EasyChair Preprint

№ 1314

In-Situ Data Reduction via Incoherent Sensing

Kai Zhang and Alireza Entezari

July 20, 2019

In-Situ Data Reduction via Incoherent Sensing*

Kai Zhang and Alireza Entezari

University of Florida, Gainesville FL 32608, USA
{zhangkai6,entezari}@ufl.edu

Abstract. We present a framework for in-situ processing of large-scale simulation data that performs a universal data reduction. Instead of direct compression of the data, we propose a different approach that can be benefit from compressed sensing (CS) theory. Unlike the direct data compression techniques where the accuracy of recovery is fixed, the proposed framework enables more accurate recovery (after in situ data reduction), with using better sparse representations, that can be learned from and optimized for the simulation data. Moreover, we discuss the practical case when the assumption of sparsity doesn't hold, the optimization-based recovery algorithm is able to recover the most important elements in the data (characterized by the best k-term approximation), despite significant reduction in the data. We provide theoretical arguments from CS theory and demonstrate experimentally the error behavior exhibited by the proposed approach compared by the best k-term approximation. These arguments, together with our experiments, support the unique feature of the proposed in-situ data reduction: the accuracy of the recovery algorithm can be improved after data reduction by learning better representations for simulation data. The proposed approach provides opportunities for developing new data reduction mechanisms in high performance computing and simulation environments.

Keywords: In-situ data reduction · Compressed sensing · Volume Rendering

1 Introduction

With the ever increasing computing power available for extreme-scale simulations, processing, storage and visualization of such large-scale data has been and continues to be an important problem hindering our ability to harness the power of extreme-scale computing ecosystems. Data reduction has become inevitable in such environments as the performance of I/O modules, constrained by physical limits, has not kept pace with the growth in computational power. Often, simulation results are discarded for a range of time-steps and compression algorithms are used for data reduction for managing the volume of the data to be stored and further processed [17, 18].

A fundamental principle leveraged in data reduction algorithms is that despite the high dimensional representation of the datasets, natural phenomena

* This project was funded in part by NSF grant IIS-1617101.

are often governed by few degrees of freedom compared to the dimension (resolution) of the ambient space they reside. This principle justifies modeling high dimensional data in low dimensional linear subspaces and nonlinear manifolds – the common theme in dimensionality reduction. Similarly the assumption that most datasets can be sparsely represented in a feature domain, is key for compression algorithms. For example, natural images, when represented in wavelet domains, exhibit very few degrees of freedom observable by the sparsity of coefficients. Compression algorithms exploit such dimensionality reduction algorithms to efficiently encode transform coefficients with low entropy. Significant compression of the data necessitates a lossy process that often involves truncation in a transform domain. Compression techniques have been explored for in-situ data processing as well as for large-scale visualization problems. Several compression-domain volume rendering techniques have been proposed in the visualization community [12, 13, 23, 24], where the decoding is combined with rendering such that data transfer rate to the GPU is minimizes. In these approaches, the data is efficiently represented, for example, by vector quantization [23], transform coding [12], dictionary learning [13], or tensor approximation [24]. During the rendering stage, only the features are transfered to the GPU and decoded on-the-fly. We discuss the merits for a different strategy than the direct compression of the data. Based on results from stable embedding and compressed sensing theory, we present arguments for using *incoherent sensing*, instead of compression, for in-situ data reduction. The key motivation behind the proposed approach is its ability to learn from data to improve the quality of recovery after the data reduction has been performed. Unlike the compression paradigm that the quality of recovery (decompression) is fixed, the proposed approach is able to improve the accuracy of recovery with learning better sparse representations from more and more data [25]. In this paper we show that even with sub-optimal sparse representations (e.g., when the assumption of sparsity fails), the recovery algorithm is able to preserve the most important elements, while significantly reducing the data size. This notion of most important elements is formally characterized by the best k-term approximation. As we demonstrate in the following section, the error in the recovery algorithm is bounded by the error committed by the best k-term approximation that we could only have obtained if we knew exactly the location and value of the most important elements. With learning better sparse representations (e.g., from wavelets to ridgelets to surfacelets or dictionary learning) more and more of the most important elements in the data are recovered from the same reduced-scale data (incoherently sensed). We also discuss practical considerations for applying incoherent sensing for large-scale data reduction, where sensing matrices can not be stored, but efficiently computed and implemented as operators.

