A Hybrid Frequency-Spatial Domain Model for Sparse Image Reconstruction in Scanning Transmission Electron Microscopy

Bintao He[1], Fa Zhang[2], Huanshui Zhang[1,*], Renmin Han[1,*]
[1]Research Center for Mathematics and Interdisciplinary Sciences, Shandong University
[2]High Performance Computer Research Center, ICT, CAS

Abstract

Scanning transmission electron microscopy (STEM) is a powerful technique in high-resolution atomic imaging of materials. Decreasing scanning time and reducing electron beam exposure with an acceptable signal-to-noise ratio are two popular research aspects when applying STEM to beam-sensitive materials. Specifically, partially sampling with fixed electron doses is one of the most important solutions, and then the lost information is restored by computational methods. Following successful applications of deep learning in image in-painting, we have developed an encoder-decoder network to reconstruct STEM images in extremely sparse sampling cases. In our model, we combine both local pixel information from convolution operators and global texture features, by applying specific filter operations on the frequency domain to acquire initial reconstruction and global structure prior. Our method can effectively restore texture structures and be robust in different sampling ratios with Poisson noise. A comprehensive study demonstrates that our method gains about 50% performance enhancement in comparison with the state-of-art methods. Code is available at https://github.com/icthrm/Sparse-Sampling-Reconstruction.

1. Introduction

Scanning Transmission Electron Microscopy(STEM) has become a powerful and successful technique in the imaging of "beam stable" materials such as inorganic crystalline samples. However, to achieve high spatial resolution better than 0.5 Å [2, 10], an order of magnitude for electron beam doses (typically in excess of $10^5 - 10^6$ $e^-/Å^2$) is necessary, in which the high-energy electron beam may burn the materials and destroy the original structures. Therefore, the imaging capability of a STEM technique at low electron dose is critical for beam-sensitive materials. By decreasing the dwell time at a pixel (scan faster) or directly reducing the electron accelerator voltage [18], we can achieve low electron dose imaging with a reduced number of electrons dashing and passing through samples in unit time [3]. However, the low electron dose configuration leads to another problem, i.e., the sparsity of the sampling signals and low signal-to-noise ratio.

Because the electron number per pixel determines the credibility of a pixel, given total electron doses, sparse sampling makes a pixel acquire more electrons, leading to a more credible pixel value. Meanwhile, for a STEM system with 200-300 keV primary beam energy, the Poisson noise dominates the noise distribution [13, 25, 14]. A one-pass sparse sampling (or partial scanning) outputs a scatter map of the true signals of a sample. Except for partial scanning, multi-pass scanning strategies are developed to further reduce electron beam damage. To obtain the structure details, restoring missing signals from the scatter map is necessary.

Traditional reconstruction methods in sparse sampling develop from basic frequency filter to compressed sensing framework. Fourier or wavelets transform combined with amplitude filter, frequency filter, and phase drift can roughly restore the repeating structures [20, 26]. But filter methods are only applicable for materials with a large range of fixed texture structures, and they can't accurately determine the material boundaries, let alone any discrepancy between the internal structures. Compressed sensing (CS) theory [8, 5, 4] offers an alternative idea to overcome the limitation. CS makes the assumption that a set of signals is able to be represented by a suitable basis in an extremely sparse form if the system satisfies several preliminaries. And if the sensing matrix obeys i.i.d Bernoulli distribution given a sparse rate, CS theory guarantees the feasibility and efficiency of image restoration. To achieve satisfying results, Traditional CS methods such as Group-based Sparse Representation (GSR) [28] and Beta Process Factor Analysis (BPFA) [29] require almost a dozen hours' execution time, which can't meet the practical need of real-time imaging.

Due to the limitation of execution speed, convolution neural network (CNN) based methods are recently proposed [16, 1, 23, 11, 24, 27, 19]. Well-trained networks

*All correspondence should be addressed to Huanshui Zhang (hszhang@sdu.edu.cn) and Renmin Han (hanrenmin@sdu.edu.cn).

have replaced the "endless" iterations with just one forward propagation, making the real-time imaging of low-dose STEM possible. These methods mainly inpaint the missing information by exploring the texture features from local patches. However, without strict theoretical guarantee, CS-based deep learning methods perform not well on inpainting problem in extremely sparse sampling cases, especially for real-world data with Poisson noise. Meanwhile, most networks adopt block-based linear mapping as initial reconstruction, which significantly limits receptive field and loses the ability of global information extraction.

Here, we propose a novel Frequency-Spatial Hybrid Network (FSHNet) to restore STEM images at an extremely low sampling rate, for example, atomic-scale STEM imaging with a sampling rate lower than 5%, which is impossible to be achieved by traditional methods. In FSHNet, the frequency domain information is filtered to ensure global similarity, and the detailed spatial domain information is captured with convolution operators to polish the local structure. By combining the global structure features from frequency filter and the local pixel information from convolution operators, FSHNet can achieve a complete structure restoration with clearer local details. Comprehensive experiments on the synthetic and real-world datasets show that our method gains $\sim$50% performance enhancement.

In summary, our main contributions are:

- A novel architecture that is able to utilize both the global structure feature and local pixel information in image inpainting.
- An approach to define a structure prior from frequency-domain to guide the inpainting.
- Introducing an adaptive non-local patches matching module to enhance image inpainting performance and alleviate irregular artifact.
- A general procedure for the simulation of STEM sampling and imaging.

2. Related Work

2.1. Traditional Compressed Sensing Algorithms

Supposing that $\boldsymbol{y} \in \mathbb{R}^{M\times 1}$ is a compressed measurements, and $\boldsymbol{\Phi} \in \mathbb{R}^{M\times N}$ ($M \ll N$) is a sampling matrix, the classical compressed sensing problem is expressed as

$$\arg\min_{\boldsymbol{x}} ||\boldsymbol{y} - \boldsymbol{\Phi x}||_2, \quad s.t \quad ||\boldsymbol{\Psi x}||_0 \leqslant S \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^{N\times 1}$ is a real-valued signal to be solved, $\boldsymbol{\Psi}$ is a mapping operator that transforms $\boldsymbol{x}$ into another space, and S is the given sparse degree. Because the solving of non-convex optimization is not trivial, Donoho et al. proposed the Basis Persuit algorithm [8], in which the loss function can be reduced into a discrepancy item and some regularization items after a simple Lagrangian multiplier transform. Many works exploited additional prior knowledge to improve CS reconstruction performance, for example, the non-local self-similarity property in natural images [7, 14] and the structure sparse property of transformed coefficients [12, 22, 28]. Although the strict mathematical proof resided in the CS theory, the hypothesis that allows CS theory be applied into an imaging system is not always exactly satisfied, and the information of image alone is not enough to restore detailed features.

