Journal of Biomedical Optics

BiomedicalOptics.SPIEDigitalLibrary.org

Fast direct reconstruction strategy of dynamic fluorescence molecular tomography using graphics processing units

Maomao Chen
Jiulou Zhang
Chuangjian Cai
Yang Gao
Jianwen Luo

SPIE.

Maomao Chen, Jiulou Zhang, Chuangjian Cai, Yang Gao, Jianwen Luo, "Fast direct reconstruction strategy of dynamic fluorescence molecular tomography using graphics processing units," *J. Biomed. Opt.* **21**(6), 066010 (2016), doi: 10.1117/1.JBO.21.6.066010.

Fast direct reconstruction strategy of dynamic fluorescence molecular tomography using graphics processing units

Maomao Chen, Jiulou Zhang, Chuangjian Cai, Yang Gao, and Jianwen Luo*
Tsinghua University, School of Medicine, Department of Biomedical Engineering, Beijing 100084, China

Abstract. Dynamic fluorescence molecular tomography (DFMT) is a valuable method to evaluate the metabolic process of contrast agents in different organs *in vivo*, and direct reconstruction methods can improve the temporal resolution of DFMT. However, challenges still remain due to the large time consumption of the direct reconstruction methods. An acceleration strategy using graphics processing units (GPU) is presented. The procedure of conjugate gradient optimization in the direct reconstruction method is programmed using the compute unified device architecture and then accelerated on GPU. Numerical simulations and *in vivo* experiments are performed to validate the feasibility of the strategy. The results demonstrate that, compared with the traditional method, the proposed strategy can reduce the time consumption by ~90% without a degradation of quality. [DOI: 10.1117/1.JBO.21.6.066010]

Keywords: image reconstruction; dynamic fluorescence molecular tomography; graphics processing units; compute unified device architecture.

Paper 160169R received Mar. 17, 2016; accepted for publication May 23, 2016; published online Jun. 14, 2016.

1 Introduction

As a rapidly developing technique for pharmacokinetics studies,[1,2] dynamic fluorescence imaging reflects the absorption, distribution, and excretion characteristics of contrast agents in the body. The reconstructed pharmacokinetic images of dynamic fluorescence molecular tomography (DFMT) can provide noninvasive and three-dimensional (3-D) monitoring of the metabolic process of fluorophores *in vivo*.[3–5] Therefore, it is promising for tumor detection, drug development, and metabolic research.[6,7]

Generally, two kinds of methods can be used to solve DFMT problems: indirect methods and direct methods. Indirect methods[3] have been developed from the conventional static fluorescence molecular tomography (FMT) procedures. It is assumed that the concentration of the fluorophores is constant during the data acquisition process of each circle. Therefore, the individual FMT image of each circle is reconstructed first, and then the metabolic parameters of each voxel can be obtained with curve fitting. Some methods have been previously proposed for FMT reconstruction. L2 regularization[8] is commonly implemented because it is simple and can be efficiently solved, while L1-based regularization[9,10] can improve the resolution of the reconstruction images especially for spars problems. However, the spatial resolution and accuracy of the images obtained with the indirect methods are relatively low because the static assumption is not suitable for the DFMT problems, especially when the concentration of the fluorophores varies fast over time. To overcome this problem, direct reconstruction methods[4,5] have been proposed to map the metabolic parameters into the acquired datasets and directly reconstruct the dynamic images from the boundary measurements in one step. By using the direct methods, the quality of the reconstructed images is improved. However, the time consumption of the direct methods is significantly increased. In a typical DFMT implementation with about 10,000 boundary measurement points, 5000 finite-element nodes, and 200 iterations, the direct methods take more than 8 h to obtain the reconstructed images.

Some methods can be used to accelerate the direct methods. Principal components analysis (PCA) has been proposed to reduce the dimension of weight matrix.[11] Wavelet[12,13] and Fourier[14] transforms have been employed to reduce the scale of the boundary measurements. However, these methods may reduce the accuracy of the reconstructed images due to the information lost by using data compression.

Parallel computation using graphics processing units (GPU) is a fast-developing technology. The computing ability of GPU is much more powerful than that of general-purpose central processing units (CPU). Combined with deeply optimized processing pipelines, hierarchical thread structures and extremely low memory latency, GPU constitutes an excellent shared memory parallel computing platform.[15] However, programming on GPU had been difficult until the compute unified device architecture (CUDA) was introduced in 2006. CUDA is a parallel computing platform and programming model and provides a software environment where developers can use a high-level programming language, such as C. CUDA-enabled GPU has been utilized in the field of fluorescence imaging, such as the acceleration of Monte Carlo algorithm,[15] modeling of time-resolved photon migration,[16] and acceleration of early-photon FMT.[17]

In this work, a fast computing strategy using GPU is proposed to accelerate the direct reconstruction algorithm of DFMT. In this strategy, the most time-consuming part of the

*Address all correspondence to: Jianwen Luo, E-mail: luo_jianwen@tsinghua.edu.cn

1083-3668/2016/$25.00 © 2016 SPIE

algorithm, i.e., iterative regularization using conjugate gradient (CG) optimization, is accelerated using CUDA. Meanwhile, to fully use the parallel computing ability of GPU, vectors of fluorophore concentration are assembled into a distribution matrix according to the order of the projection sequences in each circle. Then the forward prediction measurements are calculated in the form of matrix–matrix multiplication.

