Optimization-Inspired Cross-Attention Transformer for Compressive Sensing

Jiechong Song[1,4], Chong Mou[1], Shiqi Wang[2], Siwei Ma[3,4], Jian Zhang[1,4*]
[1]Peking University Shenzhen Graduate School, Shenzhen, China
[2]Department of Computer Science, City University of Hong Kong, China
[3]School of Computer Science, Peking University, Beijing, China
[4]Peng Cheng Laboratory, Shenzhen, China
{songjiechong, swma, zhangjian.sz} @pku.edu.cn eechongm@stu.pku.edu.cn shiqwang@cityu.edu.hk

Abstract

*By integrating certain optimization solvers with deep neural networks, deep unfolding network (DUN) with good interpretability and high performance has attracted growing attention in compressive sensing (CS). However, existing DUNs often improve the visual quality at the price of a large number of parameters and have the problem of feature information loss during iteration. In this paper, we propose an Optimization-inspired Cross-attention Transformer (OCT) module as an iterative process, leading to a lightweight **OCT**-based **U**nfolding **F**ramework (**OCTUF**) for image CS. Specifically, we design a novel Dual Cross Attention (Dual-CA) sub-module, which consists of an Inertia-Supplied Cross Attention (ISCA) block and a Projection-Guided Cross Attention (PGCA) block. ISCA block introduces multi-channel inertia forces and increases the memory effect by a cross attention mechanism between adjacent iterations. And, PGCA block achieves an enhanced information interaction, which introduces the inertia force into the gradient descent step through a cross attention block. Extensive CS experiments manifest that our OCTUF achieves superior performance compared to state-of-the-art methods while training lower complexity. Codes are available at* https://github.com/songjiechong/OCTUF.

1. Introduction

Compressive sensing (CS) is a considerable research interest from signal/image processing communities as a joint acquisition and reconstruction approach [5]. The signal is first sampled and compressed simultaneously with linear random transformations. Then, the original signal can be reconstructed from far fewer measurements than that required by Nyquist sampling rate [29, 38]. So, the two main concerns of CS are the design of the sampling matrix [7, 16] and recovering the original signal [60], and our work focuses on the latter. Meanwhile, the CS technology achieves great success in many image systems, including medical imaging [31, 45], single-pixel cameras [15, 37], wireless remote monitoring [59], and snapshot compressive imaging [4, 50, 51], because it can reduce the measurement and storage space while maintaining a reasonable reconstruction of the sparse or compressible signal.

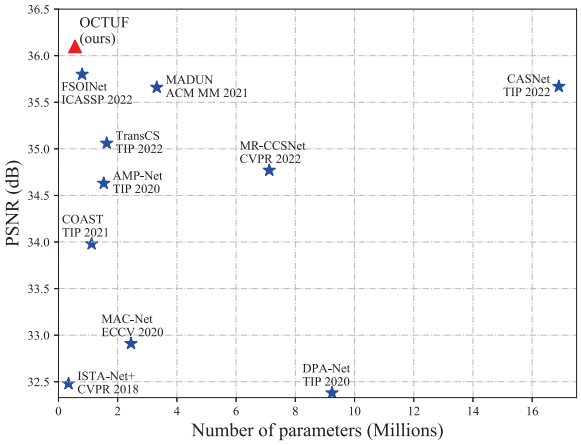


Figure 1. The PSNR (dB) performance (y-axis) of our OCTUF and some recent methods (ISTA-Net [54], DPA-Net [44], AMP-Net [60], MAC-Net [19], COAST [53], MADUN [41], CASNet [7], TransCS [39], FSOINet [10], MR-CCSNet [16]) under different parameter capacities (x-axis) on Set11 [24] dataset in the case of CS ratio $= 25\%$. Our proposed method outperforms previous methods while requiring significantly cheaper parameters.

Mathematically, a random linear measurement $\mathbf{y} \in \mathbb{R}^M$ can be formulated as $\mathbf{y} = \mathbf{\Phi x}$, where $\mathbf{x} \in \mathbb{R}^N$ is the original signal and $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ is the measurement matrix with $M \ll N$. $\frac{M}{N}$ is the CS ratio (or sampling rate). Obviously, CS reconstruction is an ill-posed inverse problem. To obtain a reliable reconstruction, the conventional CS methods

*Corresponding author. This work was supported in part by Shenzhen Research Project under Grant JCYJ20220531093215035 and Grant JSGGZD20220822095800001.

commonly solve an energy function as:

$$\arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{\Phi x}-\mathbf{y}\|_2^2+\lambda\mathcal{R}(\mathbf{x}), \tag{1}$$

where $\frac{1}{2}\|\mathbf{\Phi x}-\mathbf{y}\|_2^2$ denotes the data-fidelity term for modeling the likelihood of degradation and $\lambda\mathcal{R}(\mathbf{x})$ denotes the prior term with regularization parameter λ. For traditional model-based methods [17, 20, 26, 32, 56, 57, 64], the prior term can be the sparsifying operator corresponding to some pre-defined transform basis, such as discrete cosine transform (DCT) and wavelet [61, 62]. They enjoy the merits of strong convergence and theoretical analysis in most cases but are usually limited in high computational complexity and low adaptivity [63]. Recently, fueled by the powerful learning capacity of deep networks, several network-based CS algorithms have been proposed [24, 44]. Although network-based methods can solve CS problem adaptively with fast inferences, the architectures of most of these methods are the black box design and the advantages of traditional algorithms are not fully considered [36].

More recently, some deep unfolding networks (DUNs) with good interpretability are proposed to combine network with optimization and train a truncated unfolding inference through an end-to-end learning manner, which has become the mainstream for CS [52–55, 60]. However, existing deep unfolding algorithms usually achieve excellent performance with a large number of iterations and a huge number of parameters [41, 42], which are easily limited by storage space. Furthermore, the image-level transmission at each iteration fails to make full use of inter-stage feature information.

To address the above problems, in this paper, we propose an efficient **O**ptimization-inspired **C**ross-attention **T**ransformer (**OCT**) module as the iterative process and establish a lightweight **OCT**-based **U**nfolding **F**ramework (**OCTUF**) for image CS, as shown in Fig. 2. Our OCT module maintains maximum information flow in feature space, which consists of a Dual Cross Attention (Dual-CA) sub-module and a Feed-Forward Network (FFN) sub-module to form each iterative process. Dual-CA sub-module contains an Inertia-Supplied Cross Attention (ISCA) block and a Projection-Guided Cross Attention (PGCA) block. ISCA block calculates cross attention on adjacent iteration information and adds inertial/memory effect to the optimization algorithm. And, PGCA block uses the gradient descent step and inertial term as inputs of Cross Attention (CA) block to guide the fine fusion of channel-wise features. With the proposed techniques, OCTUF outperforms state-of-the-art CS methods with much fewer parameters, as illustrated in Fig. 1. The main contributions are summarized as follows:

- We propose a lightweight deep unfolding framework OCTUF in feature space for CS, where the optimization-inspired cross-attention Transformer (OCT) module is regarded as an iterative process.
- We design a compact Dual Cross Attention (Dual-CA) sub-module to guide the efficient multi-channel information interactions, which consists of a Projection-Guided Cross Attention (PGCA) block and an Inertia-Supplied Cross Attention (ISCA) block.
- Extensive experiments demonstrate that our proposed OCTUF outperforms existing state-of-the-art methods with cheaper computational and memory costs.

