Time of flight PET reconstruction using non-uniform update for regional recovery uniformity

Kyungsang Kim
Gordon Center for Medical Imaging, Massachusetts General Hospital and Harvard Medical School, 125 Nashua Street 6th floor, Suite 660, Boston, MA 02114, United States

Donghwan Kim
Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48105, United States

Jaewon Yang
Physics Research Laboratory, Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA 94143 USA

Georges El Fakhri
Gordon Center for Medical Imaging, Massachusetts General Hospital and Harvard Medical School, 125 Nashua Street 6th floor, Suite 660, Boston, MA 02114, United States

Youngho Seo
Physics Research Laboratory, Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA 94143 USA

Jeffrey A. Fessler
Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48105, United States

Quanzheng Li*[a]
Gordon Center for Medical Imaging, Massachusetts General Hospital and Harvard Medical School, 125 Nashua Street 6th floor, Suite 660, Boston, MA 02114, United States

Purpose: Time of flight (TOF) PET reconstruction is well known to statistically improve the image quality compared to non-TOF PET. Although TOF PET can improve the overall signal to noise ratio (SNR) of the image compared to non-TOF PET, the SNR disparity between separate regions in the reconstructed image using TOF data becomes higher than that using non-TOF data. Using the conventional ordered subset expectation maximization (OS-EM) method, the SNR in the low activity regions becomes significantly lower than in the high activity regions due to the different photon statistics of TOF bins. A uniform recovery across different SNR regions is preferred if it can yield an overall good image quality within small number of iterations in practice. To allow more uniform recovery of regions, a spatially variant update is necessary for different SNR regions.

Methods: This paper focuses on designing a spatially variant step size and proposes a TOF-PET reconstruction method that uses an non-uniform separable quadratic surrogates (NUSQS) algorithm, providing a straightforward control of spatially variant step size. To control the noise, a spatially invariant quadratic regularization is incorporated, which by itself does not theoretically affect the recovery uniformity. Nesterov's momentum method with ordered subsets (OS) is also used to accelerate the reconstruction speed. To evaluate the proposed method, an XCAT simulation phantom and clinical data from a pancreas cancer patient with full (ground truth) and 6× downsampled counts were used, where a Poisson thinning process was employed for downsampling. We selected tumor and cold regions of interest (ROIs) and compared the proposed method with the TOF-based conventional OS-EM and OS-SQS algorithms with an early stopping criterion.

Results: In computer simulation, without regularization, hot regions of OS-EM and OS-NUSQS converged similarly, but cold region of OS-EM was noisier than OS-NUSQS after 24 iterations. With regularization, although the overall speeds of OS-EM and OS-NUSQS were similar, recovery ratios of hot and cold regions reconstructed by OS-NUSQS were more uniform compared to those of the conventional OS-SQS and OS-EM. The OS-NUSQS with Nesterov's momentum converged faster than others while preserving the uniform recovery. In the clinical example, we demonstrated that the OS-NUSQS with Nesterov's momentum provides more uniform recovery ratios of hot and cold ROIs compared to OS-SQS and OS-EM. Although the cost function of all methods are equivalent, the proposed method has higher structural similarity (SSIM) values of hot and cold regions compared to other methods after 24 iterations. Furthermore, our computing time using graphics processing unit was 80× shorter than the time using quad-core CPUs.

Conclusion: This paper proposes a TOF PET reconstruction method using OS-NUSQS with Nesterov's momentum for uniform recovery of different SNR regions. In particular, the spatially non-uniform step size in the proposed method provides uniform recovery ratios of different SNR regions, and Nesterov's momentum further accelerates overall convergence while preserving uniform recovery. The computer simulation and clinical example demonstrate that the proposed method converges uniformly across ROIs. In addition, tumor contrast and SSIM of the proposed method were higher than those of conventional OS-EM and OS-SQS in early iterations.

I. INTRODUCTION

Time of flight (TOF) positron emission tomography (PET) has been developed to improve image quality and reduce image acquisition time[1,2]. The noise reduction has been studied using TOF and non-TOF data in many papers[3,4,5,6]. One of the main advantages using TOF information is to improve the signal-to-noise ratio (SNR) of the reconstructed image. The SNR of an image voxel reconstructed by filtered back-projection (FBP) is approximately[7]:

$$\mathrm{SNR} = k \cdot N^{-\frac{1}{2}} \left(\frac{T^2}{T+S+R} \right)^{\frac{1}{2}}, \tag{1}$$

where k is a physically defined constant and N is the number of voxels in an image. T is the number of true counts in a back-projected image, S and R are the numbers of scatter and random counts, respectively, in a back-projected image. The noise equivalent count rate (NEC), defined by $T^2/(T+S+R)$ with measurements, is also used as the effective sensitivity considering noise contributions, such as scatter and random events[8]. For example, in a cylinder of diameter D with uniform activity, the improvements of SNR and NEC gains of TOF PET are estimated as follows[8]:

$$\mathrm{SNR_T} = \sqrt{\frac{D}{\Delta x}} \mathrm{SNR_{n\text{-}T}}, \quad \mathrm{NEC_T} = \frac{D}{\Delta x} \mathrm{NEC_{n\text{-}T}}, \tag{2}$$

where T and n-T denote TOF and non-TOF, respectively, Δx is the full-width-half-maximum (FWHM) calculated by an intrinsic time resolution (Δt) as $\Delta x = c\Delta t/2$. Here, c is the speed of light. Therefore, TOF information improves the image quality with higher SNR and NEC[8]. In addition, Karp *et al.*[9] demonstrated that the contrast recovery coefficient (CRC) using TOF data converges faster than the CRC using non-TOF data in the maximum likelihood expectation maximization (MLEM) algorithm, and observed that CRCs using TOF data are larger than those using non-TOF data. Here, the CRC is defined as $(H-B)/B$, where H is a hot region and B is a background.

In general, due to non-uniform activity in an image, SNRs differ between regions. The standard iterative algorithms such as OS-EM[10] and OS-SQS[11] converge slowly for the low-SNR region, and regions with different SNRs converge to different recovery ratios. Although TOF PET can improve image contrast in high SNR regions, the disparity of SNRs between elements in an image becomes considerably wider by at least $\sqrt{\frac{D}{\Delta x}}$ times compared to that in non-TOF PET. For example, specific regions, such as low activity regions, can be very noisy after a few iterations, potentially degrading the overall noise distribution within an image and making it difficult to terminate an iterative method early. Therefore, development of an algorithm that provides fast and uniform recovery of ROIs without sacrificing the overall convergence speed is necessary, which can help minimize the total number of iterations needed in practice.

In iterative image reconstruction, there are two possible ways to encourage the uniform recovery for different SNR regions: 1) spatially variant regularization based on the noise statistics of data and 2) spatially non-uniform step size method for the gradient-based update. The spatially variant regularization has been used to encourage uniform image resolution[12,13,14]. Based on our knowledge, there is no spatially variant regularization for achieving uniform recovery. Due to the complexity in developing a spatially variant regularization satisfying the requirement of uniform recovery, a spatially non-uniform step size method is more straightforward to control the uniform recovery.

Algorithms using spatially variant step size have been developed to promote faster covergence by non-uniformly updating the image in iterations[15,16]. Non-homogeneous iterative coordinate descent (NH-ICD) was proposed to accelerate the conventional ICD algorithm[15]. However, NH-ICD is difficult to parallelize because each voxel is updated sequentially. Van Slambrouck and Nuyts proposed a group-wise non-uniform coordinate descent update for faster convergence speed[16], which is more amenable to parallelization than ICD-type methods, but is still limited by its group size. On the other hand, Kim *et al.* proposed a non-uniform separable quadratic surrogates (NUSQS) algorithm[17], in which the step size is spatially non-uniform and the update is parallelizable, accelerating the convergence speed. The non-uniform update in that work was used to accelerate CT reconstruction, whereas here we use a similar idea to encourage spatially uniform recovery ratio, and extend the NUSQS method for TOF PET reconstruction.

For additional acceleration, it is useful to combine the ordered subsets (OS)[10] and Nesterov's momentum method[18] that uses previous descent updates to provide momentum. Kim *et al.* exploited both OS and Nesterov's momentum for CT reconstruction[19], where the convergence speed using both OS and momentum was significantly higher than speeds of conventional methods in early iterations. Similarly, we propose a TOF-PET reconstruction exploiting the ordered subsets non-uniform separable quadratic surrogates (OS-NUSQS) algorithm with Nesterov's momentum method. To further reduce the computation time, we implemented the proposed method using a graphics processing unit (GPU), particularly forward and backward projectors for TOF reconstruction and the quadratic regularization.

To validate the proposed algorithm, we performed computer simulations using an XCAT phantom[20] under various conditions, and compared the recovery ratios of ROIs and the reconstructed image qualitatively and quantitatively. We also performed an experiment with a patient having pancreatic and other metastatic tumors. The image reconstructed after 300 iterations of one-subset version of EM using full data is used as the ground truth, and reconstructed images using data with 6× downsampled counts are used for validating the performance. Here, a Poisson thinning process is used for downsampling the counts of prompt raw data[21]. The proposed method is compared with the conventional OS-SQS and OS-EM methods after certain number of iterations. More specifically, reconstructed images are compared after 24 iterations. Note that the OS-SQS has a relatively uniform step size compared to OS-EM and OS-NUSQS, and the OS-EM is the standard method in iterative PET reconstruction, having a spatially variant update in iteration. We select both tumor and cold ROIs, and compared recovery ratios of ROIs and structural similarity (SSIM) values. Our results demonstrate that the proposed method can achieve uniform recovery ratios of ROIs, and provides a good image quality after a finite number of iterations.

This paper is organized as follows. Section II gives the problem formulation and the proposed method: OS-NUSQS and Nesterov's momentum for TOF PET reconstruction. Section III provides experimental setup details of computer simulation, clinical example, evaluation and GPU implementation. Section IV demonstrates simulation and clinical results for various aspects. Section V discusses several technical issues and Section VI concludes.

II. THEORY

II.A. Problems

We reconstruct a non-negative image $x = (x_1, ..., x_{N_v}) \in \mathbb{R}_+^{N_v}$ from a time of flight (TOF) measurement $Y = [y^1, y^2, ..., y^t, ..., y^{N_T}] \in \mathbb{Z}_+^{N_m \times N_T}$; $y^t = (y_1^t, ..., y_{N_m}^t) \in \mathbb{Z}_+^{N_m}$. N_v, N_m and N_T denote the numbers of voxels, sinogram bins and TOF time bins, respectively. Y is the number of photon counts and contains true, scatter and random coincidence events in which we assume a Poisson statistical model:

$$y_i^t \sim \text{Poisson}\{[A_t x]_i + r_i^t\}, \tag{3}$$

where y_i^t is the number of counts with the ith sinogram at the tth time bin. r_i^t is the mean value of scatter and random events[22], with the ith sinogram at the tth time bin. $A_t = (A \circ W_t)$ is the TOF system matrix at the tth time bin; $A \in \mathbb{R}_+^{N_m \times N_v}$ is the conventional system matrix that computes the line integral of a line of response (LOR), $W_t \in \mathbb{R}_+^{N_m \times N_v}$ is the Gaussian kernel along all LORs calculated by the TOF time response function, and $\circ$ is the Hadamard product. The TOF time response function is a one-dimensional Gaussian function centered at the emission position[23]. Thus, $[A_t x]_i = \sum_{j=1}^{N_v} a_{ij} w_{ij}^t x_j$ represents the Gaussian weighted line integral of a LOR for the tth time bin and w is the Gaussian weight. Here, a_{ij} denotes the probability that a pair of annihilation photons emitted from the jth voxel is detected at the ith sinogram bin, and w_{ij}^t is the Gaussian weight along the line of flight at tth time bin. Throughout the paper, we use $a_{ij}^t = a_{ij} w_{ij}^t$.

For regularized TOF PET image reconstruction, we minimize the following cost function $\Psi_o(x)$:

$$\Psi_o(x) = L(x) + R(x) \tag{4}$$

$$= \sum_{t=1}^{N_T} \sum_{i=1}^{N_m} h_i^t([A_t x]_i) + \sum_{j=1}^{N_v} \psi_j(x), \tag{5}$$

where $L(x)$ denotes the negative log-likelihood function from the Poisson statistics and $R(x)$ is a quadratic roughness regularization[24]. $h_i^t(k) = k + r_i^t - y_i^t \log(k + r_i^t)$, $\psi_j(x) = \frac{\beta}{2} \sum_{j' \in \Omega_j} \rho_{jj'} (x_j - x_{j'})^2$ and $\beta > 0$ is a regularization parameter that controls the noise variance of the reconstructed image. $\rho_{jj'}$ is the reciprocal of Euclidean distance between voxels j and j', and Ω_j is the neighbor of center voxel j. For Ω_j, we use 26 neighbor voxels in a 3-D space.

II.B. Non-uniform Separable Quadratic Surrogates for TOF PET

In TOF-PET reconstruction, because the negative log-likelihood function $L(x)$ is difficult to minimize directly, a separable quadratic surrogates (SQS) algorithm for solving $L(x)$ is widely used with a regularization for reducing noise[11,25].

First, the quadratic surrogate function of $L(x)$ is as follows:

$$L(x) \leq Q_L^{(n)}(x) \triangleq \sum_{t=1}^{N_T} \sum_{i=1}^{N_m} p_i^{t,(n)}([A_t x]_i), \tag{6}$$

where

$$p_i^{t,(n)}(k) \triangleq h_i^t(k_i^{t,(n)}) + \dot{h}_i^t(k_i^{t,(n)})(k - k_i^{t,(n)}) + \frac{v_i^{t,(n)}}{2}(k - k_i^{t,(n)})^2. \tag{7}$$

$k_i^{t,(n)} = [A_t x^{(n)}]_i$ at n-th iteration, and $v_i^{t,(n)}$ is the curvature of $p_i^{t,(n)}(k)$. The first and second derivatives of h_i^t are as follows[11]:

$$\dot{h}_i^t(k_i^{t,(n)}) = 1 - \frac{y_i^t}{k_i^{t,(n)} + r_i^t}, \tag{8}$$

$$\ddot{h}_i^t(k_i^{t,(n)}) = \frac{y_i^t}{(k_i^{t,(n)} + r_i^t)^2}. \tag{9}$$

For $v_i^{t,(n)}$, the optimal curvature[25] is

$$v_i^{t,(n)}(k^{t,(n)}) = \begin{cases} \left[2\frac{h_i^t(0) - h_i^t(k^{t,(n)}) + \dot{h}_i^t(k^{t,(n)})(k^{t,(n)})}{(k^{t,(n)})^2}\right]_+, & k^{t,(n)} > 0 \\ \left[\ddot{h}_i^t(0)\right]_+, & k^{t,(n)} = 0. \end{cases} \tag{10}$$

$[\ \ ]_+$ denotes an operator that sets negative values to zero. To reduce the computing cost, we can use the Newton curvature $\ddot{h}_i^t([A_t x^{(n)}]_i)$, or precompute as $1/\max(y_i^t, \epsilon)$, or approximately compute as $1/\max([A_t x^{(n)}]_i, \epsilon)$ (see Discussion). Here, many of TOF measurements are zero, thus we set ϵ as a small positive value.