Numerical simulations and *in vivo* mouse experiments are carried out to evaluate the performance of the proposed strategy. The results demonstrate that the proposed strategy can significantly accelerate the direct DFMT algorithm almost without any quality degradation.

2 Method

2.1 *Data Acquisition System*

A hybrid FMT/x-ray computed tomography (XCT) system[18] is used to obtain the DFMT boundary measurements. As shown in Fig. 1, a free-space and full-angle FMT system is used to acquire the fluorescence datasets, while an XCT system is used to obtain the anatomical information of the small animal, which provides structural priors for the reconstruction algorithm. For dynamic FMT problems, to monitor the metabolic process of the fluorophores, the small animal is fixed on the stage and continuously rotated for K circles, and the charge-coupled device (CCD) camera acquires S projections in each circle. Therefore, a total of $P = KS$ projections are obtained during the whole DFMT acquisition process.

2.2 *Direct Reconstruction Method for Dynamic Fluorescence Molecular Tomography Problems*

For FMT problems, the propagation of the excitation and emission light in biological tissues can be reasonably approximated by two coupled diffusion equations.[19,20] Using Green's function theory, the fluorescence signal $\Phi_m(r_d, r_s)$ detected at a point r_d due to a excitation source at r_s can be written as[21]

$$\Phi_m(r_d, r_s) = \Theta \int G_m(r_d, r) n(r) G_x(r, r_s) \mathrm{d}^3 r, \tag{1}$$

where the Green's function $G_x(r, r_s)$ stands for the light propagation from the source point r_s to an arbitrary position r inside the medium at the excitation wavelength x. The Green's function $G_m(r_d, r)$ describes the light propagation from a position r inside the medium to the detector point r_d at the emission wavelength m, and $n(r)$ denotes the fluorescent yield to be reconstructed.

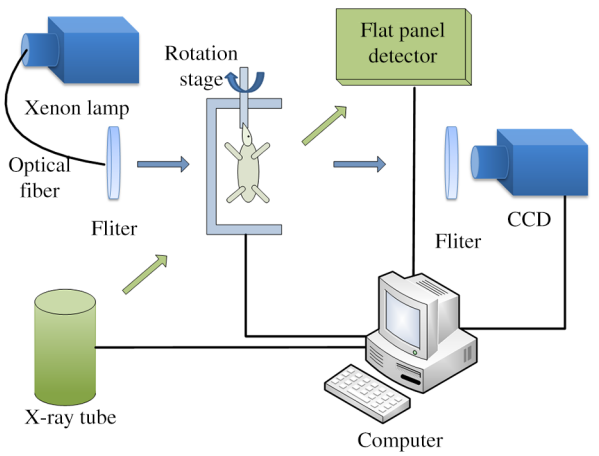

Fig. 1 Schematic of the hybrid FMT/XCT system. The FMT system is used to acquire the DFMT measurements, while the XCT system obtains the anatomical information of the small animal.

In DFMT problems, a two-compartment model is commonly used to describe the metabolic process of the fluorophores inside different organs.[22] The concentration of fluorophores $n(r, t)$ at time t can be obtained using[23]

$$n(r, t) = -A(r) \exp[-\alpha(r)t] + B(r) \exp[-\beta(r)t], \tag{2}$$

where $A(r)$ and $B(r)$ determine the zero-time concentration at position r, $\alpha(r)$, and $\beta(r)$ denote the uptake and excretion rates of the fluorophores.[23]

When the imaged 3-D volume is discretized into N voxels, the parametric images in the discrete domain can be defined as

$$X = [x_1, x_2, x_3, x_4] = [A, B, \alpha, \beta], \tag{3}$$

where each parametric image $x_u (u = 1,2,3,4)$ is given by $x_u = [x_u(r_1), \ldots, x_u(r_N)]^T$, and $r_j (j = 1, \ldots, N)$ is employed to denote the spatial locations of voxels in the discretized domain.

As demonstrated in Sec. 2.1, the measurement acquisition consists of K circles, and S projections are performed in each circle. For projection $s (s = 1, \ldots, s)$, the surface of the imaged object is orthographically projected to the fluorescence image, which is obtained from CCD camera. The CCD pixels inside the projection area are considered to be the measurement points, and the number of the measurement points is M_s. Thus, a total of $P = KS$ projections and $M = \sum_{s=1}^{s} M_s$ measurement points are obtained. Let $t_p = (p = 1, \ldots, P)$ denote the individual discrete time of projections. By combining Eqs. (1) and (2), the pharmacokinetic parameters can be mapped directly to the boundary measurements as follows:[4]

$$\begin{aligned} \Phi_m(r_{d_i}, KS_s, t_p) &= \sum_{j=1}^{N} W_s(i, j) n(r_j, t_p) \\ &= \sum_{j=1}^{N} W_s(i, j) \{-A(r_j) \exp[-\alpha(r_j) t_p] \\ &\quad + B(r_j) \exp[-\beta(r_j) t_p]\}, \end{aligned} \tag{4}$$

where W_s is the submatrix of the weight matrix W at projection $s (s = 1, \ldots, S)$, and its entries are defined as[24]

$$W_s(i, j) = \Delta V \Theta G_m(r_{d_i}, r_j) G_x(r_j, KS_s), \tag{5}$$

where is the volume of each individual voxel.