2.2. Leaning-based CS Algorithms

Dictionary learning is proposed to handle the problem where the fixed-domain methods (e.g. DCT, DWT, gradient difference) fail. The target of dictionary learning is to find an over-complete dictionary $\boldsymbol{\mathcal{D}}$ of the given images, thereby, a sparse coding vector $\boldsymbol{w}$ can well represent the raw images:

$$\boldsymbol{y} = \boldsymbol{\Phi x} = \boldsymbol{\Phi \mathcal{D} w}. \tag{2}$$

The Beta process factor analysis (BPFA) proposed by [29] is the most widely used CS method. Stevens applied it to STEM by splitting images into $B \times B$ overlapping patches and design special sensing matrix $\boldsymbol{\Phi}$ ($\boldsymbol{\Phi} = [0|e_2|0|e_4|...|e_n|0]^T$) consisting of zero rows and several selected identity matrix rows [21].

Neural network is well known for its powerful feature learning capability. Benefiting from the fast non-iterative forward inference, a lot of efforts have been made to replace the time-consuming iterative optimization in CS by deep learning approach. Mousavi proposed the stacked denoising auto-encoder to fit non-linear transform of signal vectors [16]. Kulkarni utilizes convolution neural networks to extract deep features and directly output restored image [11]. Yang [23] and Zhang [27] recently integrated CNN modules into ADMM and ISTA [6] algorithms, respectively. Deep residual reconstruction network was proposed to reconstruct a high-quality preliminary image by introducing residual module [24]. Jeffrey firstly introduced deep residual adversarial learning method in STEM sparse sampling reconstruction with spiral scanning [9]. However, similar to the previous works, it neglects the global information.

3. Methodology

3.1. Reconstruction Model

Given the sampling inputs $\{\boldsymbol{Y}_n\}$ with $\boldsymbol{Y}_n \in \mathbb{R}^{M\times M}$ and $n = 1, 2..., T$, where T is the inputted frame number, the reconstruction problem in partial scanning STEM is formulated as follows:

$$\arg\min_{\boldsymbol{X}} \sum_{n=1}^{T} \|\boldsymbol{Y}_n - \boldsymbol{\Phi}_n \odot \boldsymbol{X}\|_2 + \|\boldsymbol{\Psi X}\|_1, \tag{3}$$

where $\odot$ represent Hadamard product, $\boldsymbol{X} \in \mathbb{R}^{M\times M}$ is the unknown image to be solved, $\boldsymbol{\Phi}_n \in \mathbb{R}^{M\times M}$ is the sampling

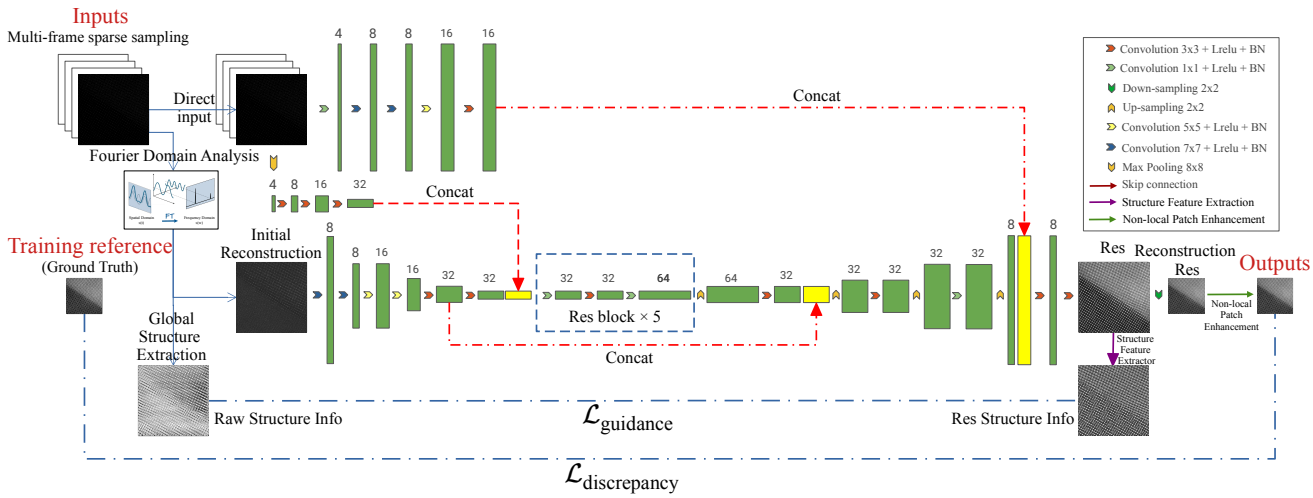

Figure 1: The architecture of FSHNet. The workflow starts with a bicubic interpolation, on which further operations are applied. The main encoder-decoder framework accepts the initial reconstruction based on low-pass filtered results, and utilizes global structure guidance prior extracted from frequency filtered results to guide training.

matrix and $\boldsymbol{\Psi X}$ denotes the transform coefficients matrix of $\boldsymbol{X}$ with respect to transform $\boldsymbol{\Psi} \in \mathbb{R}^{M\times M}$. The discrepancy item for the n^{th} sampling can be further rewritten as

$$\|\boldsymbol{Y}_n - \boldsymbol{\Phi}_n \odot \boldsymbol{X}\|_2 = \|\boldsymbol{Y}_n - \mathcal{P}(\boldsymbol{M}_n \odot \boldsymbol{X})\|_2, \quad (4)$$

where $\boldsymbol{M}_n \in \mathbb{R}^{M\times M}$ is a mask matrix consisted of 0 or 1 corresponding to the n^{th} partial sampling and $\mathcal{P}$ refers to the modulation of Poisson noise. Given the total electron dose Λ and sparsity k, the operator $\mathcal{P}$ is explicitly formulated as

$$\mathcal{P}(u) = \begin{cases} \frac{P(\lambda)\cdot u}{\lambda} & \text{if the pixel } u \text{ is selected} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where $P(\cdot)$ is the random variable subject to Poisson Distribution, and $\lambda = \Lambda/(k \cdot n^2)$.