2. Related Work

2.1. Deep Unfolding Network

The main idea of deep unfolding networks (DUNs) is that conventional iterative optimization algorithms can be implemented equivalently by a stack of recurrent trainable blocks. Such correspondence has been proposed to solve different image inverse tasks, such as denoising [11, 25], deblurring [23, 47], and demosaicking [22]. The solution is usually formulated as a bi-level optimization problem:

$$\begin{aligned}&\min_{\mathbf{\Theta}}\sum_{j=1}\mathcal{L}(\hat{\mathbf{x}}_j,\mathbf{x}_j),\\&\mathbf{s.t.}\ \hat{\mathbf{x}}_j=\arg\min_{\mathbf{x}}\frac{1}{2}\|\mathbf{\Phi x}-\mathbf{y}_j\|_2^2+\lambda\mathcal{R}(\mathbf{x}),\end{aligned} \tag{2}$$

where $\mathbf{\Theta}$ denotes the trainable parameters and $\mathcal{L}(\hat{\mathbf{x}}_j,\mathbf{x}_j)$ represents the loss function of estimated clean image $\hat{\mathbf{x}}_j$ with respect to the original image $\mathbf{x}_j$.

In the community of compressive sensing, DUN-based methods usually integrate some effective convolutional neural network (CNN) denoisers into some optimization methods, *e.g.*, proximal gradient descent (PGD) algorithm [7, 9, 10, 39, 41, 53, 54], approximate message passing (AMP) [60], and inertial proximal algorithm for nonconvex optimization (iPiano) [43]. Different optimization methods lead to different optimization-inspired DUNs. In most DUNs, the input and output of each iteration are inherently images $\mathbf{x}_j$, which seriously hamper information transmission, resulting in limited representation capability [58]. Recently, some methods [10, 36] propose the idea of combining information flow into each iteration process in feature space to enhance information transmission. However, existing solutions usually lack flexibility in dealing with channel-wise information and are beset by high model complexity. In this paper, we present an efficient solution.

2.2. Vision Transformer

Inspired by the success of Transformers [46] in natural language processing, recent researchers also extend the Transformer structure for various computer vision tasks, *e.g.*, image classification [14, 28], object detection [6, 66], segmentation [35, 48]. Transformer-based methods are also applied to image restoration tasks. PIT [8] is the first

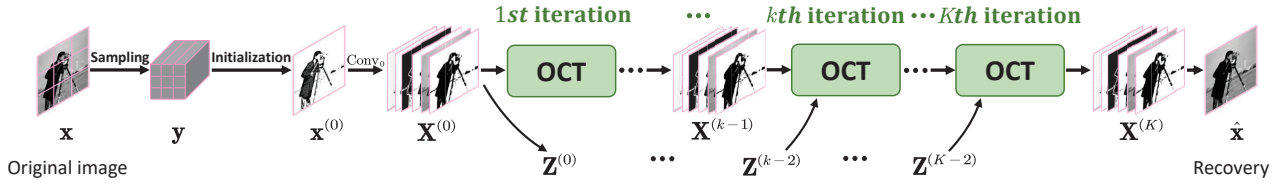


Figure 2. Architecture of our OCTUF, which consists of K iterations. $\mathbf{x}$ denotes the full-sampled image for training, $\mathbf{y}$ is the under-sampled data and $\mathbf{x}^{(0)}$ denotes the initialization. The feature $\mathbf{X}^{(k-1)}$ and $\mathbf{Z}^{(k-2)}$ are the inputs of our optimization-inspired cross-attention Transformer (OCT) module that is the kth iterative process, and $\hat{\mathbf{x}}$ is the recovered result gotten from the output $\mathbf{X}^{(K)}$ in the Kth iteration.

work to introduce Transformer to image restoration and achieves promising performance in several image restoration tasks. Subsequently, several novel designs are proposed. [27,49] utilize the Swin Transformer [28] to perform image restoration. Recently, Cai *et al*. [4] propose DAUF based on DUN structure for spectral compressive imaging where self-attention is widely adopted to build the basic Transformer block and Shen *et al*. [39] design an ISTA-based Transformer backbone for CS. These Transformers are just included in the prior term and have nothing to do with the data-fidelity term, so do not fully exploit the advantages of DUN. In this paper, we combine Transformer and DUN to build an efficient CS framework.

3. Proposed Method

3.1. Overall Architecture

The proximal gradient descent (PGD) algorithm is a well-suited approach for solving many large-scale linear inverse problems [12]. Recently, some research [2, 34] finds that such an algorithm adding the inertial term always succeeds to converge to the global optimum. Ochs *et al*. [34] propose an inertial proximal algorithm for nonconvex optimization (iPiano), which combines the gradient descent term with an inertial force. Inspired by iPiano, the whole update steps (for the kth iteration) can be expressed as:

$$\begin{aligned}\mathbf{s}^{(k)} =&\mathbf{x}^{(k-1)} - \rho^{(k)}\mathbf{\Phi}^{\top}(\mathbf{\Phi}\mathbf{x}^{(k-1)} - \mathbf{y}) \\ &+ \alpha^{(k)}(\mathbf{x}^{(k-1)} - \mathbf{x}^{(k-2)}),\end{aligned} \tag{3}$$

$$\mathbf{x}^{(k)} = \arg\min_{\mathbf{x}} \frac{1}{2}\left\|\mathbf{x} - \mathbf{s}^{(k)}\right\|_2^2 + \lambda\mathcal{R}(\mathbf{x}), \tag{4}$$

where $\mathbf{x}^{(k)}$ is the output image of the kth iteration, $\mathbf{y}$ is the sampled image, $\rho^{(k)}, \alpha^{(k)}$ are the learnable step size parameters and $\mathbf{\Phi}^{\top}$ is the transpose of the measurement matrix $\mathbf{\Phi}$.

Eq. (3) denotes the projection step, which introduces an inertial term $\alpha^{(k)}(\mathbf{x}^{(k-1)} - \mathbf{x}^{(k-2)})$ to a gradient descent term $\mathcal{F}(\mathbf{x}^{(k-1)}) = \mathbf{x}^{(k-1)} - \rho^{(k)}\mathbf{\Phi}^{\top}(\mathbf{\Phi}\mathbf{x}^{(k-1)} - \mathbf{y})$, relaxing the monotonically decreasing constraints and helping to achieve a better convergence result [33]. As mentioned previously, such traditional implementation lacks adaptability and has information loss due to image-level inter-stage transmission. To rectify these weaknesses, we propose a **D**ual **C**ross **A**ttention (**Dual-CA**) sub-module to achieve feature-level transmission by adding a multi-channel inertial force and enhancing the information interaction in the projection step. Eq. (4) is achieved by a proximal mapping step which is actually a Gaussian denoiser. Here like most DUNs, we implement it with a trainable model, *i.e.*, a **F**eed **F**orward **N**etwork (**FFN**) sub-module that is detailed in Fig. 3(e). FFN consists of two sets of LayerNorm and Feed Forward Block (FFB) with a global skip connection, where the architecture of FFB is similar to [3].

