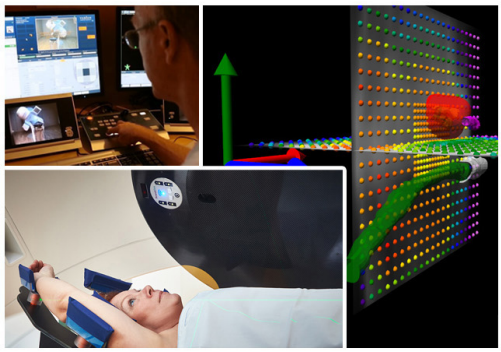
Medical Physics
The International Journal of Medical Physics Research and Practice

Accelerated fast iterative shrinkage thresholding algorithms for sparsity-regularized cone-beam CT image reconstruction

Qiaofeng Xu, Deshan Yang, Jun Tan, Alex Sawatzky, and Mark A. Anastasio

Citation: Medical Physics **43**, 1849 (2016); doi: 10.1118/1.4942812
View online: http://dx.doi.org/10.1118/1.4942812
View Table of Contents: http://scitation.aip.org/content/aapm/journal/medphys/43/4?ver=pdfcov
Published by the American Association of Physicists in Medicine



Accelerated fast iterative shrinkage thresholding algorithms for sparsity-regularized cone-beam CT image reconstruction

Qiaofeng Xu
Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, Missouri 63130

Deshan Yang
Department of Radiation Oncology, School of Medicine, Washington University in St. Louis, St. Louis, Missouri 63110

Jun Tan
Department of Radiation Oncology, The University of Texas Southwestern Medical Center, Dallas, Texas 75390

Alex Sawatzky and Mark A. Anastasio[a)]
Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, Missouri 63130

(Received 1 May 2015; revised 27 January 2016; accepted for publication 15 February 2016; published 23 March 2016)

Purpose: The development of iterative image reconstruction algorithms for cone-beam computed tomography (CBCT) remains an active and important research area. Even with hardware acceleration, the overwhelming majority of the available 3D iterative algorithms that implement nonsmooth regularizers remain computationally burdensome and have not been translated for routine use in time-sensitive applications such as image-guided radiation therapy (IGRT). In this work, two variants of the fast iterative shrinkage thresholding algorithm (FISTA) are proposed and investigated for accelerated iterative image reconstruction in CBCT.
Methods: Algorithm acceleration was achieved by replacing the original gradient-descent step in the FISTAs by a subproblem that is solved by use of the ordered subset simultaneous algebraic reconstruction technique (OS-SART). Due to the preconditioning matrix adopted in the OS-SART method, two new weighted proximal problems were introduced and corresponding fast gradient projection-type algorithms were developed for solving them. We also provided efficient numerical implementations of the proposed algorithms that exploit the massive data parallelism of multiple graphics processing units.
Results: The improved rates of convergence of the proposed algorithms were quantified in computer-simulation studies and by use of clinical projection data corresponding to an IGRT study. The accelerated FISTAs were shown to possess dramatically improved convergence properties as compared to the standard FISTAs. For example, the number of iterations to achieve a specified reconstruction error could be reduced by an order of magnitude. Volumetric images reconstructed from clinical data were produced in under 4 min.
Conclusions: The FISTA achieves a quadratic convergence rate and can therefore potentially reduce the number of iterations required to produce an image of a specified image quality as compared to first-order methods. We have proposed and investigated accelerated FISTAs for use with two nonsmooth penalty functions that will lead to further reductions in image reconstruction times while preserving image quality. Moreover, with the help of a mixed sparsity-regularization, better preservation of soft-tissue structures can be potentially obtained. The algorithms were systematically evaluated by use of computer-simulated and clinical data sets. [http://dx.doi.org/10.1118/1.4942812]

Key words: computed tomographic image reconstruction, x-ray cone-beam computed tomography, sparsity-regularized inverse problems

1. INTRODUCTION

X-ray cone-beam computed tomography (CBCT) employing a circular scanning geometry is a widely employed three-dimensional (3D) imaging modality with numerous applications that include image-guided radiation therapy (IGRT), micro-computed tomography (CT), and dental imaging, to name only a few. There exist a vast literature related to the development and application of CBCT image reconstruction methods, and we refer readers to the recent literature for representative examples.[1–8] The potential advantages of iterative algorithms over analytic algorithms are well-known and include the flexibility to incorporate physical factors in the imaging model and effectively mitigate data incompleteness and noise. The development of iterative image reconstruction algorithms that implement nonsmooth regularizers, including the total variation (TV) penalty and other sparsity-promoting forms, remain an active and

important research area.[9,10] Even with hardware acceleration, however, the overwhelming majority of the available 3D iterative algorithms that implement nonsmooth regularizers remain computationally burdensome and have not been translated for routine use in time-sensitive applications such as IGRT.

The fast iterative shrinkage thresholding algorithm (FISTA)[11,12] is a modern optimization algorithm that possesses several characteristics that are well-suited for iterative CBCT image reconstruction. However, it remains largely unexplored for this application. Because it can be employed to minimize a cost function that is specified by the sum of a smooth and convex data fidelity term and a convex but possibly nonsmooth penalty, the FISTA can be employed for penalized weighted least square (PWLS) reconstruction problems in which a TV penalty or other sparsity promoting forms are employed. Because it is based on solving a TV-proximal problem, the FISTA does not require approximate computation of the discretized TV function or the gradient discretized TV term, which most previously proposed algorithms require. The FISTA can also readily incorporate non-negativity or other bound constraints. Mathematically, it has been proven that the FISTA achieves a quadratic convergence rate. It can therefore potentially reduce the number of iterations required to produce an image of a specified image quality as compared to first-order methods such as the steepest decent method. However, because the FISTA employs a gradient-descent step, which is known to limit convergence rates in conventional algorithms, there remains an opportunity to modify it and obtain an accelerated quadratic algorithm that will lead to further reductions in image reconstruction times.

In this work, two accelerated variants of the FISTA for PWLS-based image reconstruction in CBCT are proposed. The first algorithm seeks to minimize a PWLS cost function involving a TV penalty while the second assumes a penalty formed as the sum of a TV penalty and a wavelet-sparsified ℓ_1 norm of the object. The additional wavelet-sparsified ℓ_1 norm term can potentially improve preservation of fine structures and mitigate artifacts produced by use of oversmoothing with a TV-penalty alone. These effects have been reported for different imaging modalities.[13–16] Additional details regarding use of the ℓ_1 term can be found in Sec. 2. The algorithm acceleration is obtained by replacing the original gradient-descent step by a subproblem that is solved by use of the ordered subset simultaneous algebraic reconstruction technique (OS-SART). Because of the preconditioning matrix adopted in the OS-SART, two new weighted proximal problems are introduced and fast gradient projection (FGP) algorithms are developed to solve them. We also present efficient numerical implementations of the proposed algorithms that exploit the massive data parallelism of multiple graphics processing units (GPUs).

The remainder of the paper is organized as follows. In Sec. 2, the discrete CBCT image model and the formulation of the sparsity-regularized PWLS reconstruction problems are reviewed. The standard FISTAs for solving these problems are also reviewed. Section 3 contains a detailed description the proposed accelerated forms of the FISTAs, which represents the primary contribution of this work. Section 4 addresses some implementation details regarding the proposed algorithms. The improved convergence rates of the algorithms are demonstrated and quantified by use of computer-simulated and clinical data sets in Secs. 5 and 6. Discussion and summary of the work are provided in Secs. 7 and 8.

2. BACKGROUND

2.A. Discrete imaging model for CBCT

We consider a discrete CBCT imaging model

$$\mathbf{b} = \mathbf{H}\mathbf{f}, \tag{1}$$

where $\mathbf{b} \in \mathbb{R}^M$ represents a lexicographically ordered vector describing the cone-beam projection data with M defined by the product of detector elements and number of tomographic views acquired. The vector $\mathbf{f} \in \mathbb{R}^N$ is a finite-dimensional approximation of the sought-after object function $f(\mathbf{r})$. In the studies described below, we assume without loss of generality that $\mathbf{f}$ is formed by use of voxel expansion functions. The $M \times N$ system matrix $\mathbf{H}$ represents a discrete imaging operator that maps $\mathbf{f}$ to $\mathbf{b}$. In this work, $\mathbf{H}$ is defined as a discrete approximation of a divergent beam x-ray transform. However, the presented algorithms are applicable for inversion of any linear imaging equation of the form of Eq. (1).

2.B. PWLS image reconstruction using sparsity-promoting penalties

Two PWLS estimators for CBCT image reconstruction are considered. The first estimator, hereafter referred to as the PWLS-TV estimator, is defined as

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}\geq 0} \|\mathbf{b}-\mathbf{H}\mathbf{f}\|_{\mathbf{W}}^2 + 2\lambda_{\text{tv}}\|\mathbf{f}\|_{\text{TV}}, \tag{2}$$

where $\|\cdot\|_{\text{TV}} = \|\nabla(\cdot)\|_1$ and $\|\cdot\|_1$ denote the TV and ℓ_1 norms, ∇ is a discrete 3D gradient and $\mathbf{W}$ is a diagonal weight matrix with all positive entries. The second estimator, hereafter referred to as the PWLS-TV-ℓ_1 estimator, is defined as

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}\geq 0} \|\mathbf{b}-\mathbf{H}\mathbf{f}\|_{\mathbf{W}}^2 + 2\lambda_{\text{tv}}\|\mathbf{f}\|_{\text{TV}} + 2\lambda_{\ell_1}\|\Phi\mathbf{f}\|_1, \tag{3}$$

where Φ is a sparsifying transform. In the numerical studies below, Φ was defined as a discrete Daubechies wavelet transform that involved three wavelet scales. The real-valued scalar quantities λ_{tv} and λ_{ℓ_1} are user-defined regularization parameters.

Inclusion of the ℓ_1 norm in the penalty provides the opportunity to improve image quality over use of the TV norm alone, particularly with respect to preservation of fine structures.[14–16] For example, a combined wavelet-sparsified ℓ_1-norm and TV penalty has been demonstrated to be effective in MRI and spectral x-ray CT image reconstruction problems.[13,15]

There are numerous algorithms[3,4,6–10,17–25] that have been developed to solve TV regularized image reconstruction problems. Modern optimization methods[11,12,26–33] can be applied to

a variety of reconstruction problems that employ nonsmooth regularizers. In this study, we seek to accelerate advanced FISTAs (Refs. 11 and 12) with the goal of making them feasible for time-sensitive 3D CBCT image reconstruction problems. A review of the FISTA is presented in Subsection 2.C.

2.C. FISTA for solving the PWLS-TV problem

Let

$$d(\mathbf{f}) \equiv \|\mathbf{b}-\mathbf{H}\mathbf{f}\|_{\mathbf{W}}^2 \tag{4}$$

$$g_{\text{tv}}(\mathbf{f}) \equiv 2\lambda_{\text{tv}}\|\mathbf{f}\|_{\text{TV}} + \delta_C(\mathbf{f}), \tag{5}$$

where $\mathbf{C}$ represents a proper closed convex set with nonnegative elements and δ_C is the indicator function that is defined as

$$\delta_C(\mathbf{f}) = \begin{cases} 0 & \text{if } \mathbf{f} \in C, \\ +\infty & \text{elsewhere.} \end{cases}$$

A simple flowchart of the standard FISTA (Ref. 11) that has been developed to solve Eq. (2) is provided in Algorithm I.

ALGORITHM I. FISTA-TV.

Input: $L \geq L(d(f))$—An upper bound on the Lipschitz constant of $\nabla d(\mathbf{f})$. Lipschitz constant:$\equiv 2\sigma_{\max}\{\mathbf{H}^T\mathbf{W}\mathbf{H}\}$, $\sigma_{\max}$ represents the maximum eigenvalue.
Initial Step: Take $\mathbf{e_1} = \mathbf{f_0} = \mathbf{0}$, $t_1 = 1$
for $k \leftarrow 1, n$, **do**

$$\mathbf{x}_g = \mathbf{e}_k - \frac{1}{L}\nabla d(\mathbf{e}_k) = \mathbf{e}_k - \frac{2}{L}\mathbf{H}^T\mathbf{W}(\mathbf{H}\mathbf{e}_k - \mathbf{b}) \tag{6}$$

$$\mathbf{f}_k = prox_{1/L}(g_{tv})(\mathbf{x}_g) = prox_{1/L}(2\lambda_{tv}\|\mathbf{f}\|_{TV})(\mathbf{x}_g) \tag{7}$$

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2} \tag{8}$$

$$\mathbf{e}_{k+1} = \mathbf{f}_k + \frac{t_k - 1}{t_{k+1}}(\mathbf{f}_k - \mathbf{f}_{k-1}) \tag{9}$$

end for
Output: $\mathbf{f}_n$

Its basic steps are summarized as follows. First, a gradient descent step is applied to the data fidelity $d(\mathbf{f})$ to obtain an intermediate image denoted as $\mathbf{x}_g$, as described in Eq. (6). Second, the TV-proximal problem in Eq. (7), which is defined as

$$\text{prox}_{1/L}(g_{\text{tv}})(\mathbf{x}_g) = \arg\min_{\mathbf{u}\in C}\left\{2\lambda_{\text{tv}}\|\mathbf{u}\|_{\text{TV}} + \frac{L}{2}\|\mathbf{u}-\mathbf{x}_g\|^2\right\}, \tag{10}$$

is solved by use of the FGP algorithm.[11] In the studies below, a 3D version of the FGP algorithm was employed, which is derived in Appendix A. Finally, the solution of the proximal problem is utilized to define a new image estimate that is substituted into the first step and the procedure is repeated until a convergence criterion is met.

2.D. Splitting-based FISTA for solving the PWLS-TV-ℓ_1 problem

The FISTA for solving PWLS-TV problem Eq. (2) cannot be applied directly for solving the PWLS-TV-ℓ_1 problem [Eq. (3)] because no efficient algorithms are currently available to directly solve the corresponding composite proximal problem given by

$$\text{prox}_{1/L}(g_{\text{tv}} + g_{\ell_1,\Phi})(\mathbf{x}_g) = \arg\min_{\mathbf{u}}\left\{2\lambda_{\text{tv}}\|\mathbf{u}\|_{\text{TV}} + 2\lambda_{\ell 1}\|\Phi\mathbf{u}\|_{\ell 1} + \frac{L}{2}\|\mathbf{u}-\mathbf{x}_g\|^2\right\}, \tag{11}$$

where $g_{\ell_1,\Phi}(\mathbf{u}) \equiv 2\lambda_{\ell 1}\|\Phi\mathbf{u}\|_{\ell 1}$.

To circumvent this difficulty, the composite splitting approach[26] can be employed to decompose the associated composite proximal problem into two subproximal problems. The first is associated with TV-proximal problem $\text{prox}_{1/(w_1L)}(g_{\text{tv}})(x_g)$, which is exactly the same as Eq. (10) except L is replaced by w_1L. The second is associated with the ℓ_1-proximal problem involving the sparsifying transform Φ, which can be expressed as

$$\text{prox}_{1/(w_2L)}(g_{\ell_1,\Phi})(\mathbf{x}_g) = \arg\min_{\mathbf{u}}\left\{\frac{2\lambda_{\ell 1}}{w_2}\|\Phi\mathbf{u}\|_{\ell 1} + \frac{L}{2}\|\mathbf{u}-\mathbf{x}_g\|^2\right\}. \tag{12}$$

Fortunately, when Φ corresponds to an orthogonal wavelet transform, a soft-shrinkage operator can be adopted to efficiently solve this problem.[12,34]

The two positive-valued scalars w_1 and w_2 satisfy the constraint $w_1 + w_2 = 1$.[26] In this study, we chose $w_1 = w_2 = 1/2$. After obtaining the solutions of these two proximal problems, an average of the two is computed. This proximal-average strategy has been investigated and validated previously.[13,30] In fact, under the conditions stated in Theorem 3.4 in Ref. 26, the sequence generated by this proximal-average step will converge weakly to one solution of the original composite proximal problem. A description of the splitting-based FISTA for solving Eq. (3) is given in Algorithm II.

ALGORITHM II. FISTA-TV-$\ell 1$.

Input: $L \geq L(d(f))$—An upper bound on the Lipschitz constant of $\nabla d(\mathbf{f})$. Lipschitz constant:$\equiv 2\sigma_{\max}\{\mathbf{H}^T\mathbf{W}\mathbf{H}\}$, $\sigma_{\max}$ represents the maximum eigenvalue.
Initial Step: Take $\mathbf{e_1} = \mathbf{f_0} = \mathbf{0}$, $t_1 = \mathbf{1}$, $w_1 = w_2 = 1/2$
for $k \leftarrow 1, n$ **do**

$$\mathbf{x}_g = \mathbf{e}_k - \frac{1}{L}\nabla d(\mathbf{e}_k) = \mathbf{e}_k - \frac{2}{L}\mathbf{H}^T\mathbf{W}(\mathbf{H}\mathbf{e}_k - \mathbf{b}) \tag{13}$$

$$\mathbf{f}_k^1 = prox_{1/(w_1L)}(2\lambda_{tv}\|\mathbf{f}\|_{TV})(\mathbf{x}_g) = prox_{1/L}(4\lambda_{tv}\|\mathbf{f}\|_{TV})(\mathbf{x}_g) \tag{14}$$

$$\mathbf{f}_k^2 = prox_{1/(w_2L)}(2\lambda_{\ell 1}\|\Phi\mathbf{f}\|_{\ell 1})(\mathbf{x}_g) = prox_{1/L}(4\lambda_{\ell 1}\|\Phi\mathbf{f}\|_{\ell 1})(\mathbf{x}_g) \tag{15}$$

$$\mathbf{f}_k = (\mathbf{f}_k^1 + \mathbf{f}_k^2)/2; \tag{16}$$

$$\mathbf{f}_k = project(\mathbf{f}_k, [0\ \max]); \tag{17}$$

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2} \tag{18}$$

$$\mathbf{e}_{k+1} = \mathbf{f}_k + \frac{t_k - 1}{t_{k+1}}(\mathbf{f}_k - \mathbf{f}_{k-1}) \tag{19}$$

end for
Output: $\mathbf{f}_n$

The factor of 2 in both Eqs. (14) and (15) is due to the choice of $1/w_1 = 1/w_2 = 1/2$. Additional details regarding the solution of the ℓ_1-proximal problem $\text{prox}_{1/L}(4\lambda_{\ell_1}\|\Phi\mathbf{f}\|_{\ell_1})(\mathbf{x}_g)$

in Eq. (15) by the soft thresholding operator method are provided in Appendix B. Note that Eq. (17) describes an operator that projects $\mathbf{f}_k$ into a feasible set with value range of [0 max], where the non-negativity constraint can be imposed.