Let $f(X)$ denote the forward model. The boundary measurements predicted by the forward model is given as

$$f(X, t_p) = [\Phi_m(r_{d_i}, KS_s, t_p), \ldots, \Phi_m(r_{d_{M_s}}, KS_s, t_p)], \tag{6}$$

$$f(X) = [f(X, t_1)^T, \ldots, f(X, t_P)^T]^T. \tag{7}$$

To reconstruct the parametric images in one step, by combining Eq. (7) and the conventional Tikhonov regularization method, a new objective function is obtained for the direct reconstruction method[5]

$$\Psi(X) = \|y - f(X)\|_2^2 + \lambda_1\|LA\|_2^2 + \lambda_2\|LB\|_2^2 + \lambda_3\|L\alpha\|_2^2 + \lambda_4\|L\beta\|_2^2 = \|y - f(X)\|_2^2 + \sum_{u=1}^{4}\|Lx_u\|_2^2, \quad (8)$$

where $y = [y(t_1)^T, y(t_2)^T, \ldots, y(t_p)^T]^T$ denotes the whole boundary measurement matrix assembled using measurement vectors of all the frames. $\lambda_u (u = 1,2,3,4)$ is the regularization parameter, and L is the Laplacian-type regularization matrix[25] constructed using the structural priors obtained with the XCT system. The minimization of $\Psi(X)$ can be solved using a CG scheme.[26]

2.3 *Proposed Graphics Processing Units Acceleration Strategy*

2.3.1 *Flow chart of the acceleration strategy*

The flow chart of the direct reconstruction method and the acceleration strategy are shown in Fig. 2. The entire procedure of the algorithm consists mainly of three parts, i.e., the construction of Laplacian-type regularization matrix[25] (part 1 in Fig. 2), the low dimensional estimation of the initial of parameters[27] (part 2 in Fig. 2), and the CG optimization (part 3 in Fig. 2). The most important and time-consuming part is the CG optimization. In the CG optimization, four dynamic parameters (A, B, α, and β) are iteratively optimized until the end conditions are satisfied. The iteration number of the CG optimization is usually set to be more than 100, and the iterations take more than 99% of the whole reconstruction time. Therefore, the CG optimization needs to be accelerated.

As shown in Fig. 2, the whole procedure of the optimization is programmed using CUDA, while the rest parts of the algorithm are programmed in MATLAB® 2012 (The MathWorks, Inc., Natick, Massachusetts). For each metabolic parameter, the gradient $g[\Psi(x)]$ and the conjugate searching direction $d[\psi(X)]$ of the objective function $\psi(X)$ as shown in Eq. (8) are first calculated, and then the optimal step length is determined using golden search. These steps are all parallel computed using GPU. All the datasets and parameters are transferred to and stored in GPU memories before the GC scheme is started, and the reconstructed results are read out after it is finished. The whole optimization procedure is independently and automatically operated in GPU. This design can avoid frequent data transfer between CPU and GPU, which is significantly time-consuming especially when the size of the dataset is large.

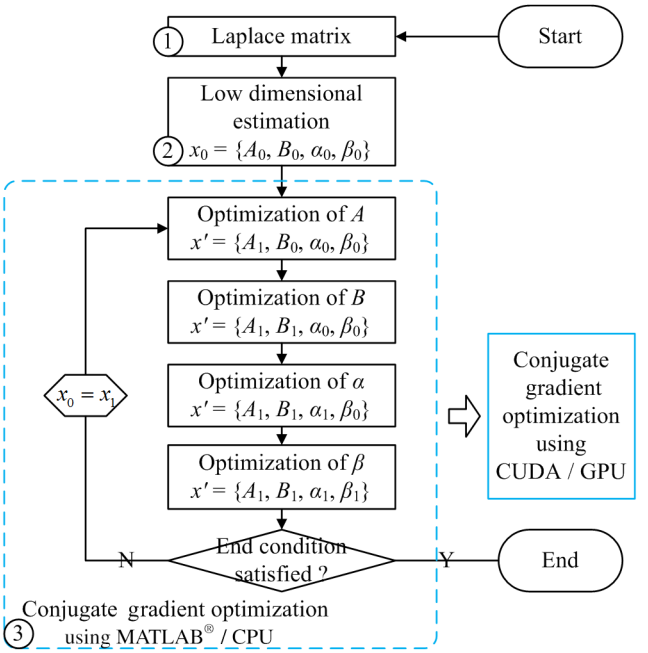

Fig. 2 Flowchart of the acceleration strategy. The whole procedure of CG optimization (part 3) is programmed and accelerated using CUDA.