Therefore, as the process in the kth iteration of **OCTUF**, our **O**ptimization-inspired **C**ross-attention **T**ransformer (**OCT**) module can be formulated as ($k \in \{1, 2, \cdots, K\}$):

$$\mathbf{S}^{(k)} = \mathcal{H}_{\text{Dual-CA}}(\mathbf{X}^{(k-1)}, \mathbf{Z}^{(k-2)}), \tag{5}$$

$$\mathbf{X}^{(k)} = \mathcal{H}_{\text{FFN}}(\mathbf{S}^{(k)}), \tag{6}$$

where $\mathbf{S}^{(k)}$, $\mathbf{X}^{(k)} \in \mathbb{R}^{H\times W\times C}$ are the outputs in the feature domain, and $\mathbf{Z}^{(k-2)} \in \mathbb{R}^{H\times W\times(C-1)}$ is obtained by clipping latter $C-1$ channels from $\mathbf{X}^{(k-2)}$. For the first iteration, the input $\mathbf{X}^{(0)}$ is generated by a 3×3 convolution ($\text{Conv}_0(\cdot)$) on the initialization $\mathbf{x}^{(0)}$, and the inertial term is not needed [34], as shown in Fig. 2. And the recovered result $\hat{\mathbf{x}}$ is gotten by splitting the first channel from $\mathbf{X}^{(K)}$.

Therefore, our proposed OCTUF can skillfully integrate the inter-stage feature-level information and achieves the perfect combination with the optimization steps.

3.2. Dual Cross Attention

To ensure maximum information flow and powerful feature correlation, we design a Dual Cross Attention (Dual-CA) sub-module to efficiently fuse information in the projection step. As shown in Fig. 3(a), to make the network more compact and maintain the potential mathematical interpretation, we finely split the input $\mathbf{X}^{(k-1)} \in \mathbb{R}^{H\times W\times C}$ into two chunks, including $\mathbf{r}^{(k-1)} \in \mathbb{R}^{H\times W\times 1}$ (from the first channel) and $\mathbf{Z}^{(k-1)} \in \mathbb{R}^{H\times W\times(C-1)}$ (from the last $C-1$ channels). $\mathbf{r}^{(k-1)}$ and $\mathbf{Z}^{(k-1)}$ are the input of the gradient descent term and the inertial term, respectively. So, to make full use of the information of the multi-channel inertial term, we design an **I**nertia-**S**upplied **C**ross **A**ttention (**ISCA**) block, yielding $\mathcal{H}_{\text{ISCA}}(\mathbf{Z}^{(k-1)}, \mathbf{Z}^{(k-2)})$. And we also propose a **P**rojection-**G**uided **C**ross **A**ttention (**PGCA**) block to perform the fusion of the gradient descent term and

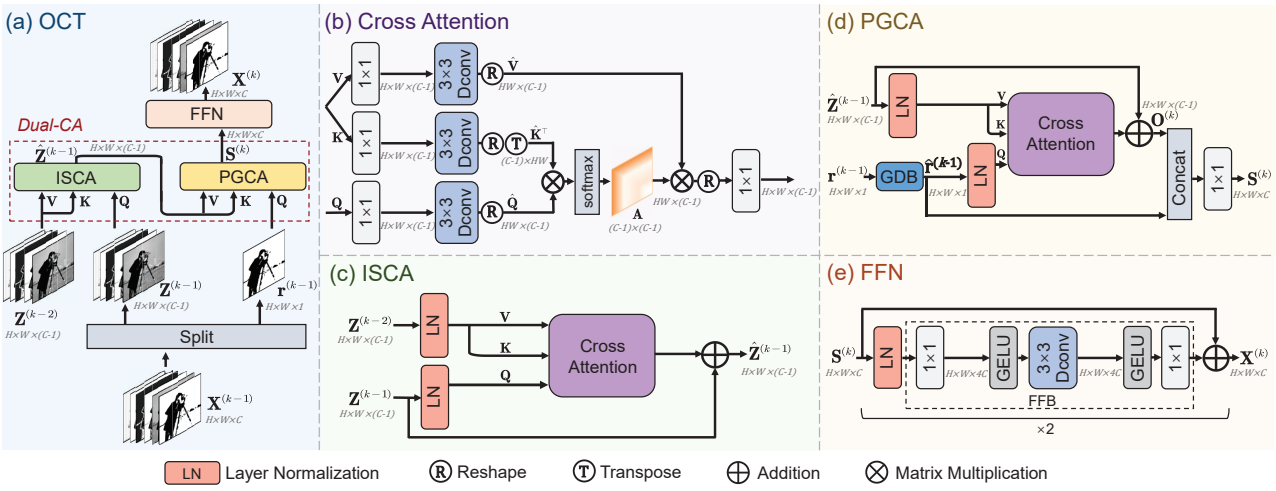


Figure 3. The architecture of Optimization-inspired Cross-attention Transformer (OCT) module. (a) OCT module consists of a Dual Cross Attention (Dual-CA) sub-module which contains an Inertia-Supplied Cross Attention (ISCA) block and a Projection-Guided Cross Attention (PGCA) block, and a Feed-Forward Network (FFN) sub-module. (b) Illustration of Cross Attention (CA) block, which is the basic component of two attention blocks. (c) ISCA block is composed of Layer Normalization (LN) and CA. (d) PGCA block is composed of Gradient Descent Block (GDB), LN, and CA. (e) FFN sub-module is composed of two sets of LN and Feed-Forward Block (FFB).

the inertial term in a more adaptive way. Therefore, our designed Dual-CA sub-module can be formulated as follows:

$$\mathbf{S}^{(k)} = \mathcal{H}_{\text{PGCA}}(\mathcal{F}(\mathbf{r}^{(k-1)}), \mathcal{H}_{\text{ISCA}}(\mathbf{Z}^{(k-1)}, \mathbf{Z}^{(k-2)})). \tag{7}$$

Among ISCA and PGCA blocks, Cross Attention (CA) plays an important role as the basic block. In the following part of Sec. 3.2, we first introduce the Cross Attention block and then present ISCA and PGCA blocks, respectively.

Cross Attention. Motivated by modeling complex relations for generating context-aware objects in multi-modal task [65], we design a Cross Attention (CA) block to aggregate the key information from the different components in the projection step of deep unfolding network, as shown in Fig. 3(b). The input $\mathbf{Q}$ comes from a different component than $\mathbf{V}$ and $\mathbf{K}$. They are first embedded by a 1×1 convolution ($\text{Conv}_{\mathbf{V,K,Q}}(\cdot)$) to obtain feature with the size being $H\times W\times(C-1)$. Then a 3×3 depth-wise convolution ($\text{Dconv}_{\mathbf{V,K,Q}}(\cdot)$) is used to encode channel-wise spatial context. Finally, a reshape operation ($\text{R}(\cdot)$) reformulates $\mathbf{V}$, $\mathbf{K}$, and $\mathbf{Q}$ into tokens $\{\hat{\mathbf{V}}, \hat{\mathbf{K}}, \hat{\mathbf{Q}}\} \in \mathbb{R}^{HW\times(C-1)}$. Therefore, this process can be defined as the following function:

$$\begin{cases} \hat{\mathbf{V}} = \text{R}(\text{Dconv}_{\mathbf{V}}(\text{Conv}_{\mathbf{V}}(\mathbf{V}))), & \text{(8a)} \\ \hat{\mathbf{K}} = \text{R}(\text{Dconv}_{\mathbf{K}}(\text{Conv}_{\mathbf{K}}(\mathbf{K}))), & \text{(8b)} \\ \hat{\mathbf{Q}} = \text{R}(\text{Dconv}_{\mathbf{Q}}(\text{Conv}_{\mathbf{Q}}(\mathbf{Q}))). & \text{(8c)} \end{cases}$$