In Sec. 3, the above algorithms are modified to form accelerated FISTAs that possess improved convergence rates that can benefit time-sensitive CBCT applications.

3. ACCELERATED FISTAS FOR IMAGE RECONSTRUCTION IN CBCT

3.A. Motivation and preconditioned ordered subsets' acceleration strategies

Algorithms I and II employ a basic gradient-descent update step,[11]

$$\mathbf{x}_g = \mathbf{e}_k - \frac{1}{L}\nabla d(\mathbf{e}_k) = \mathbf{e}_k - \frac{2}{L}\mathbf{H}^T\mathbf{W}(\mathbf{H}\mathbf{e}_k - \mathbf{b}) \tag{20}$$

to reduce the value of $d(\mathbf{f})$. The standard FISTA achieves a quadratic convergence rate described as

$$F(\mathbf{f}_k) - F(\mathbf{f}^*) \le \frac{2L\|\mathbf{f}_k - \mathbf{f}^*\|^2}{(k+1)^2}, \tag{21}$$

where $F(\cdot)$ is the objective function, k is the iteration number, $\mathbf{f}_k$ is the image estimate at the kth iteration and $\mathbf{f}^*$ is the true solution to the optimization problem. Equation (21) is established in Theorem 3.1 in Ref. 11. When the FISTA is employed for CBCT image reconstruction, the basic gradient update step in Eq. (20) will be the most time consuming step. This is because it requires computation of the forward operator $\mathbf{H}$ and the backprojection operator $\mathbf{H}^T$ for each single update of the object function estimate. Computing the action of these operators is generally computationally burdensome in CBCT due to the large amount of projection data and number of image voxels. In addition, when the Lipschitz constant L is large, the update step size $1/L$ is small in the basic gradient step, which indicates that more iterations need to be performed to minimize $d(\mathbf{f})$.

Instead of employing all of the projection data at once to compute a gradient descent step, it is well known that an intermediate solution to a least squares minimization problem can be obtained more efficiently by employing a strategy in which the estimate of the object function is updated frequently by use of ordered subsets of the projection data sequentially.[35] Such approaches can dramatically improve the convergence rate of an iterative method over classic gradient descent methods. Many advanced iterative methods that solve the least squares problem can be combined with the ordered subsets' concept to accelerate the reconstruction process.[35–39] The SART[40] is an efficient iterative algorithm for solving the least squares problem.[41,42] Therefore, we choose the OS-SART[43] to accelerate the FISTA as described below.

The OS-SART algorithm adopted in this work is now reviewed briefly. Consider Eq. (1)

$$\sum_{j=1}^{N} h_{ij} f_j = b_i, \quad i = 1,2,\ldots,M, \tag{22}$$

where f_j is jth element of the vector $\mathbf{f}$, N is the number of image voxels, M is the number of source–detector element pairs (i.e., line integrals, or "rays" recorded), and h_{ij} is the element of $\mathbf{H}$ corresponding to the ith row and jth column, which can be interpreted as a weight element that represents the contribution of the jth voxel to the ith line integral.

Consider that the projection data $\mathbf{b}$ are grouped into T subsets that are indexed by v. Let the vector $\mathbf{b}^v$ denote the projection data corresponding to the vth subset. The corresponding reduced imaging model can be expressed as $\mathbf{b}^v = \mathbf{H}_v\mathbf{f}$, where $\mathbf{H}_v$ contains a subset of the elements in the full $\mathbf{H}$. In terms of this subset notation, Eq. (22) can be rewritten as

$$\sum_{j=1}^{N} h_{ij,v} f_j = b_i^v, \quad i = 1,2,\ldots,M_v, \quad v = 1,2,\ldots,T, \tag{23}$$

where b_i^v is the ith component of $\mathbf{b}^v$, $h_{ij,v}$ is the element of $\mathbf{H}_v$ corresponding to the ith row and jth column, and M_v is the total number of rays in vth subset.

The OS-SART algorithm is composed of two substeps, a forward-correction step and a backprojection-update step. These two steps are implemented as

$$c_{i,v} = \frac{b_{i,v}^{\text{data}} - \sum_{j=1}^{N} h_{ij,v} f_{j,v-1}^k}{\sum_{j=1}^{N} h_{ij,v}}, \tag{24}$$

$$f_{j,v}^k = f_{j,v-1}^k + \gamma_v \frac{\sum_{i}^{M_v} c_{i,v} h_{ij,v}}{\sum_{i}^{M_v} h_{ij,v}}, \tag{25}$$

where $b_{i,v}^{\text{data}}$ represents the ith ray projection data in the vth subset, and $f_{j,v-1}^k$ and $f_{j,v}^k$ are the jth voxel value updated by use of the $(v-1)$th and the vth data subset at the kth iteration, respectively. Equations (24) and (25) can be expressed in a matrix–vector form[41] as

$$\mathbf{f}_v^k = \mathbf{f}_{v-1}^k - \gamma_v \mathbf{D}_v \mathbf{H}_v^T \mathbf{U}_v (\mathbf{H}_v \mathbf{f}_{v-1}^k - \mathbf{b}_v). \tag{26}$$

Here, $\mathbf{U}_v$ is a weight matrix defined as

$$\mathbf{U}_v = \text{diag}\left\{1/\sum_{j=1}^{N} h_{ij,v} \quad | \quad i = 1,2,\ldots,M_v\right\}. \tag{27}$$

Each element of $\mathbf{U}_v$ is the reciprocal of the ith ray length. The matrix $\mathbf{D}_v$ is defined as

$$\mathbf{D}_v = \text{diag}\left\{1/\sum_{i=1}^{M_v} h_{ij,v} \quad | \quad j = 1,2,\ldots,N\right\} \tag{28}$$

and can be interpreted as a preconditioning matrix. Each element of $\mathbf{D}_v$ is the reciprocal of the sum of intersection lengths of rays that intersect the jth voxel in the vth subset. The diagonal matrices $\mathbf{U}_v$ and $\mathbf{D}_v$ can be obtained as a byproduct when computing the action of the operators $\mathbf{H}_v$ and $\mathbf{H}_v^T$. Equation (26) can be interpreted as a preconditioned gradient-based scheme with sequential update strategy.

3.B. Proposed OSSF-TV and OSSF-TV-ℓ_1 algorithms

Motivated by the above observations, we propose accelerated versions of the FISTAs in which the gradient descent step is replace by an OS-SART subproblem. More specifically, Eqs. (6) and (7) in Algorithm I will be replaced by an inner loop given by

$$\begin{cases} \mathbf{e}_0^k = \mathbf{e}_T^{k-1} \\ \text{for } v = 1,\ldots,T \\ \quad \mathbf{e}_v^k = \mathbf{e}_{v-1}^k - \gamma_v \mathbf{D}_v \mathbf{H}_v'^T (\mathbf{H}_v' \mathbf{e}_{v-1}^k - \mathbf{b}_v') \\ \quad \mathbf{f}_k = \text{prox}_{\gamma_v}^{D_v^{-1}}(g_{\text{tv}}/T)(\mathbf{e}_v^k) \\ \quad \mathbf{e}_v^k = \mathbf{f}_k \\ \text{end.} \end{cases} \tag{29}$$

Here, $\mathbf{H}_v' = \mathbf{U}_v^{1/2}\mathbf{H}_v$ and $\mathbf{b}_v' = \mathbf{U}_v^{1/2}\mathbf{b}_v$. In this inner loop, due to incorporation of the preconditioning matrix $\mathbf{D}_v$, a new weighted TV-proximal problem[44] must be considered that is defined as

$$\text{prox}_{\gamma}^{D_v^{-1}}(g_{\text{tv}}/T)(\mathbf{e}_v^k) = \arg\min_{\mathbf{u}} \left\{ \frac{g_{\text{tv}}(\mathbf{u})}{T} + \frac{1}{2\gamma} \|\mathbf{u} - \mathbf{e}_v^k\|_{D_v^{-1}}^2 \right\}, \tag{30}$$

where the original penalty term $g_{\text{tv}}(\mathbf{u})$ is scaled by T, which is the total number of subsets. The reason that the term $g_{\text{tv}}(\mathbf{u})$ should be approximately scaled by T is that the subset gradient is approximately equal to $1/T$ of the original fullset gradient.[31,39] Accordingly, the effective regularization parameter can be (much) smaller for each subset depending on the size of T. Therefore, solution of the subset proximal-TV problem will generally require fewer algorithm iterations than needed to solve the original proximal-TV problem in Eq. (10).

The weighted TV-proximal problem can be efficiently solved by use of a modified version of the FGP algorithm, since the matrix D_v is diagonal. Details are provided in Appendix C. *We will refer to this accelerated version of Algorithm* I *for solving Eq. (2) as the OS-SART-FGP-TV (OSSF-TV) algorithm.* It is important to note that, because the OS-SART method is employed to reduce the value of $d(\mathbf{f})$ instead of the standard gradient-descent step, the OSSF-TV algorithm solves Eq. (2) for the case where the data fidelity weight matrix $\mathbf{W}$ is a block diagonal matrix defined as

$$\mathbf{W} = \text{diag}(\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_T), \tag{31}$$

where the matrices $\mathbf{U}_v$, $v = 1,\ldots,T$, are defined in Eq. (27). In a similar way, Eqs. (13)–(17) in Algorithm II can be replaced by the following inner loop:

$$\begin{cases} \mathbf{e}_0^k = \mathbf{e}_T^{k-1} \\ \text{for } v = 1, \ldots, T \\ \quad \mathbf{e}_v^k = \mathbf{e}_{v-1}^k - \gamma_v \mathbf{D}_v \mathbf{H}_v'^T (\mathbf{H}_v' \mathbf{e}_{v-1}^k - \mathbf{b}_v') \\ \quad \mathbf{f}^{1,k} = \text{prox}_{\gamma_v}^{D_v^{-1}}(g_{\text{tv}}/T \times 2)(\mathbf{e}_v^k) \\ \quad \mathbf{f}^{2,k} = \text{prox}_{\gamma_v}^{D_v^{-1}}(g_{l1,\Phi}/T \times 2)(\mathbf{e}_v^k) \\ \quad \mathbf{f}^k = (\mathbf{f}_v^{1,k} + \mathbf{f}_v^{2,k})/2 \\ \quad \mathbf{f}^k = \text{project}(\mathbf{f}_v^k, [0\ \max]) \\ \quad \mathbf{e}_v^k = \mathbf{f}^k \\ \text{end.} \end{cases} \tag{32}$$

Here, in addition to the weighted TV-proximal problem described above, the following weighted wavelet-based ℓ_1-proximal problem is introduced as

$$\text{prox}_{\gamma}^{D_v^{-1}}(g_{\ell_1,\Phi}/T \times 2)(\mathbf{e}_v^k) = \arg\min_{\mathbf{u}} \left\{ \frac{g_{\ell_1,\Phi}(\mathbf{u})}{T} \times 2 + \frac{1}{2\gamma_v} \|\mathbf{u} - \mathbf{e}_v^k\|_{D_v^{-1}}^2 \right\}. \tag{33}$$

To the best of our knowledge, there is no closed-form solution of Eq. (33) due to the weighted norm involved. However, the same strategy employed in the standard FGP algorithm for solving TV-proximal problem can be adopted to dualize the ℓ_1-norm, and a new FGP-type algorithm can be developed to solve the weighted wavelet-based ℓ_1-norm proximal problem. Additional details are provided in Appendix D. *We will refer to this accelerated version of Algorithm* II *for solving Eq. (3) as the OS-SART-FGP-TV-ℓ_1 (OSSF-TV-ℓ_1) algorithm.* Because the OS-SART method is employed to reduce the value of $d(\mathbf{f})$ instead of the standard gradient-descent step, the OSSF-TV-ℓ_1 algorithm solves Eq. (3) for the case where the data fidelity weight matrix $\mathbf{W}$ is given by Eq. (31).

4. IMPLEMENTATION DETAILS

4.A. Number of subsets and data accessing order

For ordered subset algorithms, the acceleration factor is approximately proportional to the number of subsets in which the entire projection data are divided for early iterations.[45] If the projection data at each view angle are defined as one subset, some conditions must be met to avoid numerical artifacts. Namely, each voxel in the reconstructed volume must be intersected by at least one ray in every subset. When a voxel is not intersected by any rays, the corresponding element of $\mathbf{D}_v$ will be zero. Therefore, this specific voxel will not be updated at this subset, which may cause inaccuracy and artifacts. To avoid this, we can either adjust the voxel size or employ more than one projection view as a subset.

Besides the number of subsets, the data-access ordering strategy can also affect the convergence speed. Several different strategies have been proposed and investigated, such as the ordering methods of sequential access, fixed angle-gap, random access,[46] prime number decomposition,[47] multilevel[48] and weighted distance.[49] In this work, the sequential access and fixed angle-gap ordering strategies were employed. Different suffixes will be appended to the

algorithm names to denote the different data access strategies employed in the OS-SART subproblem. Specifically, the first number will represent how many projections are included in one subset and the second number will denote the accessing order of the subsets. For example, the OS-SART subproblem in an OSSF-TV-1-1 algorithm treats each projection as one subset and the subsets are accessed sequentially. The OS-SART subproblem in an OSSF-TV-1-4 algorithm still treats each projection as one subset but the access order of subsets is to jump every four sequential projection views (subsets), i.e., the view angle access order is arranged as $1, 5, \ldots, (T), 2, 6, \ldots, (T), 3, 7, \ldots, (T), 4, 8, \ldots, (T)$, where T denotes the total number of subsets. In the numerical studies below, an improved version of Siddon's ray-tracing model[50] was employed to calculate the system matrix weights.

4.B. Number of iterations employed to solve the TV-proximal problem

The computation time required to solve the TV-proximal problem is much less than that required by the gradient step in this study. This is because the computational complexity of the FGP algorithm is only $\mathbf{O}(N)$.[11] The exact solution of the TV-proximal problem using one subset of data in the OSSF-TV (OSSF-TV-ℓ_1) algorithm and the standard FISTA-TV (FISTA-TV-ℓ_1) employing the entire data set have similar complexity. However, in practice, the number of iterations required to obtain an acceptable approximate solution of the TV-proximal problem for the OSSF-TV case is generally smaller than for the FISTA-TV case. For example, in a recent work[51] only one or two iterations were adopted to solve the constrained denoising problem for the ordered-subset case. We observed that the solution of the TV-proximal problem for the OSSF-TV case obtained by use of three iterations of the FGP algorithm did not differ significantly from the solution obtained by use of ten iterations. In this work, 20 iterations of the FGP algorithm were employed in the standard FISTA-TV and FISTA-TV-ℓ_1 and only three iterations were employed for each subset weighted proximal problems in our proposed OSSF-TV and OSSF-TV-ℓ_1 algorithms.

4.C. Preconditioning matrix and step size γ

It is well known that incorporating a preconditioning matrix in a gradient step can improve the convergence rate of an iterative algorithm.[52] In our case, the OS-SART algorithm implicitly incorporates a preconditioning matrix $\mathbf{D}_v$.

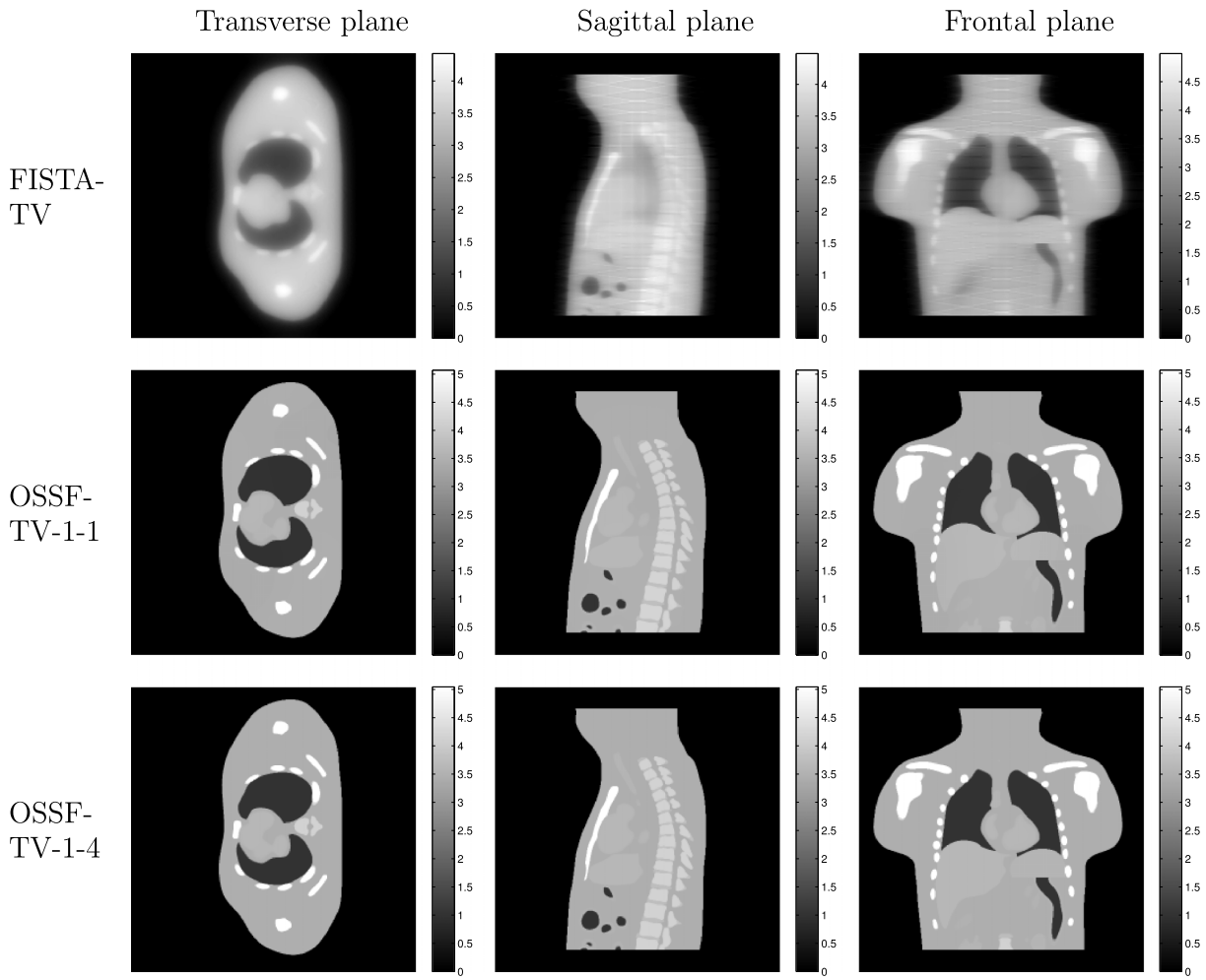


FIG. 1. NCAT numerical phantom study for the full-view (360-view) case. Examples of images reconstructed by use of the FISTA-TV (top row), OSSF-TV-1-1 (middle row), and OSSF-TV-1-4 (bottom row) algorithms are displayed. Ten algorithm iterations were employed in all cases.