We next review a separable surrogate of the quadratic surrogate function[11], which uses the following trick:

$$[A_t x]_i = \sum_{j=1}^{N_v} a_{ij}^t x_j = \sum_{j=1}^{N_v} g_{ij}^{t,(n)} \left(\frac{a_{ij}^t}{g_{ij}^{t,(n)}}(x_j - x_j^{(n)}) + [A_t x^{(n)}]_i \right), \tag{11}$$

where $g_{ij}^t = a_{ij}^t / \left(\sum_{j'=1}^{N_v} a_{ij'}^t \right)$ is a non-negative real value ($g_{ij}^t = 0$ only if $a_{ij}^t = 0$ for all i, j), and $\sum_{j=1}^{N_v} g_{ij}^{t,(n)} = 1$. By using the convexity of $p_i^{t,(n)}$, the convexity inequality can be expressed as:

$$p_i^{t,(n)}([A_t x]_i) \le \sum_{j=1}^{N_v} g_{ij}^{t,(n)} p_i^{t,(n)} \left(\frac{a_{ij}^t}{g_{ij}^{t,(n)}}(x_j - x_j^{(n)}) + [A_t x^{(n)}]_i \right). \tag{12}$$

Combined with equation (5), we have the following majorizer:

$$L(x) \le Q_L^{(n)}(x) \le \phi_L^{(n)}(x) \triangleq \sum_{t=1}^{N_T} \sum_{j=1}^{N_v} \phi_{L,j}^{t,(n)}(x_j), \tag{13}$$

where

$$\phi_{L,j}^{t,(n)}(x_j) \triangleq \sum_{i=1}^{N_m} g_{ij}^{t,(n)} p_i^{t,(n)} \left(\frac{a_{ij}^t}{g_{ij}^{t,(n)}}(x_j - x_j^{(n)}) + [A_t x^{(n)}]_i \right). \tag{14}$$

The second derivative of the surrogate function $\phi_{L,j}^{(n)}(x_j)$ is

$$d_{L,j}^{(n)} \triangleq \frac{\partial^2}{\partial x_j^2} \phi_{L,j}^{(n)}(x_j) = \sum_{t=1}^{N_T} \sum_{i=1}^{N_m} v_i^{t,(n)} (a_{ij}^t)^2 / g_{ij}^t. \tag{15}$$

We next derive a separable surrogate of the quadratic roughness regularization $R(x)$[25]:

$$\begin{aligned} R(x) &= \sum_{j=1}^{N_v} \psi_j(x) = \sum_{j=1}^{N_v} \frac{\beta}{2} \sum_{j' \in \Omega_j} \rho_{jj'} (x_j - x_{j'})^2, &(16)\\ &= \sum_{j=1}^{N_v} \sum_{j' \in \Omega_j} \frac{\beta \rho_{jj'}}{2} \left(\frac{(2x_j - x_j^{(n)} - x_{j'}^{(n)})}{2} + \frac{(-2x_{j'} + x_j^{(n)} + x_{j'}^{(n)})}{2} \right)^2, &(17)\\ &\le \sum_{j=1}^{N_v} \sum_{j' \in \Omega_j} \frac{\beta \rho_{jj'}}{4} \left((2x_j - x_j^{(n)} - x_{j'}^{(n)})^2 + (2x_{j'} - x_j^{(n)} - x_{j'}^{(n)})^2 \right), &(18)\\ &= \sum_{j=1}^{N_v} \sum_{j' \in \Omega_j} \frac{\beta \rho_{jj'}}{2} (2x_j - x_j^{(n)} - x_{j'}^{(n)})^2, &(19)\\ &\triangleq \sum_{j=1}^{N_v} \phi_{R,j}^{(n)}(x_j) = \phi_R^{(n)}(x). &(20) \end{aligned}$$

Here, we use the convexity inequality in (17) and symmetry of quadratic function ($\psi_j(k) = \psi_j(-k)$) in (18). The regularizing factor between two voxels of j and j' is computed twice when switching center and neighbor, thus we can simplify the equation for x_j in (19). First and second derivatives of $\phi_R^{(n)}(x)$ at the point $x^{(n)}$ are as follows:

$$\dot{\phi}_{R,j}^{(n)}(x^{(n)}) = 2\beta \sum_{j' \in \Omega_j} \rho_{jj'} (x_j^{(n)} - x_{j'}^{(n)}), \tag{21}$$

$$\ddot{\phi}_{R,j}^{(n)}(x) = 4\beta \sum_{j' \in \Omega_j} \rho_{jj'} = d_{R,j}. \tag{22}$$

Now, the majorizer $\Psi(x)$ is:

$$\begin{aligned} \Psi_o(x) &\le \Psi(x) = \phi_L^{(n)}(x) + \phi_R^{(n)}(x), &(23)\\ &= \sum_{t=1}^{N_T} \sum_{j=1}^{N_v} \phi_{L,j}^{t,(n)}(x_j) + \sum_{j=1}^{N_v} \phi_{R,j}^{(n)}(x_j). &(24) \end{aligned}$$

The SQS with quadratic regularization provides the voxel-wise update at each iteration as follows:

$$x_j^{(n+1)} = x_j^{(n)} - \frac{\dot{\phi}_{L,j}^{(n)}(x_j^{(n)}) + \dot{\phi}_{R,j}^{(n)}(x_j^{(n)})}{d_{L,j}^{(n)} + d_{R,j}}, \quad \forall j \in [1, ..., N_v]. \tag{25}$$

Then, the step size of SQS with quadratic regularization has this relationship[26]:

$$\Delta_j^{(n)} = x_j^{(n+1)} - x_j^{(n)} \propto \frac{1}{d_{L,j}^{(n)} + d_{R,j}}. \tag{26}$$

Because the $d_{R,j}$ does not change in iterations as shown in equation (22), the step size is mainly decided by $d_{L,j}^{(n)}$. To accelerate the SQS algorithm, a larger value of $g_{ij}^{t,(n)}$ (or equivalently smaller value of $d_{L,j}^{(n)}$) can encourage larger step size. Here, we choose $g_{ij}^{t,(n)} = \frac{a_{ij}^t u_j^{(n)}}{\sum_{j'=1}^{N_v} a_{ij'}^t u_{j'}^{(n)}}$ using the non-uniform (NU) based update factor ($u_j^{(n)}$) as follows[17,27]:

$$u_j^{(n)} = \max\{|x_j^{(n-1)} - x_j^{(n)}|, \delta\}, \tag{27}$$

where δ is a small positive value. Note that the $u_j^{(n)}$ is an approximation of the oracle non-uniform factor $|x_j^{(n)} - x_j^{(\infty)}|$. Our experimental results show that applying Gaussian filtering to noisy $u^{(n)}$ empirically provides a better approximation to the oracle non-uniform factor. The Gaussian filtering does not affect the final image quality but affects the convergence speed at early iterations. The corresponding non-uniform denominator is

$$\hat{d}_{L,j}^{(n)} = \frac{1}{u_j^{(n)}} \sum_{t=1}^{N_T} \sum_{i=1}^{N_m} v_i^{t,(n)} a_{ij}^t \left(\sum_{j'=1}^{N_v} a_{ij'}^t u_{j'}^{(n)} \right), \tag{28}$$

which leads to the NU-based update relationship of $\Delta_j^{(n)} \propto u_j^{(n)}$ that encourages the voxel-wise non-uniform step size. This is a key property that allows one to control the recovery ratio of each voxel by using non-uniform step size; Section IV demonstrates numerically that a suitable NU step size can provide similar ROI recovery ratios across iterations in TOF-PET reconstruction.

The NUSQS method exploits the surrogate function with a diagonal Hessian matrix $D^{(n)}$ of $\Psi(x)$[26]. Specifically, the jth diagonal element of $D^{(n)}$ is $D_j^{(n)} = \hat{d}_{L,j}^{(n)} + d_{R,j}$. Algorithm 1 presents the pseudo code for NUSQS (one-subset version).

Algorithm 1 NUSQS

Initialize $x^{(0)} = \mathbf{1}$ and $u^{(0)} = \mathbf{1}$.
for $n = 0, 1, ...$ **do**
  for $j = 1, 2, ..., N_v$ **do**
    $D_j^{(n)} = \hat{d}_{L,j}^{(n)} + d_{R,j}$
    $x_j^{(n+1)} = \left[x_j^{(n)} - \nabla_j \Psi(x^{(n)}) / D_j^{(n)} \right]_+$
    $u_j^{(n+1)} = \max\{|x_j^{(n+1)} - x_j^{(n)}|, \delta\}$
  end for
end for

II.C. Nesterov's Momentum and Ordered Subsets Methods

To further accelerate the convergence speed, we exploit the Nesterov's momentum and ordered subsets (OS) methods. Iterative TOF-PET reconstruction requires the forward A ($= [A_1, ..., A_T]$) and backward A^H ($= [A_1^H, ..., A_T^H]$) projection operators. We set N_s as the number of subsets and the sth-subset forward A_s ($= [A_{s1}, ..., A_{sT}]$) and backward A_s^H ($= [A_{s1}^H, ..., A_{sT}^H]$) projection operators. Subsets are equally distributed over the angular bins. The computing cost per subiteration decreases almost linearly with the number of subsets N_s.

Now, the (approximate) majorizer with ordered subsets for TOF reconstruction is

$$\Psi(x) = \sum_{s=1}^{N_s} \Psi_s(x) \quad , \tag{29}$$

where

$$\Psi_s(x) = \phi_{sL}^{(n)}(x) + \frac{1}{N_s} \phi_R^{(n)}(x) \quad , \tag{30}$$

$$= \sum_{t=1}^{N_T} \sum_{j=1}^{N_v} \phi_{sL,j}^{t,(n)}(x_j) + \frac{1}{N_s} \sum_{j=1}^{N_v} \phi_{R,j}^{(n)}(x_j), \tag{31}$$

$$\phi_{sL,j}^{t,(n)}(x_j) = \sum_{i \in \Omega_s} g_{ij}^{t,(n)} p_i^{t,(n)} \left(\frac{a_{ij}^t}{g_{ij}^{t,(n)}} (x_j - x_j^{(n)}) + [A_{st} x^{(n)}]_i \right), \tag{32}$$

$$\hat{d}_{sL,j}^{t,(n)} = \frac{1}{u_j} \sum_{i \in \Omega_s} v_i^{t,(n)} a_{ij}^t \left(\sum_{j'=1}^{N_v} a_{ij'}^t u_{j'}^{(n)} \right). \tag{33}$$

Ω_s denotes the sth subset and we evenly distribute the subsets in azimuthal bins. Algorithm 2 presents the pseudo code of OS-NUSQS. The subset balance can be approximately described as follows[10,19,26]:

$$\nabla \Psi(x) \approx N_s \nabla \Psi_s(x), \tag{34}$$

$$D^{(n)} \approx N_s D_s^{(n)}, \tag{35}$$

where $s = [1, ..., N_s]$ and D_s is a diagonal Hessian of the surrogate function of $\Psi_s(x)$ for the sth subset; $D_{sj}^{(n)}$ is a jth diagonal component of $D_s^{(n)}$.

Algorithm 2 OS-NUSQS

Initialize $x^{(0)} = \mathbf{1}$ and $u^{(0)} = \mathbf{1}$.
for $n = 0, 1, ...$ **do**
  for $s = 0, 1, ..., N_s - 1$ **do**
    $k = n \times N_s + s$
    for $j = 1, 2, ..., N_v$ **do**
      $D_{sj}^{(k)} = \hat{d}_{sL,j}^{(k)} + d_{R,j}/N_s$
      $x_j^{(k+1)} = \left[x_j^{(k)} - \nabla_j \Psi_s(x^{(k)}) / D_{sj}^{(k)} \right]_+$
      $u_j^{(k+1)} = \max\{|x_j^{(k+1)} - x_j^{(k)}|, \delta\}$
    end for
  end for
end for

Next, we consider OS-NUSQS combined with Nesterov's momentum[18,19]. Specifically, the Nesterov's momentum[18] exploits the previous descent updates for additional acceleration. By combining with the OS method, we expect the convergence speed $O(1/(nN_s)^2)$ in early iterations that is significantly faster than the speeds of NUSQS $O(1/n)$ and OS-NUSQS $O(1/(nN_s))$. This is one of the main advantages of the OS in the momentum method, which accelerates approximately $(N_s)^2$ times in early iterations[19]. Also, the momentum factor does not require additional memory because the non-uniform factor (u) can be reused in the momentum computation. In algorithm 3, the non-uniform factor (u) is calculated in each iteration; to save computation, we first calculate the variation $x_d^{(k+1)} = (x^{(k+1)} - x^{(k)})$ for momentum (line 10 in algorithm 3), and then we update $u^{(k+1)} = \max\{|x_d^{(k+1)}|, \delta\}$ (line 11 in algorithm 3).

Algorithm 3 OS-NUSQS with Nesterov's Momentum

1: Initialize $x^{(0)} = \mathbf{1}$, $z^{(0)} = \mathbf{1}$, $u^{(0)} = \mathbf{1}$ and $b_0 = \mathbf{1}$; $t_\gamma \in [0,1]$.
2: **for** $n = 0, 1, ...$ **do**
3:     **for** $s = 0, 1, ..., N_s - 1$ **do**
4:         $k = n \times N_s + s$
5:         **for** $j = 1, 2, ..., N_v$ **do**
6:             $b_{k+1} = \frac{1}{2}\left(1 + \sqrt{1 + 4b_k^2}\right)$
7:             $\gamma_k = t_\gamma \frac{b_k - 1}{b_{k+1}}$
8:             $D_{sj}^{(k)} = \hat{d}_{sL,j}^{(k)} + d_{R,j}/N_s$
9:             $x_j^{(k+1)} = \left[z_j^{(k)} - \nabla_j \Psi_s(z^{(k)})/D_{sj}^{(k)}\right]_+$
10:             $z_j^{(k+1)} = \left[x_j^{(k+1)} + \gamma_k (x_j^{(k+1)} - x_j^{(k)})\right]_+$
11:             $u_j^{(k+1)} = \max\{|x_j^{(k+1)} - x_j^{(k)}|, \delta\}$
12:         **end for**
13:     **end for**
14: **end for**

In addition, a relaxation factor t_γ in algorithm 3 is applied. Although the momentum method can practically improve the convergence speed, the OS-NUSQS with momentum algorithm can sometimes diverge due to OS[19], as will be presented in Discussion. Therefore, we set the relaxation factor, such as $t_\gamma \in [0,1]$, to avoid divergence in the proposed method.

III. EXPERIMENTAL SETUP

III.A. Computer simulation

We performed a computer simulation using an XCAT phantom[20] and the simulation geometry of a clinical Biograph mCT TOF PET/CT scanner (Siemens Medical Solutions USA, Inc.). In a TOF sinogram, the number of radial bins with a 2.005 mm pixel size and azimuthal bins are 336, and the number of time bins is 13 with a 560 ps time resolution; the radius of scanner is 427.6 mm and data was acquired with span 11. In the reconstructed image, the number of image voxels is 336×336×109 with a 2.005×2.005×2.027 mm^3 voxel size. To evaluate ROI-based convergence, we chose three hot ROIs and one cold ROI at lung, spine and liver as shown in Fig. 1. In each region, a high intensity component assumed as a tumor is added. Specifically, different tumor shapes and intensities were applied for evaluation of uniform recovery under different conditions. We put a sphere with a 7 mm radius and 0.4 intensity in lung, a sphere with a 6 mm radius and 0.3 intensity in spine and an ellipsoid with axes of length (15,7,7) mm and 0.3 intensity in liver as shown in Figs. 1(b)–(d), respectively. The muscle (background) intensity is 0.02. For ROI-based metric comparison, all ROIs were extracted by the shape of sphere with a 20 mm radius at the centers of tumors. The ground truths of tumor to muscle ratios (TMRs) are 20 in lung, 15 in spine, 15 in liver, respectively. We imposed Poisson noise to the prompt (attenuated) sinogram with signal to noise ratio (SNR) of 8 dB in which the number of counts was 3.3 $\times 10^7$. In simulation, detector-pair sensitivities, scatter and random counts were not considered.

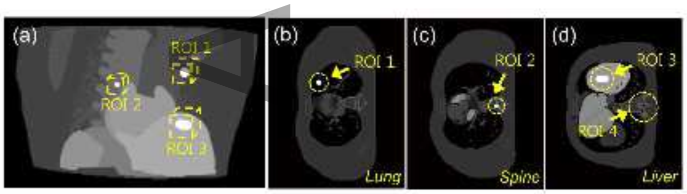


Fig. 1 XCAT phantom simulation setup using (a) three tumor ROIs with high intensity components at (b) lung, (c) spine and (d) liver. A cold ROI in (d) was used for comparison.

We also conducted experiments to evaluate the recovery ratio. We used the same regularization ($\beta = 0.2$) for all methods. To consider not only tumor but also different background regions, we selected four sphere-shape ROIs with 20 mm radius, as shown in Fig. 1. Specifically, we considered spine ROI with complex structures, lung ROI with very low intensity and liver ROI with high intensity backgrounds.