2.3.2 *Parallel computing strategy*

CUDA is a parallel computing platform and programming model invented by NVIDIA Corp. (Santa Clara, California). It is capable of sending C codes straight to GPU without complicated manipulation of the underlying GPU hardware. When using CUDA, the parallel computing functions are programmed as kernel functions, which could be operated on a GPU. However, it requires a lot of practice and experience to develop high-effective kernel functions. To solve this problem, NVIDIA provides a cuBLAS library, which is a high-effective implementation of basic linear algebra subprograms (BLAS).

In this study, the GPU method is programmed using CUDA C, and compiled using nvcc and mex compilers. The generated mex file can be invoked by MATLAB® program. Some basic linear algebra calculations, such as matrix–matrix multiplication/addition, matrix–vector multiplication, and vector–vector dot product, are programmed using cuBLAS libraries. Take matrix–matrix multiplication for example, a parallelly tiled multiplication method is used in cuBLAS library. The diagram of the tiled multiplication is shown in Fig. 3(a). The matrixes A, B and $C = A \times B$ are divided into several submatrixes. When calculating the submatrix $C_{sub} = A_{sub} \times B_{sub}$, the submatrixes A_{sub} and B_{sub} are synchronously read into the shared memories, and the elements are multiplied and accumulated parallelly in GPU threads. The block size of the submatrixes is selected based on the number of the GPU cores and the size of the GPU shared memories. The block size is set to be 16 in this study, and the number of elements for each submatrix is $16 \times 16 = 256$.

For the functions that the cuBLAS cannot support, self-defined kernel functions are used to realize parallel computing. For example, when calculating the exponential values of a

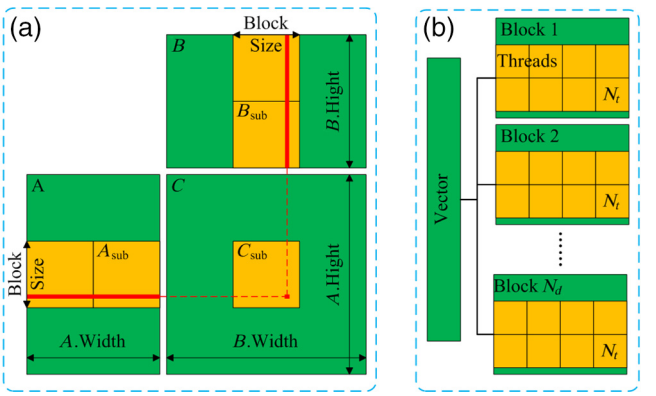

Fig. 3 Diagram of the parallel computing strategy. (a) Diagram of the tiled multiplication method for matrix–matrix multiplication. The matrixes are divided into submatrixes, and the elements of the submatrixes are parallelly calculated. (b) Diagram of the parallel computing for the exponential value of the vector. The vector is divided into N_d blocks, and each block contains N_t threads.

vector, the vector of size N_v is divided into N_d blocks, and each block contains N_t GPU threads, as shown in Fig. 3(b). In this study, the threads number of each block is 512. Therefore, the vector is divided into $N_d = N_v/512$ blocks. For each block, 512 elements of the vector are read synchronously to the shared memory, and the exponential values of the elements are calculated parallelly in the GPU threads.

2.3.3 *Predicted boundary measurements acceleration*

The CG optimization is composed of gradient calculation and golden search. During the process of golden search, the predicted boundary measurements $f(X)$ in Eq. (7) are calculated repeatedly to minimize the objective function $\Psi(X)$ in Eq. (8). As shown in Fig. 3, K circles and S projections per circle are acquired in the measurements, $n_{s,k}(s = 1, \ldots, S; k = 1, \ldots, K)$ is the $N \times 1$ fluorophore concentration distribution vector at the s'th projection in the k'th circle, and W_s is the submatrix with a size of $M_s \times N$. Thus, one predicted boundary measurement $f(X)$ needs KS matrix–vector multiplications. In a typical implementation, calculation of $f(X)$ is repeated over 7000 times, which takes more than 70% of the total reconstruction time.

Note that in Fig. 4(a), the concentration distribution vectors at the s'th projection in different circles share the same submatrix W_s. To fully utilize the advantage of parallel computation of CUDA, the KS fluorophore vectors are assembled to S fluorophore matrixes with a size of $N \times K$ according to the sequence order of the submatrix W_s. Therefore, KS matrix–vector multiplications are transformed into S matrix–matrix multiplications as shown in Fig. 4(b).

2.4 *Comparison Principal Components Analysis Acceleration Method*

PCA is a widely used method to accelerate the reconstructions of FMT problems by reducing the dimensions of the measurement datasets and the weight matrixes.[11] It linearly transforms an original set of variables into a substantially smaller set of uncorrelated variables, which can represent most of the information in the original set of variables. The cumulative percent of variance (CPV) is used to determine the number of retained principal components.[11] An appropriate CPV should be carefully selected to balance the calculation speed and the reconstruction accuracy.