2 Compressed Sensing framework

The Compressed Sensing (CS) theory, was developed by Donoho [11] and Candés [8], and has transformed many data acquisition systems, including Fourier imag-

ing (e.g., MRI, RADAR) and computed tomography (e.g., coded aperture X-ray).

Formally, let $\mathbf{x} \in \mathbb{R}^N$ be the vectorized dataset comprising N voxels, matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ ($m \ll N$) (with special properties as described in the next subsection) be the sensing matrix, and $\mathbf{y} \in \mathbb{R}^m$ the *incoherently sensed* measurements are obtained from a linear projection:

$$\mathbf{y} = \mathbf{A}\mathbf{x}. \tag{1}$$

Given the incoherently sensed data, $\mathbf{y}$, the recovery algorithm involves an optimization procedure that solves the above underdetermined linear system for $\mathbf{x}$. For simplicity of presentation, we first assume that the original dataset $\mathbf{x}$ is sparse itself (it has relatively small number of non-zero voxels). We then discuss the more realistic case in which the data is sparsified only after a transformation to a different (e.g., wavelet, Fourier) domain.

2.1 Restricted Isometry Property (RIP)

The underdetermined linear system (1) has infinite many solutions and without further knowledge it is impossible to recover the original data $\mathbf{x}$. To be able to recover $\mathbf{x}$ (or a good approximation to it) from this underdetermined system, the sensing matrix has to satisfy a condition known as Restricted Isometry Property (RIP) [8]: A matrix satisfies RIP of order k if there exists a constant δ_k (the smallest possible one) such that matrix $\mathbf{A}$ obeys

$$(1 - \delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \delta_k)\|\mathbf{x}\|_2^2 \tag{2}$$

for any k-sparse (at most k components are non-zero) vector $\mathbf{x}$.

Intuitively, inequality (2) states the energy of $\mathbf{x}$ in projection from high dimension to low dimension is mostly preserved and is distorted at most by δ_k. Hence δ_k represents the almost orthogonality of the collection of every k columns of $\mathbf{A}$. For the extreme case, if the support of vector $\mathbf{x}$ corresponds to the k columns of $\mathbf{A}$ that are orthogonal, then the distortion of $\mathbf{x}$ will be 0 which indicates that $\delta_k = 0$.

2.2 Best k-term approximation

It is demonstrated that if $\delta_{2k} < 1$, the equation (1) has a unique k-sparse solution, therefore, in this situation, only the k-sparse vector can be recovered. This also means that, in theory, by promoting the sparsity of vector $\mathbf{x}$, it's possible to solve the underdetermined linear system (1). Then we can change the problem (1) as an optimization problem (P_0):

$$\begin{aligned} &\min_{\mathbf{u}} \|\mathbf{u}\|_0, \\ &\text{subject to } \mathbf{A}\mathbf{u} = \mathbf{y}, \end{aligned} \tag{3}$$

where the ℓ_0 pseudo-norm of $\mathbf{x}$, $\|\mathbf{x}\|_0$, represents the the number of non-zero elements (sparsity) in vector $\mathbf{x}$. However, because ℓ_0 norm is non-convex, the problem (3) is turned out to be NP-hard. It implies that we can not use standard optimization algorithm to solve this. Although there are several iteration-based greedy algorithm [20,22] can be used to solve this problem and the compute time is always linear with the number of non-zeros in $\mathbf{x}$, the performance is not robust in general, especially when noise presents in the measurements. Therefore, there has to be some relaxation on this problem. Candés and Tao [5, 8] show that instead of using ℓ_0 norm, but using ℓ_1 norm (that is convex) can achieve the same result when $\delta_{2k} < \sqrt{2} - 1$. Hence, we can relax the $P0$ problem (3) to $P1$ problem:

$$\begin{aligned} &\min_{\mathbf{u}} \|\mathbf{u}\|_1, \\ &\text{subject to } \mathbf{A}\mathbf{u} = \mathbf{y}, \end{aligned} \tag{4}$$

and this is as easy as a linear programming problem which can be solved in polynomial-time. In [7], it is also proved that if the RIP condition is satistied, the solution $\mathbf{x}^*$ of (4) obeys:

$$\|\mathbf{x} - \mathbf{x}^*\|_2 \leq C_0 \frac{\|\mathbf{x} - \mathbf{x}_k\|_1}{\sqrt{k}}, \tag{5}$$

where C_0 is some well-behaved constant and $\mathbf{x}_k$ as the best k-terms approximation of $\mathbf{x}$. This indicates that the error of the recovered data is bounded by the error of the best $k-$term approximation of the ground truth. Therefore, this theory shows robust performance when the signal is sparse or has sparse representation.