III.B. Clinical example

A pancreas-focused scan was performed for 45 minutes with a TOF PET/MR (SIGNA, GE Healthcare) scanner. A bolus injection of 196.1 MBq of ^{18}F-FDG was administered. The protocol of this experiment was approved by the Institutional Review Board (IRB) of University of California, San Francisco (UCSF). The SIGNA scanner has 357×224×1981×27 (radial, angle, plane, TOF) bins with time of flight resolution of 420 ps. The scanner radius is 640 mm and the field-of-view (FOV) is 600 mm. The reconstructed image size was 256×256×89 with 2.34×2.34×2.78 mm^3 resolution.

For evaluation, we used the full dose image as the ground truth and 6× downsampled data as a measurement. The full dose image was acquired by EM (one-subset version of OS-EM) after 300 iterations. To generate 6× downsampled sinograms, a Poisson thinning process was used[21] in which coincidence events can be randomly discarded by a predetermined sampling factor. The Poisson thinning process has been applied to initial prompt data (listmode or sinogram) before random, scatter, normalization and attenuation corrections. We selected three tumor and one cold ROIs extracted by a sphere with a 12.5 mm radius. In patient data, detector-pair sensitivities (normalization), scatter and random counts were fully considered.

III.C. Evaluation

In this paper, we compared the OS-NUSQS with conventional OS-EM and OS-SQS. The step size of OS-SQS does not consider spatially different SNRs nor recovery variations. OS-EM has a spatially variant step size, however, the step size of OS-EM does not take into account the recovery variation. We compared algorithms without regularization ($\beta = 0$) and with regularization ($\beta > 0$). For regularized OS-EM algorithm, a De Pierro's EM algorithm[28] is used for comparison, in which the quadratic regularization was equivalently used as in other methods. Throughout this paper, the *iteration* in result plots denotes the sub-iteration (k), and we used 8 subsets in reconstruction.

The normalized root mean square error (NRMSE) $\frac{\sqrt{\sum_{j=1}^{N_v}(x_j^* - x_j^{(n)})^2}}{\sqrt{\sum_{j=1}^{N_v}(x_j^*)^2}}$ was computed using the ground truth image (x^*). To compare convergences of ROIs, we also calculated the ROI-based normalized root mean square difference (NRMSD) as $\frac{\sqrt{\sum_{j\in\Omega_r}(x_j^{(\infty)} - x_j^{(n)})^2}}{\sqrt{\sum_{j\in\Omega_r}(x_j^{(\infty)})^2}}$, where n is the iteration, $x^{(\infty)}$ is the converged image after 300 iterations with one-subset version of OS-EM and Ω_r is the rth ROI. The recovery ratio was measured by $\frac{\sqrt{\sum_{j\in\Omega_r}(x_j^* - x_j^{(\infty)})^2}}{\sqrt{\sum_{j\in\Omega_r}(x_j^* - x_j^{(n)})^2}}$. Here, the distance between the ground truth and the converged image ($x^* - x^{(\infty)}$) was used for calculating the recovery ratio (i.e., the recovery ratio approaches 1 with increasing iterations). Furthermore, we measure the ROI-based structural similarity (SSIM) index for comparing image quality. The SSIM is defined by $\frac{(2\mu_{r^*}\mu_r + c_1)(2\sigma_{r^*r} + c_2)}{(\mu_{r^*}^2 + \mu_r^2 + c_1)(\sigma_{r^*}^2 + \sigma_r^2 + c_2)}$, where r^* and r denote the ROIs of the ground truth and the reconstructed image, respectively. μ is the average, σ^2 is the variance and σ_{r^*r} is the covariance of intensity within the ROIs r^* and r. $c_1 = 2.5 \times 10^{-5}$ and $c_2 = 2.25 \times 10^{-4}$ were used.

III.D. Implementation

Parallel computing technologies are commonly used to speed up computations. In particular, the general purpose graphics processing unit (GPGPU) has been widely used for medical applications, such as 3-D CT and PET reconstructions[29,30,31,32]. To speed up the proposed method, we implemented TOF reconstruction using the GPU and compute unified device architecture (CUDA), similar to previous work[33]. Specifically, we implemented our TOF system model with ray-driven forward projector and a matched (transpose) back projector with a time response function that is a typical Gaussian function with FWHM based on the timing resolution. In GPU kernels of TOF forward and backward projectors, each thread corresponds to a line-of-response (LOR); thus, all time bins are updated or used in each thread. Technically, we use the *atomic* operator to avoid interference from other threads when accessing a specific address at the same time, and the 3-D linear interpolation using the *texture* memory is exploited for line integrals. A Gaussian coefficient table for the time response function is pre-calculated before the reconstruction process, and is assigned to the *constant* memory. All geometrical parameters are also assigned to the *constant* memory. The surrogate function of quadratic roughness regularization is easily parallelizable because the calculation of each voxel is independent, thus each thread in a GPU kernel can be assigned to each voxel. One main difference compared to the conventional PET reconstruction is that the denominator D in algorithms [1, 2, 3] cannot be pre-computed due to the the non-uniform factor. Thus, the proposed method requires two forward and backward projections in each iteration. To further reduce the computing cost, our code calculates $Ax^{(n)}$ and $Au^{(n)}$ at the same time in forward projection of GPU kernel that can share all geometrical computations; and the $\nabla\Psi(x)$ and D are also calculated at the same time and share the geometrical computations in backward projection of GPU kernel as similarly done by Kim *et al*[19]. This approach was 1.5 ∼ 1.7 times faster than separate calculations.

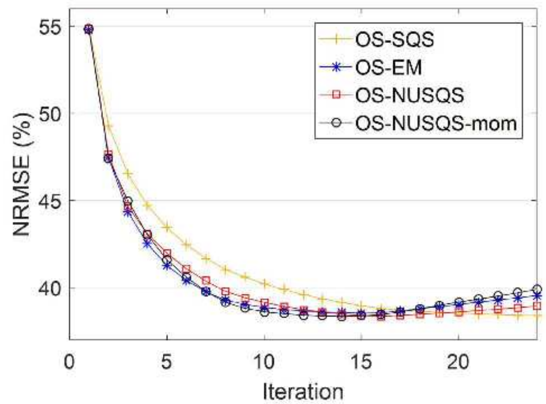


FIG. 2 Comparison of global NRMSEs for OS-SQS, OS-EM, OS-NUSQS and OS-NUSQS with momentum. $\beta = 0$ was used in simulation.

IV. RESULTS

IV.A. Computer simulation

IV.A.1. No regularization ($\beta = 0$)

First, to observe the convergence speed of different algorithms, we performed experiments without regularization, as a special case of $\beta = 0$ in Eq. (5). The conver-

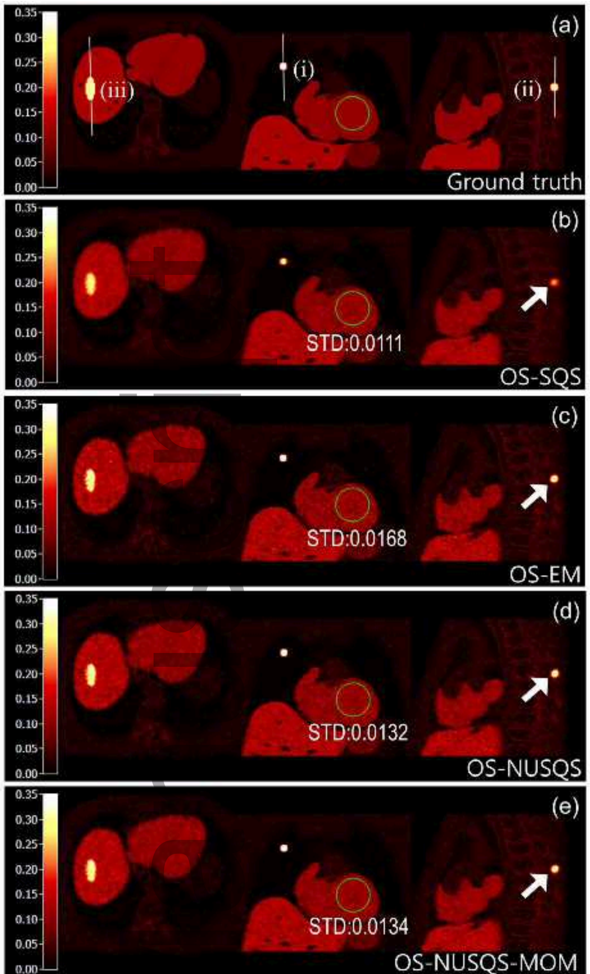


Fig. 3 (a) Ground-truth, and reconstructed images of (b) OS-SQS, (c) OS-EM, (d) OS-NUSQS and (e) OS-NUSQS with momentum at 24 iterations. Standard deviations for a flat region of interest in coronal view were compared. $\beta = 0$ was used in simulation.

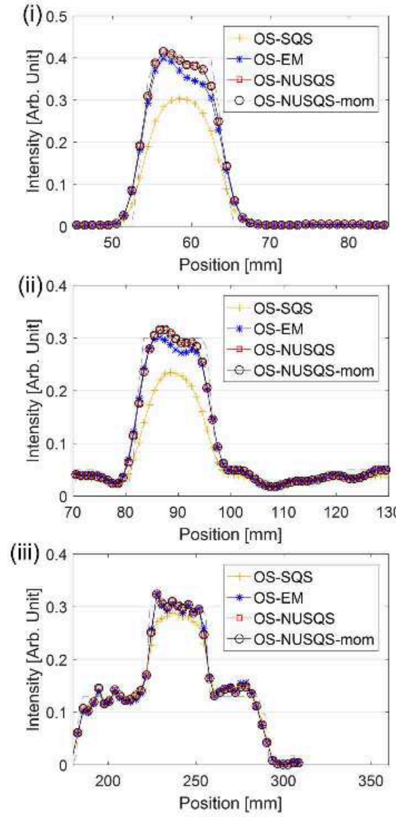


Fig. 4 Profiles of tumor ROIs in (i) lung, (ii) spine and (iii) liver as described in Fig. 3(a). Dot line is the ground truth. Profiles were measured with $\beta = 0$ at 24 iterations in simulation.

gence speed of OS-NUSQS and OS-EM were similar as shown in Fig. 2. However, the OS-EM diverged faster than the OS-NUSQS. The OS-SQS at 24 iterations did not reach the minimum NRMSE. Figure 3 shows reconstructed images of OS-SQS, OS-EM, OS-NUSQS and OS-NUSQS with momentum methods at 24 iterations. The OS-EM image was noisier than other images, particularly in the lower intensity region. We compared the standard deviations in the flat region with low intensity, a circle in coronal view as shown in Fig. 3. The standard deviations of the flat region were 0.0111, 0.0168, 0.0132 and 0.0134 for the OS-SQS, OS-EM, OS-NUSQS and OS-NUSQS with momentum methods, respectively. The reconstructed images of OS-NUSQS with and without momentum were visually similar to the image of OS-EM, and we confirmed that OS-NUSQS has reduced noise compared to OS-EM in early iterations of TOF reconstruction. The OS-SQS method did not reach convergence at 24 iterations, particularly at high intensity regions, such as tumors in lung and spine as shown in Fig. 3(b). For high intensity regions, Fig. 4 compares profiles of tumors in liver, lung and spine regions as shown in Fig. 3(a). In Fig. 4, we observed that the profiles of OS-SQS did not reach convergence after 24 iterations, and demonstrated that the contrast of OS-NUSQS was higher than that of OS-EM at 24 iterations. Although the intensity of OS-EM in liver ROI was similar to that of NUSQS, OS-EM needed more iterations for ROIs in lung and spine. Note that profiles of OS-NUSQS with and without momentum were almost identical.

IV.A.2. With regularization ($\beta > 0$)

Figure 5 compares NRMSDs for whole images of OS-SQS, OS-EM, OS-NUSQS and OS-NUSQS with momentum. The per-iteration convergence speed of OS-SQS was the slowest, and that of OS-NUSQS with momentum ($t_\gamma = 0.7$) was the fastest. In Fig. 6, NRMSDs of tumor ROIs with OS-NUSQS converged much faster than those of OS-EM. We compared tumor to muscle ratios (TMR) of OS-SQS, OS-EM, OS-NUSQS and OS-NUSQS with momentum after 24 iterations in table I. TMRs of the proposed method were slightly higher than those of OS-

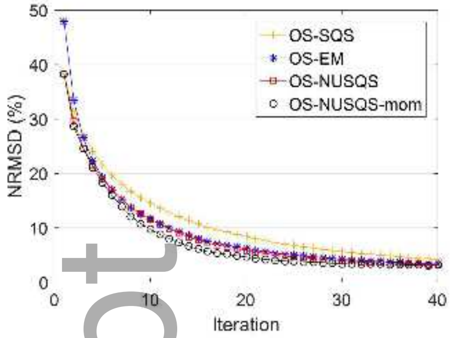


Fig. 5 Comparison of NRMSDs for OS-SQS, OS-EM, OS-NUSQS and OS-NUSQS with momentum algorithms with quadratic regularization. $\beta = 0.2$ was used in simulation.

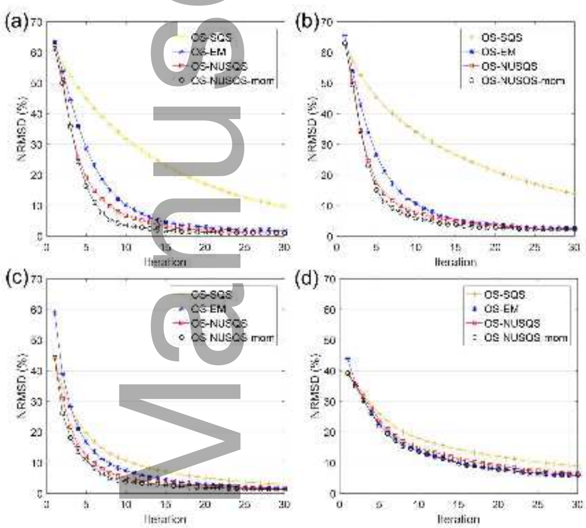


Fig. 6 Comparison of NRMSDs of OS-SQS, OS-EM, OS-NUSQS and OS-NUSQS with momentum algorithms using quadratic regularizations for tumor ROIs in (a) lung, (b) spine, (c) liver and (d) cold region as shown in Fig. 1. $\beta = 0.2$ was used in simulation.

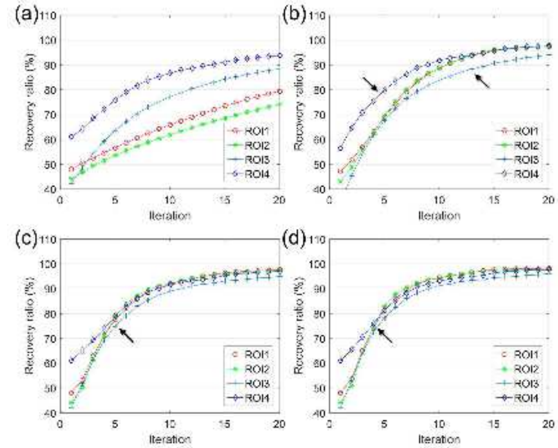


Fig. 7 ROI-based recovery ratio comparisons of (a) OS-SQS, (b) OS-EM, (c) OS-NUSQS and (d) OS-NUSQS with momentum. $\beta = 0.2$ was used in simulation.

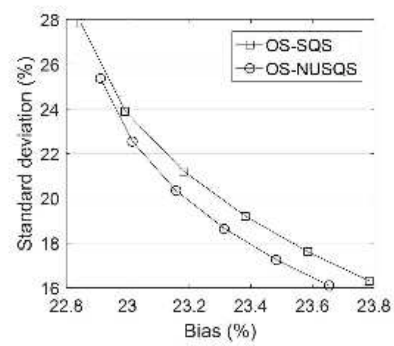


Fig. 8 Bias and standard deviation plots of OS-SQS and OS-NUSQS methods at 24 iterations. β parameters between 0.08 to 0.3 were used in simulation.

SQS and OS-EM.