Next, a transposed attention map $\mathbf{A} \in \mathbb{R}^{(C-1)\times(C-1)}$ is generated by applying softmax function to re-weight the matrix multiplication $\hat{\mathbf{K}}^\top\hat{\mathbf{Q}}$, yielding

$$\mathbf{A} = \text{Softmax}(\hat{\mathbf{K}}^\top\hat{\mathbf{Q}}), \tag{9}$$

where $\hat{\mathbf{K}}^\top$ denotes the transposed matrix of $\hat{\mathbf{K}}$. The aggregation result is calculated as $\hat{\mathbf{V}}\mathbf{A}$, which is reshaped into the features of size $\mathbb{R}^{H\times W\times(C-1)}$. Finally, we apply a 1×1 convolution $\text{Conv}_{\mathbf{A}}(\cdot)$ to enhance the feature extraction. Overall, the Cross Attention block is defined as:

$$\mathcal{G}_{\text{CA}}(\mathbf{V}, \mathbf{K}, \mathbf{Q}) = \text{Conv}_{\mathbf{A}}(\text{R}(\hat{\mathbf{V}}\mathbf{A})). \tag{10}$$

Cross Attention block helps to extract useful information via channel-wise similarity with low computational cost.

Inertia-Supplied Cross Attention. As shown in Eq. (3), the general inertial term usually adopts the simple operation by directly subtracting the adjacent iteration output, which is proved to be ineffective in the ablation of Tab. 4. To enrich the information interaction of the inertial term, we introduce a multi-channel inertial term and propose an Inertia-Supplied Cross Attention (ISCA) block. Our ISCA block consists of LayerNorm (LN) function and CA block as shown in Fig. 3(c). Specifically, we set the $(k-2)$th iteration output $\mathbf{Z}^{(k-2)}$ as *value* ($\mathbf{V}^{(k)}_{\text{ISCA}}$) and *key* ($\mathbf{K}^{(k)}_{\text{ISCA}}$), and we set the $(k-1)$th iteration output $\mathbf{Z}^{(k-1)}$ as *query* ($\mathbf{Q}^{(k)}_{\text{ISCA}}$), pass through CA block after normalization by LN function, so $\hat{\mathbf{Z}}^{(k-1)} = \mathcal{H}_{\text{ISCA}}(\mathbf{Z}^{(k-1)}, \mathbf{Z}^{(k-2)})$ as:

$$\begin{aligned} &\mathbf{V}^{(k)}_{\text{ISCA}}, \mathbf{K}^{(k)}_{\text{ISCA}}, \mathbf{Q}^{(k)}_{\text{ISCA}} = \\ &\qquad \text{LN}(\mathbf{Z}^{(k-2)}), \text{LN}(\mathbf{Z}^{(k-2)}), \text{LN}(\mathbf{Z}^{(k-1)}), \\ &\hat{\mathbf{Z}}^{(k-1)} = \mathcal{G}_{\text{CA}}(\mathbf{V}^{(k)}_{\text{ISCA}}, \mathbf{K}^{(k)}_{\text{ISCA}}, \mathbf{Q}^{(k)}_{\text{ISCA}}) + \mathbf{Z}^{(k-1)}. \end{aligned} \tag{11}$$

ISCA block adaptively learns more useful multi-channel inertial force and enhances memory effect to our network.

Table 1. Average PSNR(dB)/SSIM performance comparisons of recent deep network-based CS methods on Set11 dataset [24] with different CS ratios. The best and second-best results are highlighted in red and blue colors, respectively.

Dataset	Methods	CS Ratio					
		10%	25%	30%	40%	50%	Average
Set11	ISTA-Net$^+$ (CVPR 2018) [54]	26.58/0.8066	32.48/0.9242	33.81/0.9393	36.04/0.9581	38.06/0.9706	33.39/0.9197
	DPA-Net (TIP 2020) [44]	27.66/0.8530	32.38/0.9311	33.35/0.9425	35.21/0.9580	36.80/0.9685	33.08/0.9306
	AMP-Net (TIP 2020) [60]	29.40/0.8779	34.63/0.9481	36.03/0.9586	38.28/0.9715	40.34/0.9804	35.74/0.9473
	MAC-Net (ECCV 2020) [9]	27.68/0.8182	32.91/0.9244	33.96/0.9372	35.94/0.9560	37.67/0.9668	33.63/0.9205
	COAST (TIP 2021) [53]	28.74/0.8619	33.98/0.9407	35.11/0.9505	37.11/0.9646	38.94/0.9744	34.78/0.9384
	MADUN (ACM MM 2021) [41]	29.91/0.8986	35.66/0.9601	36.94/0.9676	39.15/0.9772	40.77/0.9832	36.48/0.9573
	CASNet (TIP 2022) [7]	30.36/0.9014	35.67/0.9591	36.92/0.9662	39.04/0.9760	40.93/0.9826	36.58/0.9571
	TransCS (TIP 2022) [39]	29.54/0.8877	35.06/0.9548	35.62/0.9588	38.46/0.9737	40.49/0.9815	35.83/0.9513
	FSOINet (ICASSP 2022) [10]	30.46/0.9023	35.80/0.9595	37.00/0.9665	39.14/0.9764	41.08/0.9832	36.70/0.9576
	MR-CCSNet (CVPR 2022) [16]	-/-	34.77/0.9546	-/-	-/-	40.73/0.9828	-/-
	OCTUF (Ours)	30.70/0.9030	36.10/0.9604	37.21/0.9673	39.41/0.9773	41.34/0.9838	36.95/0.9584
	OCTUF$^+$ (Ours)	30.73/0.9036	36.10/0.9607	37.32/0.9676	39.43/0.9774	41.35/0.9838	36.99/0.9586

Projection-Guided Cross Attention. In order to adaptively combine the gradient descent term and the inertial term, we introduce a Projection-Guided Cross Attention (PGCA) block in Fig. 3(d). Similar to ISCA block, PGCA block captures rich feature information based on channel-wise similarity. Specifically, given $\mathbf{X}^{(k-1)}$, the input of gradient descent term is gotten by its first channel (*i.e.*, $\mathbf{r}^{(k-1)}$). So, the calculation of the term has the following expression:

$$\hat{\mathbf{r}}^{(k-1)} = \mathbf{r}^{(k-1)} - \rho^{(k)}\mathbf{\Phi}^\top(\mathbf{\Phi}\mathbf{r}^{(k-1)} - \mathbf{y}). \tag{12}$$