2.3 Sparsifying Transformations

Most of the natural signals are not necessarily sparse in canonical space domain, however, lots of transform bases can sparsely represent signals. For example, DCT, wavelet, curvelet, surfacelet, dictionary learning and deep learning based non-linear transform can lead many kinds of signals to sparse or compressible representations, that is to say most of the transform coefficients are zero or very close to zero.

Let denote $\boldsymbol{\Psi} \in \mathbb{R}^{N \times N}$ as the transform basis, and $\mathbf{z} = \boldsymbol{\Psi}\mathbf{u}$ denote the sparse representation of variable $\mathbf{u}$. Then the optimization problem (4) can be reformulated as:

$$\begin{aligned} &\min_{\mathbf{u}} \|\mathbf{z}\|_1, \\ &\text{subject to } \mathbf{A}\mathbf{u} = \mathbf{y}, \end{aligned} \tag{6}$$

this provides us a convenience that instead of regularizing the sparsity of signal itself, the signal can be recovered by promoting the sparsity in any "well-chosen"

sparsifying domain. Since the measurement vector $\mathbf{y}$ here is obtained only via the sensing matrix $\mathbf{A}$ without any priori knowledge of $\mathbf{x}$ and transform basis $\boldsymbol{\Psi}$. This implies that we can choose any superior basis or dictionary learned by some learning algorithms as $\boldsymbol{\Psi}$ *after in-situ data reduction* producing $\mathbf{y}$.

2.4 Practical RIP Matrices

Because the sensing matrix has to satisfy the RIP, randomly sensing is naturally suitable in such case. In this section, we will introduce three deeply analyzed random matrices .

Gaussian Matrix [10] Assume that the sensing matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ satisfies that the entries are i.i.d.(independently and identically distributed) and drawn from the normal distribution with mean 0 and variance $1/m$, then if

$$k = \mathcal{O}(\frac{m}{\log \frac{N}{m}}), \tag{7}$$

then with high probability $(1 - O(e^{-\gamma N}))$ with some $\gamma > 0$, the signal can be recovered by solving P_1 problem.

Bernoulli Matrix [9] If entries of $\mathbf{A}$ are i.i.d. and drawn from symmetric Bernoulli distribution $(P(A_{i,j} = \pm 1/\sqrt{m}) = 1/2)$, then if k obeys condition (7), the same result as Gaussian Matrix can be achieved.

Partial Fourier Matrix [21] Suppose $\mathbf{A}$ is a partial Fourier matrix by uniformly randomly selecting m rows from an N by N Fourier matrix, then if

$$k = \mathcal{O}((\frac{m}{\log N})^4), \tag{8}$$

with overwhelming probability (the probability decays exponentially in m), we can get same result as Gaussian matrix.

3 Data Reduction vis CS

For in-situ processing of large scale data by the incoherent sensing framework, the extremely large size of the sensing matrix $\mathbf{A}$ is impractical to be stored into system memory, and the performance of the I/O module limitting the bandwidth also becomes the bottleneck for the data reduction process. Fortunately, all of the "good" sensing matrices discussed in section 2.4 have the superiority that is able to be computed parallelly on-the-fly. This results in the incoherent sensing naturally fit for large scale in-situ processing.

For aspect of implementation, the matrix $\mathbf{A}$ can be represented by the indices of the frequencies when the random DFT or DCT are selected as sensing

matrix and just requires as few as $\mathcal{O}(m)$ bytes storage space. If Gaussian or Bernoulli measurements are used in the application, only seeds need to be loaded into memory to generate sensing matrix. For the sparse presentation, when the transform basis $\boldsymbol{\Psi}$ is determined, the there are also fast and patented parallel algorithms to perform the domain transformations on-the-fly (e.g. FFT [14], FWT [3], FCT [4, 19]), therefore, only small bandwidth is needed in domain transform as well. At last, the only relatively large scale data has to be transferred is the measurements vector $\mathbf{y}$. However, as described above, in the CS framework, only m measurements, which is much less than dimensionality of $\mathbf{x}$, needed to recover the original data $\mathbf{x}$.