Figure 7 compares the recovery ratios of OS-SQS, OS-EM, OS-NUSQS and OS-NUSQS with momentum for four ROIs. In general, similar ratios of ROIs illustrate the uniform recovery of ROIs. We observed that the OS-NUSQS and OS-NUSQS with momentum methods show uniform recovery ratios after 5 iterations for both hot and cold ROIs. In Fig. 8, the bias and standard deviation plots of OS-SQS and OS-NUSQS methods at 24 iteration were compared using 10 simulations (same SNR) with random Poisson noise. Fig. 8 used β parameters between 0.08 to 0.3 for both OS-SQS and OS-NUSQS, which demonstrates that the non-uniform method can improve the image quantitatively in early iteration. Here, the bias and standard deviation values of OS-EM, OS-NUSQS and OS-NUSQS with momentum were almost the same after 24 iterations. Note that after convergence with sufficient number of iterations, bias and standard deviation values of OS-SQS and OS-NUSQS are approximately equivalent since they solve the same optimization problem.

In the proposed method, the voxel-wise non-uniform factor ($u_j^{(n)} = \max\{|x_j^{(n)} - x_j^{(n-1)}|, \delta\}$) is calculated at each iteration. We used Gaussian filtering on the non-uniform factor to provide a better approximation of the oracle non-uniform factor in our experiment, which lead to higher convergence speed at early iterations. To observe the relationship between noise levels and the optimal FWHM of Gaussian filtering, we compared NRMSEs of reconstructed images using the non-uniform factors without and with Gaussian filtering with FWHM of 2, 4 and 6 mm, as shown in Fig. 9. Two noise levels with total counts of (a) 1.98×10^7 and 4.62×10^7 were used. Note that the total photon counts used in the simulation was 3.3×10^7. The non-uniform factor using

TMR	OS-SQS	OS-EM	OS-NUSQS	OS-NUSQS-mom
Lung	13.7	14.4	14.5	14.5
Spine	10.7	11.7	11.8	11.8
Liver	12.7	12.7	12.8	12.8

Table I Tumor to muscle ratios of (a) OS-SQS, (b) OS-EM, (c) OS-NUSQS and (d) OS-NUSQS with momentum. $\beta = 0.2$ was used in simulation.

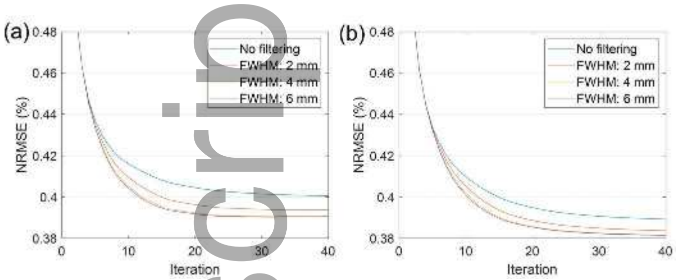


Fig. 9 NRMSE comparison without and with Gaussian filtering of FWHM 2, 4, 6 mm for different noise levels with total counts of (a) 1.98×10^7 and (b) 4.62×10^7. $\beta = 0.2$ was used in simulation.

Gaussian filtering showed fast decrease of NRMSEs at early iterations. The performance with Gaussian filtering of FWHM larger than 4 mm was almost the same. We observed that the FWHM of Gaussian filtering was not highly related to the noise level, thus an FWHM of 4 mm was used for Gaussian filtering in non-uniform based algorithms.

IV.B. Clinical example

To evaluate the recovery ratio with converged images, we performed reconstructions with quadratic regularization having $\beta = 0.03$ and relaxation factor of 0.5 for Nesterov's momentum. We will additionally discuss effects of parameters such as the number of subsets, relaxation factor in Section V.

Figure 10 shows the reconstructed images of OS-SQS, OS-EM, OS-NUSQS and OS-NUSQS with momentum methods at 24 iterations. Figure 10(a) is the ground truth of full dose EM image after 300 iterations, and Fig. 10(b) is the converged EM image of 6× downsampled data after 300 iterations. OS-SQS was not fully converged in Fig. 10(c) (see arrow), however, other methods converged after 24 iterations.

Figure 11 compares the recovery ratios of OS-SQS, OS-EM, OS-NUSQS and OS-NUSQS with momentum for four ROIs. The OS-NUSQS and OS-NUSQS with momentum methods show uniform recovery ratios after 24 iterations for both hot and cold ROIs. In Fig. 12, we also compared the recovery images of OS-SQS, OS-EM, OS-NUSQS and OS-NUSQS with momentum, where the voxel-wise recovery ratio was calculated at 24 iterations and the saggital view in Fig. 10 was used. Because of the

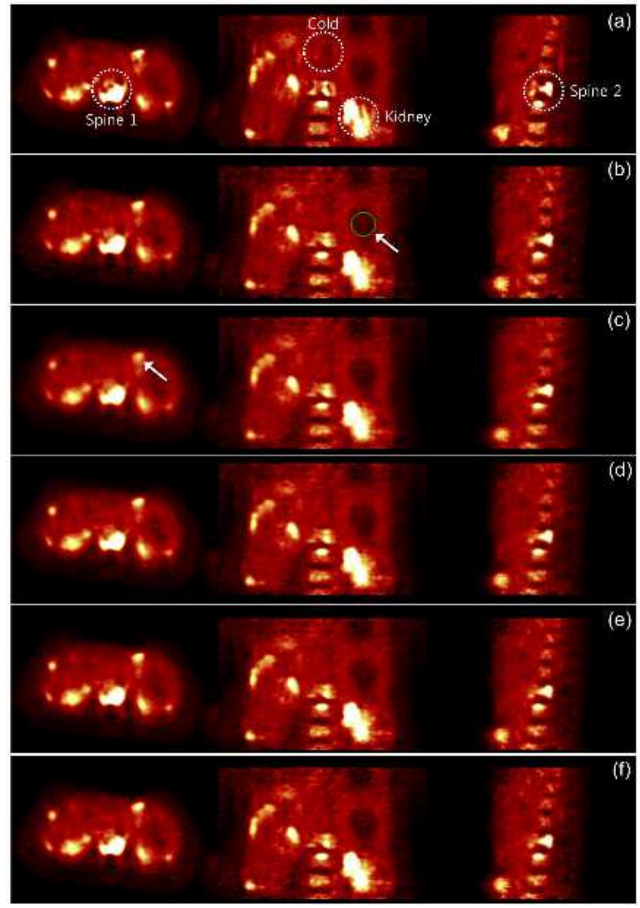


Fig. 10 (a) Ground truth of full dose EM image, and reconstructed images of (b) converged OS-EM image with 300 iterations, (c) OS-SQS, (d) OS-EM, (e) OS-NUSQS and (f) OS-NUSQS with momentum. (b)–(f) used 6× downsampled data. Three hot ROIs (spine 1, kidney, spine 2) were extracted at centers of tumors and a cold region was also extracted. $\beta = 0.03$ was used with patient data.

high noise of recovery image, a Gaussian filtering with FWHM 2.5 mm was additionally applied only for visualization in Fig. 12. Boundaries and inner regions in recovery images of OS-SQS and OS-EM were not uniform at 24 iterations. Although the recovery ratios of specific ROIs of OS-NUSQS and OS-NUSQS with momentum were similar, the recovery image of OS-NUSQS with momentum was more uniform than the recovery image of OS-NUSQS.

For image quality comparison, we compared ROI-based SSIM values as shown in table II. The SSIMs of NUSQS combining with or without momentum were the highest for all ROIs. Here, SSIM considers mean (bias) and variance (noise). We confirmed that the mean values of hot regions of OS-EM, OS-NUSQS, OS-NUSQS with momentum were almost identical. We additionally compared biases for a cold region in Fig. 10(b) (see arrow), where biases of OS-SQS, OS-EM, OS-NUSQS and OS-NUSQS with momentum were 0.8%, 0.6%, 0.09% and 0.12%. We confirmed that the OS-NUSQS has faster

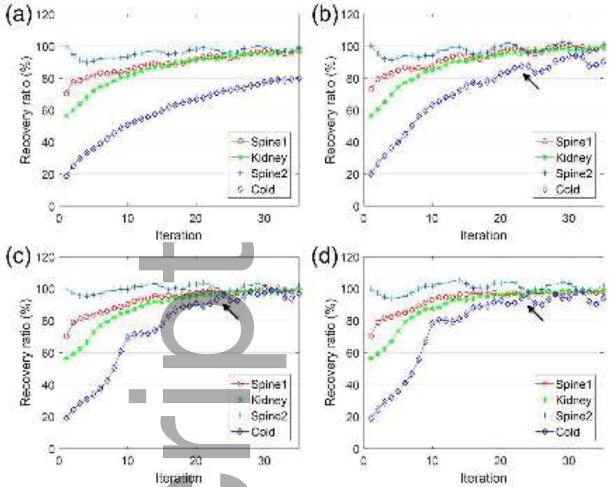


FIG. 11 ROI-based recovery ratio comparisons of (a) OS-SQS, (b) OS-EM, (c) OS-NUSQS and (d) OS-NUSQS with momentum. $\beta = 0.03$ was used with patient data.

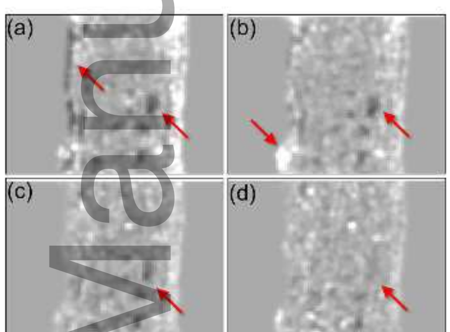


FIG. 12 Comparison of recovery images of (a) OS-SQS, (b) OS-EM, (c) OS-NUSQS and (d) OS-NUSQS with momentum at 24 iterations. The saggital view of Fig. 10 was used. The intensity window is [0,1.5].

convergence of lower SNR region compared to OS-EM.

Although the momentum method can increase the computing speed while preserving the uniform recovery, the reconstructed image can diverge when combined with ordered subsets, thus the relaxation factor was incorporated into the momentum method. We performed an empirical comparison of relaxation factor, effect of subsets and β, as similarly done by Berker *et al*[34]. Figures 13(a–c) show NRMSE comparisons for various number of subsets of 1, 8 and 16 with relaxation factors (t_γ) of 0 (without momentum), 0.3, 0.5 and 0.7, where the patient data and fixed $\beta = 0.03$ were used. Without ordered subsets in Fig. 13(a), the OS-NUSQS with momentum decreased NRMSE monotonically with relaxation factors less than 0.7. With 8 or 16 subsets, NRMSEs diverge with high relaxation factors (see arrows in Figs. 13(c) and (d)). Figure 13(d) shows NRMSE comparison for various β with 16 subsets and relaxation factor of 0.7. In our observations, the proper hyper-parameter and relaxation factor for 16 subsets were $\beta = 0.05$ and $t_\gamma = 0.5$, respectively. Similarly, the proper parameters for 8 subsets were $\beta = 0.03$ and $t_\gamma = 0.5$. The results indicated that higher β is required for higher relaxation factors to enable the convergence, however, different βs converge to different solutions.

SSIM	OS-SQS	OS-EM	OS-NUSQS	OS-NUSQS-mom
Spine 1	0.841	0.850	**0.852**	0.840
Kidney	0.727	0.741	**0.748**	0.724
Spine 2	0.688	0.690	**0.694**	0.683
Cold	0.942	0.953	0.963	**0.982**

TABLE II ROI-based SSIM comparisons of OS-SQS, OS-EM, OS-NUSQS and OS-NUSQS with momentum. $\beta = 0.03$ was used with patient data.

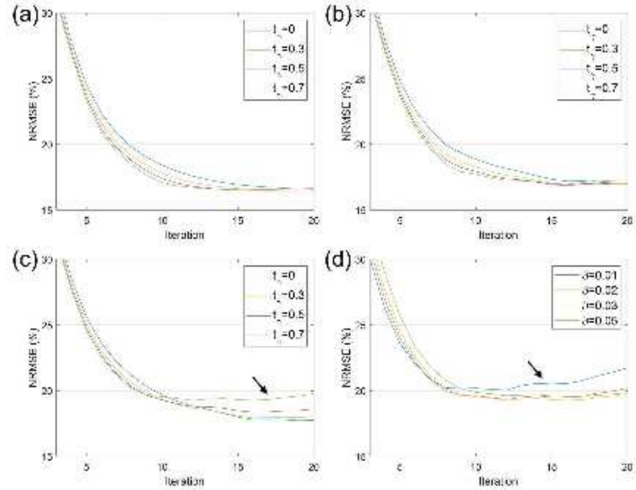


FIG. 13 Comparison of NRMSEs using different number of subsets of (a) 1, (b) 8 and (c) 16 combining with relaxation factors (t_γ) of 0 (without momentum), 0.3, 0.5 and 0.7, and (d) NRMSEs of $\beta = 0.01$, 0.02, 0.03 and 0.05 with 16 subsets and relaxation factor of 0.7.

IV.C. Execution time

In our iterative algorithms, the most time-consuming operations are forward and backward projectors. Table III compares the computing time using quad-core CPU with 3.5 GHz, 48-core CPU server with 2.4 GHz and GPU (Geforce GTX 1080, Nvidia). Here, GE SIGNA TOF data without subset was used. For implementation, we used OpenMP for CPU and CUDA for GPU, specifically, 4-cores with 3.5 GHz of a personal computer, 48-cores with 2.4 GHz of a cluster server and Nvidia Geforce GTX 1080 were compared. We observed that the overall acceleration is about 80×.

We also compared computing time (sec) of OS-EM, OS-NUSQS and OS-NUSQS with momentum, which in-

	Forward	Backward	Acceleration
CPU(4)	1920	2063	1.0×
CPU(48)	368	480	5.0×
GPU	25	25.1	80.0×
	OS-EM	OS-NUSQS	OS-NUSQS+mom
GPU	14.1	15.8	16.0

TABLE III Computing time (sec) of TOF forward and backward projectors without subsets using CPU and GPU. Specifically, 4-cores with 3.5 GHz of a personal computer, 48-cores with 2.4 GHz of a cluster server and Nvidia Geforce GTX 1080 for GPU implementation were compared. Computing time (sec) of OS-EM, OS-NUSQS and OS-NUSQS with momentum were compared. GPU time indicates one sub-iteration (one of 8 subsets) computing time including TOF forward and backward projectors and quadratic penalty.

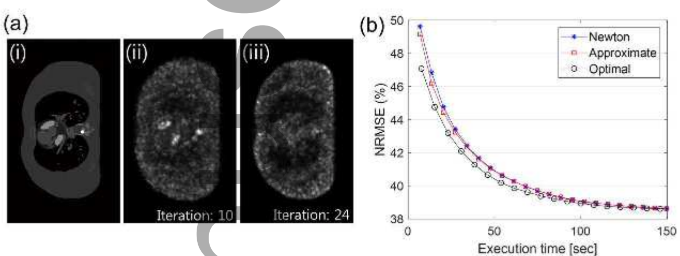


FIG. 14 (a)(i) Ground truth image and non-uniform factors at iterations of (ii) 10 and (iii) 24, and (b) comparison of NRMSEs of the reconstructed images over time using the Newton, approximate and optimal curvatures. $\beta = 0.2$ was used in simulation.

dicates one sub-iteration time of 8 subsets including TOF forward and backward projectors and quadratic penalty. We observed that the OS-EM was 12% faster than the OS-NUSQS.

V. DISCUSSION

In Fig. 14(a), we observed variations of the non-uniform factor (u in Eq. (27)) by iterations. In early iterations, high intensity regions have larger non-uniform factors, indicating larger step size. After reaching similar recovery ratios, we observed that the step size in non-uniform factor became more uniform, which means non-uniform updates yield uniform recovery in early iterations. Figure 14(b) compares the Newton curvature ($\ddot{h}_i^t([A_t x^{(n)}]_i)$), approximate curvature ($1/\max([A_t x^{(n)}]_i, \epsilon)$) and optimal curvature in equation (10). Although the optimal curvature is the fastest at early iterations, optimal curvature requires additional computations when $[A_t x^{(n)}]_i = 0$. Because the NRMSE of optimal curvature became similar to NRMSEs of others after 10 iterations and the approximate curvature can be directly calculated by re-projected values in each iteration, we used the approximate curvature in this paper. Note that although the approximate curvature does not hold inequalities of Eqs. (6), (12), (13) and (23) in general, we can reduce the computing time (13% in our implementation).