Forming $\mathbf{D}_v$ does not involve additional computations since each element will be readily determined by the elements of $\mathbf{H}_v$ via Eq. (28). Other preconditioning matrices that may yield improved numerical properties can be employed for this purpose. However, to do so generally requires computing additional products of $\mathbf{H}_v$ and $\mathbf{H}_v^T$.

Mathematically, the spectral radius of $\mathbf{D}\mathbf{H}^T\mathbf{U}\mathbf{H}$ is less than or equal to 1.[41,42] Since the subset balance condition[51] is generally satisfied for CBCT, the spectral radius of $\mathbf{D}_v\mathbf{H}_v^T\mathbf{U}_v\mathbf{H}_v$ can also be treated as less or equal to 1. The step size γ_v should satisfy the condition $0 \le \gamma_v \le 2/\rho(\mathbf{D}_v\mathbf{H}_v^T\mathbf{U}_v\mathbf{H}_v)$ to ensure convergence, where the notation $\rho(\mathbf{X})$ denotes the spectral radius of a matrix $\mathbf{X}$. Accordingly, a general choice for γ_{v+1} should satisfy $0 < \gamma_{v+1} < 2$. The step size γ_v can also be optimized to achieve the maximum decrease for each subset. However, this would involve additional computations that would increase reconstruction times. For simplicity, we utilized a fixed step size $\gamma = 1/2$ for all subsets in this study. Because the convergence rate at the first few iterations is proportional to the number of subsets, which was relatively large, we did not need to employ a very aggressive step size to achieve rapid convergence.

4.D. Specification of PWLS estimators

As described in Sec. 3.B, the OSSF-TV and OSSF-TV-ℓ_1 algorithms seek to minimize Eqs. (2) and (3) for the case where the data fidelity weight matrix $\mathbf{W}$ is defined according to Eq. (31). When implementing the standard FISTAs given by Algorithms I and II, the weight matrix $\mathbf{W}$ was also defined according to Eq. (31). In this way, the standard and accelerated FISTAs sought to minimize the same objective functions and their convergence properties could be compared directly. Although not presented, we also implemented the standard FISTAs for the case where the data fidelity terms in the objective functions were unweighted. The general observations regarding the relative convergence rates of the standard and accelerated algorithms described below persisted for those cases.

4.E. GPU implementations

Highly efficient parallel implementations of the OSSF-TV and OSSF-TV-ℓ_1 algorithms that can utilize a single or multiple GPUs are presented in Appendix E.

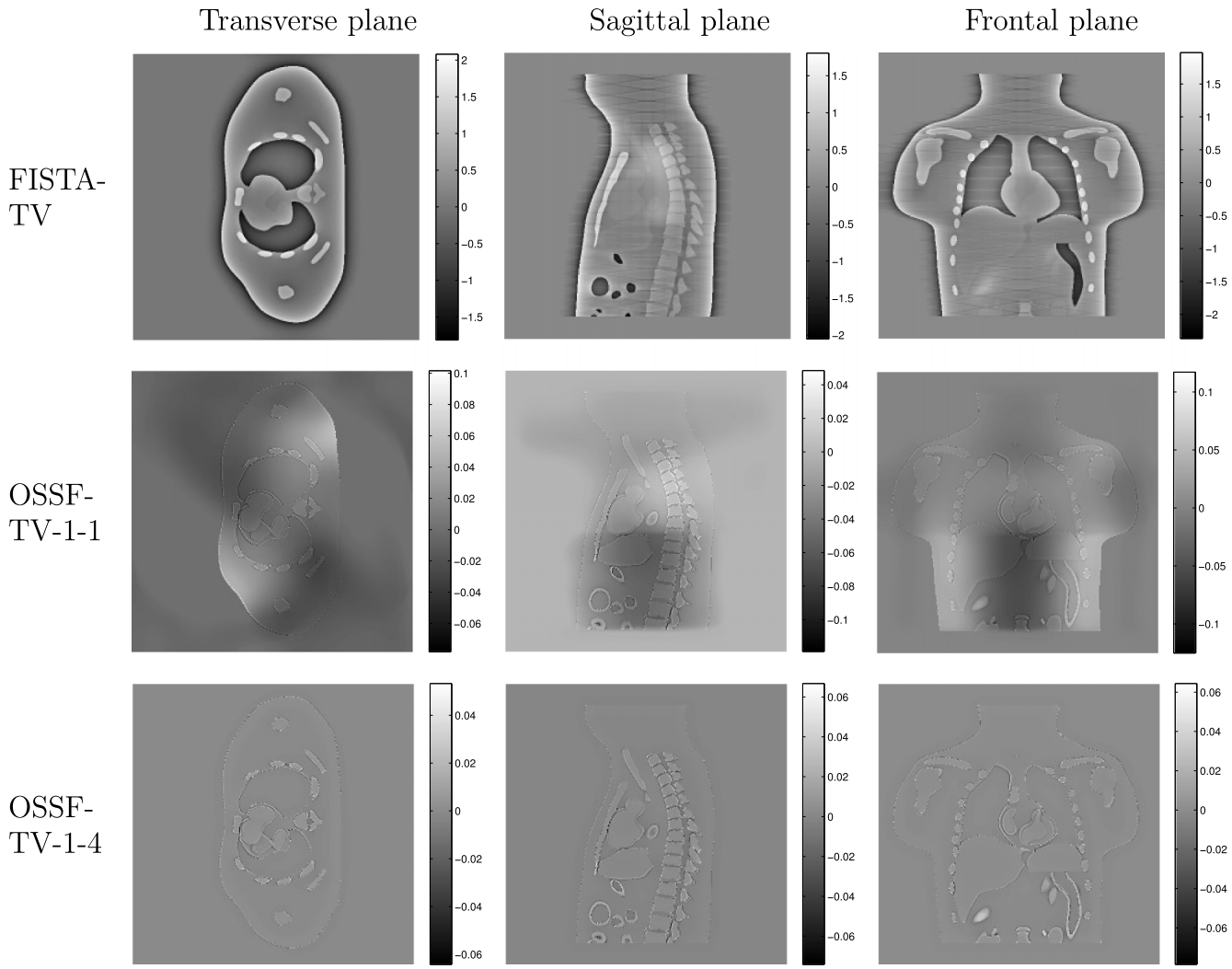


FIG. 2. Difference images corresponding to the images displayed in Fig. 1.

5. COMPUTER-SIMULATION STUDIES

Computer-simulation studies were conducted to validate the proposed reconstruction algorithms and quantify their improvements in convergence rates as compared with the standard FISTAs.

5.A. Numerical phantom and simulated projection data

A NCAT phantom[53] was adopted that contained 256×256 ×256 voxels of dimension 0.5 mm. A circular CBCT imaging geometry with a source-to-rotation center distance of 50 cm and source-to-detector distance of 150 cm was employed. A flat detector of size 20 × 20 cm was assumed that possessed 512×512 elements. At each of 360 tomographic view angles that were uniformly spaced over a 2π angular range, CBCT projection data were computed numerically by use of the system matrix described below. The projection data produced in this way did not contain modeling errors; however, such errors and other real-world data inconsistencies are contained in the experimental data studies in Sec. 6. The simulated projection data were contaminated by use of a simplified additive noise model in which the noise at each detector pixel was specified as a Poisson random variable. The mean and variance of this random variable were specified as 3% of the noiseless intensity data at each detector pixel. Although this is an approximate noise model, its use will not affect our conclusions regarding convergence rates, which is our singular focus. The complete set of projection data containing all 360 views will be referred to as the "full-view" data, while an angularly subsampled version containing 45 equally spaced views.

5.B. Full-view case: Reconstructed images and corresponding difference maps

Images reconstructed from the full-view noisy projection data by use of the standard FISTA-TV and the proposed OSSF-TV-1-1 and OSSF-TV-1-4 algorithms are shown in Fig. 1. All algorithms were terminated after 10 iterations and utilized the same regularization parameter λ_{tv}. Images reconstructed by use of the FISTA-TV algorithm (first row in Fig. 1) have a significantly blurred appearance, indicating that additional

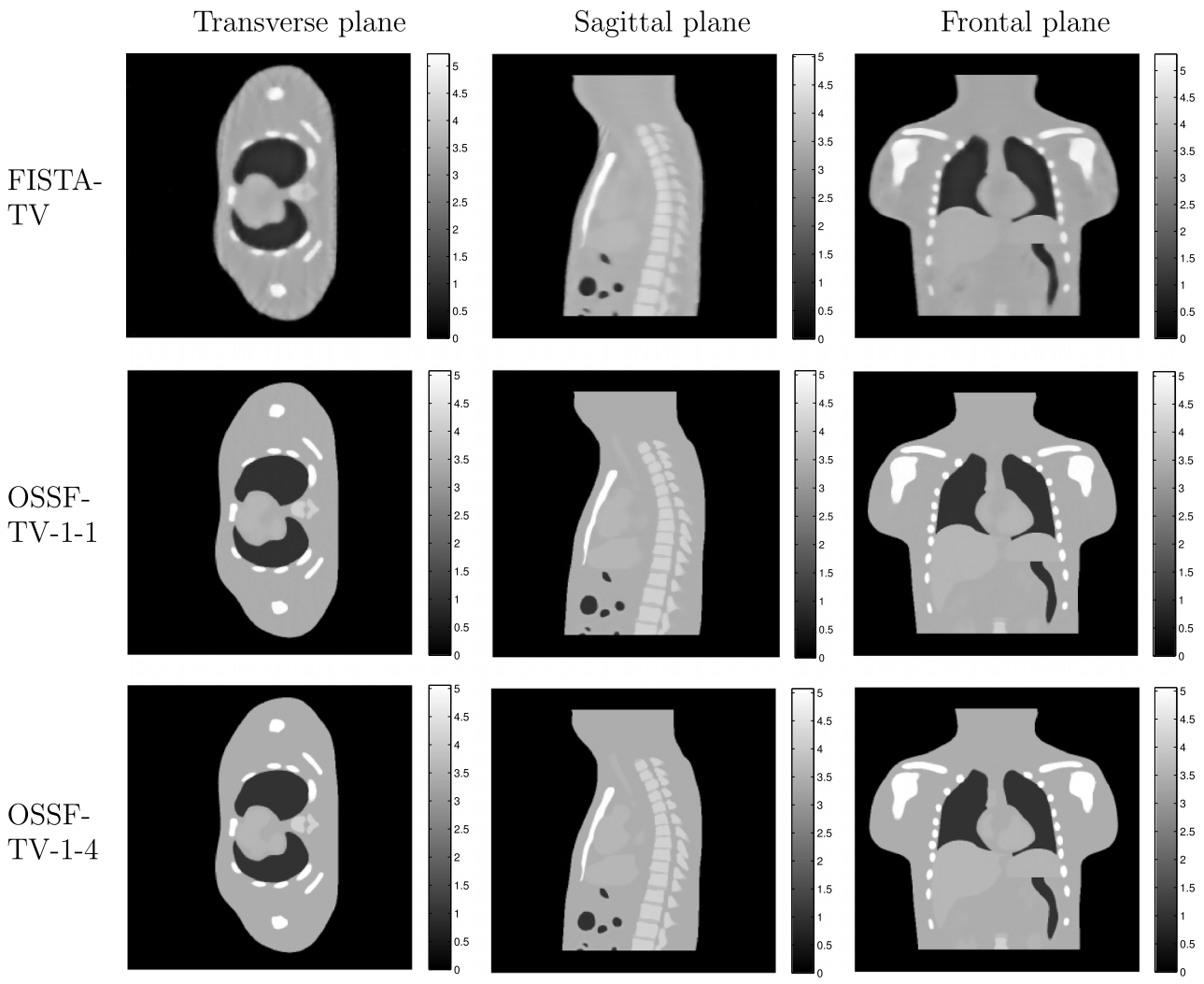


FIG. 3. NCAT numerical phantom study for the sparse-view (45 view) case. Examples of images reconstructed by use of the FISTA-TV (top row), OSSF-TV-1-1 (middle row), and OSSF-TV-1-4 (bottom row) algorithms are displayed. 30 algorithm iterations were employed in all cases.

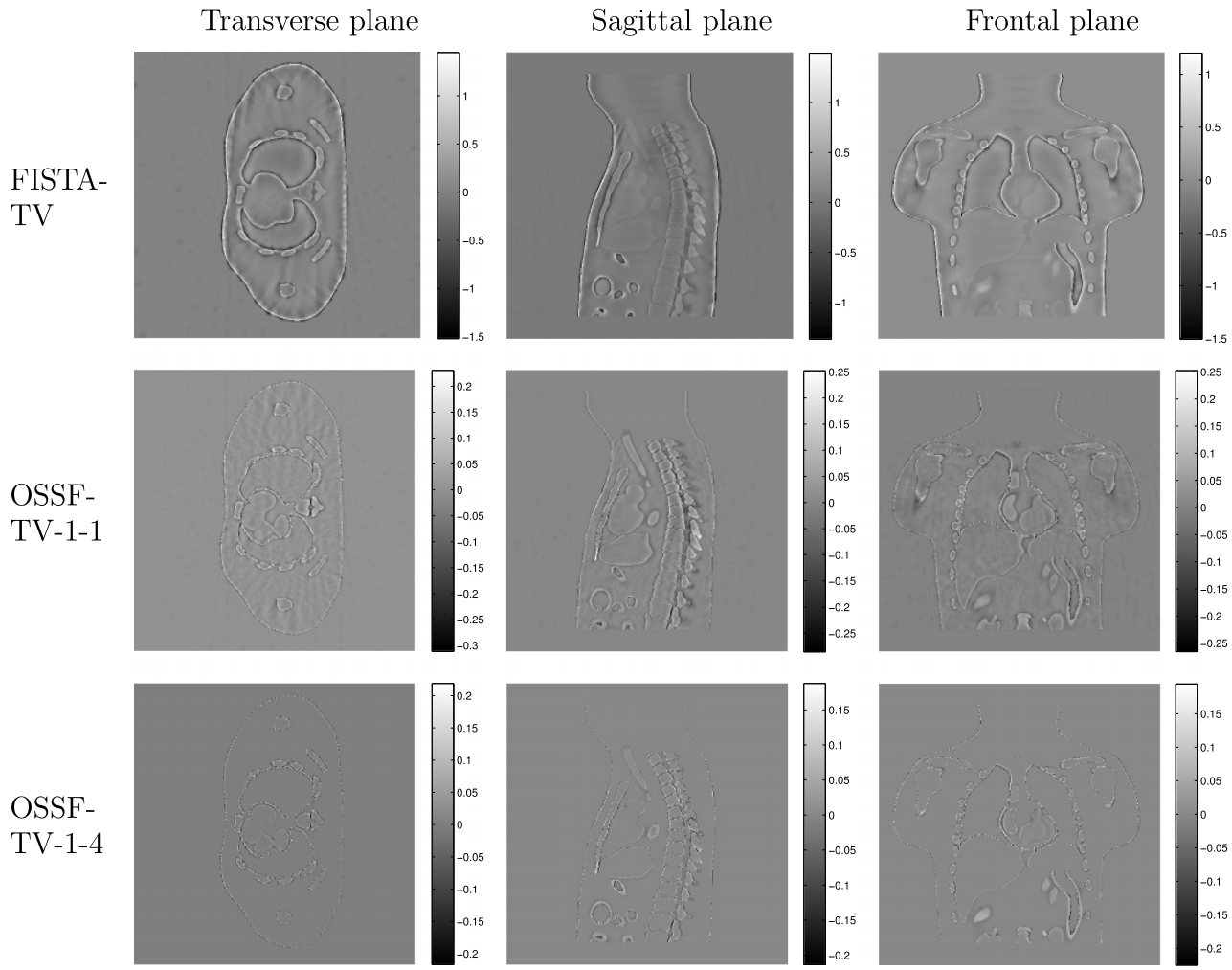


FIG. 4. Difference images corresponding to the images displayed in Fig. 3.

algorithm iterations are required to recover sharp boundaries and accurate pixel values. However, the images reconstructed by use of the OSSF-TV-1-1 and OSSF-TV-1-4 algorithms contain sharp structures with high contrast, despite the use of only 10 iterations in the algorithms. To quantitatively examine the reconstructed images, difference images produced by subtracting the reconstructed images from the true phantom are shown in Fig. 2. The maximum magnitudes of the difference images obtained by the standard FISTA-TV algorithm are nearly two orders larger than those obtained from the proposed OSSF-TV algorithms. Moreover, the difference images reveal that the OSSF-TV-1-1 algorithm can produce certain view-

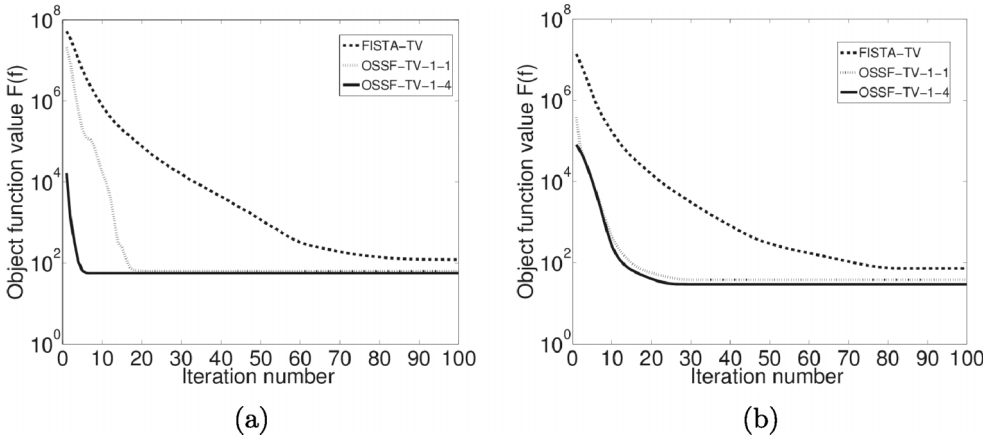


FIG. 5. Convergence analysis: Plots of the objective function value as a function of iteration number for the FISTA-TV and OSSF-TV algorithms for the (a) full-view (360-view) case and (b) few-view (45-view) case.