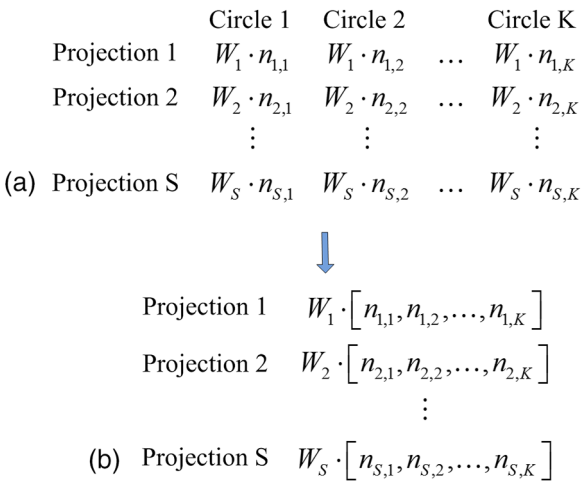

Fig. 4 Schematic of the predicted boundary measurements acceleration strategy. (a) The traditional method requires KS matrix–vector multiplications to obtain the predicted boundary measurements. (b) The acceleration strategy assembles the fluorophore vectors to S matrixes and requires only S matrix–matrix multiplications to obtain the predicted boundary measurements.

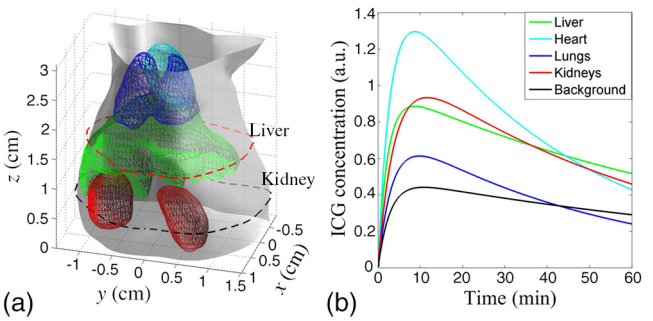

Fig. 5 Numerical simulations. (a) The 3-D Digimouse model used in the simulations. The mouse torso from the neck to the bottom of the kidneys is selected as the investigated region, with a total length of 3.1 cm. (b) ICG concentration curves simulating the metabolic process of ICG in different organs.

Table 1 Optical properties and pharmacokinetic parameters setups in different regions.

Regions	μ_a (cm^{-1})	μ_s' (cm^{-1})	A (a.u.)	B (a.u.)	α (min^{-1})	β (min^{-1})
Heart	0.350	23	1.7	1.7	0.330	0.023
Liver	0.500	13	1.0	1.0	0.435	0.011
Lungs	0.250	30	0.8	0.8	0.296	0.020
Kidneys	0.175	20	1.2	1.2	0.254	0.016
Background	0.300	10	0.5	0.5	0.348	0.009

Table 2 Original and PCA compressed data sizes for simulations.

	Voxels N	Measurements M		Compression ratio
		Original	PCA	
Group 1				
Case 1	5651	15,620	2229	7.0
Case 2	6968	15,620	1933	8.0
Case 3	9096	15,620	1440	10.8
Case 4	12,768	15,620	1030	15.2
Group 2				
Case 5	10,793	7730	696	11.1
Case 6	10,793	9443	803	11.8
Case 7	10,793	11,476	997	11.5
Case 8	10,793	12,861	1047	12.2
Average				10.9

For FMT problems, the optimal CPV can be reasonably set to be about 0.9 according to the previous study.[11] However, for DFMT problems, the optimal CPVs vary dramatically among different experimental cases. To obtain a proper balance between the time consumption and the reconstruction quality, the CPVs for the numerical simulations and the *in vivo* experiments are empirically selected to be 0.7 and 0.995, respectively.

Table 3 Time consumptions and acceleration ratios of the GPU and PCA methods for simulations.

	Time (s)			Acceleration ratio	
	Traditional	GPU	PCA	GPU	PCA
Group 1					
Case 1	22,135	1485	2861	14.9	7.7
Case 2	25,996	1867	3032	13.9	8.6
Case 3	33,225	2491	3453	13.3	9.6
Case 4	42,448	3646	4176	11.6	10.2
Group 2					
Case 5	18,135	1826	3393	9.9	5.3
Case 6	22,057	2060	3498	10.7	6.3
Case 7	26,557	2343	3654	11.3	7.3
Case 8	29,682	2566	3708	11.6	8.0
Average				12.2	7.9

2.5 *Evaluation Method*

The time consumptions of the traditional method, the proposed GPU method, and the PCA method are compared and the acceleration ratios are calculated. The relative difference (RD) is calculated to evaluate the differences between the results reconstructed by the acceleration methods and the traditional method, and it is defined as[11,13,17]

$$RD = sqrt(\|X_{\text{fast}} - X_{\text{traditional}}\|/X_{\text{traditional}}), \tag{9}$$

where $X_{\text{traditional}}$ is the entire 3-D parametric images reconstructed using the traditional method. X_{fast} denotes the 3-D reconstruction results obtained from the GPU- or PCA-based acceleration method.