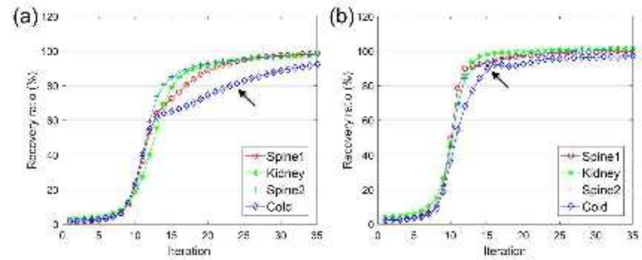


FIG. 15 Recovery ratio comparisons of (a) OS-EM and (b) OS-NUSQS. Here, the same initial image far from the ground truth was used in both (a) and (b). $\beta = 0.03$ was used with patient data.

In Figs. 7 and 11, the uniformity of recovery ratio was achieved at different number of iterations and we observed that the fast reaching of uniformity depends on the distance between the initial value and the converged value in reconstruction. We used 1 as an initial value for all voxels in both simulated data and real data in which the initial value of the cold region in the simulation was closer to the converged value and thus the uniformity was achieved faster. To investigate the impact of the initial value, particularly when it is very far from the solution, we compared recovery ratios of OS-EM and OS-NUSQS using different initial values in Fig. 15. We used a back-projected image as an initial image and same ROIs in Fig. 11 to compute the recovery ratio. Fig. 15(a) showed that the cold region of OS-EM converged slower than hot regions. The uniformity of recovery ratios of OS-NUSQS was achieved around 15 iterations in Fig. 15(b) compared to 24 iterations in Fig. 11(c), which demonstrated that the fast reaching of uniformity depends on the initial value.

The momentum method can speed up convergence in early iterations particularly when it is combined with the ordered subset approach. However, because the gradient at each iteration is updated from a subset of whole data, the algorithm can diverge. Fig. 13 demonstrated that the relaxation factor can achieve relatively faster convergence speed compared to non-momentum method. The selection of the relaxation factor is still an open question. We empirically selected the optimal relaxation factor and the number of subsets, and will further investigate the relationship of optimal parameters and noise.

VI. CONCLUSIONS

In conclusion, we derived an ordered subset based non-uniform separable quadratic surrogates (OS-NUSQS) with the Nesterov's momentum method for TOF PET reconstruction. The spatially non-uniform step size in the proposed method provided uniform recovery ratios of different SNR regions. The computer simulation and clinical example showed that the proposed method converged uniformly regardless of hot and cold ROIs, and the contrast and SSIM were higher at early iterations using the proposed method than using conventional OS-EM and OS-SQS methods. Furthermore, our GPU implementation was able to achieve 80× acceleration compared to the implementation using 4-core CPU.

ACKNOWLEDGMENT

This research was supported by the National Institute of Health (NIH) P41 EB022544-6814 and NIH AG052653. The authors have no conflicts to disclose.

[a] li.quanzheng@mgh.harvard.edu

[1] C. C. Watson, "An evaluation of image noise variance for time-of-flight PET," IEEE Transactions on Nuclear Science **54**, 1639–1647 (2007)

[2] M. Conti, "State of the art and challenges of time-of-flight PET," Physica Medica **25**, 1–11 (2009)

[3] D. L. Snyder, L. J. Thomas, and M. M. Ter-Pogossian, "A matheematical model for positron emission tomography systems having time-of-flight measurements," IEEE Transactions on Nuclear Science **3**, 3575–3583 (1981)

[4] T. Tomitani, "Image reconstruction and noise evaluation in photon time-of-flight assisted positron emission tomography," IEEE Transactions on Nuclear Science **28**, 4581–4589 (1981)

[5] M. Yamamoto, D. C. Ficke, and M. M. Ter-Pogossian, "Experimental assessment of the gain achieved by the utilization of time-of-flight information in a positron emission tomograph (Super PETT I)," IEEE Transactions on Medical Imaging **1**, 187–192 (1982)

[6] T. F. Budinger, "Time-of-flight positron emission tomography: status relative to conventional PET," Journal of Nuclear Medicine **24**, 73–78 (1983)

[7] S. Strother, M. Casey, and E. Hoffman, "Measuring PET scanner sensitivity: relating countrates to image signal-to-noise ratios using noise equivalents counts," IEEE Transactions on Nuclear Science **37**, 783–788 (1990)

[8] M. Conti, "Focus on time-of-flight PET: the benefits of improved time resolution," European Journal of Nuclear Medicine and Molecular Imaging **38**, 1147–1157 (2011)

[9] J. S. Karp, S. Surti, M. E. Daube-Witherspoon, and G. Muehllehner, "Benefit of time-of-flight in PET: experimental and clinical results," Journal of Nuclear Medicine **49**, 462–470 (2008)

[10] H. M. Hudson and R. S. Larkin, "Accelerated image reconstruction using ordered subsets of projection data," IEEE Transactions on Medical Imaging **13**, 601–609 (1994)

[11] S. Ahn and J. A. Fessler, "Globally convergent image reconstruction for emission tomography using relaxed ordered subsets algorithms," IEEE Transactions on Medical Imaging **22**, 613–626 (2003)

[12] J. Qi and R. M. Leahy, "A theoretical study of the contrast recovery and variance of MAP reconstructions from PET data," IEEE Transactions on Medical Imaging **18**, 293–305 (1999)

[13] Q. Li, B. Bai, S. Cho, A. Smith, and R. Leahy, "Count independent resolution and its calibration," in *International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine* (2009) pp. 223–226

[14] J. H. Cho and J. A. Fessler, "Regularization designs for uniform spatial resolution and noise properties in statistical image reconstruction for 3-D X-ray CT," IEEE Transactions on Medical Imaging **34**, 678–689 (2015)

[15] Z. Yu, J. Thibault, C. A. Bouman, K. D. Sauer, and J. Hsieh, "Fast model-based X-ray CT reconstruction using spatially nonhomogeneous ICD optimization," IEEE Transactions on Image Processing **20**, 161–175 (2011)

[16] K. Van Slambrouck and J. Nuyts, "Reconstruction scheme for accelerated maximum likelihood reconstruction: The patchwork structure," IEEE Transactions on Nuclear Science **61**, 173–181 (2014)

[17] 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 Transactions on Medical Imaging **32**, 1965–1978 (2013)

[18] Y. Nesterov, "A method of solving a convex programming problem with convergence rate O (1/k2)," in *Soviet Mathematics Doklady*, Vol. 27 (1983) pp. 372–376

[19] D. Kim, S. Ramani, and J. A. Fessler, "Combining ordered subsets and momentum for accelerated X-ray CT image reconstruction," IEEE Transactions on Medical Imaging **34**, 167–178 (2015)

[20] W. Segars, M. Mahesh, T. Beck, E. Frey, and B. Tsui, "Realistic CT simulation using the 4D XCAT phantom," Medical Physics **35**, 3800–3808 (2008)

[21] K. Kim, D. Wu, K. Gong, J. Dutta, J. H. Kim, Y. D. Son, H. K. Kim, G. E. Fakhri, and Q. Li, "Penalized PET Reconstruction Using Deep Learning Prior and Local Linear Fitting," IEEE Transactions on Medical Imaging **37**, 1478–1487 (2018)

[22] J. L. Humm, A. Rosenfeld, and A. Del Guerra, "From PET detectors to PET scanners," European Journal of Nuclear Medicine and Molecular Imaging **30**, 1574–1597 (2003)

[23] S. Vandenberghe and J. Karp, "Rebinning and reconstruction techniques for 3D TOF-PET," Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **569**, 421–424 (2006)

[24] J. Qi, R. M. Leahy, S. R. Cherry, A. Chatziioannou, and T. H. Farquhar, "High-resolution 3D Bayesian image reconstruction using the microPET small-animal scanner," Physics in Medicine and Biology **43**, 1001–1013 (1998)

[25] J. A. Fessler and H. Erdogan, "A paraboloidal surrogates algorithm for convergent penalized-likelihood emission image reconstruction," in *Nuclear Science Symposium*, Vol. 2 (IEEE, 1998) pp. 1132–1135

[26] H. Erdogan and J. A. Fessler, "Ordered subsets algorithms for transmission tomography," Physics in Medicine and Biology **44**, 2835–2851 (1999)

[27] D. Kim and J. A. Fessler, "Parallelizable algorithms for X-ray CT image reconstruction with spatially non-uniform updates," The Second International Conference on Image Formation in X-ray Computed Tomography , 33–36 (2012)

[28] A. R. De Pierro, "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography," IEEE Transactions on Medical Imaging **14**, 132–137 (1994)

[29] G. Pratx, G. Chinn, F. Habte, P. Olcott, and C. Levin, "Fully 3-D list-mode OSEM accelerated by graphics processing units," in *Nuclear Science Symposium and Medical Imaging Conference*, Vol. 4 (IEEE, 2006) pp. 2196–2202

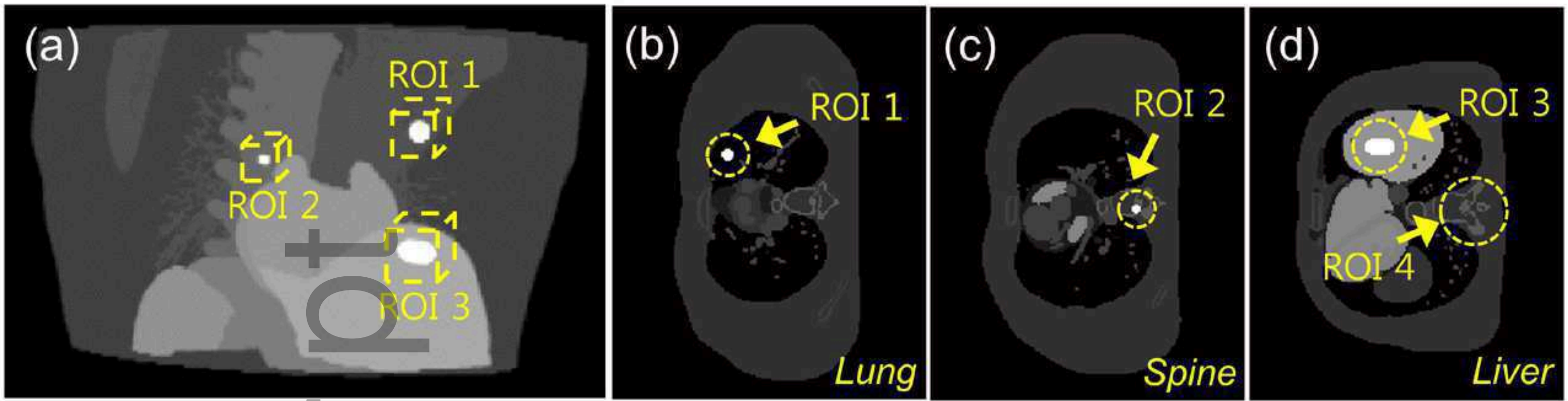
[30] B. Bai and A. Smith, "Fast 3D iterative reconstruction of PET images using PC graphics hardware," in *IEEE Nuclear Science Symposium and Medical Imaging Conference*, Vol. 5 (IEEE, 2006) pp. 2787–2790

[31] K. Kim and J. Ye, "Fully 3D iterative scatter-corrected OSEM for HRRT PET using a GPU," Physics in Medicine and Biology **56**, 4991–5009 (2011)

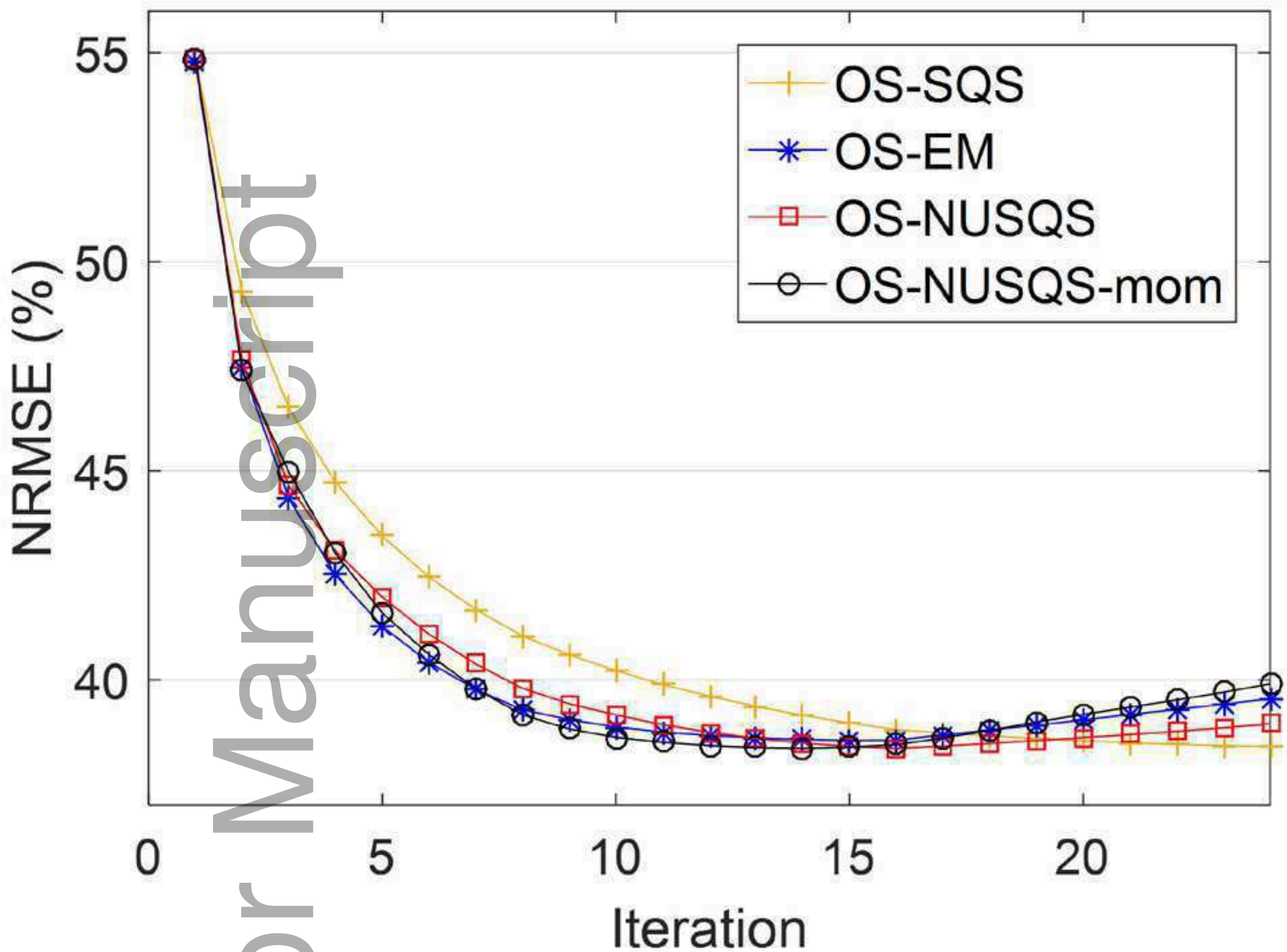
[32] K. Kim, Y. D. Son, Y. Bresler, Z. H. Cho, J. B. Ra, and J. C. Ye, "Dynamic PET reconstruction using temporal patch-based low rank penalty for ROI-based brain kinetic analysis," Physics in Medicine and Biology **60**, 2019–2046 (2015)

[33] J.-y. Cui, G. Pratx, S. Prevrhal, and C. S. Levin, "Fully 3D list-mode time-of-flight PET image reconstruction on GPUs using CUDA," Medical Physics **38**, 6775–6786 (2011)