4 Experiments

In this section, we show the accuracy of incoherent sensing framework for in-situ data reduction. Theoretically, we present the error bound as statement in (5) to show the robustness of incoherent sensing methods. Pratically, we visualize the reconstruction of the large scale volumetric dataset. In addition, we compare our volumetric data reduction by using our incoherent sensing with commonly used data reduction algorithms including run-length encoding and downsampling. We exhibit that our incoherent sensing method is comparable or superior to those methods. At last, We argue that the random behavior of the incoherent sensing that requires no priori knowledge in terms of the data can be regarded as a universal encoder. By only using the small mount of reusable "code" (sensing data) and a sparse transform basis, this alternative data reduction method can flexibly refine the reconstruction and, in turn, improve the quality of the visualization.

4.1 ℓ_2 norm bound

In this section, we present the experiment showing that the error of the recovered data is bounded by the best $k-$term approximation. We demonstrate that with small set of random Gaussian measurements, the recovered data achieved by solving equation (4) will always obey the bound (5).

We randomly generate the data vector $\mathbf{x}$ with dimension $N = 200$ and fix the sparse level (the number of non-zeros) as $k = 30$. Then we generate the Gaussian sensing matrix $\mathbf{A} \in \mathbb{R}^{m\times 200}$ with mean 0 and variance $1/m$, and vary the number of samples, m, from 30 to 100. For the optimization, we choose the FISTA [1] algorithm which is fastest algorithm compare to other iterative algorithms in this case. In order to get the average performance, for each m, we reconstruct the signal 100 times and regenerate the sensing matrix for each reconstruction. Figure 4.1 shows the results of our experiments. Obviously, the ℓ_2 norm of the recovered error is bounded by the error introduced by best $k-$term approximation and also it should be noticed that with increasing the number of measurements, the error keeps decreasing, which is reasonable because the larger number of measurements gurantees the accuracy of reconstruction.

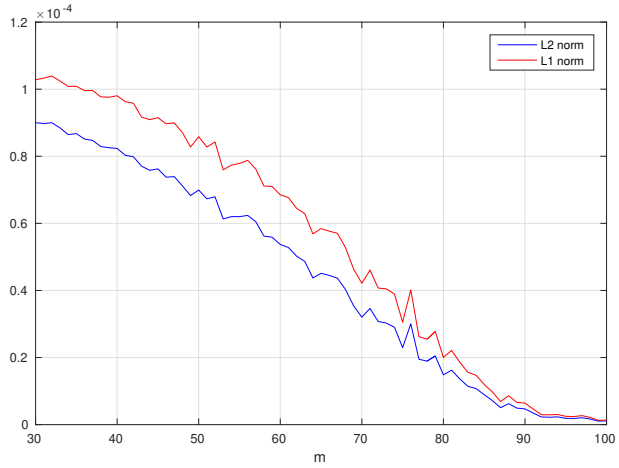


Fig. 1. The blue line represents the reconstruction error, which is $\|\mathbf{x} - \mathbf{x}^*\|_2$ where $\mathbf{x}^*$ is the solution for equation (4), the red line shows the error of best k-term approximation which is $\|\mathbf{x} - \mathbf{x}_k\|_1/\sqrt{k}$.

4.2 Volumetric Dataset

As stated in section 2.3, due to the fact that most of nature signals or data are not necessarily sparse in canonical space domain, in order to increase the accuracy of the reconstruction, seeking the transform for sparse representation of the data is important. The common choice of the tranform basis is wavelet (JPEG2000 image format) or discrete cosine transform (DCT)(JPG image format), however, DCT provides worse compression performance and wavelet is unable to efficiently approximate curve or surface which are usually presented in volumetric dataset. In this experiment, we will choose several geometric extensions of wavelet comprised of curvelets [6], shearlets [15] and surfacelets [16] as the transform bases and compare their performance with other data reduction methods (e.g. run-length encoding and downsampling). For the evaluation of reconstruction accuracy, we use the signal to noise ratio (SNR) which is measured logarithmic scale (dB) over the entire volumetric data as the metric.