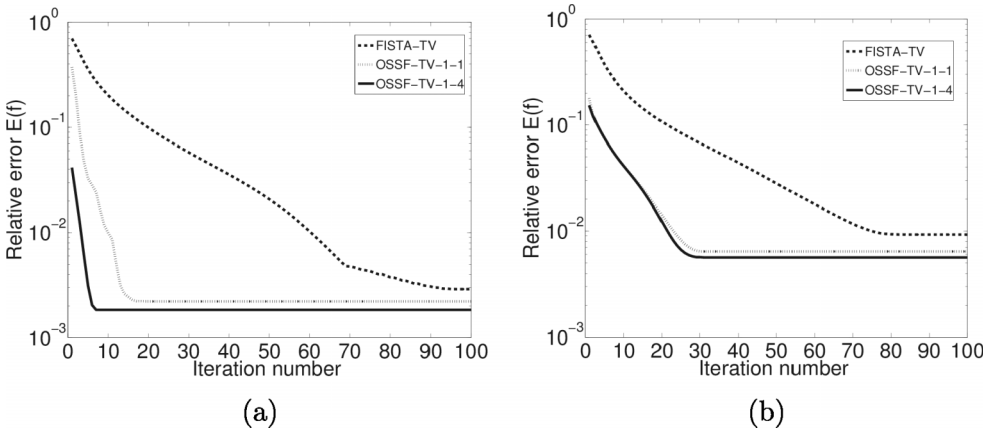


FIG. 6. Accuracy analysis: Plots of the image RE as a function of iteration number for the FISTA-TV and OSSF-TV algorithms for the (a) full-view case and (b) few-view case.

dependent shading patterns when the tomographic views are densely sampled. This is because -1-1 data access and update strategy can introduce "over-shooting" in some directions under such circumstances.[49]

5.C. Sparse-view case: Reconstructed images and corresponding difference maps

Images reconstructed from the sparse-view noisy projection data by use of the standard FISTA-TV and the proposed OSSF-TV-1-1 and OSSF-TV-1-4 algorithms and the associated difference images are shown in Figs. 3 and 4. All algorithms were terminated after 30 iterations and utilized the same regularization parameter λ_{tv}. As in the full-view case, images reconstructed by use of the OSSF-TV-1-1 and OSSF-TV-1-4 algorithms both display sharper boundaries and higher contrast compared to the image reconstructed by use of the FISTA-TV algorithm. The maximum magnitude of the difference images corresponding to the standard FISTA-TV algorithm is an order of magnitude larger than those corresponding to the OSSF-

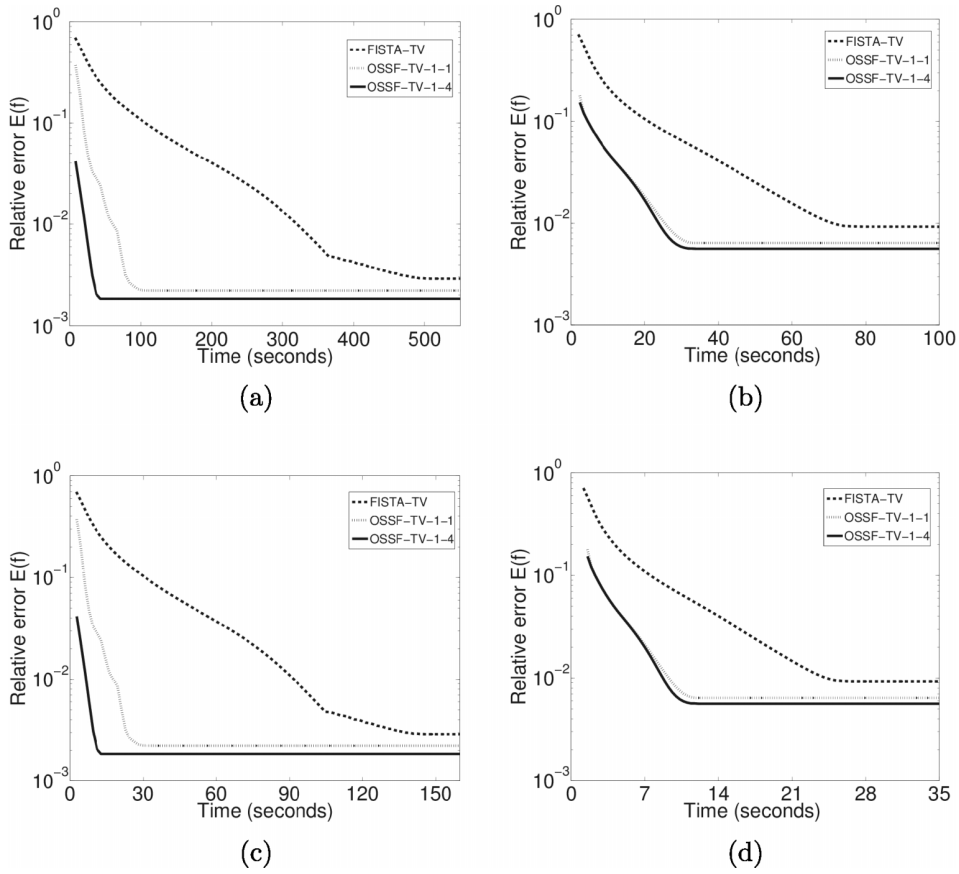


FIG. 7. Plots of image RE as a function of reconstruction time for (a) full-view case with one GPU, (b) few-view case with one GPU, (c) full-view case with four GPUs, and (d) few-view case with four GPUs.

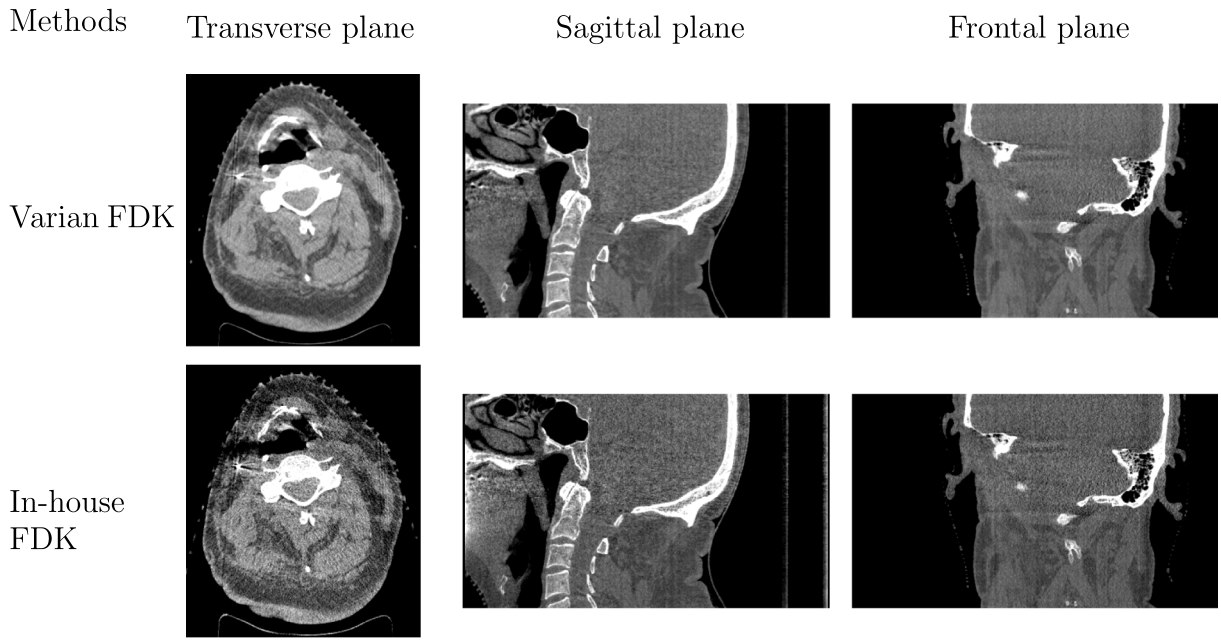


FIG. 8. Example images reconstructed by FDK algorithms to convey projection data quality. First row: Images reconstructed by the commercial Varian software. Second row: Images reconstructed by our research-based FDK algorithm with a simple ramp filter. The transverse images are shown in a soft-tissue window [−300 200] HU. The sagittal and frontal images are shown in a display window [−500 800] HU.

TV algorithms. Unlike in the full-view case, the OSSF-TV-1-4 algorithm shows a similar performance to the OSSF-TV-1-1 algorithm due to the sparse nature of the projection data.

5.D. Convergence and accuracy curves

To quantify the improvement in convergence rate yielded by the OSSF-TV algorithm, the objective function values $F(f) = \|\mathbf{b}-\mathbf{H}\mathbf{f}_{\text{recon}}\|^2_{\mathbf{W}} + 2\lambda_{\text{tv}}\|\mathbf{f}_{\text{recon}}\|_{\text{TV}}$ were plotted as a function of iteration number. The curves corresponding to the FISTA-TV OSSF-TV-1-1 and OSSF-TV-1-4 algorithms are displayed in Fig. 5 for both the full- and sparse-view cases. For the full-view case [Fig. 5(a)], both the OSSF-TV-1-4 and the OSSF-TV-1-1 algorithms yielded a more rapid decay in the objective function values than the FISTA-TV algorithm. Specifically, the OSSF-TV-1-4 curve suggests that convergence has been approximately achieved by around five to six iterations in the full-view case. Even for the OSSF-TV-1-1 curve, only 18 iterations were required to achieve this. On the other hand, the curve corresponding to the FISTA-TV algorithm

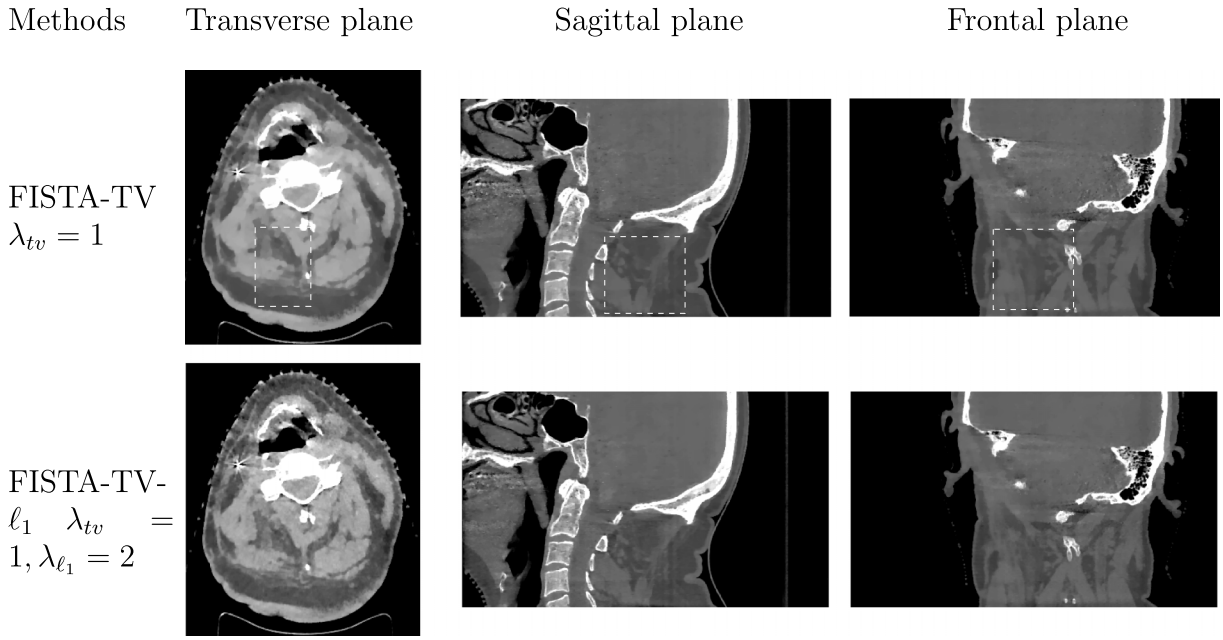


FIG. 9. Reference images reconstructed by the standard FISTA-TV (top row) and FISTA-TV-ℓ_1 (bottom row). In both cases, the algorithms were run to convergence. Three boxes with dashed while lines represent three extracted ROIs for comparison studies in Sec. 6.C.

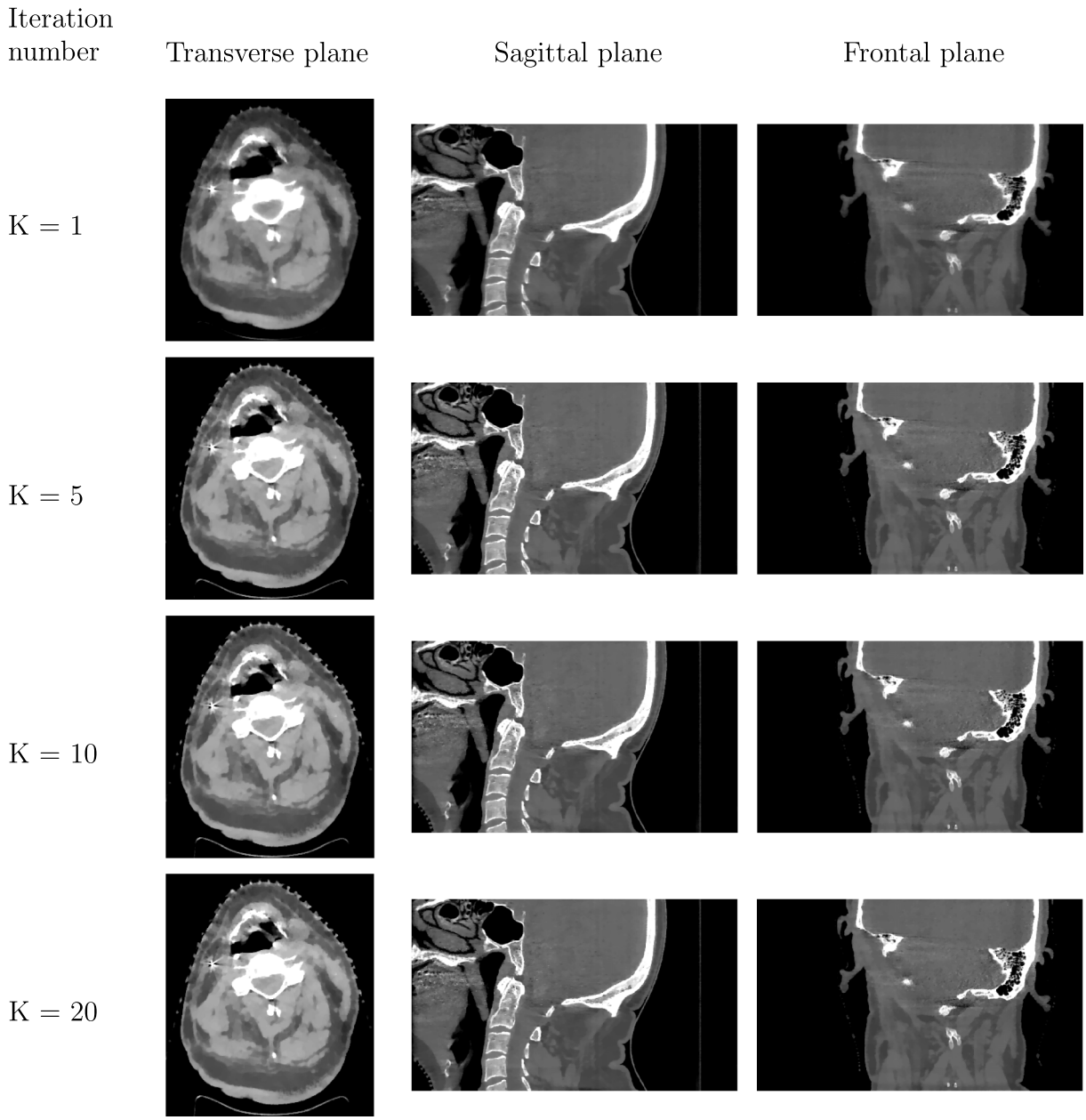


FIG. 10. Reconstructed images corresponding to different iteration numbers (K) for the OSSF-TV algorithm. First column: Transverse plane with soft-tissue display window [−300 200] HU. Second column: Sagittal plane with display window [−500 800] HU. Third column: Frontal plane with display window [−500 800] HU.

indicates that the algorithm requires more than one hundred iterations to achieve approximate convergence for the full-view case. Similar observations regarding the relative convergence rates of the algorithms were obtained for the sparse-view case shown in Fig. 5(b).

The relative error (RE) defined by $E(f) = \|f^{\text{recon}} - f^{\text{true}}\|_2 / \|f^{\text{true}}\|_2$, where f^{recon} and f^{true} denote the reconstructed and true phantom image, respectively, was also computed and plotted as a function of iteration number for the three algorithms in Fig. 6. The relative behavior of the RE curves is similar to the objective function curves described above. The small values of REs indicate that the solution of the optimization problem is close to the true phantom. The above results corroborate our claim that the OSSF-TV algorithms possess superior convergence rates as compared to the standard FISTA-TV algorithm while maintaining reconstruction accuracy. We have also verified that the OSSF-TV-ℓ_1 algorithm outperforms the FISTA-TV-ℓ_1 algorithm in a similar way.