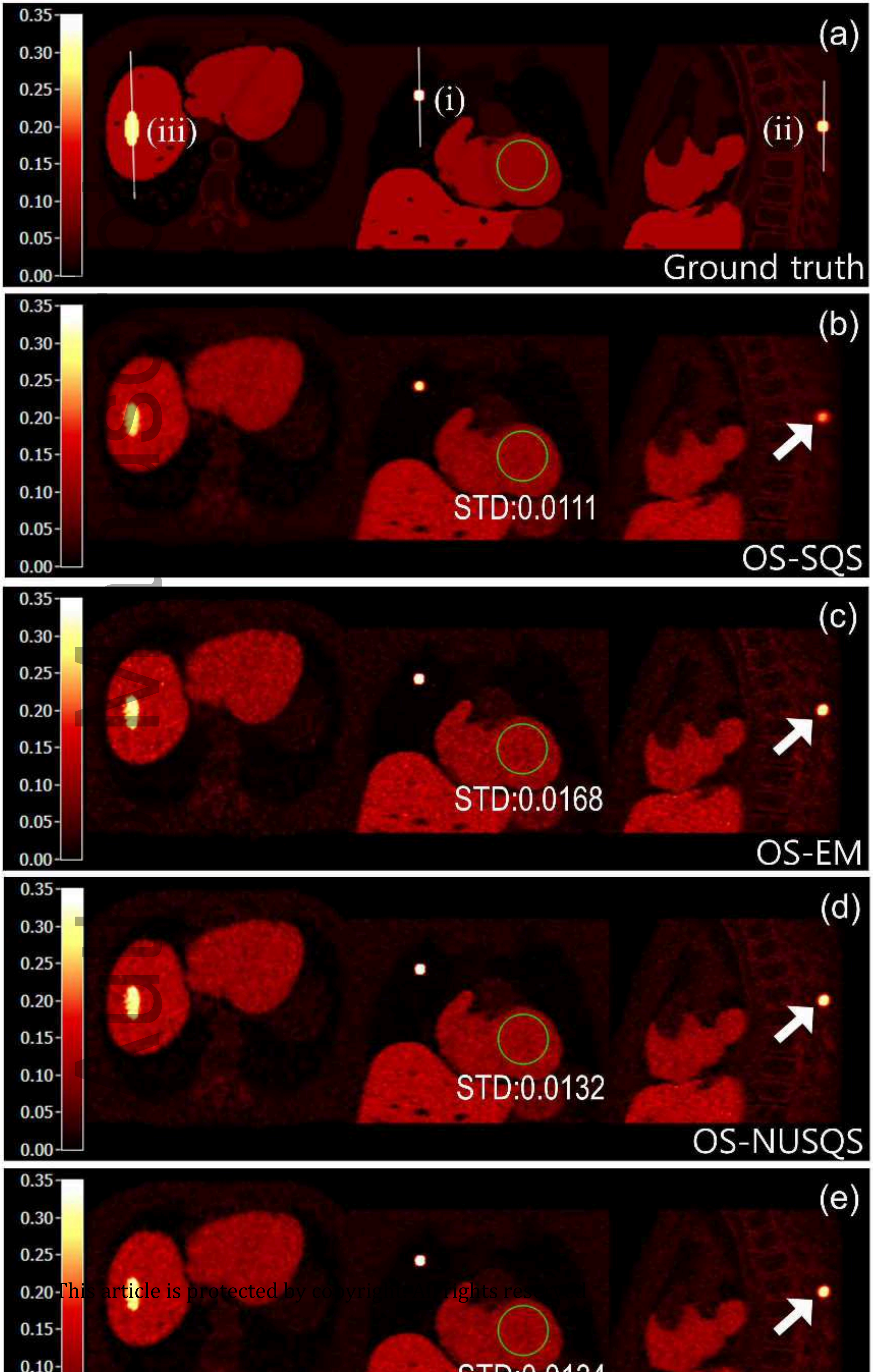
[34] Y. Berker, J. S. Karp, and V. Schulz, "Numerical algorithms for scatter-to-attenuation reconstruction in PET: Empirical comparison of convergence, acceleration, and the effect of subsets," IEEE Transactions on Radiation and Plasma Medical Sciences **1**, 426–434 (2017)

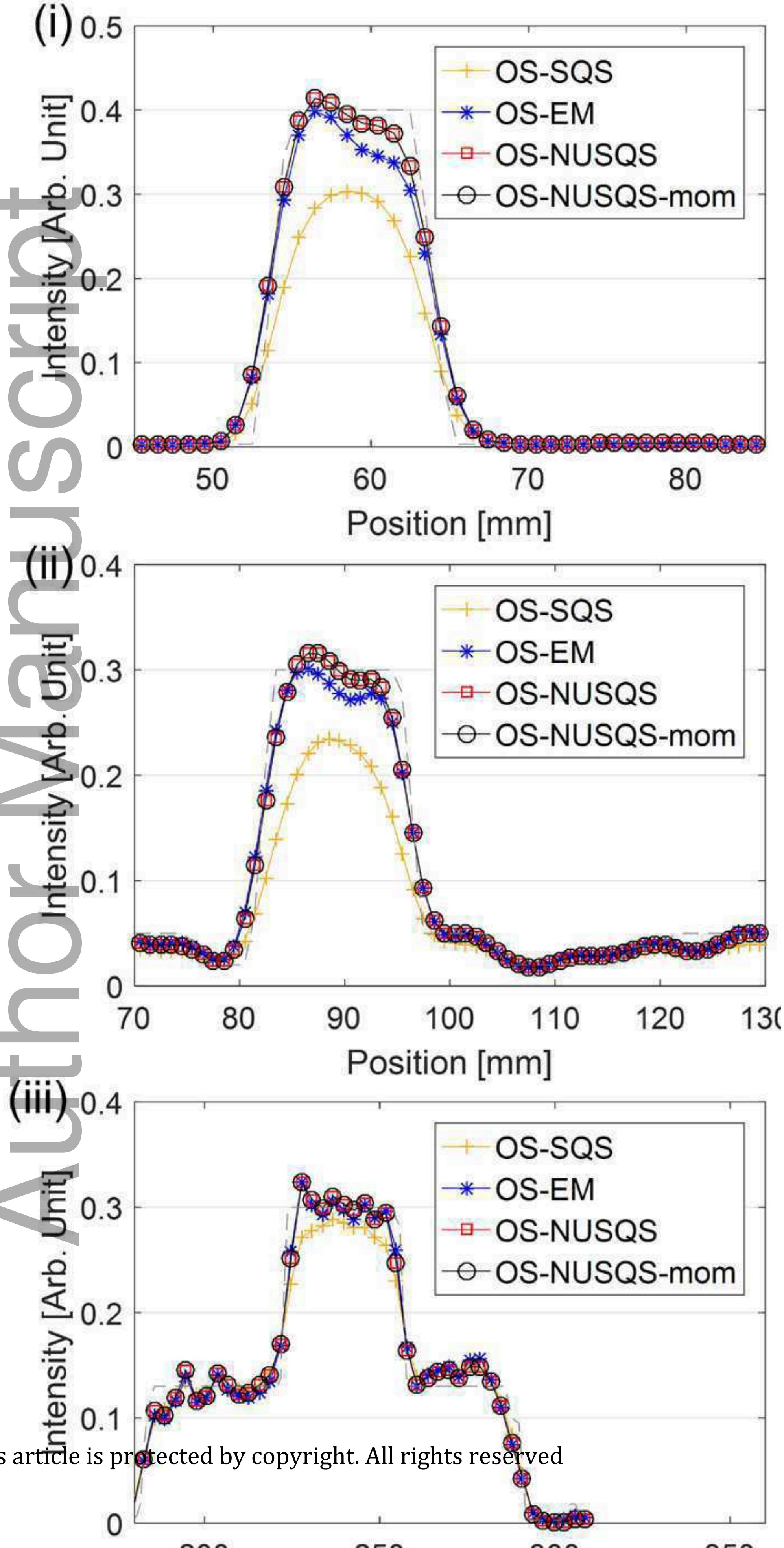


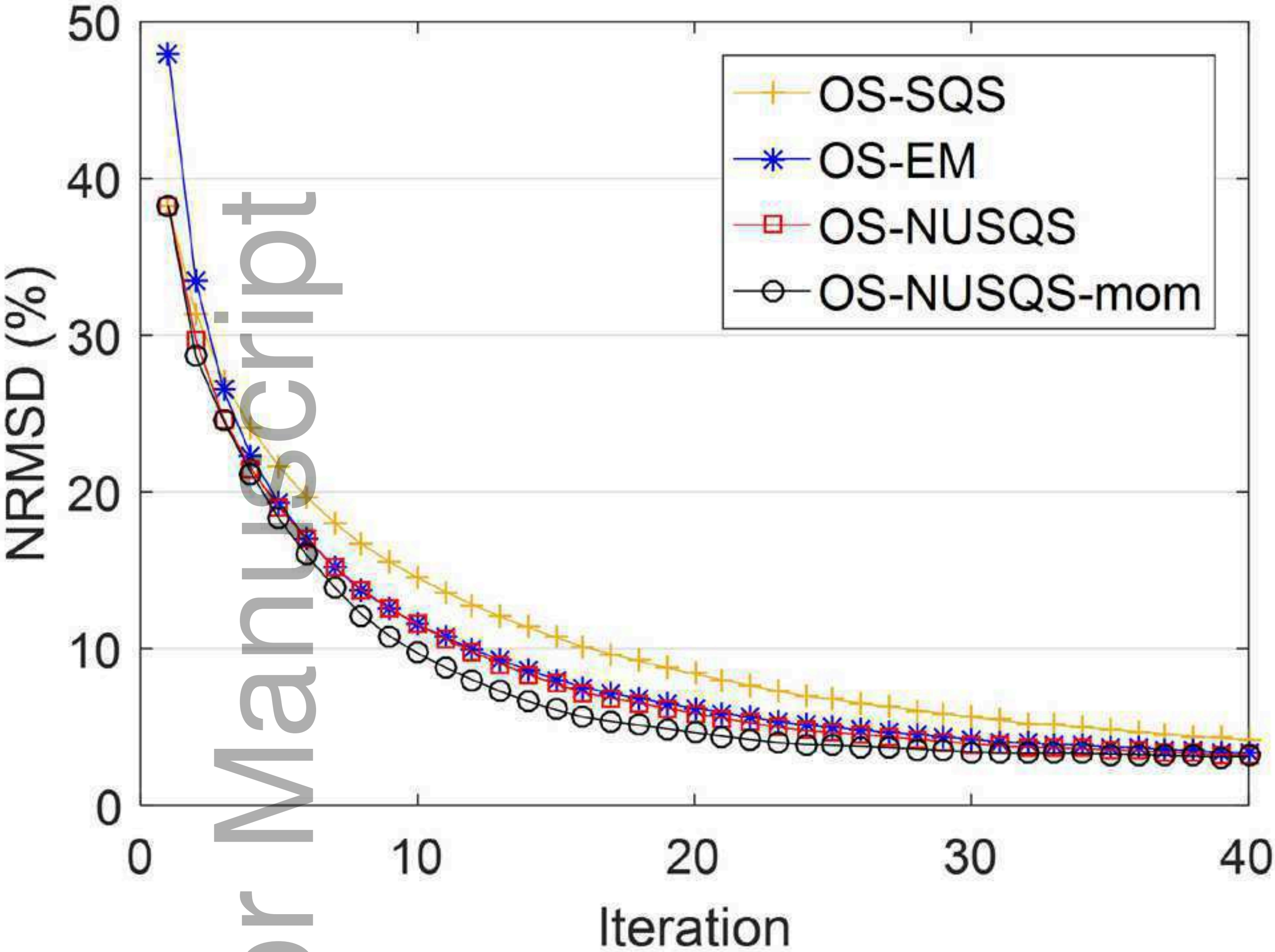
mp_13321_f1.eps



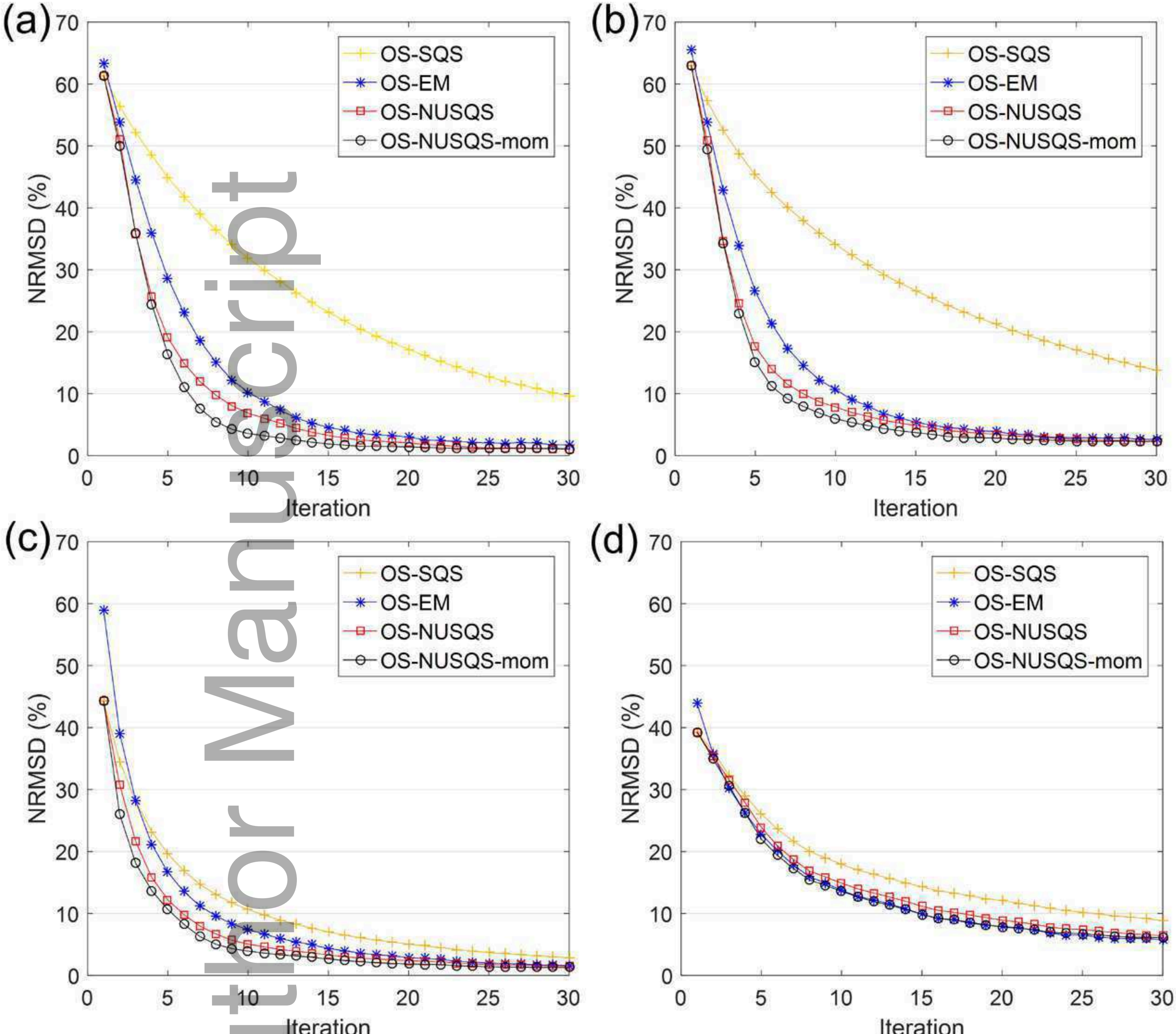
mp_13321_f2.eps


0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
(a)
(i)
(ii)
(iii)
Ground truth
(b)
STD:0.0111
OS-SQS
(c)
STD:0.0168
OS-EM
(d)
STD:0.0132
OS-NUSQS
(e)