The regularization item $\|\Psi \boldsymbol{X}\|_1$ is extended into two components in our model, to reflect the structure features serving as image reconstruction guidance and the inherent characteristics residing in the image.

Motivated by the successful application of frequency filter in low-dose image restoration, the global structure information could be discovered from the frequency domain. Here, we designed the regularization of structure features as

$$Re_s = \text{Diff}(\mathcal{S}(\sum_{n=1}^{T} \mathcal{F}(\boldsymbol{Y}_n)), \boldsymbol{X}), \quad (6)$$

where $\mathcal{F}$ is the Fourier transform, $\mathcal{S}$ is an operator to extract structure guidance information. The goodness of $\boldsymbol{X}$ is judged by the Diff$(\cdot)$ function in terms of structure features.

Meanwhile, considering the inherent property of the image itself, another regularization item is defined as

$$Re_i = \mathcal{R}(\boldsymbol{X}), \quad (7)$$

where $\mathcal{R}$ is selected to reflect the sparsity of transformed coefficients and local smoothness.

Thus, the reconstruction model for partial scanning STEM is extended as

$$\arg\min_{\boldsymbol{x}} \sum_{n=1}^{T} \|\boldsymbol{Y}_n - \mathcal{P}(\boldsymbol{M}_n \odot \boldsymbol{X})\|_2 + Re_s + Re_i. \quad (8)$$

3.2. Network Architecture

A Frequency-Spatial Hybrid Network (FSHNet) is designed according to the reconstruction model, based on an encoder-decoder architecture (shown in Figure 1). Different from the previous CS methods that clip image into a set of small patches, FSHNet accepts the complete image as input, for the better retainment of global information. The basic convolution block in our net consists of a Conv2D layer, a Leaky-Relu layer (k=0.2) and a batch norm layer. The up-sampling block in the decoder architecture adopts a 2× nearest interpolation, with the last up-sampling block utilizing a 2× pixel shuffle operation to shrink channel number.

In FSHNet, the inputted partial sampling data is duplicated and fed into two parallel processes, of which one directly shuffles the sparse sampling data in spatial domain with a 4-layer CNN, and the other makes a Fourier domain analysis (subsection 3.3.1) on the integrated Fourier Transform (FT) maps of the inputs. The Fourier domain analysis will reconstruct an initial STEM image and extract its global structure features. Then, the initial reconstruction is fed into the encoder layers to reduce data dimension, and combined with the shuffled sparse sampling data. Finally, under the guidance of global structure features, the decoder layers will output a fine-reconstructed STEM image, following with a non-local module (subsection 3.3.2) to suppress the effect of Poisson noise.

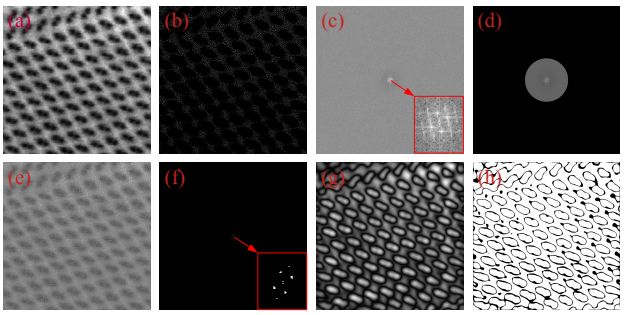

Figure 2: An illustration of global structure feature extraction (the zero-frequency is centralized for visualization). **(a)** Groundtruth. **(b)** A frame of partial sampling data. **(c)** Integrated FT map. **(d)** Frequency-filtered FT map. **(e)** Initial reconstruction. **(f)** Amplitude-filtered FT map. **(g)** Global structure features extraction. **(h)** Extracted feature after Laplace of Gaussian filter.

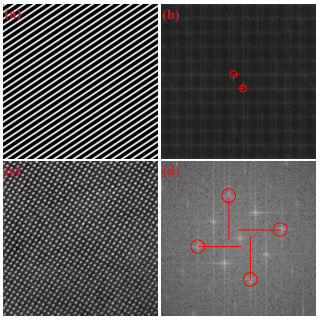

Figure 3: Adaptive patch sizes determination. **(a)** is an image plotted by $f(i,j) = \sin(\frac{48\pi \cdot i}{512} + \frac{32\pi \cdot j}{512})$. **(b)** is the FT map of (a). **(c)** is an example of STEM image. **(d)** is the FT of (c). The patch size used in non-local matching is determined by the maximum horizontal and vertical component of the selected peak, which is marked in red.

3.3. Main Component

3.3.1 Fourier domain analysis

Essentially, the Fourier transform looks for repeating structures. Thus, compared with spatial domain analysis, frequency domain analysis could provide more information for STEM image. Figure 2 gives out an example of the Fourier operations carried out on an FSHNet's input.

Basic concepts. Given an FT coefficient map $\boldsymbol{A} = [a_{ij}] \in \mathbb{R}^{M\times M}$ and a mask matrix $\boldsymbol{M} = [m_{ij}]$, the Fourier mask operation is defined as

$$\mathcal{M}(\boldsymbol{M}, \boldsymbol{A}) = \boldsymbol{M} \odot \boldsymbol{A}. \tag{9}$$

Here, a low-pass filter $\mathcal{H}_{fre}(\boldsymbol{A}, r) = \mathcal{M}(\boldsymbol{M}_{fre}, \boldsymbol{A})$ and an top-k amplitude filter $\mathcal{H}_{amp}(\boldsymbol{A}, h) = \mathcal{M}(\boldsymbol{M}_{amp}, \boldsymbol{A})$ are defined based on the Fourier mask operation. The mask matrix M_{fre} in $\mathcal{H}_{fre}(\cdot)$ is set with $m_{ij} = 1$ if and only if $i^2 + j^2 \leq r^2$, where r is a threshold to truncate frequency. The mask matrix M_{amp} in $\mathcal{H}_{amp}(\cdot)$ is set with $m_{ij} = 1$ if and only if $|a_{ij}|$ ranks top h of all the amplitude.

