized methods are limited in resolution and performance, and previous block-based methods are impractical. In this paper, we propose a deep unfolding network with hybrid-attention Transformer on Kronecker SPI model, dubbed HATNet, to realize practicality and SOTA performance. By unrolling the computation graph of tensor ISTA, HATNet addresses SPI reconstruction problem through two alternative modules: efficient tensor gradient descent and hybrid-attention Transformer (HAT) based deep denoising. By virtue of Kronecker SPI, HATNet can efficiently reduce the computational costs, GPU memory, and inference time by replacing a regular large measurement matrix with tow small matrices in the gradient decent projection. Toward deep denoising module, we propose HAT to aggregate high- and low-frequency information in spatial dimensions and recalibrate global information along channel dimension. Overall, the proposed method has a potential of improving real-world SPI cameras and take one significant step towards real-world computational imaging applications [46].

Acknowledgement

This work was supported by National Natural Science Foundation of China (62271414), Science Fund for Distinguished Young Scholars of Zhejiang Province (LR23F010001), Research Center for Industries of the Future (RCIF) at Westlake University and and the Key Project of Westlake Institute for Optoelectronics (Grant No. 2023GD007).

References

[1] Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011. 5

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, 2021. 3

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009. 2

[4] Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009. 2

[5] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006. 1

[6] Stanley H Chan, Xiran Wang, and Omar A Elgendy. Plug-and-play admm for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1):84–98, 2016. 3

[7] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 3

[8] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004. 2

[9] Weisheng Dong, Guangming Shi, Xin Li, Yi Ma, and Feng Huang. Compressive sensing via nonlocal low-rank regularization. *IEEE Transactions on Image Processing*, 23(8): 3618–3632, 2014. 2

[10] David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. 1

[11] Marco F Duarte and Richard G Baraniuk. Kronecker compressive sensing. *IEEE Transactions on Image Processing*, 21(2):494–504, 2011. 2, 3

[12] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021. 3

[13] Mário AT Figueiredo, Robert D Nowak, and Stephen J Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, 2007. 2

[14] Tom Goldstein, Brendan O'Donoghue, Simon Setzer, and Richard Baraniuk. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014. 2

[15] Joel Greenberg, Kalyani Krishnamurthy, and David Brady. Compressive single-pixel snapshot x-ray diffraction imaging. *Optics letters*, 39(1):111–114, 2014. 1

[16] Justin P Haldar, Diego Hernando, and Zhi-Pei Liang. Compressed-sensing mri with random encoding. *IEEE transactions on Medical Imaging*, 30(4):893–903, 2010. 1

[17] Lihan He and Lawrence Carin. Exploiting structure in wavelet-based bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 57(9):3488–3497, 2009. 2

[18] YuHang He, YiYi Huang, ZhiRong Zeng, YiFei Li, JunHao Tan, LiMing Chen, LingAn Wu, MingFei Li, BaoGang Quan, SongLin Wang, et al. Single-pixel imaging with neutrons. *Science Bulletin*, 66(2):133–138, 2021. 1

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. 5

[20] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34:14745–14758, 2021. 3

[21] Yookyung Kim, Mariappan S Nadar, and Ali Bilgin. Compressed sensing using a gaussian scale mixtures model in wavelet domain. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2010. 2

[22] Kuldeep Kulkarni, Suhas Lohit, Pavan Turaga, Ronan Kerviche, and Amit Ashok. Reconnet: Non-iterative reconstruction of images from compressively sensed measurements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 449–458, 2016. 2, 3, 6, 7, 8

[23] Chengbo Li, Wotao Yin, Hong Jiang, and Yin Zhang. An efficient augmented lagrangian method with applications to total variation minimization. *Computational Optimization and Applications*, 56:507–530, 2013. 2

[24] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1833–1844, 2021. 3

[25] Yang Liu, Xin Yuan, Jinli Suo, David J. Brady, and Qionghai Dai. Rank minimization for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):2990–3006, 2019. 2

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 3, 5

[27] Benjamin Lochocki, Adrian Gambín, Silvestre Manzanera, Esther Irles, Enrique Tajahuerce, Jesus Lancis, and Pablo Artal. Single pixel camera ophthalmoscope. *Optica*, 3(10): 1056–1059, 2016. 1