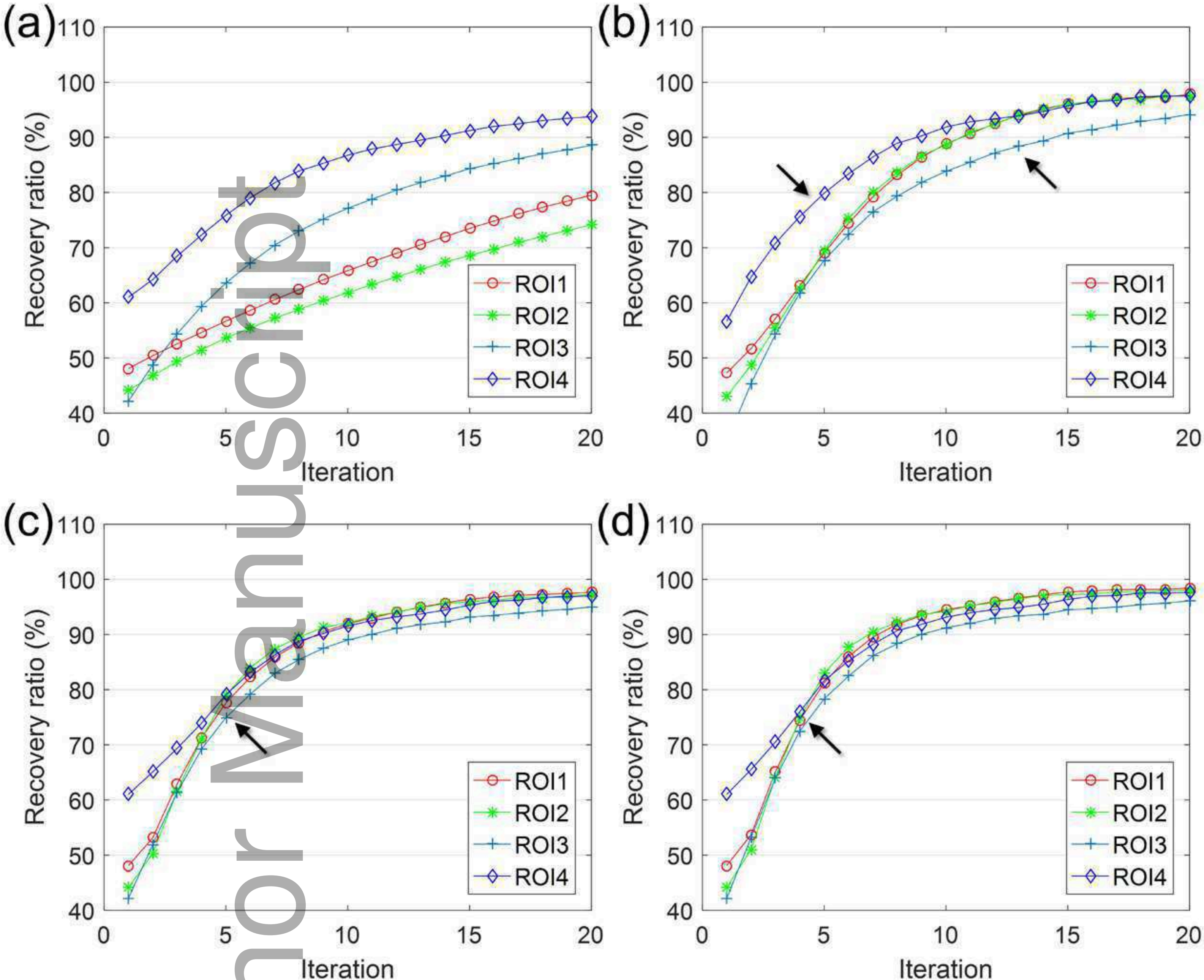

(i)
Intensity [Arb. Unit]
Position [mm]
OS-SQS
OS-EM
OS-NUSQS
OS-NUSQS-mom
(ii)
Intensity [Arb. Unit]
Position [mm]
OS-SQS
OS-EM
OS-NUSQS
OS-NUSQS-mom
(iii)
Intensity [Arb. Unit]
OS-SQS
OS-EM
OS-NUSQS
OS-NUSQS-mom