Initial Reconstruction. Though simple linear mapping is widely-used for initial reconstruction in traditional methods, it is not applicable in the case of extremely sparse sampling. In FSHNet, the initial reconstruction is generated by applying a low-pass filter on the integrated FT map of the inputted multi-frame data, i.e.,

$$\boldsymbol{I} = \mathcal{F}^{-1}(\mathcal{H}_{fre}(\sum_{n=1}^{T} \mathcal{F}(\boldsymbol{Y}_n), r)) \tag{10}$$

where r is a relatively large threshold for frequency truncation. Considering the elements in $\boldsymbol{I} = [a_{ij}]$ and $\boldsymbol{Y} = \sum_{n=1}^{T} \boldsymbol{Y}_n = [b_{ij}]$, we further let

$$a_{ij} = \begin{cases} a_{ij} & \text{if } b_{ij} = 0 \\ \frac{1}{2}(a_{ij} + b_{ij}) & \text{otherwise.} \end{cases} \tag{11}$$

to better interpret the pixels of the realistic sampling.

Global Structure Features Extraction. Low-frequency signals determine the main structure of images and strong amplitude signals dominate the image fluctuation. The fluctuation can be approximately regarded as the edge information in an image. However, due to the extremely sparsity of partial scanning STEM image, the edge information is impossible to be extracted in spatial domain. Given the inputs $\{\boldsymbol{Y}_n\}$, an amplitude-frequency joint filter is devised to extract the global structure, which is formulated as

$$\mathcal{S}(\mathcal{F}(\boldsymbol{Y})) = \mathcal{F}^{-1}(\mathcal{H}_{fre}(\mathcal{H}_{amp}(\sum_{n=1}^{T} \mathcal{F}(\boldsymbol{Y}_n), k), r_g)) \tag{12}$$

where k and r_g are parameters related to the imaging system. By applying the amplitude filter on the FTs and setting a relative small threshold, the attention of FSHNet is able to be maintained on the main structure of the image.

3.3.2 Non-local Patch Enhancement.

To suppress the effect of Poisson noise, a structure refining strategy is designed based on non-local patch matching and weighting. For a patch $\boldsymbol{P} = [p_{ij}] \in \mathbb{R}^{W\times W}$ locating in the local region $\mathcal{Q}$, its value is updated by

$$\boldsymbol{P} = \sum_{\boldsymbol{P}_n \in \mathcal{Q}} \left(\frac{\exp^{-\text{MSE}(\boldsymbol{P},\boldsymbol{P}_n)/h^2}}{\sum_{\boldsymbol{P}_n \in \mathcal{Q}} \exp^{-\text{MSE}(\boldsymbol{P},\boldsymbol{P}_n)/h^2}} \cdot \boldsymbol{P}_n\right), \tag{13}$$

where $\text{MSE}(\cdot)$ is the function to calcuate the mean squared error between two patches and h is a bandwidth parameter.

The patch size W is adaptively determined by selecting the k strongest amplitude peak within a maximum frequency range r_p and calculating their maximum horizontal and vertical component, i.e.,

$$W = \max\{i \vee j | \forall f_{ij} \in \boldsymbol{F} \wedge f_{ij} \neq 0\}, \tag{14}$$

where $\boldsymbol{F} = [f_{ij}] = \mathcal{H}_{amp}(\mathcal{H}_{fre}(\sum_{n=1}^{T} \mathcal{F}(\boldsymbol{Y}_n), r_p), k)$. Figure 3 gives out an example of patch size determination.

3.4. Loss Function

FSHNet accepts a multi-frame sparse sampling $\{\boldsymbol{Y}_n\}_{n=1}^T$ as input and outputs a reconstructed image $\boldsymbol{X}$. Given a training pair $\{\{\boldsymbol{Y}_n\}_{n=1}^T, \boldsymbol{Z}\}$, with $\boldsymbol{Z}$ representing the reference ground truth, FSHNet tunes the model by minimizing a combined loss function, i.e.,

$$\mathcal{L} = \mathcal{L}_{dis} + \lambda_1 \mathcal{L}_{gui} + \lambda_2 \mathcal{L}_{reg} \tag{15}$$

with

$$\begin{cases} \mathcal{L}_{dis} = \text{MSE}(\boldsymbol{Z}, \boldsymbol{X}) \\ \mathcal{L}_{gui} = \text{MSE}(\mathcal{T}_{edge}(\mathcal{S}(\mathcal{F}(\boldsymbol{Y}))), \mathcal{T}_{edge}(\boldsymbol{X})) \\ \mathcal{L}_{reg} = \mathcal{TV}(\boldsymbol{X}) \end{cases}, \tag{16}$$

where $\mathcal{L}_{dis}$ is a discrepancy loss to make fundamental comparison, $\mathcal{L}_{gui}$ is a structure loss based on the result of Global Structure Features Extraction and $\mathcal{L}_{reg}$ is a regularization term of total variation. Here, $\mathcal{T}_{edge}$ operator is a Laplace of Gaussian (LoG) filter combined with a binary mask, i.e.,

$$\mathcal{T}_{edge}(\cdot) = \boldsymbol{M} \odot \mathcal{T}_{LoG}(\cdot). \tag{17}$$

The element in matrix $\boldsymbol{M}$ is set 0 if and only if the corresponding value x in the image ≤ 0. By this, the network will pay more attention to the structure information.

4. Experiments

FSHNet is compared with the dictionary learning method BPFA, the deap learning method CSNet and ReconNet. Here, BPFA is well known as the most accurate CS method [15], while CSNet and ReconNet are the state-of-art CNN-based methods [19, 11]. These methods are tested on both synthetic datasets and real-world datasets[1].