Hydrogen The ground truth of Hydrogen dataset is a volumetric dataset with resolution $128 \times 128 \times 128$. Our experiment (from [25]) compares the volumetric incoherent sensing framework with dataset reconstructed from interpolation of downsampled dataset. For the interpolation, we use both linear and cubic spline as the filters. The NESTA [2] algorithm is used to solve the P_1 problem (6). For the measurements, it randomly chooses $m = 12.5\%N$ columns of discrete cosine transform (DCT) measurements. For downsampling case, the sampling rate is set to a factor of two for each dimension, resulting in rate $\rho = 12.5\%$.

Both of these two techniques stores the same size of data (Measurements size in incoherene sensing is 12.5% of dataset size, and sampling rate in downsampling is also 12.5%). Figure 2 shows the volume rendering images of the CS reconstruction as well as interpolation from downsampling dataset. It's obvious that CS reconstruction can achieve higher SNR than downsampling method. We can also observe that changing sparsifying domain can improve the CS reconstruct performance, and surfacelets yield the best result. It verifies that better sparse representations is able to improve the accuracy of recovery.

Supernova The supernova datasets is a volumetric dataset with resolution of $432 \times 432 \times 432$. Still we use the 12.5% measurements and downsampling rate is set to $\rho = 12.5\%$. The visualizations are shown in Figure 3. The result is similar to Hydrogen dataset. We also observe that CS reconstruction tends to smooth the dataset, because the CS recovers the most significant coefficients (usually low frequency) while keeps other coefficients (usually high frequency) as zeros.

5 Conclusion

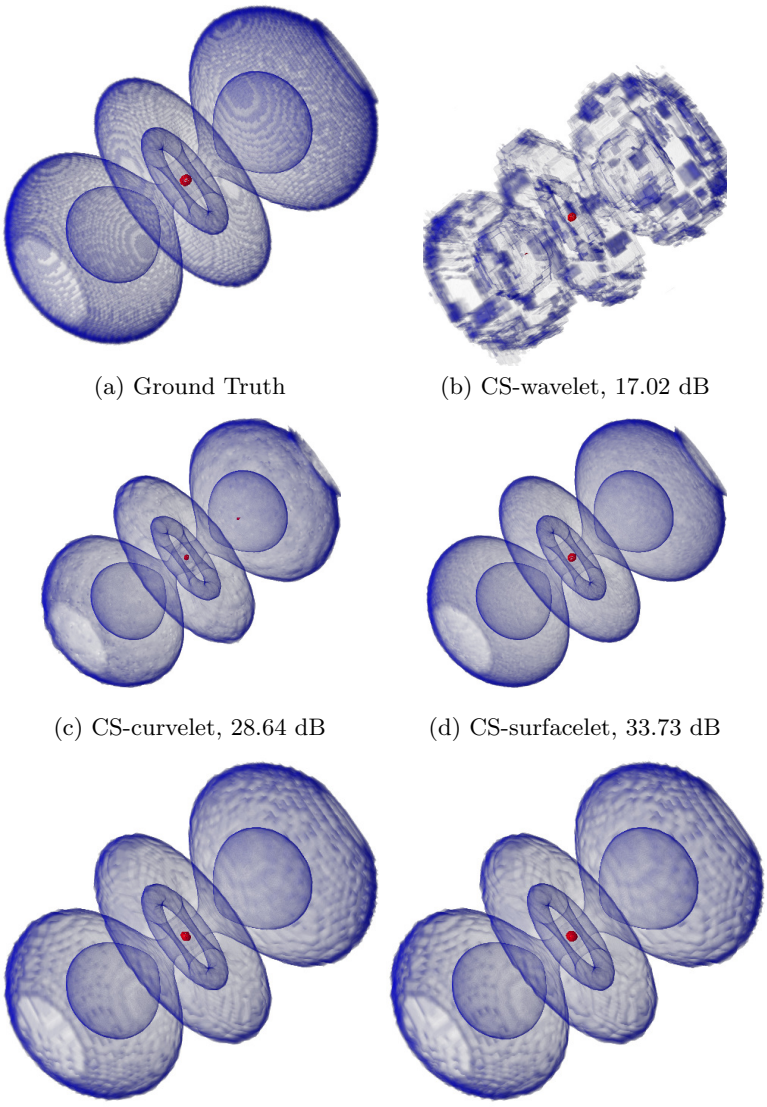
We present the merits of using incoherent sensing, originally introduced in compressed sensing, for in-situ processing and reduction of large-scale simulation data from theoretical and practical aspects. This approach is able to learn from data to improve the quality of recovery after the data reduction has been performed. With carefully choosing better sparse representations the recovery algorithm can achieve higher accuracy as it is able to recover the most important elements in the data. Only a little measurements needed in the process leads to significant decrease in the data storage and saves the bandwidth. The universality of random measurements requiring no priori knowledge in terms of signal also shows the attractive utility of incoherent sensing. In the future work, we will study on the learning-based or customer-designed sparse representations of volumetric data.