The traditional method, the proposed GPU acceleration strategy, and the PCA acceleration method are all performed in MATLAB® 2012 (The MathWorks, Inc., Natick, Massachusetts) on a PC workstation with Intel® Core™ i7-4770 CPU at 3.40 GHz and 12 GB RAM. The PC workstation is embedded with an NVDIA GTX 750 Ti graphics card with 640 CUDA cores and 2 GB video memories.

3 Experiments and Results

3.1 *Numerical Simulations*

3.1.1 *Simulation setups*

Numerical simulations are performed to validate the performance of the acceleration strategy. A Digimouse atlas[28] shown in Fig. 5(a) is employed to construct a 3-D simulation model, which includes four kinds of organs: heart, lungs, liver, and kidneys. Different optical properties are assigned to these organs to constitute a heterogeneous model, as presented in Table 1.[29] The Digimouse model is discretized into N mesh voxels using COMSOL Multiphysics 3.5 (COMSOL Inc., Stockholm, Sweden).

Table 4 RDs of the results obtained with the GPU and PCA methods for simulations.

	GPU				PCA			
	A	B	α	β	A	B	α	β
Group 1								
Case 1	0.003	0.002	0.003	0.002	0.038	0.003	0.006	0.003
Case 2	0.009	0.001	0.002	0.001	0.020	0.003	0.021	0.003
Case 3	0.053	0.003	0.012	0.004	0.048	0.002	0.013	0.003
Case 4	0.011	0.003	0.003	0.004	0.042	0.002	0.007	0.004
Group 2								
Case 5	0.017	0.002	0.011	0.001	0.047	0.003	0.014	0.003
Case 6	0.003	0.003	0.013	0.002	0.031	0.003	0.014	0.006
Case 7	0.005	0.002	0.008	0.003	0.013	0.001	0.006	0.002
Case 8	0.004	0.001	0.008	0.001	0.009	0.002	0.007	0.003
Average	0.013	0.002	0.008	0.002	0.031	0.003	0.011	0.003

Figure 5(b) shows indocyanine green (ICG) concentration curves, which simulate the metabolic processes of ICG in different organs and tissues. The curves are obtained according to Eq. (2), and the corresponding pharmacokinetic parameters are listed in Table 1.[5]

In the simulations, the atlas model is suspended on a rotation stage and continuously rotated for 60 circles. Each circle takes 1 min and acquires 24 projections with an angular increment of 15 deg. During the process of the measurement, the ICG concentrations of different organs vary in each projection according to the metabolic curves shown in Fig. 5(b).

Two parameters determine the size of the datasets and the time consumption of the reconstruction, i.e., the number of the discretized voxels N and the number of the boundary measurements M. Two groups of datasets are made to evaluate the effect of these two parameters on the reconstruction speed, and each group contained four cases. In group one, M is constant and N varies in different cases, while in group two, N is constant and M varies. The parameter setup in each case is listed in Table 2. To compare with the GPU method, the PCA method is implemented to reduce the size of the measurements M, and the compression threshold CPVs are set to be 0.7 for all the simulation cases.

3.1.2 *Simulation results*

The original and the PCA compressed data sizes of the numerical simulations are listed in Table 2. For the eight simulation cases, the original data sizes of the measurements range from 7730 to 15,620, while the compressed data sizes are between