4.1. Synthetic Data Generator

Generally, a real-world STEM image is usually degenerated by electronic noise, and the true sample structure of the STEM image is inaccessible. Nevertheless, the ground-truth is very important in model training and method comparison. Here, to prepare the sampling-clear training pairs, a synthetic data generator is designed, whose workflow is shown in Figure 4. 800 real-world STEM images of crystalline structures (512×512 pixels) are prepared for synthetic data generation. As shown in Figure 4, firstly, these images are smoothed by a Gaussian filter and an iterative median filter, to prepare a clean ground-truth dataset. Then, the filtered images are upsampled by a factor of 2, with a bicubic interpolation, to better simulate the sampling process. Finally, the upsampled images are sparse sampled by a block random scanning strategy, which makes random sampling in a block level, for example, random sampling 1 pixel within an 8×8 pixel block. Meanwhile, following Eq. 5, Poisson noise is applied to the sparse sampled images, to better simulate the real sampling process.

[1]Owing to space limitations, Section 4 here only presents partial results. More detailed results are provided in Supplementary Materials.

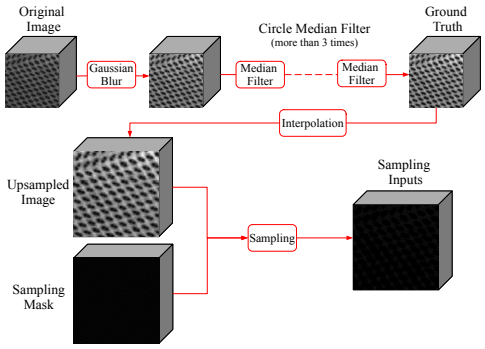

Figure 4: The workflow to generate synthetic data.

In this work, for each image in the clean ground-truth dataset, we sample four times to generate a multi-frame synthetic input. We have produced 6 of sparse sampling datasets under 6 different sampling ratios, i.e., 1.56%, 3.125%, 5%, 6.25%, 8% and 9.375%.

4.2. Network training details

FSHNet is implemented by PyTorch [17] and all the experiments are trained on the NVIDIA RTX 3080 GPU. The batch size is set to 4 with $4 \times 1024 \times 1024$ input size during total 100 epochs. The optimizer of model is Adam with default parameters and the learning rate is 0.0001. For Eq. 13, the control parameter are set as $h = 0.5$ if the sampling ratio 3%, otherwise $h = 1$. For the initial reconstruction, the max frequency domain threshold is set as $r_i = 150$. For the global structure features extraction, the max frequency domain threshold is set as $r_g = 100$ and the number of selected amplitude component is set as $k = 200$. For the non-local patch enhancement, the region of interest is set as $\mathcal{R} = 128 \times 128$ and the stride step is set to the quarter of patch sizes. The parameters in the loss function (Eq. 15) takes $\lambda_1 = 0.2$ and $\lambda_2 = 0.01$.

4.3. Evaluation on synthetic data

We challenged our method and the other three reconstruction methods on 45 synthetic data. To best exhibit the difference between these methods, we selected the synthetic data with a variety of structures under different sampling ratio. For BPFA, CSNet and ReconNet, we tried to reproduce the best results with the publicly released codes. To assess the performance, we calculated the PSNR/SSIM of the results for each synthetic data, whose average values are summarized in Table 1. Judging from Table 1, it can be found that all CNN-based methods perform better than the traditional CS method. Nevertheless, our method outperforms the other methods in terms of PSNR and SSIM.

Table 1: PSNR(dB)/SSIM results of methods on simulated datasets

Sampling ratio \ Method	Dose per Pixel	BPFA	CSNet	ReconNet	FSHNet
1.56%	10	9.973/0.357	18.28/0.638	15.70/0.472	25.22/0.818
	25	10.22/0.377	17.80/0.625	16.01/0.531	25.38/0.836
3.125%	10	11.03/0.458	18.38/0.652	15.98/0.506	25.85/0.847
	25	11.16/0.480	18.47/0.655	16.58/0.553	25.67/0.854
5%	10	11.26/0.485	18.74/0.664	15.98/0.506	26.08/0.856
	25	11.40/0.503	18.75/0.665	16.49/0.555	26.67/0.868
6.25%	10	11.14/0.486	18.44/0.657	15.79/0.505	26.45/0.867
	25	11.53/0.521	18.85/0.668	16.59/0.558	27.55/0.875
8%	10	11.29/0.491	18.77/0.664	16.39/0.527	26.34/0.874
	25	11.70/0.529	18.93/0.669	16.44/0.557	27.85/0.885
9.375%	10	11.27/0.487	18.77/0.662	16.44/0.527	27.24/0.882
	25	12.00/0.540	19.02/0.671	16.58/0.560	27.90/0.885

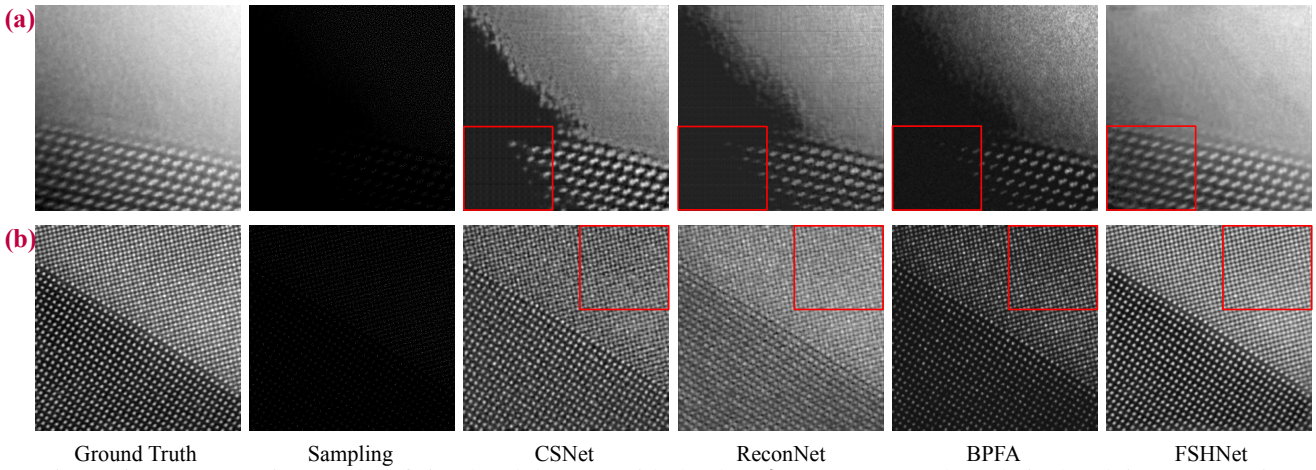

Figure 5: Reconstruction results of simulated datasets with the dose $\lambda = 10e^{-}$ per selected pixel and 5% sparse ratio.