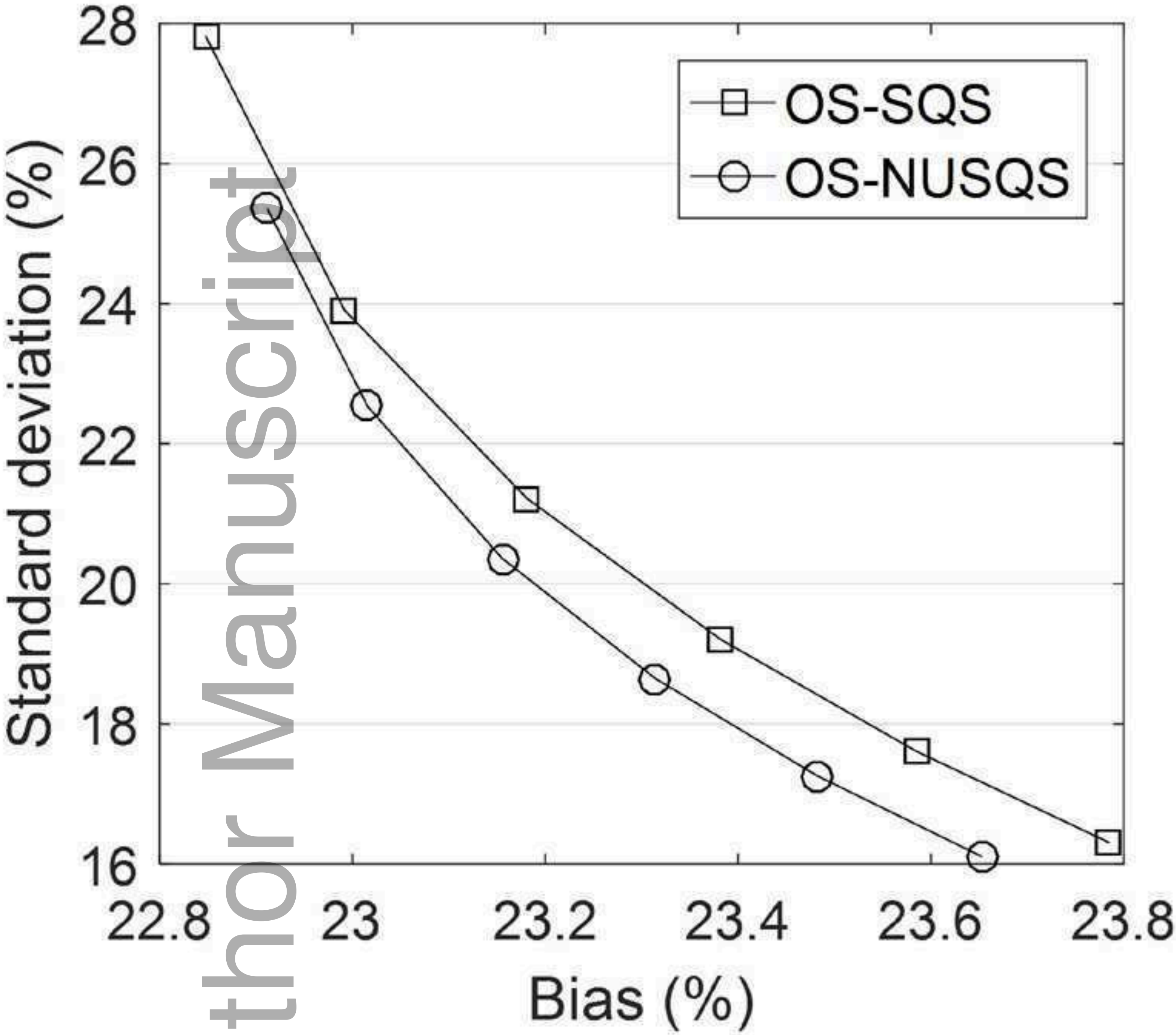
mp_13321_f5.eps



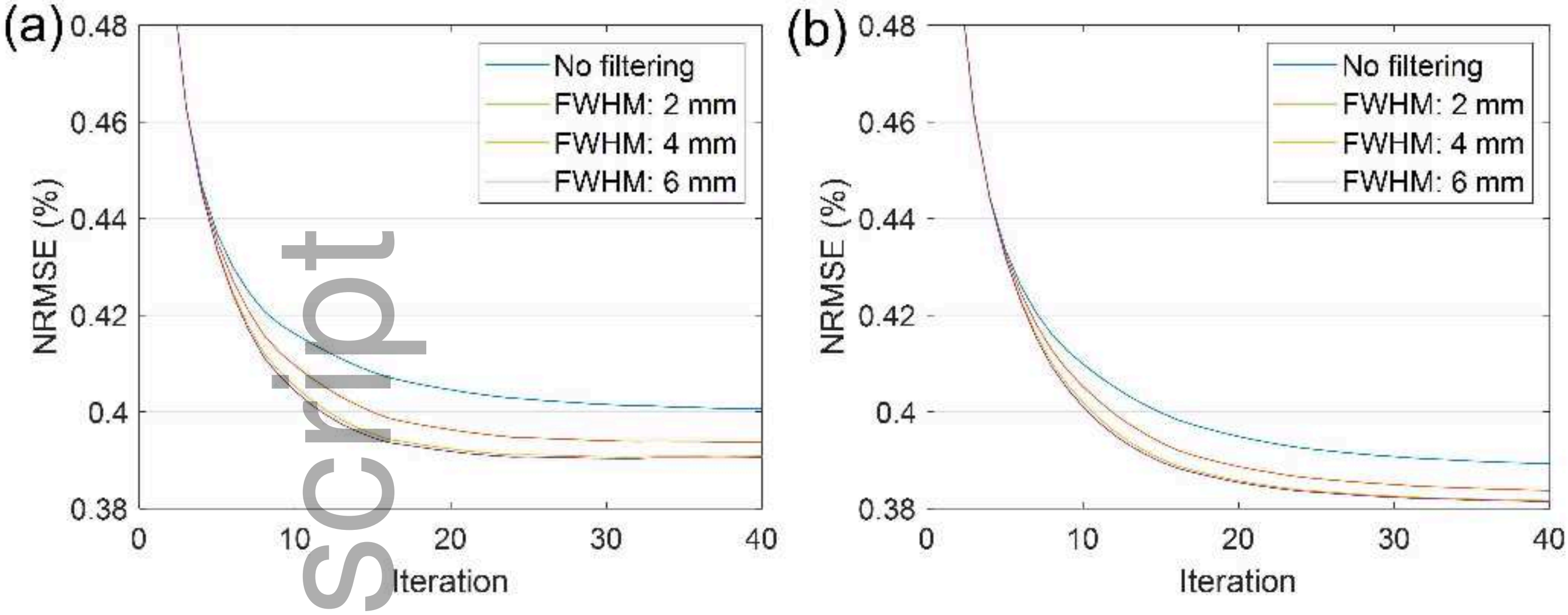
mp_13321_f6.eps



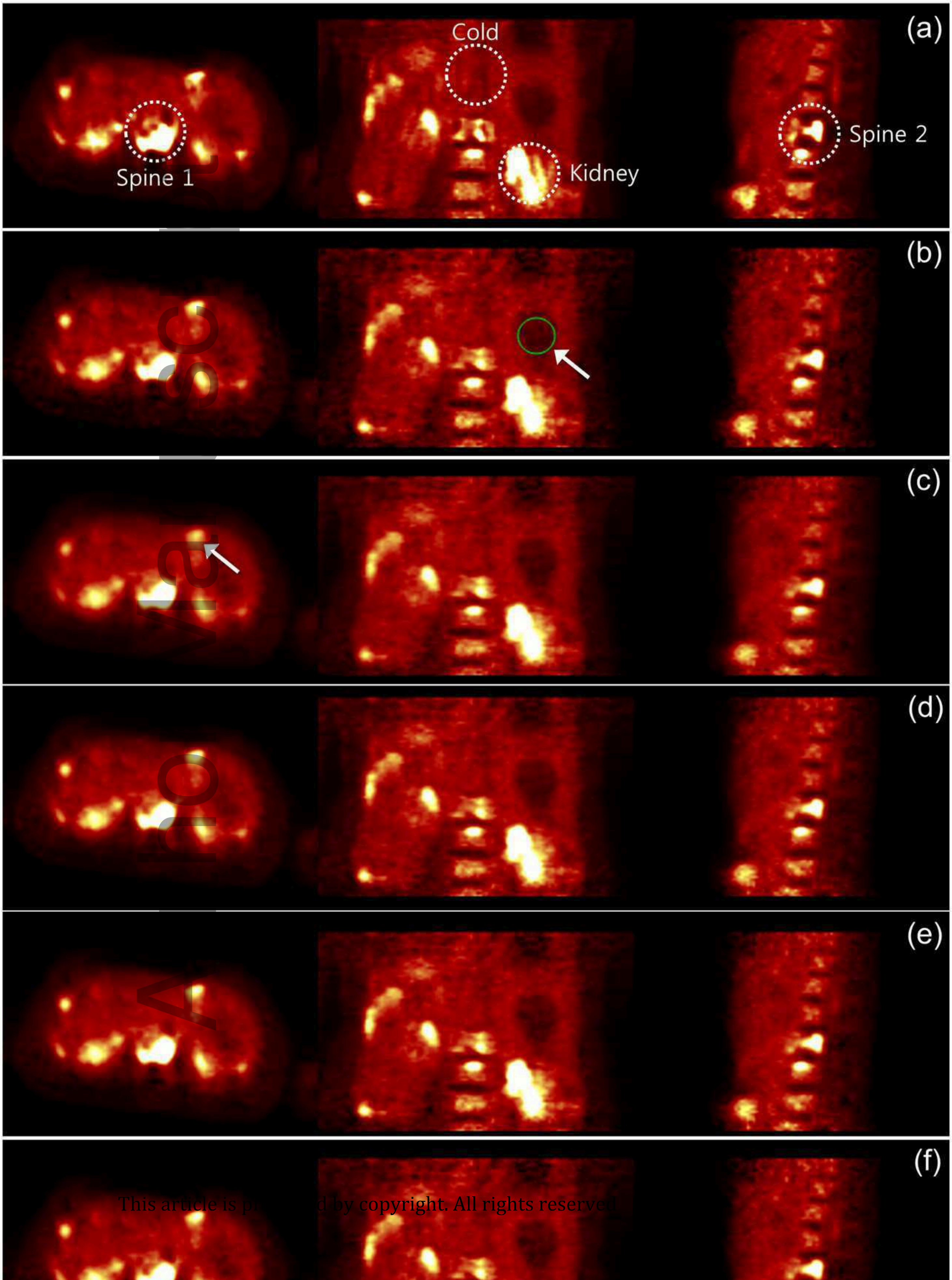
mp_13321_f7.eps

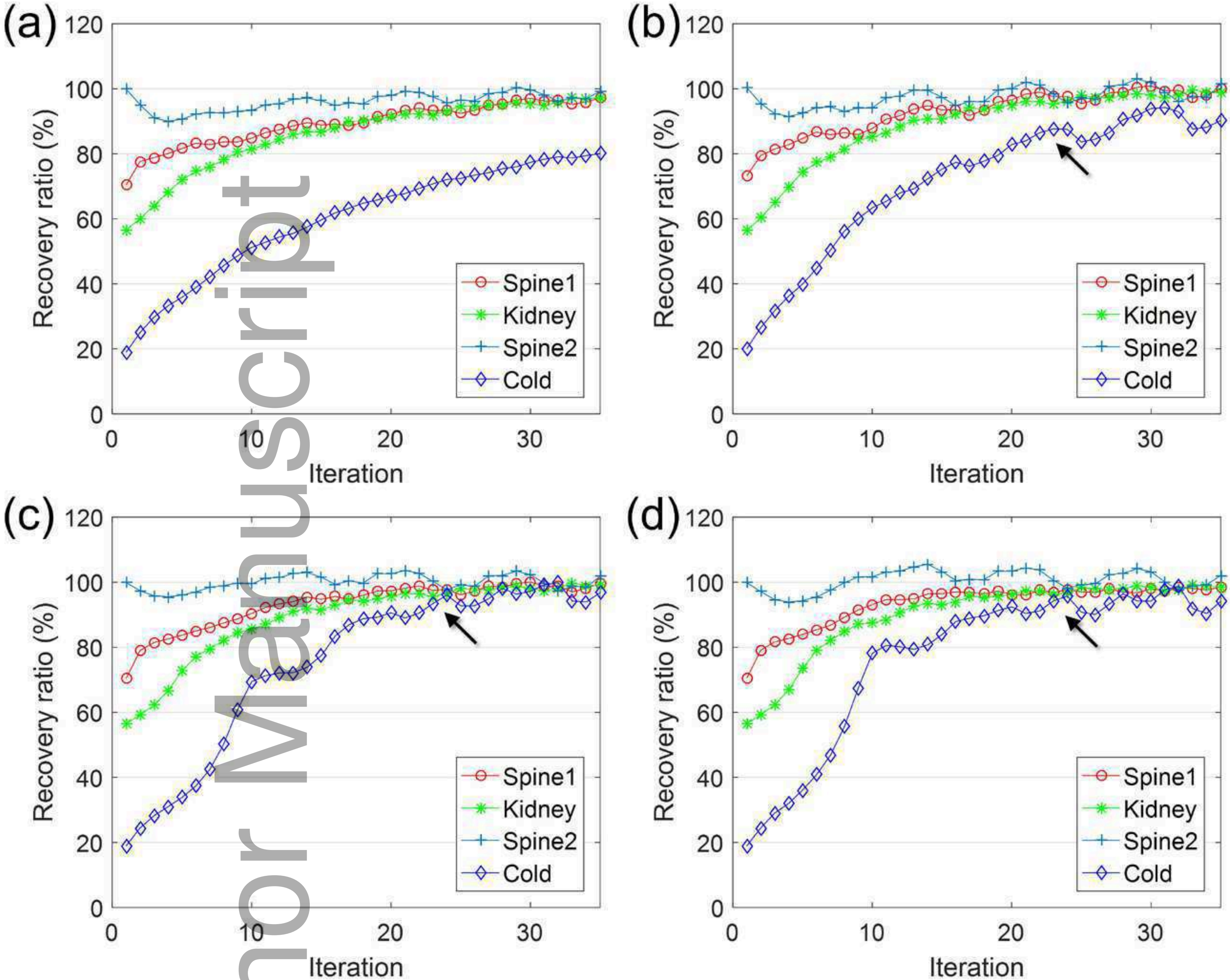


mp_13321_f8.eps

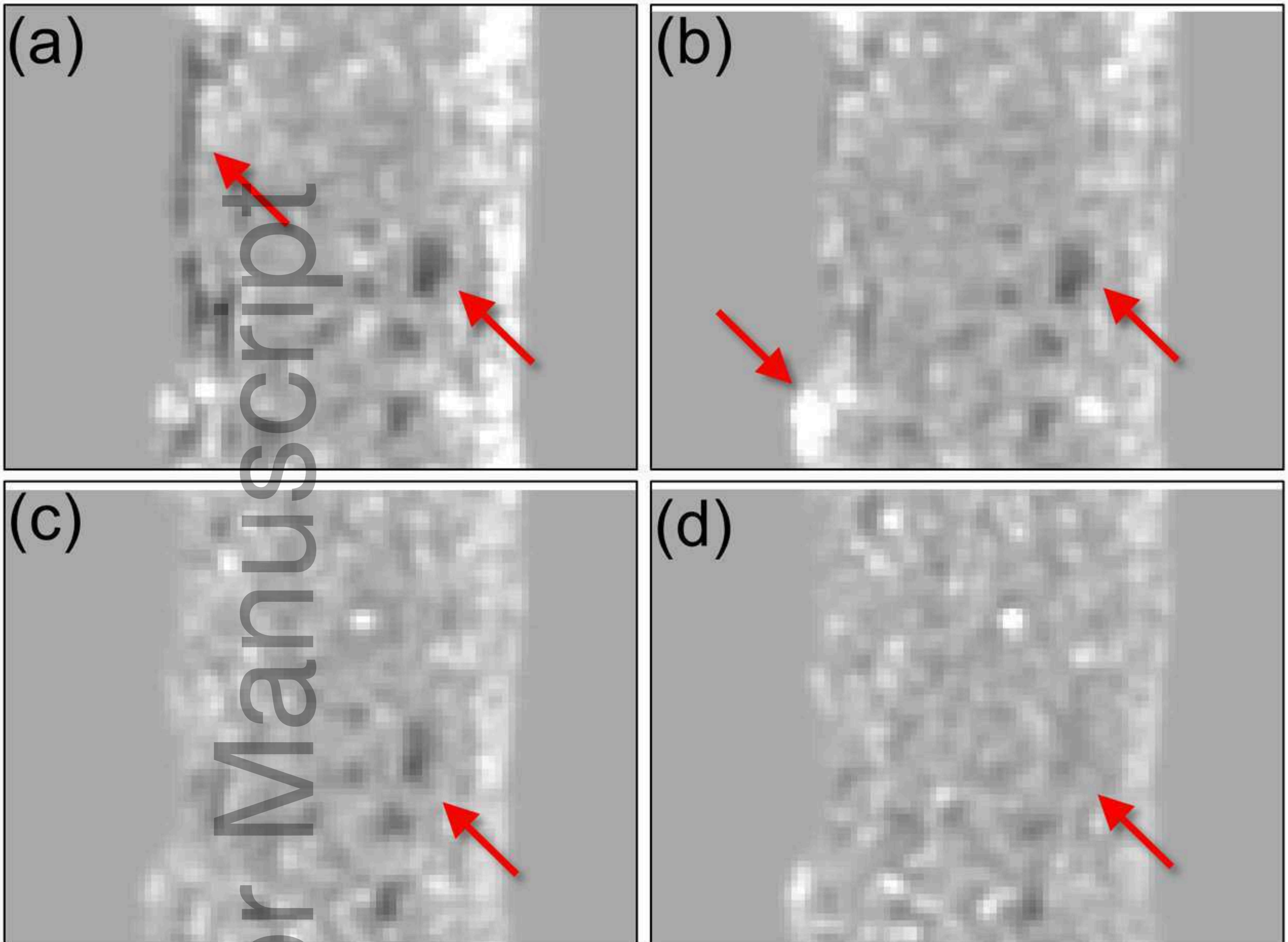


mp_13321_f9.eps

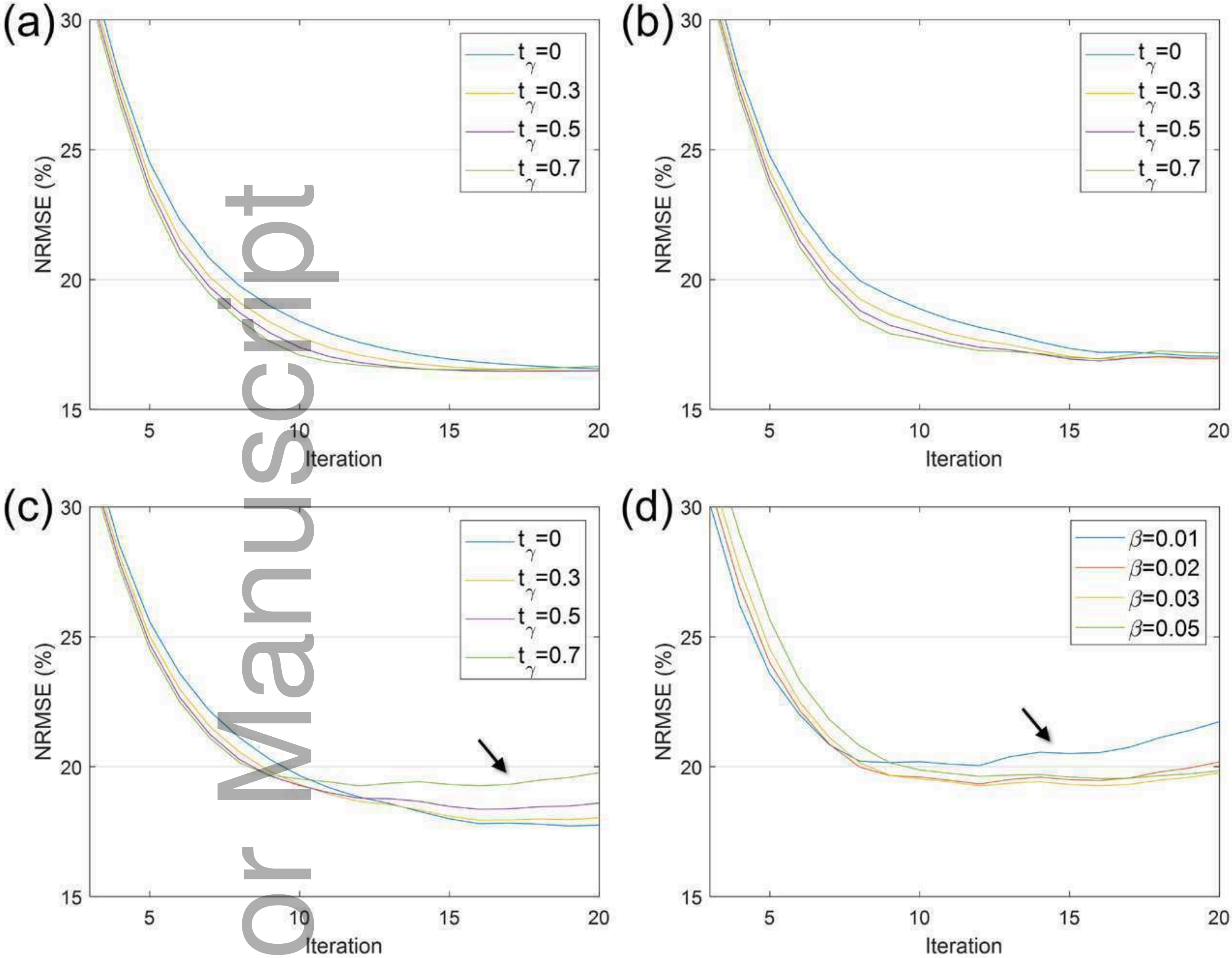

(a)
Cold
Spine 1
Kidney
Spine 2
(b)
(c)
(d)
(e)
(f)



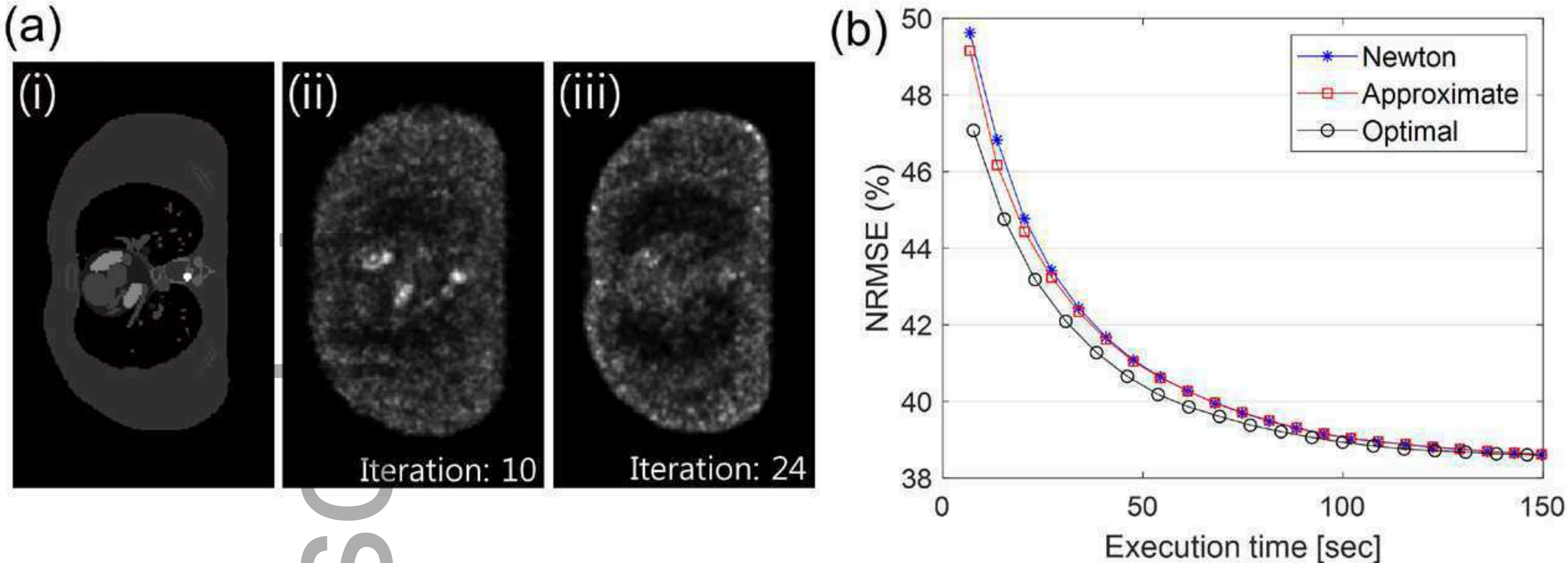
mp_13321_f11.eps



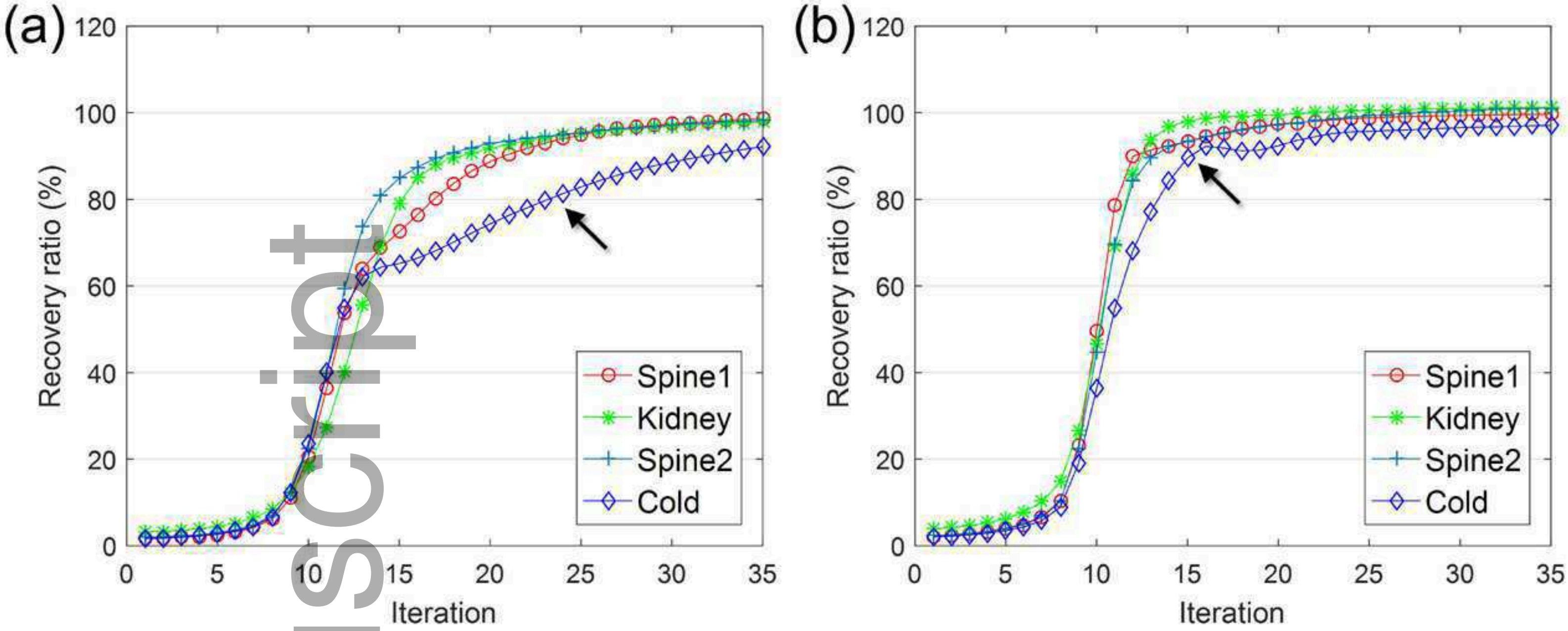
mp_13321_f12.eps



mp_13321_f13.eps



mp_13321_f14.eps



mp_13321_f15.eps