Next, $\hat{\mathbf{r}}^{(k-1)}$ and the ISCA output $\hat{\mathbf{Z}}^{(k-1)}$ pass through LayerNorm function and CA block, yielding

$$\begin{aligned}\mathbf{V}^{(k)}_{\text{PGCA}},\ \mathbf{K}^{(k)}_{\text{PGCA}},\ \mathbf{Q}^{(k)}_{\text{PGCA}} &= \\ \text{LN}(\hat{\mathbf{Z}}^{(k-1)}), \text{LN}(\hat{\mathbf{Z}}^{(k-1)}), \text{LN}(\hat{\mathbf{r}}^{(k-1)}),& \\ \mathbf{O}^{(k)} = \mathcal{G}_{\text{CA}}(\mathbf{V}^{(k)}_{\text{PGCA}}, \mathbf{K}^{(k)}_{\text{PGCA}}, \mathbf{Q}^{(k)}_{\text{PGCA}}) + \hat{\mathbf{Z}}^{(k-1)}&.\end{aligned} \tag{13}$$

It is worth noting that $\hat{\mathbf{r}}^{(k-1)}$ generates feature maps with $C-1$ channels by $\text{Conv}_{\mathbf{Q}}(\cdot)$ of Eq. (8a) to enrich the feature expression. Finally, $\mathbf{O}^{(k)}$ and $\hat{\mathbf{r}}^{(k-1)}$ are concatenated, reshaped to match the original channel dimensions and mixed with a 1×1 convolution ($\text{Conv}_{\mathbf{O}}(\cdot)$):

$$\mathbf{S}^{(k)} = \text{Conv}_{\mathbf{O}}(\text{Concat}(\mathbf{O}^{(k)}, \hat{\mathbf{r}}^{(k-1)})). \tag{14}$$

PGCA not only inherits the advantage of Eq. (3) but also tactfully achieves the multi-channel feature fusion between the gradient descent term and the inertial term.

3.3. Loss Function

Given a set of full-sampled images $\{\mathbf{x}_j\}_{j=1}^{N_a}$ and some sampling patterns with the specific sampling rate, the compressed measurements can be obtained by $\mathbf{y}_j = \mathbf{\Phi}\mathbf{x}_j$, producing the train data pairs $\{(\mathbf{y}_j, \mathbf{x}_j)\}_{j=1}^{N_a}$. Our model takes $\mathbf{y}_j$ as input and generates the reconstruction result $\hat{\mathbf{x}}_j$ as output. We employ the MSE loss function with respect to $\mathbf{x}_j$ and $\hat{\mathbf{x}}_j$ as following shows:

$$\mathcal{L}(\mathbf{\Theta}) = \frac{1}{NN_a}\sum_{j=1}^{N_a}\|\mathbf{x}_j - \hat{\mathbf{x}}_j\|_2^2, \tag{15}$$

where N_a and N represent the number of the training images and the size of each image respectively. $\mathbf{\Theta}$ denotes the learnable parameter set of our proposed OCTUF and can be formulated as $\mathbf{\Theta} = \{\mathbf{\Phi}, \text{Conv}_0(\cdot)\} \bigcup \{\mathcal{H}^{(k)}_{\text{Dual-CA}}(\cdot), \mathcal{H}^{(k)}_{\text{FFN}}(\cdot)\}_{k=1}^{K}$.

4. Experiments

4.1. Implementaion Details

For training, we use 400 images from the training and test dataset of BSD500 dataset [1]. The training images are cropped to 89600 patches of 96×96 pixel size with data augmentation following [40]. For a given CS ratio $\frac{M}{N}$, the corresponding learnable measurement matrix $\mathbf{\Phi}$ is constructed by a convolution layer with the kernel size of $M\times1\times\sqrt{N}\times\sqrt{N}$ to sample the original image $\mathbf{x}$. And then, we utilize the transpose convolution whose kernel weight is the sampling matrix to obtain initialization $\mathbf{x}^{(0)}$.

For the network parameters, the block size is 32, *i.e.* $N = 1,024$, the default batch size is 16, the default number of feature maps C is 32 and the learnable parameter $\rho^{(k)}$ is initialized to 0.5. We use Adam [21] optimizer to train the network with the initial learning rate, which decreased to 5×10^{-5} through 100 epochs using the cosine annealing strategy [10, 30] and the warm-up epochs are 3. For testing, we utilize two widely-used benchmark datasets, yielding Set11 [24] and Urban100 [13]. Color images are processed in the YCbCr space and evaluated on the Y channel. Two common-used image assessment criteria, Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM), are adopted to evaluate the reconstruction results.

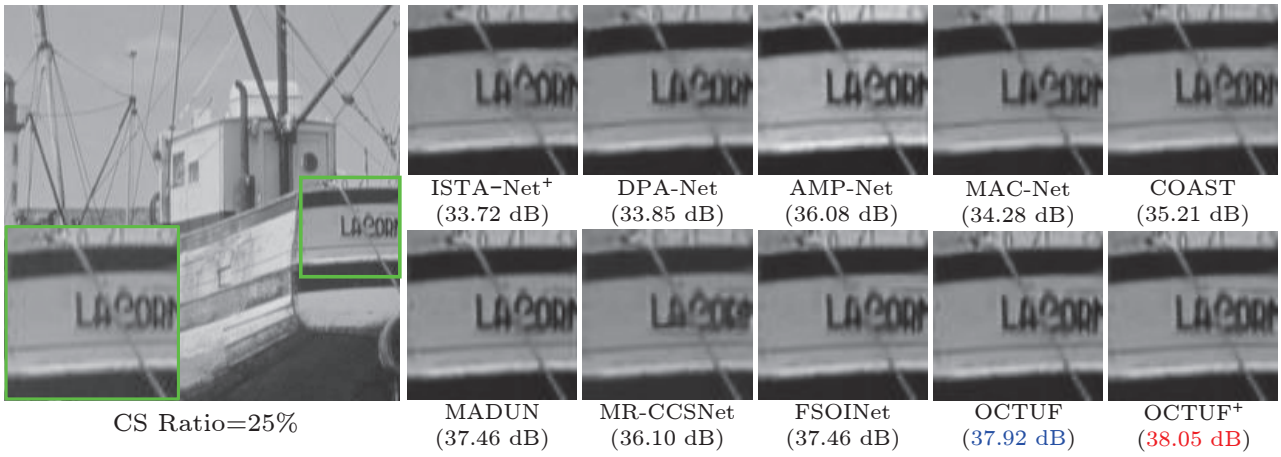


Figure 4. Comparisons on recovering an image from Set11 dataset [24] in the case of CS ratio = 25%.

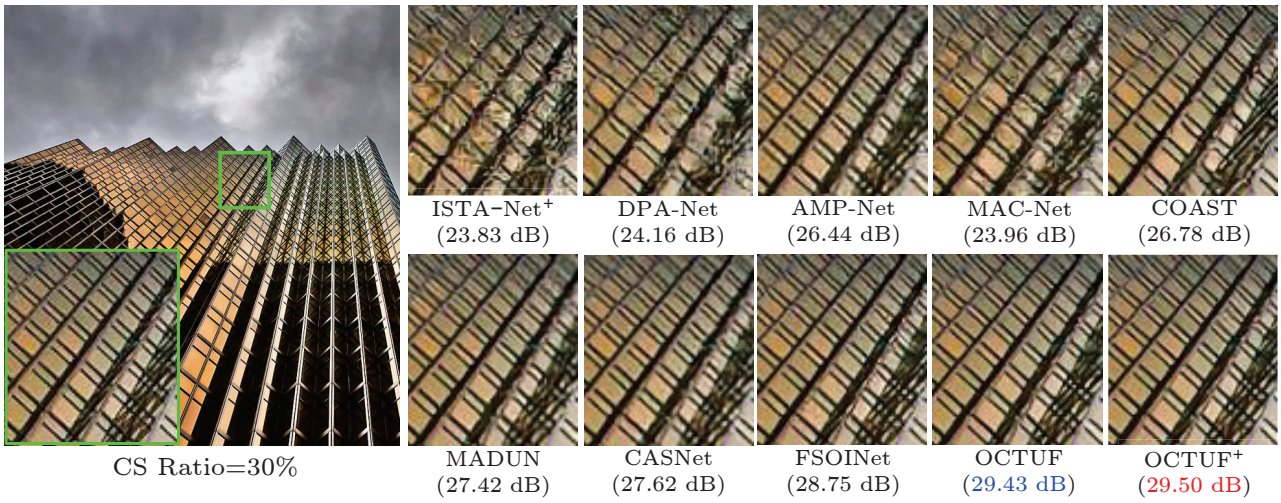


Figure 5. Comparisons on recovering an image from Urban100 dataset [13] in the case of CS ratio = 30%.