When the electron dose is below a certain level, the true signal may lost in sampling process under the disturbance of Poisson noise (the left bottom in Figure 5a). On the other hand, when there are different periodic structures overlapped, the smaller structure may be blurred after inpainting (the right top in Figure 5b). Without frequency information, the methods like CSNet, ReconNet and BPFA were almost impossible to restore the complete structures, while FSHNet correctly reconstructed the structure. Here, interested reader are referred to Supplementary Materials for more detailed information.

4.4. Evaluation on real-world data

We further challenged our method on real-world data. Figure 6 shows an atom level microstructure of $NiTiO_3$, which is taken under microscope at 200kV and recorded on a High Angle Annular Dark Field (HAADF) detector at a dwell time of 60 μs. Similar with previous results, BPFA, CSNet and ReconNet restore blurred and aliasing structure, while our method inpaints most clear and complete result, within which the particle size and arrangement can be easily identified. Figure 7 shows the Annular Bright Field (ABF) image reconstruction of $SrTiO_3$ at 200kV. Our method produces a relatively clear boundary of the particles, while the other CNN-based methods only restore blurring ones with specious structure.

Table 2: Average execution time (in seconds) per image and number of parameters for the compared methods

Method	Runtime (s)	Number of parameters
CSNet	0.4824	428936
ReconNet	0.5518	40578
FSHNet	0.3788	161761
BPFA	74465	—†

† Because the BPFA is a dictionary learning method, its number of parameters is not comparable with the deep learning methods.

4.5. Model parameters and execution speed

Table 2 summarizes the average execution time of the above experiments and the number of model parameters for each method. All the deep learning based methods are run

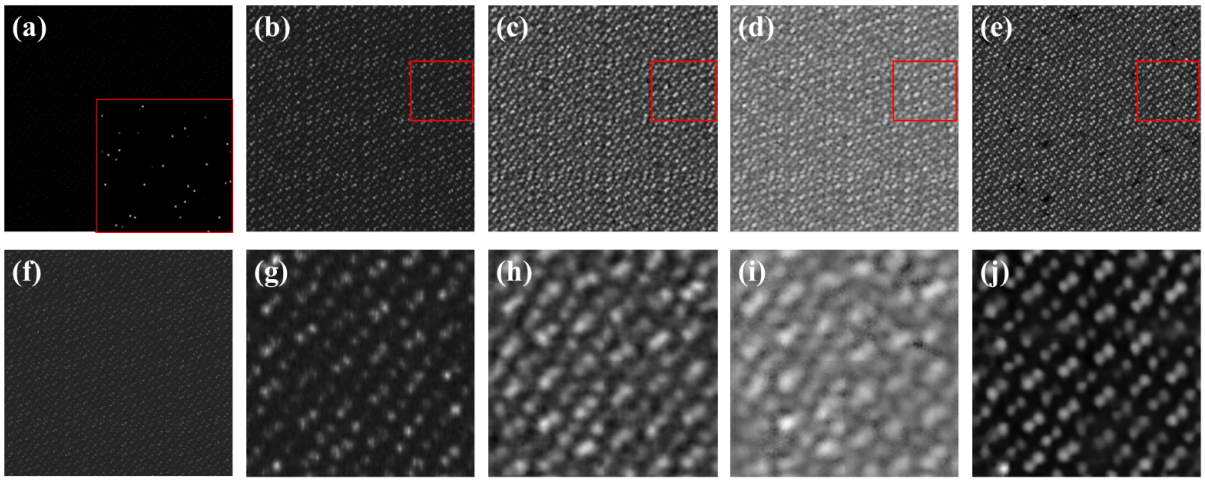

Figure 6: Reconstruction of $NiTiO_3$ under 5% sampling raio. **(a)** Sampling result; **(f)** Frequency filtered result; **(b)**, **(c)**, **(d)**, **(e)** Reconstruction result from BPFA, CSNet, ReconNet and FSHNet, respectively; **(g)**, **(h)**, **(i)**, **(j)** Local enlargement.

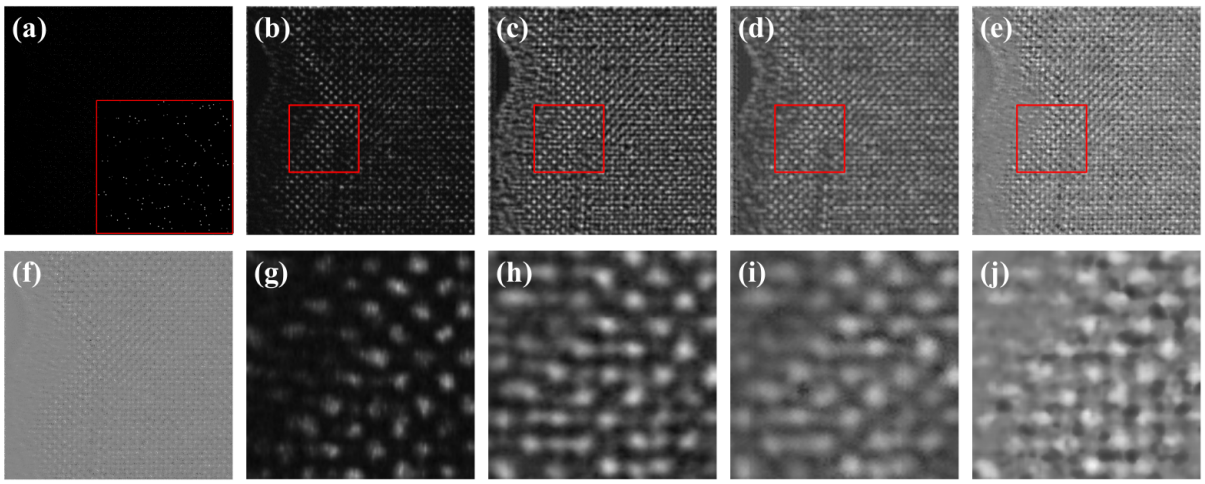

Figure 7: Reconstruction of $SrTiO_3$ under 5% sampling raio. **(a)** Sampling result; **(f)** Frequency filtered result; **(b)**, **(c)**, **(d)**, **(e)** Reconstruction result from BPFA, CSNet, ReconNet and FSHNet, respectively; **(g)**, **(h)**, **(i)**, **(j)** Local enlargement.