[28] Meng Lyu, Wei Wang, Hao Wang, Haichao Wang, Guowei Li, Ni Chen, and Guohai Situ. Deep-learning-based ghost imaging. *Scientific reports*, 7(1):17865, 2017. 2

[29] Shiqian Ma, Wotao Yin, Yin Zhang, and Amit Chakraborty. An efficient algorithm for compressed mr imaging using total

variation and wavelets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 2

[30] Tim Meinhardt, Michael Moller, Caner Hazirbas, and Daniel Cremers. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1781–1790, 2017. 3

[31] Ziyi Meng, Xin Yuan, and Shirin Jalali. Deep unfolding for snapshot compressive imaging. *International Journal of Computer Vision*, 131(11):2933–2958, 2023. 2, 3

[32] Chris Metzler, Ali Mousavi, and Richard Baraniuk. Learned d-amp: Principled neural network based compressive image recovery. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 3

[33] Christopher A Metzler, Arian Maleki, and Richard G Baraniuk. From denoising to compressed sensing. *IEEE Transactions on Information Theory*, 62(9):5117–5144, 2016. 2

[34] Chong Mou, Qian Wang, and Jian Zhang. Deep generalized unfolding networks for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17399–17410, 2022. 2, 3, 5, 6, 8

[35] Zizheng Pan, Jianfei Cai, and Bohan Zhuang. Fast vision transformers with hilo attention. *Advances in Neural Information Processing Systems*, 35:14541–14554, 2022. 3, 5

[36] Lee C Potter, Emre Ertin, Jason T Parker, and Müjdat Cetin. Sparsity and compressed sensing in radar imaging. *Proceedings of the IEEE*, 98(6):1006–1020, 2010. 2

[37] Gang Qu, Xiangfeng Meng, Yongkai Yin, and Xiulun Yang. A demosaicing method for compressive color single-pixel imaging based on a generative adversarial network. *Optics and Lasers in Engineering*, 155:107053, 2022. 5, 6

[38] Gang Qu, Xiangfeng Meng, Yongkai Yin, and Xiulun Yang. A demosaicing method for compressive color single-pixel imaging based on a generative adversarial network. *Optics and Lasers in Engineering*, 155:107053, 2022. 7, 8

[39] Minghe Shen, Hongping Gan, Chao Ning, Yi Hua, and Tao Zhang. Transcs: a transformer-based hybrid architecture for image compressed sensing. *IEEE Transactions on Image Processing*, 31:6991–7005, 2022. 2, 3, 5, 6

[40] Wuzhen Shi, Feng Jiang, Shaohui Liu, and Debin Zhao. Image compressed sensing using convolutional neural network. *IEEE Transactions on Image Processing*, 29:375–388, 2019. 6

[41] Wuzhen Shi, Feng Jiang, Shaohui Liu, and Debin Zhao. Scalable convolutional neural network for image compressed sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12290–12299, 2019. 2, 6

[42] Jiechong Song, Bin Chen, and Jian Zhang. Memory-augmented deep unfolding network for compressive sensing. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4249–4258, 2021. 2, 3, 5, 6

[43] Jiechong Song, Bin Chen, and Jian Zhang. Dynamic path-controllable deep unfolding network for compressive sensing. *IEEE Transactions on Image Processing*, 2023.

[44] Jiechong Song, Chong Mou, Shiqi Wang, Siwei Ma, and Jian Zhang. Optimization-inspired cross-attention transformer for compressive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6174–6184, 2023. 2, 3, 5, 6, 8

[45] Baoqing Sun, Matthew P Edgar, Richard Bowman, Liberty E Vittert, Stuart Welsh, Adrian Bowman, and Miles J Padgett. 3d computational imaging with single-pixel detectors. *Science*, 340(6134):844–847, 2013. 1

[46] Jinli Suo, Weihang Zhang, Jin Gong, Xin Yuan, David J Brady, and Qionghai Dai. Computational imaging and artificial intelligence: The next revolution of mobile vision. *Proceedings of the IEEE*, 2023. 8

[47] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996. 2

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[49] Pedro G Vaz, Daniela Amaral, LF Requicha Ferreira, Miguel Morgado, and João Cardoso. Image quality of compressive single-pixel imaging using different hadamard orderings. *Optics Express*, 28(8):11666–11681, 2020. 5, 6