We also show the number of parameters and the computation cost (including the computations of the convolution, the fully-connected layer and matrix multiplication) measured in floating-point operations per second (FLOPs).

4.2. Qualitative Evaluation

We compare our proposed methods with recent representative CS reconstruction methods. The average PSNR/SSIM reconstruction performances on Set11 dataset [24] with respect to five CS ratios are summarized in Tab. 1. For our OCTUF, we set the iteration number as 10 and set the initial learning rate as 5×10^{-4}. To further improve the model performance, we also present a plus version, namely OCTUF$^+$, whose iteration number is 16 and initial rate is 2×10^{-4}. From Tab. 1, one can observe that our OCTUF and OCTUF$^+$ outperform all the other competing methods in PSNR and SSIM across all the cases. On average, OCTUF$^+$ outperforms ISTA-Net$^+$ [54], DPA-Net [44], AMP-Net [60], MAC-Net [19], COAST [53], MADUN [41], CASNet [7], TransCS [39] and FSOINet [10] by 3.60 dB, 3.91 dB, 1.25 dB, 3.36 dB, 2.21 dB, 0.51 dB, 0.41 dB, 1.16 dB and 0.29 dB in terms of PSNR on Set11 dataset, respectively. In addition, the average SSIM of OCTUF$^+$ can be improved 0.0389, 0.0280, 0.0113, 0.0381, 0.0202, 0.0013, 0.0015, 0.0073 and 0.0010, respectively. Fig. 4 further show the visual comparisons on challenging images when CS ratio is 25%, which can be seen that our OCTUFs can recover much clear edge information than other methods.

Furthermore, in Tab. 2, we compare OCTUFs with other methods on Urban100 dataset [13] that contains more high-resolution images and incorporates more abundant image distributions. It shows that both OCTUF and OCTUF$^+$ achieve a better reconstruction quality at all sampling ratios. Fig. 5 presents the visual comparisons on challenging images. Our OCTUF and OCTUF$^+$ generate images that are visually pleasant and faithful to the groundtruth. It

Table 2. Average PSNR(dB)/SSIM performance comparisons of recent deep network-based CS methods on Urban100 dataset [13] with different CS ratios. The best and second best results are highlighted in red and blue colors, respectively.

Dataset	Methods	CS Ratio					
		10%	25%	30%	40%	50%	Average
Urban100	ISTA-Net$^+$ (CVPR 2018) [54]	23.61/0.7238	28.93/0.8840	30.21/0.9079	32.43/0.9377	34.43/0.9571	29.92/0.8821
	DPA-Net (TIP 2020) [44]	24.55/0.7841	28.80/0.8944	29.47/0.9034	31.09/0.9311	32.08/0.9447	29.20/0.8915
	AMP-Net (TIP 2020) [60]	26.04/0.8151	30.89/0.9202	32.19/0.9365	34.37/0.9578	36.33/0.9712	31.96/0.9202
	MAC-Net (ECCV 2020) [9]	24.21/0.7445	28.79/0.8798	29.99/0.9017	31.94/0.9272	34.03/0.9513	29.79/0.8809
	COAST (TIP 2021) [53]	25.94/0.8035	31.10/0.9168	32.23/0.9321	34.22/0.9530	35.99/0.9665	31.90/0.9144
	MADUN (ACM MM 2021) [41]	27.13/0.8393	32.54/0.9347	33.77/0.9472	35.80/0.9633	37.75/0.9746	33.40/0.9318
	CASNet (TIP 2022) [7]	27.46/0.8616	32.20/0.9396	33.37/0.9511	35.48/0.9669	37.45/0.9777	33.19/0.9394
	TransCS (TIP 2022) [39]	26.72/0.8413	31.72/0.9330	31.95/0.9483	35.22/0.9648	37.20/0.9761	32.56/0.9327
	FSOINet (ICASSP 2022) [10]	27.53/0.8627	32.62/0.9430	33.84/0.9540	35.93/0.9688	37.80/0.9777	33.54/0.9412
	OCTUF (Ours)	27.79/0.8621	32.99/0.9445	34.21/0.9555	36.25/0.9669	38.29/0.9797	33.91/0.9423
	OCTUF$^+$ (Ours)	27.92/0.8652	33.08/0.9453	34.27/0.9559	36.31/0.9700	38.28/0.9795	33.97/0.9432

Table 3. Ablation study of our approach on Set11 dataset [24] in the case of CS ratio $= 50\%$. The best performance is labeled in bold.

Cases	Dual-CA	FFN	LayerNorm	Learning rate	PSNR(dB)	SSIM	Parameters
(a)	-	-	-	5e-4(warmup)	38.25	0.9759	0.72 M
(b)	-	√	√	5e-4(warmup)	38.96	0.9783	0.72 M
(c)	√	-	√	5e-4(warmup)	41.16	0.9834	0.82 M
(d)	√	√	-	5e-4(warmup)	41.21	0.9834	0.82 M
(e)	√	√	√	1e-4(fix)	41.17	0.9834	0.82 M
(f)	√	√	√	2e-4(fix)	41.27	0.9836	0.82 M
(g)	√	√	√	5e-4(fix)	41.30	0.9837	0.82 M
OCTUF	√	√	√	5e-4(warmup)	**41.34**	**0.9838**	0.82 M

should be noted that the images on Urban100 dataset do not satisfy a special constraint of MR-CCSNet [16] that all the image sizes must be divisible by 4, so the performance of MR-CCSNet is only presented on Set11 dataset.

4.3. Ablation Study

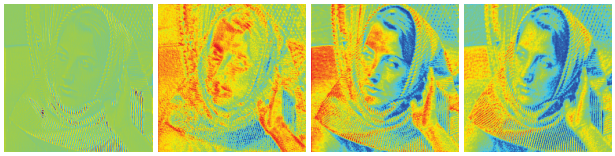
In this part, we conduct ablation studies on Set11 dataset for our OCTUF whose iteration number is 10.