on the system same in network training and the BPFA is run on a system with Intel Core i9-9980 CPU. Here, it can be found that FSHNet has moderate parameters but runs fastest, while BPFA is the slowest. An execution time less than 0.5 seconds is possible for real-time data process.

4.6. Robust Study

In practice, sparse sampling in STEM is usually taken under an uncertain sampling ratio. Sometimes, one block may has more or fewer sampling pixels than configured, which makes the robustness of a learning method very important. Here, we trained the learning model with 5% sampling ratio dataset but tested the model with the data taken under different sampling ratios, of which results (average value of the test set) are demonstrated in Table 3. Though the performance of all the methods have enhanced with the increasing of sampling ratio, the performance of FSHNet improves the most.

4.7. Ablation Studies

We tested how the net behaves without non-local patch enhancement, structure prior and initial reconstruction. The numerical results are summarized in Table 4 and visual results are shown in Figure 8. Here, all the model are trained with the dataset under 5% sampling. The "init-recon" in Table 4 is referred to "initial reconstruction". More detailed information can be found in Supplementary Materials.

Study on non-local module. The second column of Table 4 shows that the method without non-local patch enhancement has a drop on PSNR and SSIM. The gap of PSNR shrinks from 1.4 dB to 0.3 dB and the gap of SSIM from 0.07 to 0.01 with the increase of sampling ratio. According to the comparison of Figure 8b&c, non-local patch matching is an effective way to remove individual noise and enhance the structure, especially in the red box marked area.

Study on structure prior. The Figure 8d shows that, losing the structure prior, our method barely restores an

Table 3: PSNR(dB)/SSIM results of robust study (train on 5% sparse ratio, dose $\lambda = 10e^{-}$ per selected pixel)

Sampling ratio \ Method	BPFA	CSNet	ReconNet	FSHNet
1.56%	9.973/0.357	17.80/0.616	15.59/0.508	21.22/0.726
3.125%	11.03/0.458	18.42/0.649	15.91/0.522	25.39/0.845
5%	11.23/0.485	18.62/0.658	16.24/0.527	26.72/0.866
6.25%	11.14/0.486	18.74/0.664	16.18/0.528	26.84/0.867
8%	11.29/0.491	18.81/0.666	15.87/0.520	26.92/0.873
9.375%	11.27/0.487	18.80/0.668	16.06/0.525	26.80/0.873

Table 4: PSNR(dB)/SSIM results of ablation studies (train on 5% sparse ratio, dose $\lambda = 10e^{-}$ per selected pixel)

Sampling ratio \ Method	FSHNet	FSHNet without non-local patch	FSHNet without structure prior	FSHNet without frequency init-recon	FSHNet with linear spatial init-recon
1.56%	21.22/0.726	19.97/0.657	20.83/0.657	21.05/0.663	19.02/0.634
3.125%	25.39/0.845	23.89/0.802	24.71/0.79	22.10/0.700	21.02/0.757
5%	26.72/0.866	25.86/0.845	25.84/0.838	22.07/0.708	22.07/0.793
6.25%	26.84/0.867	26.54/0.860	26.02/0.853	22.18/0.711	22.13/0.807
8%	26.92/0.873	26.67/0.864	26.14/0.861	22.28/0.711	22.42/0.812
9.375%	26.80/0.872	26.96/0.867	26.08/0.866	22.22/0.712	22.70/0.815

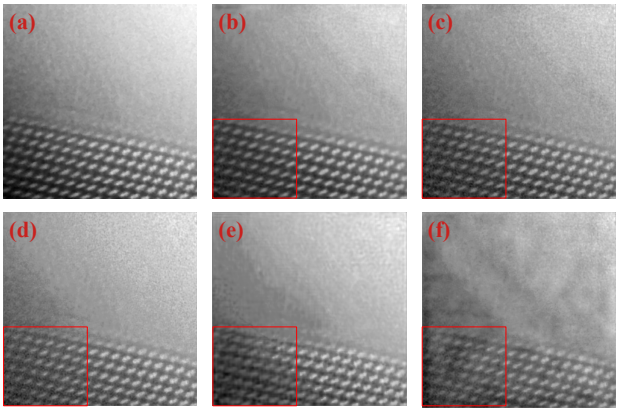

Figure 8: Visual results of ablation studies. **(a)** Ground truth; **(b)** FSHNet reconstruction; **(c)** Reconstruction result from FSHNet without non-local patch enhancement; **(d)** Reconstruction result from FSHNet without structure prior; **(e)** Reconstruction result from FSHNet without initial reconstruction; **(f)** Reconstruction result from FSHNet with linear initial reconstruction.

rough structure and blurs edge of particles, even resulting in structural overlapping. The oval structures selected in red box have collapsed into conical or punctuate ones. Numerical results in the third column of Table 4 further demonstrates that the structure prior helps recover better structure.