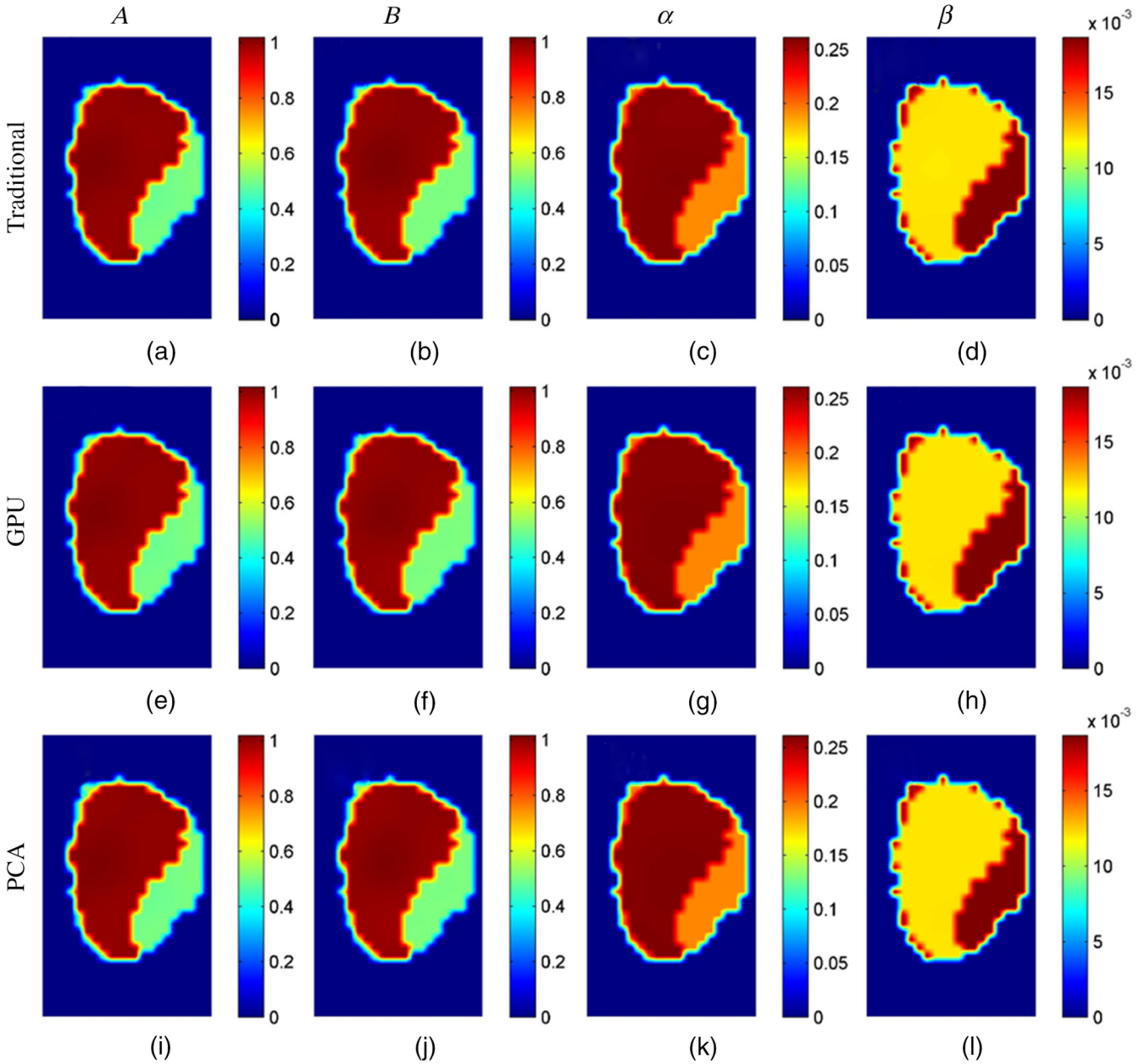

Fig. 6 Cross-sectional parametric images for simulation case 4 in the region of liver corresponding to the red dashed line in Fig. 5(a). (a)–(d) The results reconstructed with the traditional method. (e)–(h) The results reconstructed with the GPU method. (i)–(l) The results reconstructed with the PCA method.

696 and 2229. The maximum and minimum compression ratios are 7.0 and 15.2, respectively. According to the results, using PCA compression, the data sizes of the measurements are reduced on average to about 10% of the original data sizes.

The time consumptions of the traditional, GPU, and PCA methods are listed in Table 3, and the acceleration ratios are calculated. The time consumption of the traditional method ranges between 18,135 and 42,448 s, which is considered to be too long, especially when the reconstruction is operated repeatedly to obtain the optimal regularization parameters. The time costs of the GPU method are between 1485 and 3646 s, while the time costs of the PCA method are between 2861 and 4176. The average acceleration ratios of the GPU and PCA methods are 12.2 and 7.9, respectively. It indicates that the acceleration performance of the GPU method is better than the PCA method.

The RD*s* of the parametric images obtained from the GPU and PCA methods are listed in Table 4. The maximum average RD for the GPU and PCA methods are 0.013 and 0.031, respectively. According to the results demonstrated in Tables 3 and 4, a comparison can be made between the GPU and PCA methods for the numerical simulations. When the CPV value is set to be 0.7, the computational speed of the GPU method is faster than that of the PCA method, while the reconstruction quality of the GPU method is better than that of the PCA method.

Figure 6 shows the cross-sectional images of the reconstruction results obtained by the traditional, GPU, and