Break-down Ablation. We first conduct a break-down ablation experiment in the case of CS ratio $= 50\%$ to investigate the effect of each component towards higher performance. The results are listed in Tab. 3. Case (a) is our baseline which contains ResBlock [18] with a similar number of parameters as OCT module. When we successively apply our FFN and Dual-CA sub-modules respectively, namely Cases (b) and (c), the model achieves 0.71 dB and 2.91 dB improvements. And the model can greatly enhance 3.09 dB gains with little storage place when both sub-modules are used together. We also discuss the effect of the LayerNorm (LN) function in Case (d), which addresses that our OCTUF achieves better performance with the LayerNorm function. Note that without "LayerNorm" represents removing all LN from our OCTUF. What is more, we train our models with different learning rates as seen from Cases (e), (f), and (g). "fix" denotes that the learning rate is not changed during training, and "warmup" denotes that the training strategy is the same with our work as shown in Sec. 4.1. Our proposed OCTUF has a stable training process with large learning rates and meanwhile, we use the "warmup" strategy to improve its anytime performance when training.

Table 4. Ablation of Dual-CA sub-module on Set11 dataset [24] when CS ratio is 30%. "IF" denotes the inertial force achieved by the easy way and "FD" denotes the enhanced iterative process in the feature domain. The best PSNR(dB) is labeled in bold.

Cases	FFN	GDB	IF	FD	PGCA	ISCA	PSNR
(a)	√	-	-	-	-	-	34.59
(b)	√	√	-	-	-	-	35.93
(c)	√	√	-	√	-	-	36.82
(d)	√	√	√	√	-	-	36.83
(e)	√	√	-	√	-	√	37.13
(f)	√	-	-	√	√	-	37.08
OCTUF	√	-	-	√	√	√	**37.21**

Dual Cross Attention. We also do elaborate ablation experiments on the components of Dual-CA sub-module in the case of CS ratio $= 30\%$ in Tab. 4, where "IF" denotes the inertial force achieved by a simple way similar to Eq. (3) and "FD" represents that the overall iteration process is achieved in feature domain. Case (b) achieves 1.34 dB improvement compared with Case (a), which proves the superiority of DUN compared with the structure that only



w/o Dual-CA w/o PGCA w/o ISCA OCTUF

Figure 6. Visual analysis of the feature map in the fifth iteration of our proposed OCTUF. It shows that both ISCA and PGCA blocks pay more high-fidelity attention to details.

Table 5. Ablation of Feed-Forward Network on Set11 dataset [24] when CS ratio = 30%. The best performance is labeled in bold.

Method	Baseline	LN+FFB	LN+2×FFB	Ours
PSNR(dB)	37.05	37.04	37.12	**37.21**
SSIM	0.9663	0.9664	0.9672	**0.9673**
Parameters	0.61 M	0.52 M	0.61 M	0.61 M

contains a neural network. Then, the performance can continuously improve by 0.89 dB after using "FD" shown in Case (c). We also conduct fine contrast experiments for the inertial force in Cases (c), (d), and (e), demonstrating that our ISCA block can more fully play the role of inertia force. And as shown in Cases (e)(f), applying PGCA and ISCA blocks can get better performance. Our proposed Dual-CA sub-module takes into account the gradient descent algorithm and the inertial force, allocates the different channels reasonably, and gives full play to the structural characteristics. Tab. 4 should be also noted that "GDB" is included in "PGCA" so it is not selected when "PGCA" is selected. Moreover, to intuitively show the advantages of Dual-CA sub-module, we visualize the feature map in the fifth iteration for four cases. The result in Fig. 6 presents that both our ISCA and PGCA blocks pay more high-fidelity attention to the detailed contents and structural textures.

Feed-Forward Network. As shown in Fig. 3 (e), our proposed Feed-Forward Network (FFN) sub-module consists of two groups of LayerNorm and Feed-Forward Block (FFB) with a global skip connection. We do ablations to investigate the effects of the group number and LayerNorm (LN) number in Tab. 5. "Baseline" is the same setting with Case (c) of Tab. 3, "LN+FFB" denotes one group, and "LN+2×FFB" denotes that Norm is only added to the first group. Therefore, as can be seen from the table, our proposed method achieves the best performance.

4.4. Complexity Analysis

The computation cost and model size are important in many practical applications. Tab. 6 provides the comparisons of the parameters, the model size, and FLOPs for reconstructing a 256×256 image when CS ratio is 10%. CASNet [7] designs a complex sampling network and reconstruction network, which has a large number of parameters and computational overhead. Our OCTUFs have the same parameters/cost for the sampling process with MADUN and

Table 6. Comparison of the parameters, the model size and FLOPs for reconstructing a 256×256 image in the case of CS ratio = 10%. The best performance is labeled in bold.

Method	MADUN	CASNet	FSOINet	OCTUF	OCTUF$^+$
Params.(M)	3.14	16.90	0.64	**0.40**	0.58
Size(MB)	11.9	66.3	7.8	**5.2**	7.5
FLOPs(G)	419.2	13391.5	266.6	**189.3**	294.6

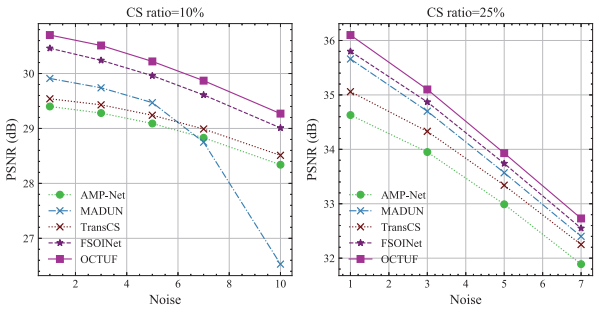


Figure 7. Comparison of robustness to Gaussian noise.

FSOINet, but use fewer parameters and less computation burden to produce much sharper recovered images.

4.5. Sensitivity to Noise

In the real application, the imaging model may be affected by noise and currently, no real open dataset is suitable for such CS reconstruction methods. So to test the robustness of our designed OCTUF, we first add the Gaussian noise with different noise levels to the original images for Set11 dataset. Then, OCTUF and the other methods take the noisy images as input, and sample and recover the reconstruction images when CS ratios are 10% and 25%. Fig. 7 shows the plots of PSNR values of all methods versus various standard variances noise. It can be seen that our OCTUF possesses strong robustness to noise corruption.

5. Conclusion

In this paper, we propose a novel optimization-inspired cross-attention Transformer (OCT) module as an iteration, leading to a lightweight OCT-based unfolding framework (OCTUF) for CS. Specifically, we present a Dual Cross Attention (Dual-CA) sub-module, which contains an inertia-supplied cross attention (ISCA) block and a projection-guided cross attention (PGCA) block as the projection step in iterative optimization. ISCA block precisely helps to achieve a good convergence by introducing a feature-level inertial term. PGCA block utilizes a cross attention mechanism to fuse the gradient descent term and the inertial term while ensuring the maximum information flow. Extensive experiments show that our OCTUF achieves superior performance compared to state-of-the-art methods with lower complexity. In the future, we will extend our OCTUF to other image inverse problems and video applications.