5.E. Reconstruction time by using GPUs

Additional studies were conducted to quantify image reconstruction times. RE curves as a function of reconstruction time are plotted in Fig. 7 for the case when a single GPU [Figs. 7(a) and 7(b)] or four GPUs [Figs. 7(c) and 7(d)] were employed in the implementation. For the single-GPU case with full-view data [Fig. 7(a)], the OSSF-TV-1-4 algorithm required only 33 s to reach the approximate convergence point. With the

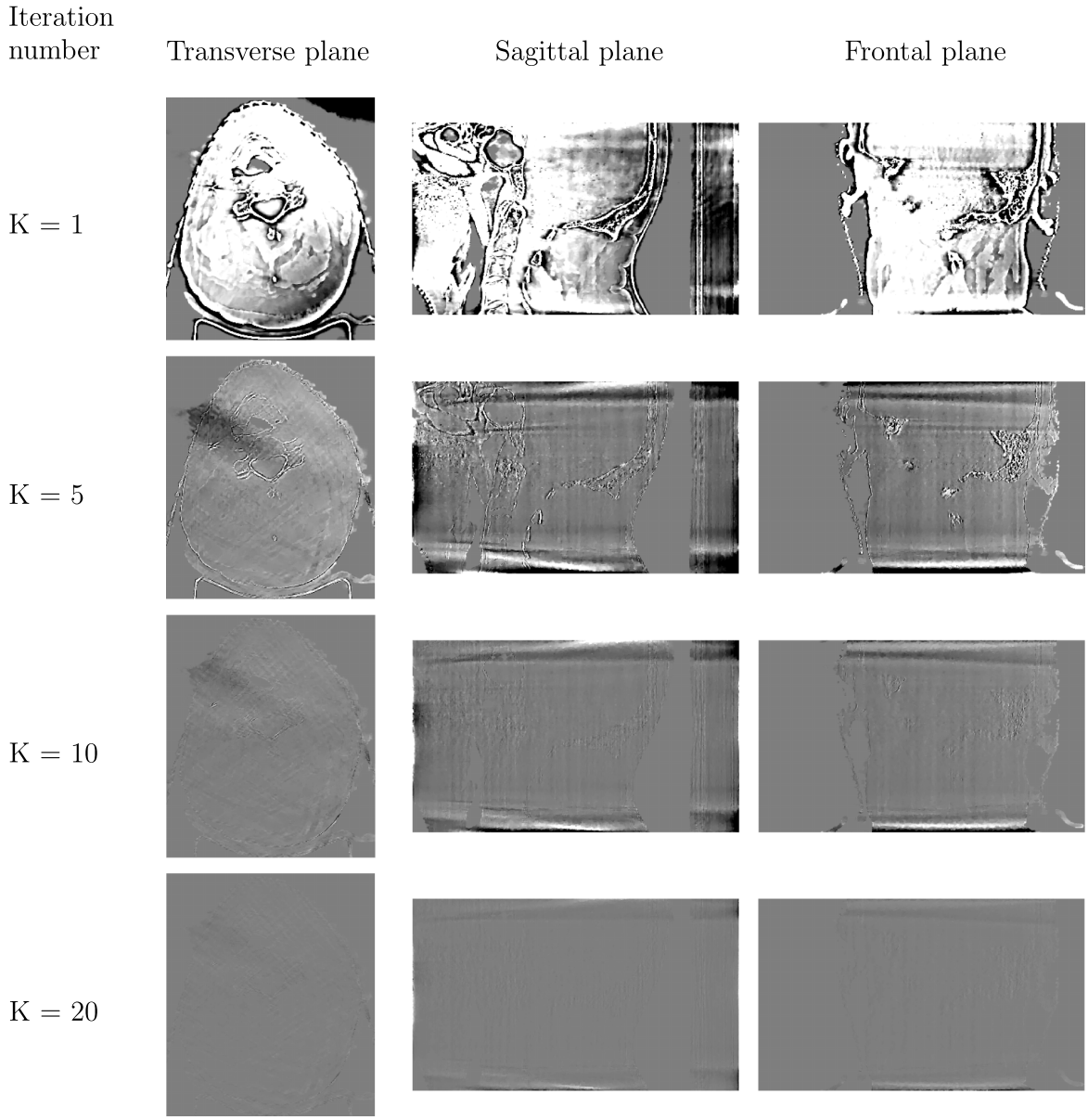


FIG. 11. Difference images corresponding to different iteration numbers (K) for the OSSF-TV algorithm. The display window was [−25 25] HU.

sparse-view data, it required approximately 34 s. However, it should be noted that diagnostically useful images may be produced by the algorithm before this degree of convergence is obtained. With the four-GPU implementation, the OSSF-TV-1-4 algorithm required 10 s to converge with the full-view data and 11 s for the sparse-view case. These results are consistent with the claim in Appendix E that the multi-GPU implementations of the OSSF-TV algorithm will provide a speed-up over the single-GPU implementation that is linearly proportional to the number of GPUs employed.

6. INVESTIGATION OF ALGORITHM PERFORMANCE USING CLINICAL DATA

The rapid convergence rates of the OSSF-TV and OSSF-TV-ℓ_1 algorithms were corroborated by use of clinical CBCT projection data. Because the OSSF-TV and OSSF-TV-ℓ_1 algorithms both employed the -1-4 data accessing strategy in these studies, the suffix -1-4 to the algorithm names is omitted below.

6.A. Experimental data and image reconstruction

Previously acquired circular CBCT projection data corresponding to a head-and-neck cancer patient were obtained under an IRB-approved IGRT study. The data were acquired by use of a kilovoltage (kV) On-Board Imager (OBI) on a Varian TrueBeam radiation therapy treatment machine (Varian Medical System, USA). The source-to-axis distance (SAD) and detector-to-axis distances were 100 and 50 cm, respectively. A flat panel detector of size 30 cm (768 rows) × 40 cm (1024 columns) was employed. Additional details regarding the imaging hardware are described elsewhere.[54] The data set was comprised of 364 uniformly spaced projections that spanned an angular range of approximately 200°.

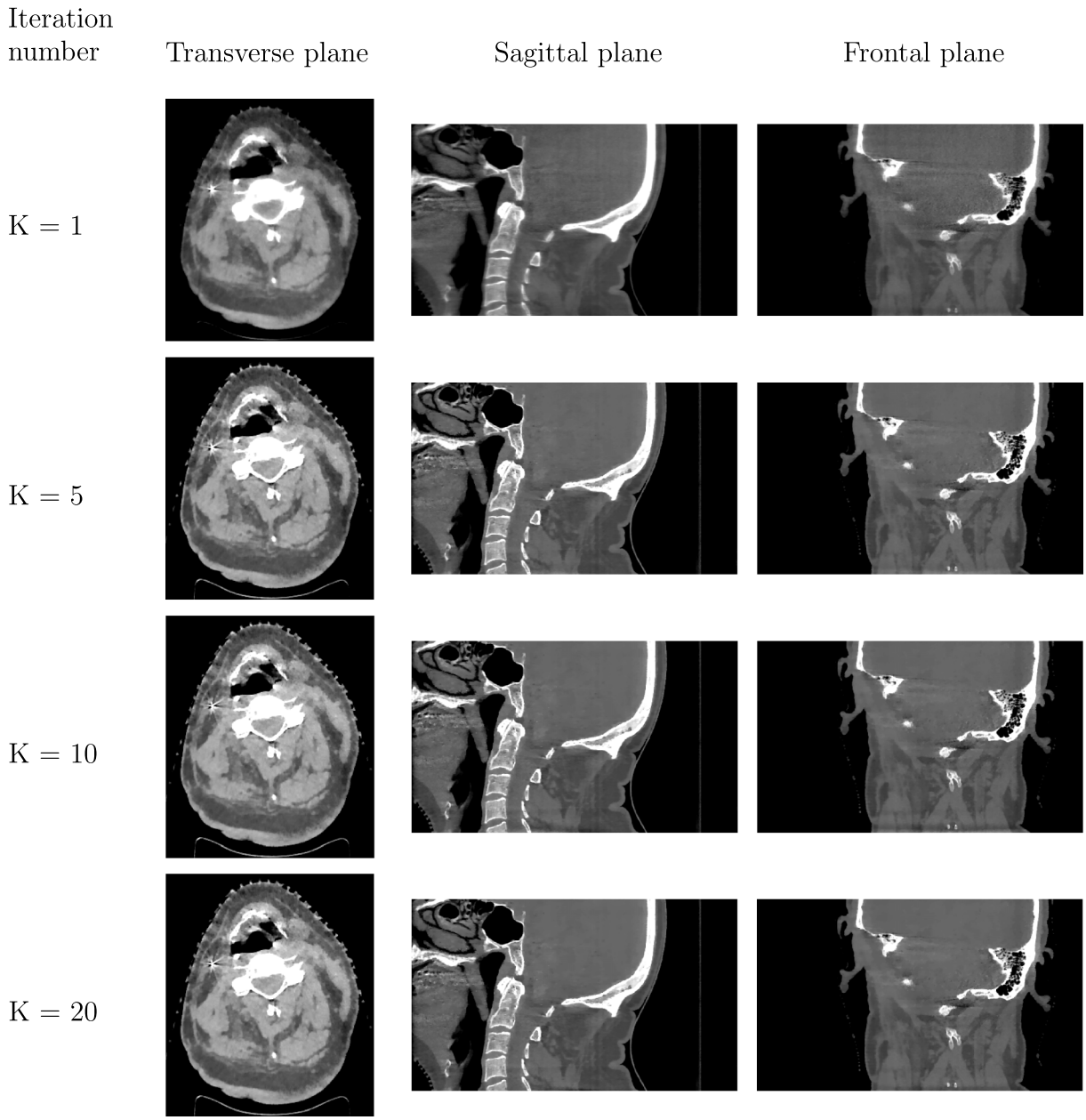


FIG. 12. Reconstructed images corresponding to different iteration numbers (K) for the OSSF-TV-ℓ_1 algorithm. First column: Transverse plane with soft-tissue display window [−300 200] HU. Second column: Sagittal plane with display window [−500 800] HU. Third column: Frontal plane with display window [−500 800] HU.

The acquired raw projection data were subjected to five preprocessing steps: scatter correction, air normalization, bowtie filtration, beam-hardening correction, and logarithm transformation, as described in a previous study.[54] Examples of images reconstructed by use of the clinically employed Varian software package and our own FDK algorithm with a simple ramp filter are shown in Fig. 8. These images are presented to give the reader a qualitative sense of the data quality.

Reference images were computed by running the standard FISTA-TV and FISTA-TV-ℓ_1 algorithms thousands of times until the values of the object function did not change up to the single precision floating point accuracy. The value of the regularization parameter λ_{tv} was set at 1.0 for both the PWLS-TV and PWLS-TV-ℓ_1 estimators and λ_{ℓ_1} was set at 2.0 for PWLS-TV-ℓ_1 estimator. These images, shown in Fig. 9, were employed to evaluate the accuracy of the OSSF-TV and OSSF-TV-ℓ_1 algorithms, respectively.

In the implementations of the OSSF-TV and OSSF-TV-ℓ_1 algorithms, the multi-GPU scheme described in Appendix E with four NVIDIA Tesla K40 GPUs was employed. All reconstructed images were of dimension 512×512×379 (slices) with a voxel dimension of 0.512 mm.

6.B. Demonstration of rapid convergence rate with clinical data

A series of images corresponding to three orthogonal planes through the volumetric images reconstructed by use of the

OSSF-TV algorithm at iteration numbers $K = 1$, 5, 10, and 20 are shown in Fig. 10. These results reveal that the visual appearances of the images after the 5th iteration do not considerably vary. This observation is consistent with the behavior of the difference images corresponding to the three planes that are displayed in Fig. 11. The difference images were produced by subtracting the OSSF-TV reconstructed images from the reference image produced by the standard FISTA-TV algorithm that was run to convergence (i.e., no change in objective function value to machine precision). The difference images reveal that homogeneous tissue regions have been accurately recovered by the 5th iteration. The tissue interfaces and small bone features have been accurately recovered by the 10th iteration. By the 20th iteration, the maximum values of the difference images were only a few HUs (at the boundaries) reflecting that the reconstructed image is nearly identical to the reference image.

The same images reconstructed by use of the OSSF-TV-ℓ_1 algorithm at iteration numbers $K = 1$, 5, 10, and 20 are shown in Fig. 12. The corresponding difference images are displayed in Fig. 13. The observations described above regarding the rapid convergence rate of the OSSF-TV algorithm were found to also describe the behavior of the OSSF-TV-ℓ_1 algorithm. The corresponding RE curves are shown in Fig. 14. Figure 14(b) demonstrates that a volumetric CBCT image (close to convergence) can be reconstructed from a clinical data set in under 4 minutes (240 s) with the proposed multi-GPU scheme implementation by use of four NVIDIA K40 GPUs. This reconstruction time can be reduced readily by use of additional and/or more powerful GPU cards. For example, by simply switching to NVIDIA K80 GPUs, the reconstruction time can be expected to be approximately cut in half (approximately 2 minutes). Moreover, it should be noted that the zero-image was employed to initialize all algorithms in this study. The use

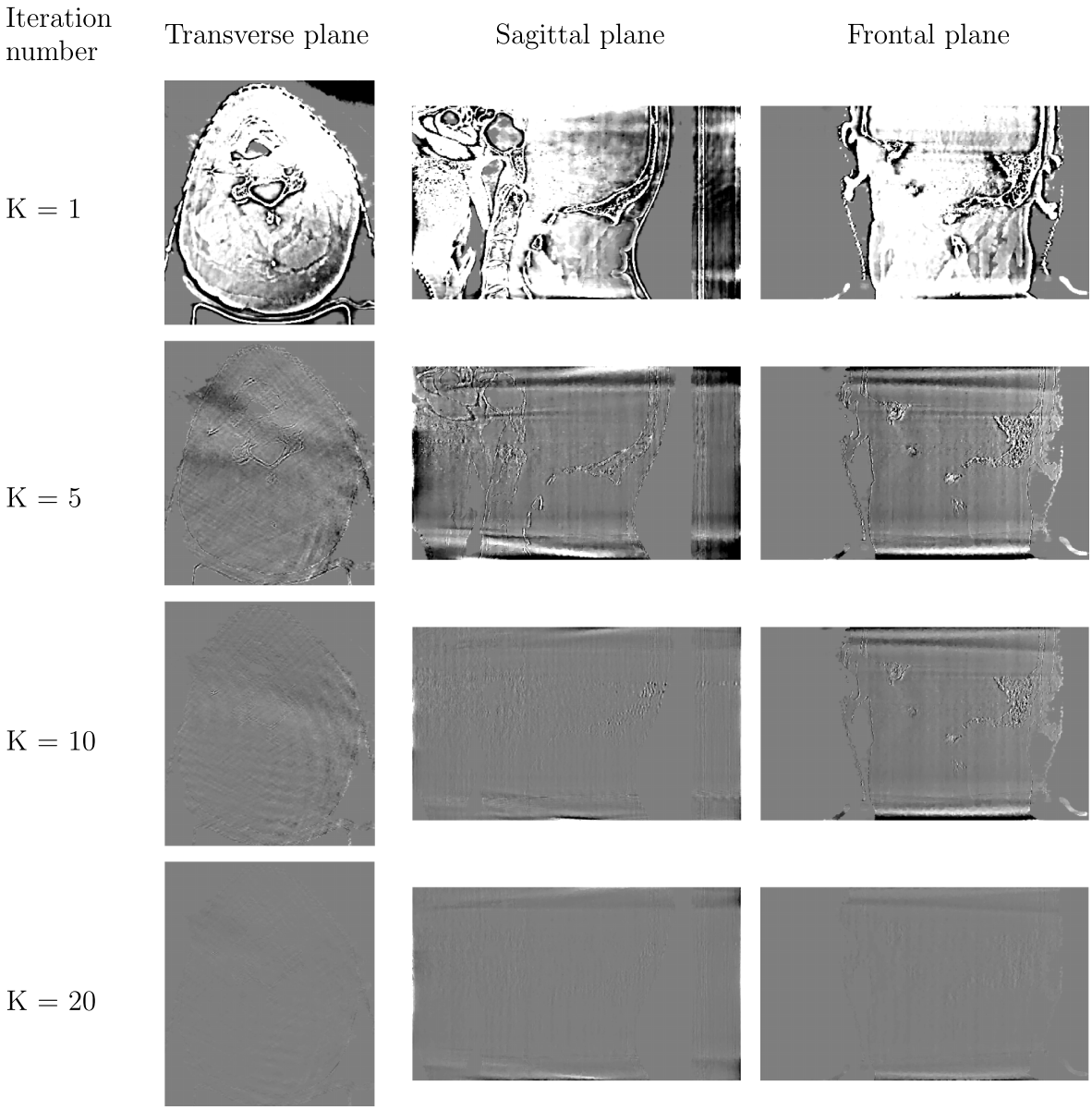


FIG. 13. Difference images corresponding to different iteration numbers (K) for the OSSF-TV-ℓ_1 algorithm. The display window was [−25 25] HU.

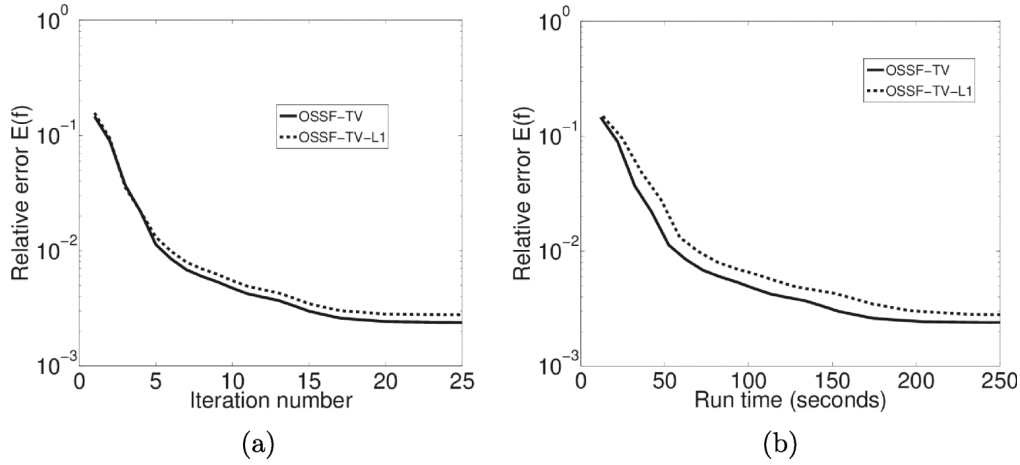


FIG. 14. (a) Plots of the REs as a function of iteration number for images reconstructed by use of the OSSF-TV (solid) and OSSF-TV-$\ell 1$ (dashed) algorithms. (b) Plots of the REs as a function of computation time for images reconstructed by use of OSSF-TV (solid) and OSSF-TV-$\ell 1$ (dashed) algorithms. Four GPUs were employed in the implementation as described in the text.

of more accurate initial guesses can result in further reductions in reconstruction times.

6.C. Demonstration of the effect of the additional wavelet-sparsified ℓ_1-norm penalty

As mentioned previously, it has been suggested that the use of a wavelet-sparsified ℓ_1-norm penalty in combination with a TV penalty can potentially mitigate patch-like artifacts and improve certain measures of image quality. In order to demonstrate this, three region-of-interests (ROIs) were extracted from the FISTA-TV- and FISTA-TV-ℓ_1- produced reference images in Fig. 9. The corresponding three ROIs were also extracted from images in Fig. 8 that were produced by use of the Varian FDK algorithm. The locations of the three ROIs are indicated by the white boxes in Fig. 9. The results produced by use of the FISTA-TV and FISTA-TV-ℓ_1 are displayed in the first and second rows of Fig. 15, while the results produced by use of the Varian FDK algorithm are displayed in the third row of Fig. 15. The soft-tissue display window was [−300 200] HU. A visual inspection of these images suggests that more details and small structures are present in the images reconstructed by use of the FISTA-TV-ℓ_1 compared with those produced by use of the FISTA-TV, while the images produced by use of FISTA-TV contain significantly more blurring but lower noise level. The soft-tissue structures presented in FISTA-TV-ℓ_1 images are also present in the images produced by use of the Varian FDK algorithm. However, the latter images appear to contain higher noise levels. As described previously, the regularization parameter $\lambda_{\text{tv}} = 1.0$ was kept the same for both FISTA-TV and FISTA-TV-ℓ_1. While the image produced by use of the FISTA-TV could be potentially improved through additional tuning of λ_{tv}, our preliminary observations support the conjecture that the combination of the TV and ℓ_1-norm penalties may provide the opportunity to enhance image quality over use of the TV penalty alone. This was the motivation for developing accelerated algorithms for solving the PWLS-TV-ℓ_1 estimation problem in this work. In addition, the wavelet transform can be replaced by another sparsity promoting transformation such as the curvelet[55,56] and ridgelet[57,58] transforms. In those cases, new ℓ_1-proximal problems may need to be introduced.