References

1. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences **2**(1), 183–202 (2009)
2. Becker, S., Bobin, J., Candès, E.J.: Nesta: A fast and accurate first-order method for sparse recovery. SIAM Journal on Imaging Sciences **4**(1), 1–39 (2011)
3. Beylkin, G., Coifman, R., Rokhlin, V.: Fast wavelet transforms and numerical algorithms i. Communications on pure and applied mathematics **44**(2), 141–183 (1991)
4. Candés, E., Demanet, L., Donoho, D., Ying, L.: Fast discrete curvelet transforms. Multiscale Modeling & Simulation **5**(3), 861–899 (2006)
5. Candés, E.J.: The restricted isometry property and its implications for compressed sensing. Comptes rendus mathematique **346**(9-10), 589–592 (2008)

6. Candés, E.J., Donoho, D.L.: Curvelets: A surprisingly effective nonadaptive representation for objects with edges. Tech. rep., Stanford Univ Ca Dept of Statistics (2000)
7. Candés, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Communications on pure and applied mathematics **59**(8), 1207–1223 (2006)
8. Candés, E.J., Tao, T.: Decoding by linear programming. IEEE transactions on information theory **51**(12), 4203–4215 (2005)
9. Davidson, K.R., Szarek, S.J.: Local operator theory, random matrices and banach spaces. Handbook of the geometry of Banach spaces **1**(317-366), 131 (2001)
10. Donoho, D., Tanner, J.: Counting faces of randomly projected polytopes when the projection radically lowers dimension. Journal of the American Mathematical Society **22**(1), 1–53 (2009)
11. Donoho, D.L.: Compressed sensing. IEEE Transactions on information theory **52**(4), 1289–1306 (2006)
12. Fout, N., Ma, K.L.: Transform coding for hardware-accelerated volume rendering. IEEE Transactions on Visualization and Computer Graphics **13**(6), 1600–1607 (2007)
13. Gobbetti, E., Iglesias Guitián, J.A., Marton, F.: Covra: A compression-domain output-sensitive volume rendering architecture based on a sparse representation of voxel blocks. In: Computer Graphics Forum. vol. 31, pp. 1315–1324. Wiley Online Library (2012)
14. Govindaraju, N.K., Lloyd, B., Dotsenko, Y., Smith, B., Manferdelli, J.: High performance discrete fourier transforms on graphics processors. In: Proceedings of the 2008 ACM/IEEE conference on Supercomputing. p. 2. IEEE Press (2008)
15. Kutyniok, G., Labate, D.: Shearlets: Multiscale analysis for multivariate data. Springer Science & Business Media (2012)
16. Lu, Y.M., Do, M.N.: Multidimensional directional filter banks and surfacelets. IEEE Transactions on Image Processing **16**(4), 918–931 (2007)
17. Ma, K.L.: In situ visualization at extreme scale: Challenges and opportunities. IEEE Computer Graphics and Applications **29**(6), 14–19 (2009)
18. Ma, K.L., Wang, C., Yu, H., Tikhonova, A.: In-situ processing and visualization for ultrascale simulations. In: Journal of Physics: Conference Series. vol. 78, p. 012043. IOP Publishing (2007)
19. Motamedi, M., Sobhieh, S., Motamedi, S.A., Rezaie, A.H.: An ultra-fast, optimized and massively-parallelized curvelet transform algorithm on gp-gpus. In: 2013 21st Iranian Conference on Electrical Engineering (ICEE). pp. 1–6. IEEE (2013)
20. Needell, D., Tropp, J.A.: Cosamp: Iterative signal recovery from incomplete and inaccurate samples. Applied and computational harmonic analysis **26**(3), 301–321 (2009)
21. Rudelson, M., Vershynin, R.: Sparse reconstruction by convex relaxation: Fourier and gaussian measurements. In: 2006 40th Annual Conference on Information Sciences and Systems. pp. 207–212. IEEE (2006)
22. Sahoo, S.K., Makur, A.: Signal recovery from random measurements via extended orthogonal matching pursuit. IEEE Transactions on Signal Processing **63**(10), 2572–2581 (2015)
23. Schneider, J., Westermann, R.: Compression domain volume rendering. In: IEEE Visualization, 2003. VIS 2003. pp. 293–300. IEEE (2003)
24. Suter, S.K., Guitian, J.A.I., Marton, F., Agus, M., Elsener, A., Zollikofer, C.P., Gopi, M., Gobbetti, E., Pajarola, R.: Interactive multiscale tensor reconstruction