[50] Fei Wang, Chenglong Wang, Chenjin Deng, Shensheng Han, and Guohai Situ. Single-pixel imaging using physics enhanced deep learning. *Photonics Research*, 10(1):104–110, 2022. 7, 8

[51] Ping Wang and Xin Yuan. Saunet: Spatial-attention unfolding network for image compressive sensing. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5099–5108, 2023. 2, 3, 5, 6

[52] Ping Wang, Lishun Wang, Mu Qiao, and Xin Yuan. Full-resolution and full-dynamic-range coded aperture compressive temporal imaging. *Optics Letters*, 48(18):4813–4816, 2023. 1

[53] Ping Wang, Lishun Wang, and Xin Yuan. Deep optics for video snapshot compressive imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10646–10656, 2023. 1

[54] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 5

[55] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022. 3

[56] Yibo Xu, Liyang Lu, Vishwanath Saragadam, and Kevin F Kelly. A compressive hyperspectral video imaging system using a single-pixel detector. *Nature Communications*, 15(1): 1456, 2024. 1

[57] Junfeng Yang and Yin Zhang. Alternating direction algorithms for ℓ_1-problems in compressive sensing. *SIAM Journal on Scientific Computing*, 33(1):250–278, 2011. 2

[58] Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Admm-csnet: A deep learning approach for image compressive sensing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3):521–538, 2018. 2

[59] Hantao Yao, Feng Dai, Shiliang Zhang, Yongdong Zhang, Qi Tian, and Changsheng Xu. Dr2-net: Deep residual reconstruction network for image compressive sensing. *Neurocomputing*, 359:483–493, 2019. 2

[60] Ting Yao, Yehao Li, Yingwei Pan, Yu Wang, Xiao-Ping Zhang, and Tao Mei. Dual vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 2023. 3

[61] WenKai Yu. Super sub-nyquist single-pixel imaging by means of cake-cutting hadamard basis sort. *Sensors*, 19(19):4122, 2019. 5, 6

[62] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International Conference on Image Processing*, pages 2539–2543, 2016. 2, 6

[63] Xin Yuan and Raziel Haimi-Cohen. Image compression based on compressive sensing: End-to-end comparison with jpeg. *IEEE Transactions on Multimedia*, 22(11):2889–2904, 2020. 1

[64] Xin Yuan and Yunchen Pu. Parallel lensless compressive imaging via deep convolutional neural networks. *Optics Express*, 26(2):1962–1977, 2018.

[65] Xin Yuan, Hong Jiang, Gang Huang, and Paul A. Wilford. Slope: Shrinkage of local overlapping patches estimator for lensless compressive imaging. *IEEE Sensors Journal*, 16(22): 8091–8102, 2016. 1

[66] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 3

[67] Zhiyuan Zha, Xin Yuan, Bihan Wen, Jiantao Zhou, Jiachao Zhang, and Ce Zhu. A benchmark for sparse coding: When group sparsity meets rank minimization. *IEEE Transactions on Image Processing*, 29:5094–5109, 2020. 2

[68] Zhiyuan Zha, Xin Yuan, Bihan Wen, Jiantao Zhou, and Ce Zhu. Group sparsity residual constraint with non-local priors for image restoration. *IEEE Transactions on Image Processing*, 29:8960–8975, 2020. 2

[69] Jian Zhang and Bernard Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1828–1837, 2018. 2, 3, 6

[70] Jian Zhang, Debin Zhao, and Wen Gao. Group-based sparse representation for image restoration. *IEEE Transactions on Image Processing*, 23(8):3336–3351, 2014. 2

[71] Jian Zhang, Chen Zhao, and Wen Gao. Optimization-inspired compact deep compressive sensing. *IEEE Journal of Selected Topics in Signal Processing*, 14(4):765–774, 2020. 2, 3, 6

[72] Zhonghao Zhang, Yipeng Liu, Jiani Liu, Fei Wen, and Ce Zhu. Amp-net: Denoising-based deep unfolding for compressive image sensing. *IEEE Transactions on Image Processing*, 30: 1487–1500, 2020. 2, 3, 5, 6

[73] Xingchen Zhao, Xiaoyu Nie, Zhenhuan Yi, Tao Peng, and Marlan O Scully. Imaging through scattering media via spatial–temporal encoded pattern illumination. *Photonics Research*, 10(7):1689–1694, 2022. 1

[74] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2