7. DISCUSSION

7.A. Convergence rates compared to some recently proposed algorithms

The convergence rates of several recently reported CBCT reconstruction algorithms that solve PWLS-TV optimization problem are discussed below. The gradient projection Barzilai–Borwein (GP-BB) method[8] was employed to produce a RE curve corresponding to use of 40 uniformly spaced fan-beam projections of a 2D Shepp–Logan phantom. This revealed that the GP-BB method required approximately 20 iterations to force the RE down to 10%. (See Fig. 5 in Ref. 8.) Another recent work[25] employed an unknown-parameter Nesterov (UPN) method, which was interpreted as an improved version of the GP-BB method.[7] However, in studies involving 66 projections of a 2D Shepp–Logan phantom, the algorithm was reported to require more than 60 iterations to force the RE down to 10%. (See Fig. 2 in Ref. 25.) Another method[4] was employed to produce a RE curve in a study that utilized 40 uniformly distributed CBCT projections of an NCAT phantom. The method was reported to require approximately 50 iterations to achieve a 10% RE value. (See Fig. 6 in Ref. 4.)

To emphasize the superior convergence rates of the accelerated FISTAs proposed in this work, recall that the OSSF-TV algorithm required only three iterations to reduce the RE to 10% in the NCAT phantom study involving 45 CBCT projections. Moreover, the RE values decreased to 1% after only 22 iterations [Fig. 6(b)]. It is interesting to note that, for the same phantom study, even the standard FISTA-TV method required only 23 iterations to reach a RE value of 10%, which indicates it possesses a faster convergence rate than the algorithms mentioned above. This observation is consistent with the fact that the FISTA possesses a quadratic convergence rate while the algorithms highlighted above, as well as other recently proposed algorithms,[3,5] possess first-order convergence rates.

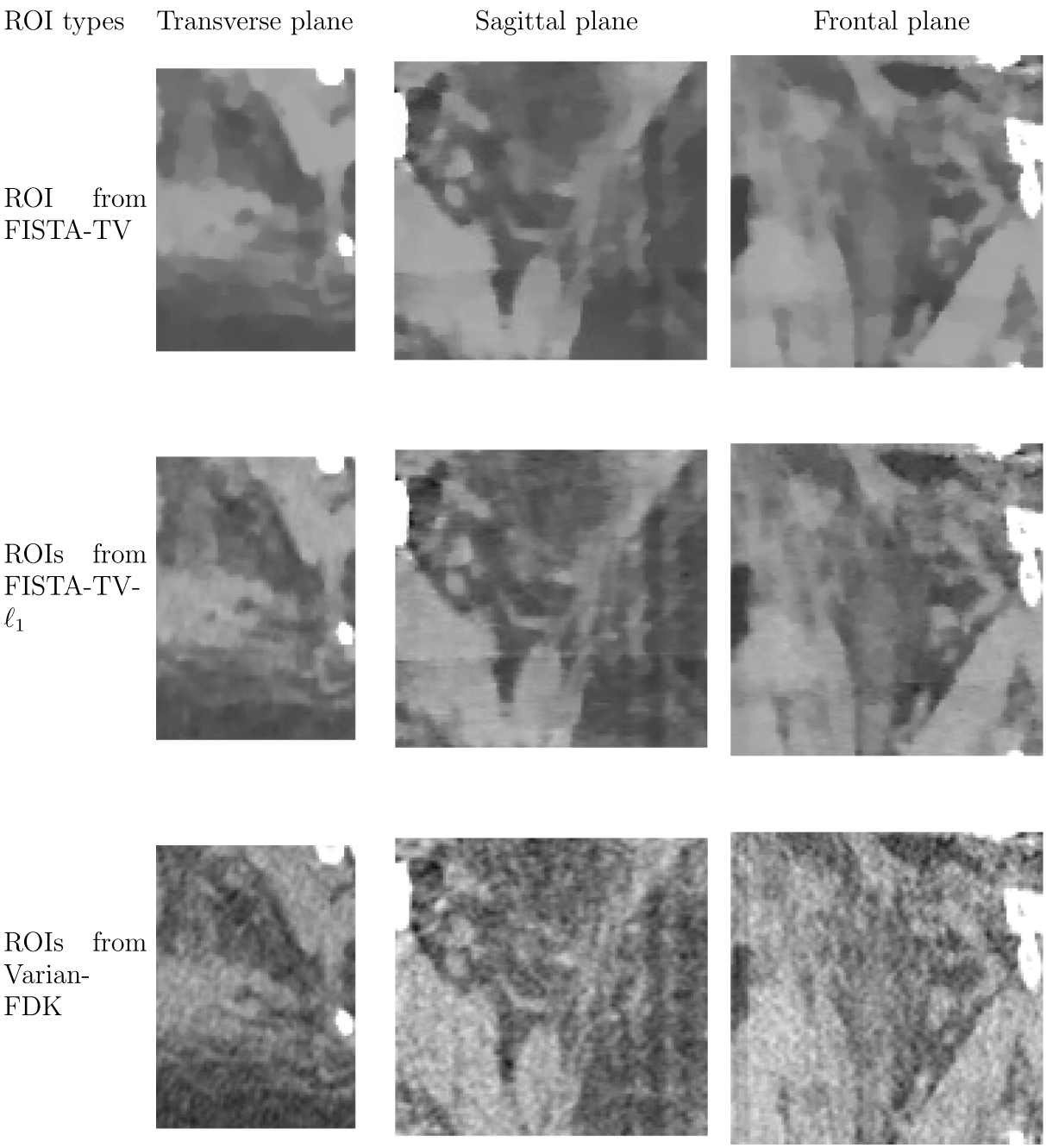


FIG. 15. Three ROIs were selected from the FISTA-TV (first row) and FISTA-TV-ℓ_1 (second-row) reference images in Fig. 9. The corresponding ROIs reconstructed by use of the Varian FDK algorithm in Fig. 8 are displayed in the bottom row. All images are displayed in soft-tissue window [−300 200] HU.

These results suggest that, even though there were differences in the measurement data utilized in the different studies, it is highly likely that the proposed accelerated algorithms possess a significant performance advantage over the previously reported works that were mentioned.

7.B. Other modern algorithms that employ OS strategies

A preliminary version of this work was presented at the 2012 AAPM meeting.[22] Since then, other groups have explored similar ideas. For example, the OS-type strategy has been incorporated with other optimization methods to form accelerated algorithms[31,59] for helical CT. In Ref. 59 an augmented Lagrangian method is combined with an ordered subsets' approach to solve a penalized weighted least square (PWLS) problem with Tikhonov regularization. A different work[31] employed an ordered subsets' strategy to accelerate a nonuniform separable quadratic surrogate algorithm (NU-SQS) that solves a PWLS problem with Tikhonov regularization. Another OS-based algorithm[51] has been proposed to accelerate CT image reconstruction. In that work, a deterministic downward continuation method was employed to determine the step size in their proposed OS-LALM algorithm.[51] However, the problem

formulation and optimization methods employed in those works are different from those employed to develop the OSSF-TV and OSSF-TV-ℓ_1 algorithms. A systematic comparison between those works and the proposed OSSF-TV or OSSF-TV-ℓ_1 algorithms remains a task for future study.

8. SUMMARY AND CONCLUSION

The FISTA is a modern optimization algorithm that possesses a quadratic convergence rate and is suitable for minimizing PWLS cost functions that contain nonsmooth penalties. In this work, accelerated variants of the FISTA were proposed and investigated for CBCT image reconstruction. Algorithm acceleration was achieved by replacing the gradient-descent step in the standard FISTAs by an OS-SART subproblem, which is essentially one type of preconditioned gradient-based scheme combined with the ordered subset concept. Because of the adopted preconditioning matrices, two weighted proximal problems corresponding to TV penalty and wavelet-based ℓ_1 penalty were introduced and solved by FGP-type algorithms.

The developed numerical framework will allow researchers to design their own preconditioning matrices and adapt the proposed FGP-type algorithms to solve the corresponding weighted proximal problems. Moreover, the proposed approach can be applied readily to accelerate FISTAs that solve PWLS reconstruction problems that utilize alternative sparsity-promoting penalty forms than the ones considered in this work. By use of computer-simulated CBCT data, it was verified that the OSSF-TV and OSSF-TV-ℓ_1 algorithms possessed significantly greater convergence rates than the corresponding standard FISTAs. The rapid convergence properties of the algorithms were verified further by use of clinical CBCT data.

A reconstruction algorithm that possesses a rapid convergent rate can potentially produce a diagnostically useful image in fewer iterations than an algorithm that possesses a slower convergence rate. However, a rapid convergence rate does not necessarily translate into shortened reconstruction times. This depends on how efficiently each iteration can be computed. In order to reduce image reconstruction times in practice, we developed efficient GPU implementations of the proposed algorithms that utilize either a single or multiple GPUs. When multiple GPUs are employed, we demonstrated that the reduction in reconstruction time over the single GPU implementation is approximately linear with the number of GPUs employed. The rapid convergence rates of the algorithms coupled with efficient GPU implementations may make them suitable for certain time-sensitive clinical applications.

The topic of optimizing image quality has intentionally not been addressed in this paper, as our singular focus has been on the development of accelerated iterative image reconstruction algorithms for CBCT. Our results suggest that images reconstructed by use of the accelerated FISTAs will have an accuracy that is comparable to those reconstructed by use of the standard FISTA. How to specify the penalty form and regularization parameters in a PWLS estimator that are most appropriate for a particular diagnostic task is beyond the scope of this study. However, because the developed algorithms can drastically reduce image reconstruction times, they can facilitate the systematic investigation of such issues.

The formulations of the reconstruction problems in this work have not explicitly exploited information regarding the statistical properties of the projection data. Incorporating statistical information can potentially improve image quality[15,33] in, for example, low-dose imaging applications.[15] Fortunately, the proposed OSSF-TV and OSSF-TV-ℓ_1 algorithms can be generalized readily to exploit statistical information. Specifically, the weight matrix $\mathbf{W}$ can be defined as the inverse covariance matrix of the data.[15,33,60] In this case, the geometrically motivated weight matrix simply needs to be replaced by the corresponding statistical weight matrix and the preconditioning matrix in Eq. (27) should be appropriately redefined. More iterations may be needed to solve the weighted proximal problems for each subset when the matrix $\mathbf{D}$ has a very large dynamic range. Another way to incorporate the statistical weights into the SART (or OS-SART) method has been reported by Gregor and Fessler.[61] In that work, the authors combined a diagonal statistical weight matrix and geometry weight matrix to form a new diagonal weight matrix and a new preconditioning matrix for a SART algorithm. Besides CBCT, the proposed algorithms can also be explored for other CT imaging applications such as helical CBCT. The investigation of these topics can be pursued in future studies.

ACKNOWLEDGMENTS

This work was supported in part by NIH Award Nos. EB009715 and EB010049 and NSF Award No. CBET 1263988.

APPENDIX A: DESCRIPTION OF THE FGP ALGORITHM TO SOLVE THE STANDARD 3D TV-PROXIMAL PROBLEM

Below, the 3D FGP algorithm for solving the standard TV-proximal problem is described. Without loss of generality, we assume that $g_{\text{tv}}(\mathbf{u}) = c_1\lambda_{\text{tv}}\|\mathbf{u}\|_{\text{TV}}$, where c_1 is a positive constant. Therefore,

$$\text{prox}_{1/L}(g_{\text{tv}})(\mathbf{x}_g) := \arg\min_{\mathbf{u}}\left\{c_1\lambda_{\text{tv}}\|\mathbf{u}\|_{\text{TV}} + \frac{L}{2}\|\mathbf{u}-\mathbf{x}_g\|^2\right\}, \quad \text{(A1)}$$

which is equivalent to the minimization problem

$$\hat{\mathbf{u}} := \arg\min_{\mathbf{u}\in C}\left\{\|\mathbf{u}-\mathbf{x}_g\|^2 + 2\alpha\|\mathbf{u}\|_{\text{TV}}\right\}, \quad \text{(A2)}$$

where $\alpha = c_1\lambda_{\text{tv}}/L$. It has been demonstrated[11] that the FGP method can efficiently solve the above problem in 2D case. Algorithm III describes the extension of the 2D FGP algorithm to 3D, for use with CBCT image reconstruction. The 3D FGP algorithm structure is very similar to that of the 2D FGP algorithm. Because a third dimension is added, the Lipschitz constant becomes larger in 3D FGP algorithm. Following the derivation of Lemma 4.2 in Ref. 11, the factor 12 is employed in Eq. (A3) of the 3D FGP algorithm instead of the factor 8 employed in the original 2D FGP algorithm.

ALGORITHM III. FGP algorithm to solve the TV-proximal problem in Eq. (A2).

Input: $\mathbf{x}_g$, α

Output: $\hat{\mathbf{u}}$—An optimal solution of Eq. (A1) (up to a tolerance).

Step 0. Take $(\mathbf{r}^1,\mathbf{s}^1,\mathbf{t}^1)=(\mathbf{o}^0,\mathbf{p}^0,\mathbf{q}^0)=[\mathbf{0}_{(m-1)\times n\times l},\mathbf{0}_{m\times(n-1)\times l},\mathbf{0}_{m\times n\times(l-1)}]$

for $k\leftarrow 1,K$ **do**

$$(\mathbf{o}^k,\mathbf{p}^k,\mathbf{q}^k)=P_{\mathcal{P}}\left[(\mathbf{r}^k,\mathbf{s}^k,\mathbf{t}^k)+\frac{1}{12\alpha}\mathcal{L}^T(P_C[\mathbf{x}_g-\alpha\mathcal{L}(\mathbf{r}^k,\mathbf{s}^k,\mathbf{t}^k)])\right] \quad \text{(A3)}$$

$$t_{k+1}=\frac{1+\sqrt{1+4t_k^2}}{2} \quad \text{(A4)}$$

$$(\mathbf{r}^{k+1},\mathbf{s}^{k+1},\mathbf{t}^{k+1})=(\mathbf{o}^k,\mathbf{p}^k,\mathbf{q}^k)+\left(\frac{t_k-1}{t_{k+1}}\right)(\mathbf{o}^k-\mathbf{o}^{k-1},\mathbf{p}^k-\mathbf{p}^{k-1},\mathbf{q}^k-\mathbf{q}^{k-1}) \quad \text{(A5)}$$

end for

Set $\mathbf{f}_K=P_C[\mathbf{x}_g-\alpha\mathcal{L}(\mathbf{o}^K,\mathbf{p}^K,\mathbf{q}^K)]$

The relevant operators are explicitly defined as follows.

- The linear operator $\mathcal{L}:\mathbb{R}^{(m-1)\times n\times l}\times\mathbb{R}^{m\times(n-1)\times l}\times\mathbb{R}^{m\times n\times(l-1)}\rightarrow\mathbb{R}^{m\times n\times l}$ is a classical discretized variant of divergence operator with Neumann boundary conditions, which can be defined as

$$(\mathcal{L}(\mathbf{r},\mathbf{s},\mathbf{t}))_{i,j,h}=[\mathbf{r}]_{i,j,h}-[\mathbf{r}]_{i-1,j,h}+[\mathbf{s}]_{i,j,h}-[\mathbf{s}]_{i,j-1,h}+[\mathbf{t}]_{i,j,h}-[\mathbf{t}]_{i,j,h-1},$$
$$i=1,\ldots,m,\quad j=1,\ldots,n,\quad h=1,\ldots,l$$

  where we assume that $[\mathbf{r}]_{0,j,h}=[\mathbf{r}]_{m,j,h}=[\mathbf{s}]_{i,0,h}=[\mathbf{s}]_{i,n,h}=[\mathbf{t}]_{i,j,0}=[\mathbf{t}]_{i,j,l}\equiv 0$, for every $i=1,\ldots,m$ and $j=1,\ldots,n$ and $h=1,\ldots,l$. In our CBCT case, the values of m, n, and l represent the dimensions of the 3D discrete object.
- P_C is an orthogonal projection operator onto the convex feasible set C. In our CBCT case, we consider the operator P_C is a non-negativity constraint

$$P_C[\mathbf{x}]=\max\{\mathbf{0},\mathbf{x}\},$$

  where $\mathbf{x}$ is an arbitrary input matrix and max applies on the vector or matrix $\mathbf{x}$ in an element-wise way.
- The operator $\mathcal{L}^T$: $\mathbb{R}^{m\times n\times l}\rightarrow\mathbb{R}^{(m-1)\times n\times l}\times\mathbb{R}^{m\times(n-1)\times l}\times\mathbb{R}^{m\times n\times(l-1)}$, which is the adjoint of $\mathcal{L}$, a negative gradient operator, given by

$$\mathcal{L}^T(\mathbf{x})=(\mathbf{r},\mathbf{s},\mathbf{t}),$$

  where $\mathbf{r}\in\mathbb{R}^{(m-1)\times n\times l}$, $\mathbf{s}\in\mathbb{R}^{m\times(n-1)\times l}$, and $\mathbf{t}\in\mathbb{R}^{m\times n\times(l-1)}$ are the matrices defined by

$$[\mathbf{r}]_{i,j,h}=[\mathbf{x}]_{i,j,h}-[\mathbf{x}]_{i+1,j,h},$$
$$i=1,\ldots,m-1,j=1,\ldots,n,h=1,\ldots,l$$
$$[\mathbf{s}]_{i,j,h}=[\mathbf{x}]_{i,j,h}-[\mathbf{x}]_{i,j+1,h},$$
$$i=1,\ldots,m,j=1,\ldots,n-1,h=1,\ldots,l$$
$$[\mathbf{t}]_{i,j,h}=[\mathbf{x}]_{i,j,h}-[\mathbf{x}]_{i,j,h+1},$$
$$i=1,\ldots,m,j=1,\ldots,n,h=1,\ldots,l-1.$$

- The operator $P_{\mathcal{P}}$: $\mathbb{R}^{(m-1)\times n\times l}\times\mathbb{R}^{m\times(n-1)\times l}\times\mathbb{R}^{m\times n\times(l-1)}\rightarrow\mathbb{R}^{(m-1)\times n\times l}\times\mathbb{R}^{m\times(n-1)\times l}\times\mathbb{R}^{m\times n\times(l-1)}$, which is a projection operator onto the set $\mathcal{P}$ such as

$$P_{\mathcal{P}}(\mathbf{r},\mathbf{s},\mathbf{t})=(\mathbf{o},\mathbf{p},\mathbf{q})$$

where $\mathbf{r},\mathbf{s},\mathbf{t}$ and $\mathbf{o},\mathbf{p},\mathbf{q}$ denote the input and output matrices, respectively, are the matrices defined by

$$[\mathbf{o}]_{i,j,h}=\frac{\mathbf{r}_{i,j,h}}{\max\{1,\sqrt{[\mathbf{r}]_{i,j,h}^2+[\mathbf{s}]_{i,j,h}^2+[\mathbf{t}]_{i,j,h}^2}\}},$$
$$i=1,\ldots,m-1,j=1,\ldots,n,h=1,\ldots,l$$
$$[\mathbf{p}]_{i,j,h}=\frac{\mathbf{s}_{i,j,h}}{\max\{1,\sqrt{[\mathbf{r}]_{i,j,h}^2+[\mathbf{s}]_{i,j,h}^2+[\mathbf{t}]_{i,j,h}^2}\}},$$
$$i=1,\ldots,m,j=1,\ldots,n-1,h=1,\ldots,l$$
$$[\mathbf{q}]_{i,j,h}=\frac{\mathbf{t}_{i,j,h}}{\max\{1,\sqrt{[\mathbf{r}]_{i,j,h}^2+[\mathbf{s}]_{i,j,h}^2+[\mathbf{t}]_{i,j,h}^2}\}},$$
$$i=1,\ldots,m,j=1,\ldots,n,h=1,\ldots,l-1,$$

where we have $[\mathbf{r}]_{m,j,h}=[\mathbf{s}]_{i,n,h}=[\mathbf{t}]_{i,j,l}\equiv 0$.