References

[1] Pablo Arbelaez, 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, 2010. 5

[2] Radu Ioan Boţ, Ernö Robert Csetnek, and Szilárd Csaba László. An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions. *EURO Journal on Computational Optimization*, 4(1):3–25, 2016. 3

[3] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[4] Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Degradation-aware unfolding half-shuffle Transformer for spectral compressive imaging. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3

[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] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with Transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[7] Bin Chen and Jian Zhang. Content-aware scalable deep compressed sensing. *IEEE Transactions on Image Processing*, 31:5412–5426, 2022. 1, 2, 5, 6, 7, 8

[8] 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. 2

[9] Jiwei Chen, Yubao Sun, Qingshan Liu, and Rui Huang. Learning memory augmented cascading network for compressed sensing of images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 5, 7

[10] Wenjun Chen, Chunling Yang, and Xin Yang. FSOINET: feature-space optimization-inspired network for image compressive sensing. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 1, 2, 5, 6, 7

[11] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1256–1272, 2016. 2

[12] Patrick L Combettes and Valérie R Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005. 3

[13] Weisheng Dong, Peiyao Wang, Wotao Yin, Guangming Shi, Fangfang Wu, and Xiaotong Lu. Denoising prior driven deep neural network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2305–2318, 2018. 5, 6, 7

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 2

[15] Marco F Duarte, Mark A Davenport, Dharmpal Takhar, Jason N Laska, Ting Sun, Kevin F Kelly, and Richard G Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):83–91, 2008. 1

[16] Zi-En Fan, Feng Lian, and Jia-Ni Quan. Global sensing and measurements reuse for image compressed sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 5, 7

[17] Xinwei Gao, Jian Zhang, Wenbin Che, Xiaopeng Fan, and Debin Zhao. Block-based compressive sensing coding of natural images by local structural measurement matrix. In *Proceedings of Data Compression Conference (DCC)*, 2015. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[19] Sheena A Josselyn and Susumu Tonegawa. Memory engrams: Recalling the past and imagining the future. *Science*, 367(6473), 2020. 1, 6

[20] 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

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 5

[22] Filippos Kokkinos and Stamatios Lefkimmiatis. Deep image demosaicking using a cascade of convolutional residual denoising networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[23] Jakob Kruse, Carsten Rother, and Uwe Schmidt. Learning to push the limits of efficient FFT-based image deconvolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[24] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 5, 6, 7, 8

[25] Stamatios Lefkimmiatis. Non-local color image denoising with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[26] 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(3):507–530, 2013. 2

[27] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[28] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[29] Antoine Liutkus, David Martina, Sébastien Popoff, Gilles Chardon, Ori Katz, Geoffroy Lerosey, Sylvain Gigan, Laurent Daudet, and Igor Carron. Imaging with nature: Compressive imaging using a multiply scattering medium. *Scientific Reports*, 4:5552, 2014. 1

[30] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 5

[31] Michael Lustig, David Donoho, and John M Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007. 1

[32] 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

[33] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003. 3

[34] Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock. iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014. 3

[35] Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, and Luc Soler. U-Net Transformer: Self and cross attention for medical image segmentation. In *Proceedings of the International Workshop on Machine Learning in Medical Imaging*, 2021. 2

[36] Chao Ren, Xiaohai He, Chuncheng Wang, and Zhibo Zhao. Adaptive consistency prior based deep network for image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[37] Florian Rousset, Nicolas Ducros, Andrea Farina, Gianluca Valentini, Cosimo D'Andrea, and Françoise Peyrin. Adaptive basis scan by wavelet prediction for single-pixel imaging. *IEEE Transactions on Computational Imaging*, 3(1):36–46, 2016. 1

[38] Aswin C Sankaranarayanan, Christoph Studer, and Richard G Baraniuk. CS-MUVI: Video compressive sensing for spatial-multiplexing cameras. In *Proceedings of the IEEE International Conference on Computational Photography (ICCP)*, 2012. 1

[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*, 2022. 1, 2, 3, 5, 6, 7

[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. 5

[41] Jiechong Song, Bin Chen, and Jian Zhang. Memory-augmented deep unfolding network for compressive sensing. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2021. 1, 2, 5, 6, 7

[42] Jiechong Song, Bin Chen, and Jian Zhang. Deep memory-augmented proximal unrolling network for compressive sensing. *International Journal of Computer Vision*, pages 1–20, 2023. 2

[43] Yueming Su and Qiusheng Lian. iPiano-Net: Nonconvex optimization inspired multi-scale reconstruction network for compressed sensing. *Signal Processing: Image Communication*, 89:115989, 2020. 2

[44] Yubao Sun, Jiwei Chen, Qingshan Liu, Bo Liu, and Guodong Guo. Dual-path attention network for compressed sensing image reconstruction. *IEEE Transactions on Image Processing*, 29:9482–9495, 2020. 1, 2, 5, 6, 7

[45] T. P. Szczykutowicz and G. Chen. Dual energy CT using slow kVp switching acquisition and prior image constrained compressed sensing. *Physics in Medicine & Biology*, 55(21):6411, 2010. 1

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 2

[47] Haixin Wang, Tianhao Zhang, Muzhi Yu, Jinan Sun, Wei Ye, Chen Wang, and Shikun Zhang. Stacking networks dynamically for image restoration based on the plug-and-play framework. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[48] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with Transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[49] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[50] Zhuoyuan Wu, Jian Zhang, and Chong Mou. Dense deep unfolding network with 3D-CNN prior for snapshot compressive sensing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[51] Zhuoyuan Wu, Zhenyu Zhang, Jiechong Song, and Jian Zhang. Spatial-temporal synergic prior driven unfolding network for snapshot compressive imaging. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2021. 1

[52] Di You, Jingfen Xie, and Jian Zhang. ISTA-Net^{++}: Flexible deep unfolding network for compressive sensing. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2021. 2

[53] Di You, Jian Zhang, Jingfen Xie, Bin Chen, and Siwei Ma. COAST: Controllable arbitrary-sampling network for compressive sensing. *IEEE Transactions on Image Processing*, 30:6066–6080, 2021. 1, 2, 5, 6, 7

[54] Jian Zhang and Bernard Ghanem. ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 5, 6, 7

[55] 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

[56] Jian Zhang, Chen Zhao, Debin Zhao, and Wen Gao. Image compressive sensing recovery using adaptively learned sparsifying basis via L0 minimization. *Signal Processing*, 103:114–126, 2014. 2

[57] 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

[58] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[59] Zhilin Zhang, Tzyy-Ping Jung, Scott Makeig, and Bhaskar D Rao. Compressed sensing for energy-efficient wireless telemonitoring of noninvasive fetal ECG via block sparse bayesian learning. *IEEE Transactions on Biomedical Engineering*, 60(2):300–309, 2012. 1

[60] 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. 1, 2, 5, 6, 7

[61] Chen Zhao, Siwei Ma, and Wen Gao. Image compressive-sensing recovery using structured laplacian sparsity in DCT domain and multi-hypothesis prediction. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2014. 2

[62] Chen Zhao, Siwei Ma, Jian Zhang, Ruiqin Xiong, and Wen Gao. Video compressive sensing reconstruction via reweighted residual sparsity. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(6):1182–1195, 2016. 2

[63] Chen Zhao, Jian Zhang, Siwei Ma, and Wen Gao. Nonconvex Lp nuclear norm based ADMM framework for compressed sensing. In *Proceedings of Data Compression Conference (DCC)*, 2016. 2

[64] Chen Zhao, Jian Zhang, Ronggang Wang, and Wen Gao. CREAM: CNN-REgularized ADMM framework for compressive-sensed image reconstruction. *IEEE Access*, 6:76838–76853, 2018. 2

[65] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 4

[66] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 2