for multiresolution volume visualization. IEEE Transactions on Visualization and Computer Graphics **17**(12), 2135–2143 (2011)
25. Xu, X., Sakhaee, E., Entezari, A.: Volumetric data reduction in a compressed sensing framework. In: Computer Graphics Forum. vol. 33, pp. 111–120. Wiley Online Library (2014)

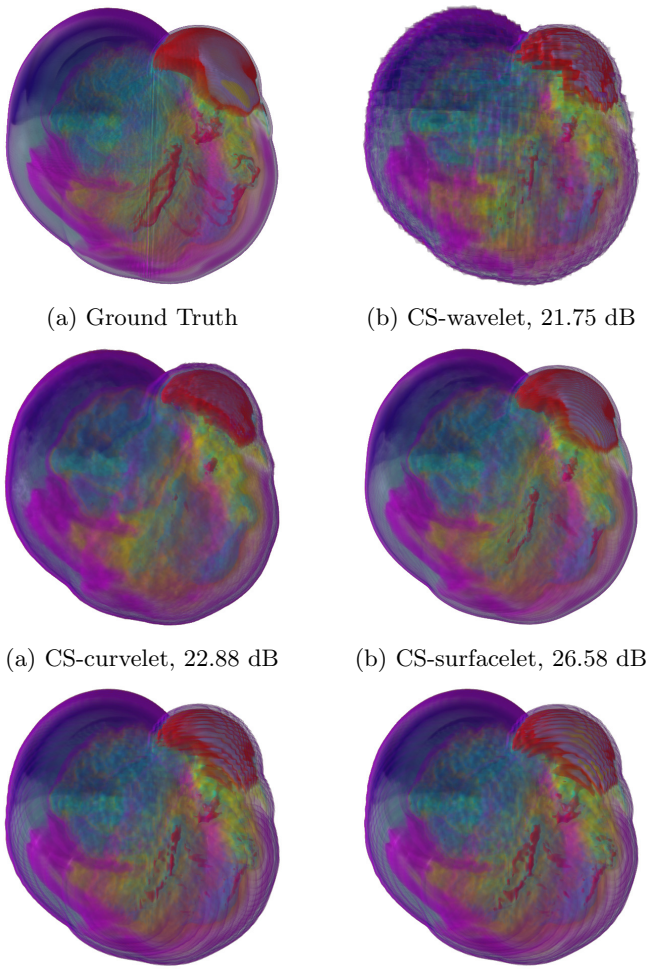


(a) Ground Truth (b) CS-wavelet, 17.02 dB

(c) CS-curvelet, 28.64 dB (d) CS-surfacelet, 33.73 dB

(e) Downsample-linear, 29.78 dB (f) Downsample-cubic, 30.43 dB

Fig. 2. Hydrogen dataset: Sparse approximation and Downsample with sample rate $\rho = 12.5\%$.



(a) Ground Truth (b) CS-wavelet, 21.75 dB

(a) CS-curvelet, 22.88 dB (b) CS-surfacelet, 26.58 dB

(c) Downsample-linear, 25.07 dB (d) Downsample-cubic, 25.64 dB

Fig. 3. Supernova dataset: Sparse approximation and Downsample with sample rate $\rho = 12.5\%$.