APPENDIX B: DESCRIPTION OF WAVELET-BASED ℓ_1 PROXIMAL PROBLEM

Without loss of generality, consider that $g_{\ell_1,\Phi}(\mathbf{u})=c_2\lambda_{\ell_1}\|\Phi\mathbf{u}\|_1$, in which c_2 is a positive constant and Φ is a 3D discrete Daubechies wavelet transform operator. In this case, the relevant proximal problem is defined as

$$\text{prox}_{1/L}(g_{\ell_1,\Phi})(\mathbf{x}_g):=\arg\min_{\mathbf{u}}\left\{c_2\lambda_{\ell 1}\|\Phi\mathbf{u}\|_1+\frac{L}{2}\|\mathbf{u}-\mathbf{x}_g\|^2\right\}, \quad \text{(B1)}$$

which is equivalent to the minimization

$$\hat{\mathbf{u}}:=\arg\min_{\mathbf{u}}\{\|\mathbf{u}-\mathbf{x}_g\|^2+\beta\|\Phi\mathbf{u}\|_1\}, \quad \text{(B2)}$$

where $\beta=2c_2\lambda_{\ell 1}/L$. Since the Daubechies wavelets are orthogonal, Eq. (B2) is equivalent to the following minimization problem

$$\hat{\mathbf{u}}:=\arg\min_{\mathbf{u}}\{\|\Phi\mathbf{u}-\Phi\mathbf{x}_g\|^2+\beta\|\Phi\mathbf{u}\|_1\}, \quad \text{(B3)}$$

or

$$\hat{\tilde{\mathbf{u}}}:=\arg\min_{\tilde{\mathbf{u}}}\{\|\tilde{\mathbf{u}}-\tilde{\mathbf{x}}_g\|^2+\beta\|\tilde{\mathbf{u}}\|_1\}, \quad \text{(B4)}$$

where $\tilde{u}$ and $\tilde{\mathbf{x}}_g$ represent the wavelet coefficients of $\mathbf{u}$ and $\mathbf{x}_g$, respectively. It has been demonstrated[12,34] that the iterative shrinkage thresholding algorithm (ISTA) can readily solve this problem by employing an operator

$$\hat{\tilde{\mathbf{u}}}=\mathcal{T}_\beta(\tilde{\mathbf{x}}_g),$$

where $\mathcal{T}_\beta$ is defined as

$$\mathcal{T}_\beta(\tilde{\mathbf{x}}_g)=(\|\tilde{\mathbf{x}}_g\|-\beta)_+\text{sgn}(\tilde{\mathbf{x}}_g),$$

where $(\cdot)_+$ returns the argument if it is positive and returns zero otherwise, the function sgn returns the sign of $(\cdot)$ and all operations are performed in an element-wise way. The ISTA requires only one computation of the discrete wavelet transform of $\mathbf{x}_g$ to obtain the wavelet coefficients $\tilde{\mathbf{x}}_g$, a shrinkage-thresholding operator to get the solution and followed by an inverse discrete wavelet transform. Both of these operations can be computed efficiently.

APPENDIX C: DESCRIPTION OF A MODIFIED FGP ALGORITHM TO SOLVE THE WEIGHTED TV-PROXIMAL PROBLEM IN EQ. (30)

Below, a modified FGP algorithm is described to solve the weighted TV-proximal problem given by

$$\begin{aligned}\text{prox}_{\gamma_v}^{D_v^{-1}}(g_{\text{tv}}/T)(\mathbf{e}_v^k) &= \arg\min_{\mathbf{u}}\left\{\frac{g_{\text{tv}}(\mathbf{u})}{T}+\frac{1}{2\gamma}\|\mathbf{u}-\mathbf{e}_v^k\|^2_{D_v^{-1}}\right\}\\ &= \arg\min_{\mathbf{u}}\left\{\|\mathbf{u}-\mathbf{e}_v^k\|^2_{D_v^{-1}}+\frac{2\gamma_v}{T}c_1\lambda_{\text{tv}}\|\mathbf{u}\|_{\text{TV}}\right\}.\end{aligned} \tag{C1}$$

To consider the effect of the weighted matrix D_v^{-1}, Eq. (A3) in Algorithm III needs to be modified[32] as

$$\begin{aligned}(\mathbf{o}^k,\mathbf{p}^k,\mathbf{q}^k) = P_{\mathcal{P}}\Big[&(\mathbf{r}^k,\mathbf{s}^k,\mathbf{t}^k)+\frac{1}{12\alpha'\max(D_v)}\\ &\times\mathcal{L}^T(P_C[\mathbf{x}_g-\alpha' D_v\mathcal{L}(\mathbf{r}^k,\mathbf{s}^k,\mathbf{t}^k)])\Big],\end{aligned} \tag{C2}$$

where $\alpha'=c_1\lambda_{\text{tv}}\gamma_v/T$, and $\max(D_v)$ is the maximum element of the matrix D_v. Finally, the solution after Kth iteration will be set to $\mathbf{f}_K=P_C[\mathbf{x}_g-\alpha' D_v\mathcal{L}(\mathbf{o}^K,\mathbf{p}^K,\mathbf{q}^K)]$. The symbols and other equations remain the same as in Algorithm III.

APPENDIX D: DESCRIPTION OF A FGP-TYPE ALGORITHM TO SOLVE THE WAVELET-BASED WEIGHTED ℓ_1-PROXIMAL PROBLEM IN EQ. (33)

As mentioned in Sec. 3.B, the global soft-thresholding operator method cannot be applied to solve the wavelet-based weighted ℓ_1-proximal problem in Eq. (33). However, the FGP-type algorithm can be still adopted to solve this weighted proximal problem in an iterative way. The main idea behind the FGP-type algorithm to solve TV-proximal problem is to adopt a dual approach. The similar dual equivalence strategy can also be adopted to solve the weighted wavelet-based ℓ_1-proximal problem, which can be expressed as

$$\begin{aligned}&\text{prox}_{\gamma_v}^{D_v^{-1}}(g_{\ell_1,\Phi}/T\times 2)(\mathbf{e}_v^k)\\ &\quad=\arg\min_{\mathbf{u}}\left\{\|\mathbf{u}-\mathbf{e}_v^k\|^2_{D_v^{-1}}+\frac{4c_2\lambda_{\ell_1}\gamma_v}{T}\|\Phi\mathbf{u}\|_1\right\}\\ &\quad=\arg\min_{\mathbf{u}}\{\|\mathbf{u}-\mathbf{e}_v^k\|^2_{D_v^{-1}}+2\beta'\|\Phi\mathbf{u}\|_1\},\end{aligned} \tag{D1}$$

where $\beta'=2c_2\lambda_{\ell_1}\gamma_v/T$. To dualize the wavelet-based ℓ_1 norm, we can have

$$\|\Phi\mathbf{u}\|_{\ell_1}=\max_{w\in W}\langle\Phi\mathbf{u},w\rangle=\max_{w\in W}\langle w,\Phi^T\mathbf{u}\rangle, \tag{D2}$$

where $W=\{w:\|w\|_\infty\le 1\}$. To better understand the purpose of this dual approach, we recall that the TV penalty can be dualized as

$$\|\mathbf{u}\|_{\text{TV}}=\|\nabla\mathbf{u}\|_{\ell_1}=\max_{z\in Z}\langle\nabla\mathbf{u},\mathbf{z}\rangle=\max_{z\in Z}-\langle\mathbf{z},\nabla\cdot\mathbf{u}\rangle, \tag{D3}$$

where ∇ indicates the gradient operator, $\nabla\cdot$ represents the divergence operator, the matrix $\mathbf{z}$ has the same structure as $\nabla\mathbf{u}$, and the set Z is given by $Z=\{z:\|z\|_\infty\le 1\}$. We can connect the gradient operator ∇ and the divergence operator $\nabla\cdot$ to the operator $\mathcal{L}^T$ and $\mathcal{L}$ adopted in Eq. (A3) for the standard FGP algorithm. In fact, we have $\mathcal{L}^T=-\nabla$ and $\mathcal{L}=\nabla\cdot$ in the standard FGP algorithm that solves the TV-proximal problem in Appendix A. By employing the same strategy and comparing Eqs. (D2) and (D3), we can define two new operators $\mathcal{L}'=-\Phi^T$ and $\mathcal{L}'^T=-\Phi$ to adopt the FGP-type algorithm to solve the wavelet-based weighted ℓ_1 proximal problem. Therefore, we can modify Eq. (C2) by incorporating the two new operators $\mathcal{L}'$ and $\mathcal{L}'^T$ to solve the wavelet-based weighted ℓ_1-proximal problem and the modified formula is given by

$$\mathbf{w}_2^k=P_{\mathcal{P}}\left[\mathbf{w}_1^k+\frac{1}{\beta'\max(D_v)}\mathcal{L}'^T(P_C[\mathbf{x}_g-\beta' D_v\mathcal{L}'(\mathbf{w}_1^k)])\right], \tag{D4}$$

where the matrices $\mathbf{w}_1^k$ and $\mathbf{w}_2^k$ have the same structures as $\Phi\mathbf{u}$ and other symbols and operators have similar meanings as in Algorithm III. Finally, the solution after Kth iteration will be set to $\mathbf{f}_K=P_C[\mathbf{x}_g-\beta' D_v\mathcal{L}'(\mathbf{w}^K)]$.

APPENDIX E: HARDWARE ACCELERATION EMPLOYING SINGLE AND MULTIPLE GPUS

Single- and multi-GPU implementations of the OSSF-TV algorithm are described below. Although not presented, the implementations for the OSSF-TV-ℓ_1 algorithm are essential similar to those of the OSSF-TV algorithm.

1. Single GPU implementation of the OSSF-TV algorithm for CBCT

All implementation in this work were based on NVIDIA Tesla K40 GPUs, each of which has 2880 processing cores and 12GB of RAM. Figure 16 describes the basic structure of the single-GPU implementation of the OSSF-TV algorithm.

Specific details are as follows.

- **Projection data $\mathbf{b}^{\text{data}}$**: The projection data $\mathbf{b}^{\text{data}}$ are transferred into the GPU global memory from the host memory. If the GPU global memory allows, the projection data should be transferred into GPUs at one time instead of multiple transfers.
- **OSSF-TV**
  - — *Projection–correction step (one kernel function in GPU)*: For the vth subset, each thread distributed by the GPU is employed to compute one element of the corrective matrix $c_{i,v}$ according to the Eq. (24). In order to accelerate this step, the 3D discrete object matrix $\mathbf{f}$ should be stored in texture memory. The corresponding values of $h_{ij,v}$ can be calculated independently by use of a previous proposed method[50] in each thread. The calculated 2D corrective matrix $c_{i,v}$ for the vth subset is located in GPU global memory, which will be employed in the following backprojection-update step.
  - — *Backprojection-update step (one kernel function in GPU)*: For the vth projection data subset, each thread independently updates one specific voxel of the 3D volume $\mathbf{f}$ from the previous obtained 2D corrective matrix $c_{i,v}$ according to Eq. (25). To update each

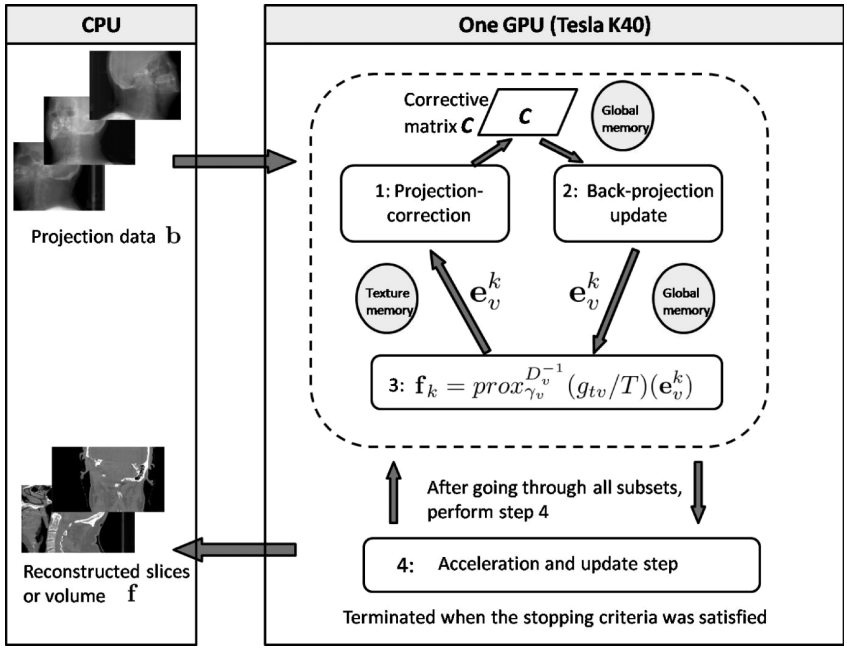


FIG. 16. A simple flowchart of the OSSF-TV algorithm with single GPU approach.

element f_j independently, a key step is to find the indices of the ray that intersects the jth voxel. This can be accomplished by projecting the eight vertices of the jth voxel onto the detector plane to determine the potential intersection range. For each subset, the object matrix $\mathbf{f}$ is updated once, which is employed in the following step of solving weighted TV-proximal problem.

— *Weighted TV-proximal problem (one kernel function in GPU):* As described previously, the four operators $\mathcal{L}$, P_C, $\mathcal{P}^P$, and $\mathcal{L}^T$ operate in a element-wise manner in Eq. (C2). This indicates that each element of ($\mathbf{r}$,$\mathbf{s}$,$\mathbf{t}$) and ($\mathbf{o}$,$\mathbf{p}$,$\mathbf{q}$) can be updated independently by use of a GPU thread. Accordingly, implementations of Eqs. (C2) can efficiently exploit GPU parallelism because only simple and independent arithmetic operations are required by each thread that generally cause few memory conflicts. Note that only three auxiliary vectors $\mathbf{r}^k$, $\mathbf{s}^k$, $\mathbf{t}^k$ need to be stored in the GPU global memory and each vector approximately has the same size as the 3D matrix $\mathbf{f}$.

2. Multi-GPU implementation of the OSSF-TV algorithm for CBCT

To further reduce the computation time, a multi-GPU scheme is proposed in this section. We assume four NVIDIA K40s are employed and demonstrate some basic rules and possible arrangements. A flowchart is shown in Fig. 17. The original 3D vector $\mathbf{f}$ is divided into four equal subvolumes that are distributed among four GPUs ($id = 0,1,2,3$), respectively. Details regarding the multi-GPU implementation are as follows.

- **Projection data $\mathbf{b}^{\text{data}}$:** The projection data $\mathbf{b}^{\text{data}}$ are transferred to the global memories of the four GPUs.
- **OSSF-TV:** The projection–correction step is divided into two individual substeps as follows:
  — *Projection step (one kernel function in GPUs):* Each thread launched by the idth GPU ($id = 0,1,2,3$) simultaneously computes one ray integral through the idth subvolume as

$$b_{i,v}^{id} = \sum_{j=1}^{N/4} h_{ij,v}^{id} f_{j,v-1}^{id}, \quad i = 1,2,\ldots,M/T;\ id = 0,1,2,3,$$

$$rl_{i,v}^{id} = \sum_{j=1}^{N/4} h_{ij,v}^{id}, \qquad i = 1,2,\ldots,M/T;\ id = 0,1,2,3,$$

  where the superscript id indicates the subvolume, $h_{ij,v}^{id}$ represents the contribution from the jth voxel in the idth subvolume to the ith ray, and $b_{i,v}^{id}$ represents the ith ray integral through the idth subvolume. The notation $f_j^{id,v-1}$ specifies the jth voxel value in the idth subvolume at the $(v-1)$th update in one full OS-SART iteration, and $rl_{i,v}^{id}$ represents the length of the ith ray that intersected the idth subvolume.
  — *Correction step (one kernel in GPUs):* In order to calculate the 2D corrective matrix $c_{i,v}$, the ray integrals and ray lengths calculated from all subvolumes are summed to obtain the values for the full volume as follows:

$$b_{i,v} = b_{i,v}^0 + b_{i,v}^1 + b_{i,v}^2 + b_{i,v}^3,$$
$$rl_{i,v} = rl_{i,v}^0 + rl_{i,v}^1 + rl_{i,v}^2 + rl_{i,v}^3, \quad i = 1,2,\ldots,M/T.$$

  Therefore, the 2D corrective matrix $c_{i,v}$ is computed as

$$c_{i,v} = \frac{b_{i,v}^{\text{data}} - b_{i,v}}{rl_{i,v}}, \qquad i = 1,2,\ldots,M/T.$$

  Next, the $c_{i,v}$ are copied to the global memory of all GPUs to prepare the last backprojection-update step.