Study on initial reconstruction. Without initial reconstruction, our method just simply fills the blank region with repeating structures, resulting in quite naive textures (Figure 8e). The fourth column of Table 4 shows that the gap of PSNR sharply increase from 0.07dB to 4.6dB and the gap of SSIM from 0.06 to 0.16 with the increase of sampling ratio. Moreover, we try to replace the initial reconstruction of frequency domain by a spatial linear initial reconstruction, to study the benefits of frequency domain initial reconstruction. The corresponding Fourier operation layer in FSHNet is replaced by a linear combination layer. Figure 8f shows that FSHnet with linear initial reconstruction has inpaint rough structure but still lose lots of detailed information, resulting in a degenerated numerical results in Table 4. Furthermore, the FSHNet with linear initial reconstruction owns 4364257 model parameters, which is more than ten-fold that of the original FSHNet. Thus, it can be concluded that the linear initial reconstruction is inadequate for large-size STEM image inpainting.

5. Conclusion

In this paper, we have proposed a CNN-based reconstruction model for partial scanning STEM, which utilizes frequency domain information to restore periodic structure of STEM crystalline images in extremely sparse sampling (lower than 5%). Our method utilizes both the global structure features and local smooth information, achieving a complete structure restoration with clearer details.

6. Acknowledgements

This research is supported by the National Key Research and Development Program of China (No. 2017YFA0504702 and No. 2020YFA0712401), and the NSFC projects Grants (62072280, 61932018, 62072441).

References

[1] A. Adler, D. Boublil, and M. Zibulevsky. Block-based compressed sensing of images via deep learning. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2017. 1

[2] P. E. Batson, N. Dellby, and O. L. Krivanek. Sub-ångstrom resolution using aberration corrected electron optics. *Nature*, pages 617–620, 2002. 1

[3] Benjamin Berkels, Peter Binev, Douglas A. Blom, Wolfgang Dahmen, Robert C. Sharpley, and Thomas Vogt. Optimized imaging using non-rigid registration. *Ultramicroscopy*, 138:46 – 56, 2014. 1

[4] E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006. 1

[5] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006. 1

[6] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004. 2

[7] W. Dong, G. Shi, X. Li, Y. Ma, and F. Huang. Compressive sensing via nonlocal low-rank regularization. *IEEE Transactions on Image Processing*, 23(8):3618–3632, 2014. 2

[8] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. 1, 2

[9] Jeffrey M. Ede, Beanland, and Richard. Partial scanning transmission electron microscopy with deep learning. *Scientific Reports*, page 8332, 2020. 2

[10] M. Haider, S. Uhlemann, E. Schwan, and et al. Electron microscopy image enhanced. *Nature*, pages 768–769, 1998. 1

[11] 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)*, June 2016. 1, 2, 5

[12] Lu Gan. Block compressed sensing of natural images. In *2007 15th International Conference on Digital Signal Processing*, pages 403–406, 2007. 2

[13] Niklas Mevenkamp, Peter Binev, Wolfgang Dahmen, Paul M. Voyles, Andrew B. Yankovich, and Benjamin Berkels. *Advanced Structural and Chemical Imaging*, 1(3):1, 2015. 1

[14] Niklas Mevenkamp, Andrew B. Yankovich, Paul M. Voyles, and Benjamin Berkels. Non-local Means for Scanning Transmission Electron Microscopy Images and Poisson Noise based on Adaptive Periodic Similarity Search and Patch Regularization. In Jan Bender, Arjan Kuijper, Tatiana von Landesberger, Holger Theisel, and Philipp Urban, editors, *Vision, Modeling & Visualization*. The Eurographics Association, 2014. 1, 2

[15] Etienne Monier, Thomas Oberlin, Nathalie Brun, Xiaoyan Li, Marcel Tencé, and Nicolas Dobigeon. Fast reconstruction of atomic-scale stem-eels images from sparse sampling. *Ultramicroscopy*, 215:112993, 2020. 5

[16] A. Mousavi, A. B. Patel, and R. G. Baraniuk. A deep learning approach to structured signal recovery. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1336–1343, 2015. 1, 2

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, 12 2019. 5

[18] H. Sawada, T. Sasaki, F. Hosokawa, and K. Suenaga. Atomic-resolution stem imaging of graphene at low voltage of 30 kv with resolution enhancement by using large convergence angle. *Phys. Rev. Lett.*, 114:166102, Apr 2015. 1

[19] W. Shi, F. Jiang, S. Liu, and D. Zhao. Image compressed sensing using convolutional neural network. *IEEE Transactions on Image Processing*, 29:375–388, 2020. 1, 5

[20] A. Stevens, L. Luzi, H. Yang, L. Kovarik, B. L. Mehdi, A. Liyu, M. E. Gehm, and N. D. Browning. A sub-sampled approach to extremely low-dose stem. *Applied Physics Letters*, 112(4):043104, 2018. 1

[21] Andrew Stevens, Hao Yang, Lawrence Carin, Ilke Arslan, and Nigel D. Browning. The potential for Bayesian compressive sensing to significantly reduce electron dose in high-resolution STEM images. *Microscopy*, 63(1):41–51, 10 2013. 2

[22] Sungkwang Mun and J. E. Fowler. Block compressed sensing of images using directional transforms. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 3021–3024, 2009. 2

[23] yan yang, Jian Sun, Huibin Li, and Zongben Xu. Deep admm-net for compressive sensing mri. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 10–18. Curran Associates, Inc., 2016. 1, 2

[24] 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. 1, 2

[25] B. Zhang, J. M. Fadili, and J. Starck. Wavelets, ridgelets, and curvelets for poisson noise removal. *IEEE Transactions on Image Processing*, 17(7):1093–1108, 2008. 1

[26] Daliang Zhang, Yihan Zhu, Lingmei Liu, Xiangrong Ying, Chia-En Hsiung, Rachid Sougrat, Kun Li, and Yu Han. Atomic-resolution transmission electron microscopy of electron beam–sensitive crystalline materials. *Science*, 359(6376):675–679, 2018. 1

[27] 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)*, June 2018. 1, 2

[28] J. Zhang, D. Zhao, and W. Gao. Group-based sparse representation for image restoration. *IEEE Transactions on Image Processing*, 23(8):3336–3351, 2014. 1, 2

[29] Mingyuan Zhou, Haojun Chen, Lu Ren, Guillermo Sapiro, Lawrence Carin, and John Paisley. Non-parametric bayesian dictionary learning for sparse image representations. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22, pages 2295–2303. Curran Associates, Inc., 2009. 1, 2