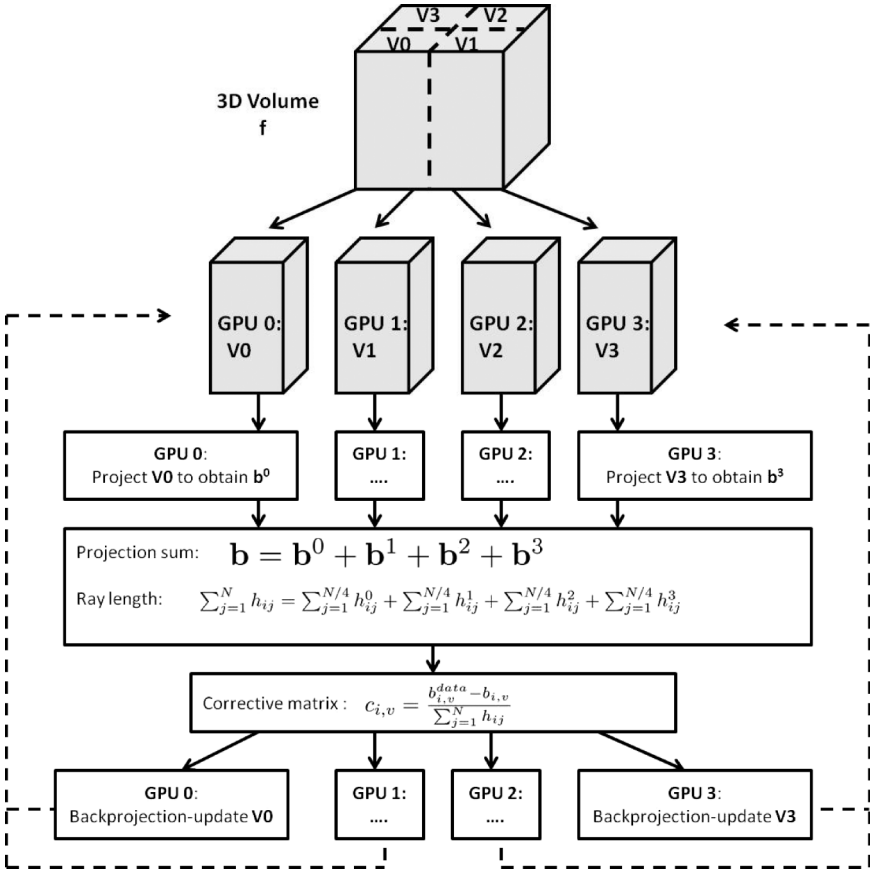


FIG. 17. One simple scheme of four GPUs implementation for OS-SART part in proposed OSSF-TV and OSSF-TV-ℓ_1 algorithms.

— *Backprojection-update step (one kernel in GPUs):* Similar rules and strategies described in the single GPU implementation are also applicable here. Each thread launched by the idth GPU is employed to independently update one voxel in the idth subvolume as

$$f_{j,v}^{id,k} = f_{j,v-1}^{id,k} + \gamma \frac{\sum_{i\in\phi^{id}} (c_{i,v-1}) h_{ij,v}^{id}}{\sum_{i\in\phi^{id}} h_{ij,v}^{id}},$$

$j = 1,2,\ldots,N/4.$

The above equation is essential the same as Eq. (25), but it is performed in each subvolume by each corresponding GPU. During each view-update, four GPUs are synchronized for the projection and backprojection-update step since all four ray integrals $b_{i,v}^{id}$ are required in the correction step. Four subvolumes are updated once in their respective GPUs for each subset before being employed into the next step of solving the weighted TV-proximal problem.

— *Weighted TV-proximal problem:* In the OS-SART computation, an intermediate solution $\mathbf{x_g}$ to minimization problem $d(\mathbf{f})$ is obtained and stored as four subvolumes $\mathbf{x}_g^0$, $\mathbf{x}_g^1$, $\mathbf{x}_g^2$, $\mathbf{x}_g^3$ in four GPUs. Because of the element-wise property of four operators in Eq. (C2), it can be independently executed for each subvolume in each GPU concurrently. Except when the operators $\mathcal{L}$ and $\mathcal{L}^T$ act on elements of ($\mathbf{r}$,$\mathbf{s}$,$\mathbf{t}$) and $\mathbf{x}_g$ located at the boundaries between two subvolumes, each GPU only needs to access its own memory. However, the number of such elements is small when compared to the number elements inside each subvolume.

The computation time for the projection operation when four GPUs is employed is approximately one quarter of the time required by single GPU implementation for the same sized reconstruction volume. This is because both the computing time for calculating ray integrals through one subvolume in each thread and the number of nonzero ray integral threads in each GPU would be approximately one half of those in single GPU case. The computation time for the backprojection-update step with four GPUs will also be approximately one quarter of the time required by one single GPU, since each GPU only updated one subvolume data, whose size was only one quarter of the original volume. Moreover, the computation time for the correction step is negligible, since only simple arithmetic operations are involved for small 2D matrices.

In addition, during one full OSSF-TV iteration, the idth subvolume always remains in the memory of the idth GPU. There is no need to frequently transfer large vectors between GPUs, which minimizes communication times. Moreover, when solving the TV-proximal problem, since each GPU

executed for one quarter of the data, the computation time required by four GPUs is approximately one quarter of the time required by a single GPU. The overhead and communication time between GPUs in this operation is minimal. Accordingly, the time reduction factor by adopting a multi-GPU scheme to solve the proposed OSSF-TV algorithm (and OSSF-TV-$\ell 1$ algorithm) is approximately equal to the number of GPUs employed. The above observations and conclusions generalize to the case where more than four GPUs are employed. This feature is highly attractive and suggests that reconstruction times can be readily reduced by using additional GPUs.

[a]Electronic mail: anastasio@wustl.edu

[1]J. Bian, J. H. Siewerdsen, X. Han, E. Y. Sidky, J. L. Prince, C. A. Pelizzari, and X. Pan, "Evaluation of sparse-view reconstruction from flat-panel-detector cone-beam CT," Phys. Med. Biol. **55**, 6575–6599 (2010).
[2]X. Han, J. Bian, E. L. Ritman, E. Y. Sidky, and X. Pan, "Optimization-based reconstruction of sparse images from few-view projections," Phys. Med. Biol. **57**, 5245–5273 (2012).
[3]K. Choi, J. Wang, L. Zhu, T.-S. Suh, S. Boyd, and L. Xing, "Compressed sensing based cone-beam computed tomography reconstruction with a first-order method," Med. Phys. **37**, 5113–5125 (2010).
[4]X. Jia, Y. Lou, J. Lewis, R. Li, X. Gu, C. Men, W. Y. Song, and S. B. Jiang, "GPU-based fast low-dose cone beam CT reconstruction via total variation," J. X-Ray Sci. Technol. **19**, 139–154 (2011).
[5]J. H. Jørgensen, T. L. Jensen, P. C. Hansen, S. H. Jensen, E. Y. Sidky, and X. Pan, "Accelerated gradient methods for total-variation-based CT image reconstruction," preprint arXiv:1105.4002.
[6]L. Ritschl, F. Bergner, C. Fleischmann, and M. Kachelrieß, "Improved total variation-based CT image reconstruction applied to clinical data," Phys. Med. Biol. **56**, 1545–1561 (2011).
[7]T. Niu and L. Zhu, "Accelerated barrier optimization compressed sensing (ABOCS) reconstruction for cone-beam CT: Phantom studies," Med. Phys. **39**, 4588–4598 (2012).
[8]J. C. Park, B. Song, J. S. Kim, S. H. Park, H. K. Kim, Z. Liu, T. S. Suh, and W. Y. Song, "Fast compressed sensing-based CBCT reconstruction using Barzilai-Borwein formulation for application to on-line IGRT," Med. Phys. **39**, 1207–1217 (2012).
[9]E. Y. Sidky, C.-M. Kao, and X. Pan, "Accurate image reconstruction from few-views and limited-angle data in divergent-beam CT," J. X-Ray Sci. Technol. **14**, 119–139 (2006).
[10]E. Y. Sidky and X. Pan, "Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization," Phys. Med. Biol. **53**, 4777–4807 (2008).
[11]A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," IEEE Trans. Image Process. **18**, 2419–2434 (2009).
[12]A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," SIAM J. Imaging Sci. **2**, 183–202 (2009).
[13]J. Huang, S. Zhang, and D. Metaxas, "Efficient MR image reconstruction for compressed MR imaging," Med. Image Anal. **15**, 670–679 (2011).
[14]J. Dutta, S. Ahn, C. Li, S. R. Cherry, and R. M. Leahy, "Joint l1 and total variation regularization for fluorescence molecular tomography," Phys. Med. Biol. **57**, 1459–1476 (2012).
[15]Q. Xu, A. Sawatzky, M. Anastasio, and O. C. Schirra, "Sparsity-regularized image reconstruction of decomposed K-edge data in spectral CT," Phys. Med. Biol. **59**, N65–N79 (2014).
[16]Z. Zhu, K. Wahid, P. Babyn, D. Cooper, I. Pratt, and Y. Carter, "Improved compressed sensing-based algorithm for sparse-view CT image reconstruction," Comput. Math. Methods Med. **2013**, 1–15.
[17]G.-H. Chen, J. Tang, and S. Leng, "Prior image constrained compressed sensing (PICCS): A method to accurately reconstruct dynamic CT images from highly undersampled projection data sets," Med. Phys. **35**, 660–663 (2008).
[18]J. Tang, B. E. Nett, and G.-H. Chen, "Performance comparison between total variation (TV)-based compressed sensing and statistical iterative reconstruction algorithms," Phys. Med. Biol. **54**, 5781–5804 (2009).
[19]M. Defrise, C. Vanhove, and X. Liu, "An algorithm for total variation regularization in high-dimensional linear problems," Inverse Probl. **27**, 065002 (2011).
[20]M. Persson, D. Bone, and H. Elmqvist, "Total variation norm for three-dimensional iterative reconstruction in limited view angle tomography," Phys. Med. Biol. **46**, 853–866 (2001).
[21]M. Li, H. Yang, and H. Kudo, "An accurate iterative reconstruction algorithm for sparse objects: Application to 3D blood vessel reconstruction from a limited number of projections," Phys. Med. Biol. **47**, 2599–2609 (2002).
[22]Q. Xu, D. Yang, J. Tan, and M. Anastasio, "SU-F-BRCD-09: Total variation (TV) based fast convergent iterative CBCT reconstruction with GPU acceleration," Med. Phys. **39**, 3857 (2012).
[23]Q. Xu, E. Y. Sidky, X. Pan, M. Stampanoni, P. Modregger, and M. A. Anastasio, "Investigation of discrete imaging models and iterative image reconstruction in differential x-ray phase-contrast tomography," Opt. Express **20**, 10724–10749 (2012).
[24]A. M. Zysk, A. B. Garson, Q. Xu, E. M. Brey, W. Zhou, J. G. Brankov, M. N. Wernick, J. R. Kuszak, and M. A. Anastasio, "Nondestructive volumetric imaging of tissue microstructure with benchtop x-ray phase-contrast tomography and critical point drying," Biomed. Opt. Express **3**, 1924–1932 (2012).
[25]T. Niu, X. Ye, Q. Fruhauf, M. Petrongolo, and L. Zhu, "Accelerated barrier optimization compressed sensing (ABOCS) for CT reconstruction with improved convergence," Phys. Med. Biol. **59**, 1801–1814 (2014).
[26]P. L. Combettes and J. C. Pesquet, "A proximal decomposition method for solving convex variational inverse problems," Inverse Probl. **24**, 065014 (2008).
[27]Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Advances in Neural Information Processing systems 24* (Curran Associates, Inc., 2011), pp. 612–620.
[28]X. Wang and X. Yuan, "The linearized alternating direction method of multipliers for dantzig selector," SIAM J. Sci. Comput. **34**, A2792–A2811 (2012).
[29]X. Ren and Z. Lin, "Linearized alternating direction method with adaptive penalty and warm starts for fast solving transform invariant low-rank textures," Int. J. Comput. Vision **104**, 1–14 (2013).
[30]Y.-L. Yu, "Better approximation and faster algorithm using the proximal average," in *Advances in Neural Information Processing systems 26* (Curran Associates, Inc., 2013), pp. 458–466.
[31]D. Kim, D. Pal, J. B. Thibault, and J. A. Fessler, "Accelerating ordered subsets image reconstruction for x-ray CT using spatially nonuniform optimization transfer," IEEE Trans. Med. Imaging **32**, 1965–1978 (2013).
[32]A. Sawatzky, "Performance of first-order algorithms for TV penalized weighted least-squares denoising problem," in *Image and Signal Processing*, Lecture Notes in Computer Science (Springer, Berlin, Germany, 2014), Vol. 8509, pp. 340–349.
[33]A. Sawatzky, Q. Xu, C. Schirra, and M. Anastasio, "Proximal ADMM for multi-channel image reconstruction in spectral x-ray CT," IEEE Trans. Med. Imaging **33**, 1657–1668 (2014).
[34]I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," Commun. Pure Appl. Math. **57**, 1413–1457 (2004).
[35]H. Malcolm Hudson and R. S. Larkin, "Accelerated image reconstruction using ordered subsets of projection data," IEEE Trans. Med. Imaging **13**, 601–609 (1994).
[36]S. H. Manglos, G. M. Gagne, A. Krol, F. D. Thomas, and R. Narayanaswamy, "Transmission maximum-likelihood reconstruction with ordered subsets for cone beam CT," Phys. Med. Biol. **40**, 1225–1241 (1995).
[37]J. Nuyts, B. De Man, P. Dupont, M. Defrise, P. Suetens, and L. Mortelmans, "Iterative reconstruction for helical CT: A simulation study," Phys. Med. Biol. **43**, 729–737 (1998).
[38]C. Kamphuis and F. J. Beekman, "Accelerated iterative transmission CT reconstruction using an ordered subsets convex algorithm," IEEE Trans. Med. Imaging **17**, 1101–1105 (1998).
[39]H. Erdogan and J. A. Fessler, "Ordered subsets algorithms for transmission tomography," Phys. Med. Biol. **44**, 2835–2851 (1999).
[40]A. H. Andersen and A. C. Kak, "Simultaneous algebraic reconstruction technique (SART): A superior implementation of the art algorithm," Ultrason. Imaging **6**, 81–94 (1984).

[41]Y. Censor and T. Elfving, "Block-iterative algorithms with diagonally scaled oblique projections for the linear feasibility problem," SIAM J. Matrix Anal. Appl. **24**, 40–58 (2002).
[42]M. Jiang and G. Wang, "Convergence studies on iterative algorithms for image reconstruction," IEEE Trans. Med. Imaging **22**, 569–579 (2003).
[43]G. Wang and M. Jiang, "Ordered-subset simultaneous algebraic reconstruction techniques (OS-SART)," J. X-Ray Sci. Technol. **12**, 169–178 (2004).
[44]E. Chouzenoux, J.-C. Pesquet, and A. Repetti, "Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function," J. Optim. Theory Appl. **162**(1), 107–132 (2014).
[45]F. J. Beekman and C. Kamphuis, "Ordered subset reconstruction for x-ray CT," Phys. Med. Biol. **46**, 1835–1844 (2001).
[46]M. C. van Dijke, "Iterative methods in image reconstruction," Ph.D. dissertation, Utrecht University, The Netherlands, 1992.
[47]G. T. Herman and L. B. Meyer, "Algebraic reconstruction techniques can be made computationally efficient [positron emission tomography application]," IEEE Trans. Med. Imaging **12**, 600–609 (1993).
[48]H. Guan and R. Gordon, "A projection access order for speedy convergence of ART (algebraic reconstruction technique): A multilevel scheme for computed tomography," Phys. Med. Biol. **39**, 2005–2022 (1999).
[49]K. Mueller, R. Yagel, and J. Fredrick Cornhill, "The weighted-distance scheme: A globally optimizing projection ordering method for ART," IEEE Trans. Med. Imaging **16**, 223–230 (1997).
[50]F. Jacobs, E. Sundermann, B. De Sutter, M. Christiaens, and I. Lemahieu, "A fast algorithm to calculate the exact radiological path through a pixel or voxel space," J. Comput. Inf. Technol. **6**, 89–94 (1998).
[51]H. Nien and J. A Fessler, "Fast x-ray CT image reconstruction using a linearized augmented Lagrangian method with ordered subsets," IEEE Trans. Med. Imaging **34**, 388–399 (2015).
[52]J. A. Fessler and S. D. Booth, "Conjugate-gradient preconditioning methods for shift-variant PET image reconstruction," IEEE Trans. Image Process. **8**, 688–699 (1999).
[53]A. I. Veress, W. P. Segars, J. A. Weiss, B. M. W. Tsui, and G. T. Gullberg, "Normal and pathological NCAT image and phantom data based on physiologically realistic left ventricle finite-element models," IEEE Trans. Med. Imaging **25**, 1604–1616 (2006).
[54]J. Tan, H. H. Li, E. Klein, H. Li, P. Parikh, and D. Yang, "Physical phantom studies of helical cone-beam CT with exact reconstruction," Med. Phys. **39**, 4695–4704 (2012).
[55]E. Candes and L. Demanet, "Curvelets and Fourier integral operators," C. R. Math. **336**, 395–398 (2003).
[56]J.-L. Starck, F. Murtagh, and J. M. Fadili, *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity* (Cambridge University Press, Cambridge, United Kingdom, 2010).
[57]E. J. Candes, "Ridgelets: Theory and applications," Ph.D. thesis, Stanford University, 1998.
[58]E. J. Candès and D. L. Donoho, "Ridgelets: A key to higher-dimensional intermittency?," Philos. Trans. R. Soc., A **357**, 2495–2509 (1999).
[59]H. Nien and J. A. Fessler, "Combining augmented Lagrangian method with ordered subsets for x-ray CT image reconstruction," in *Proceedings of International Meeting on Fully 3D Image Reconstruction in Radiology and Nuclear Medicine, Lake Tahoe, CA* (2013), pp. 280–283.
[60]Y. Zou, D. Shi, and A. Zamyatin, "Weighted simultaneous algebraic reconstruction techniques," in *11th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine, Potsdam, Germany* (2011), p. 157.
[61]J. Gregor and J. Fessler, "Comparison of SIRT and SQS for regularized weighted least squares image reconstruction," IEEE Trans. Comput. Imaging **1**, 44–55 (2015).