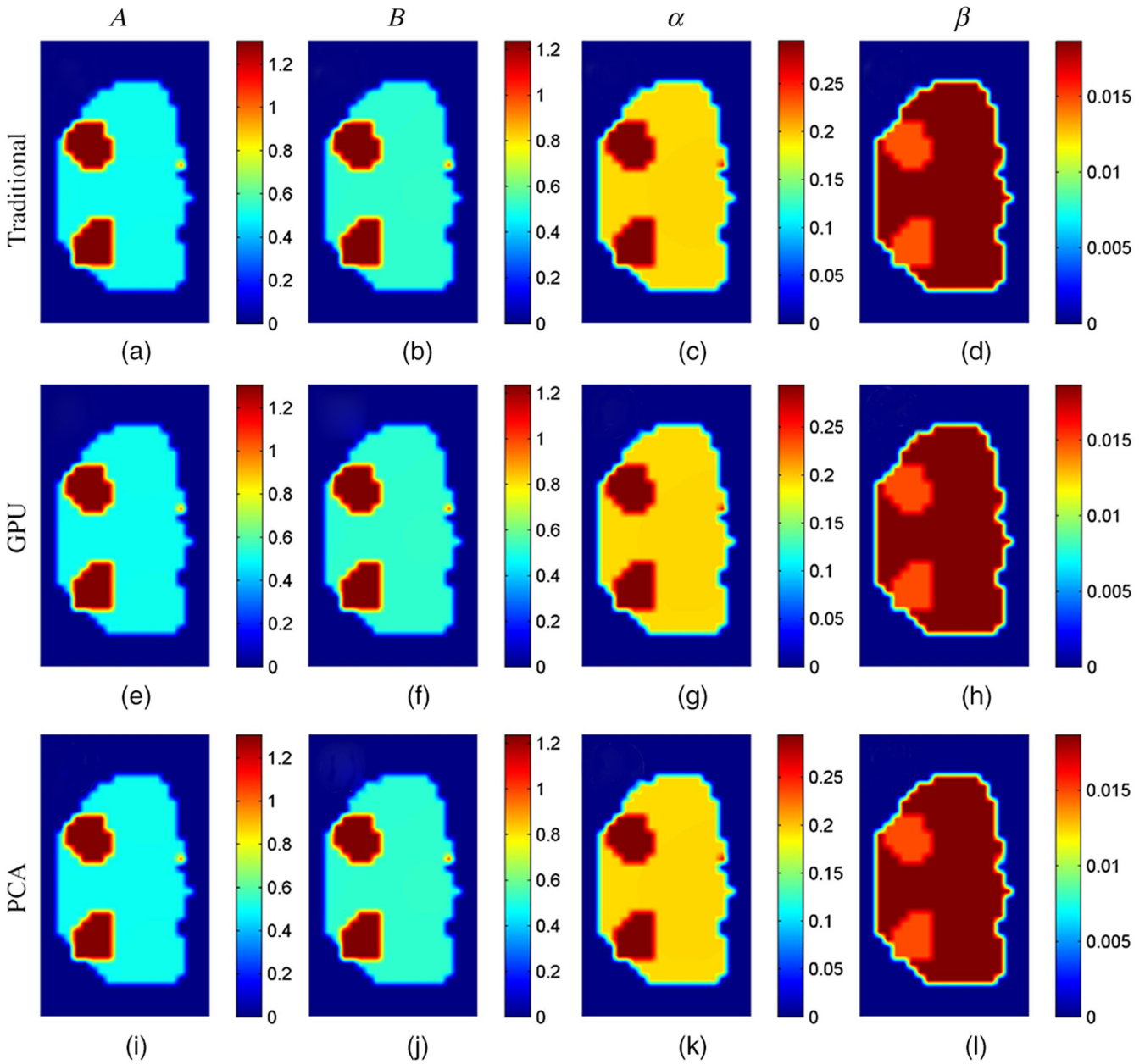

Fig. 7 Cross-sectional parametric images for simulation case 4 in the region of kidneys corresponding to the black dashed line in Fig. 5(a). (a)–(d) The results reconstructed with the traditional method. (e)–(h) The results reconstructed with the GPU method. (i)–(l) The results reconstructed with the PCA method.

PCA methods in simulation case 4, and the images are taken from the liver region of the Digimouse corresponding to the red dashed line in Fig. 5(a). Figure 7 shows the cross-sectional images in the kidney region corresponding to the black dashed line in Fig. 5(a). There are no obvious differences between the reconstructed results. The same findings can be obtained in all other cases in the simulations (not shown).

3.2 *In Vivo Experiment*

3.2.1 *Experimental setups*

In vivo experiments are conducted based on a hybrid FMT/XCT system described in Sec. 2.1. A 300 W Xenon lamp (MAX-302, Asahi Spectra, Torrance, California) is employed as the excitation source. A fiber is attached to the lamp to generate a line-shaped excitation source with a length of 4 cm. The excitation light is filtered through a 770 ± 6-nm bandpass excitation filter (XBPA770, Asahi Spectra, Torrance, California), and the power density of the excitation light is 0.03 $\mathrm{mW/cm^2}$. On the opposite side of the excitation source, the emitted ICG fluorescence is filtered with an 840 ± 6-nm bandpass emission filter (FF01-840/12-25, Semrock, Rochester, New York) and detected by a 512×512 pixel, -70°C cooled CCD camera (iXon DU-897, Andor Technologies, Belfast, Northern Ireland, United Kingdoms). The exposure time of the CCD camera is 1 s, and the CCD binning is set to be 512×512.

A healthy BALB/c nude mouse with an age of about 8 weeks is fixed on the rotation stage and anesthetized. A bolus of ICG (0.1 mL, 50 μg/mL) is injected via the tail vain. During the DFMT measurements acquisition, the mouse is continuously rotated for 50 circles ($K = 50$) with an angular increment of 15 deg. Therefore, 24 projections ($S = 24$) are obtained in each circle, and a total of 1200 projections ($P = KS = 50 \times 244$) are acquired for the entire DFMT measurement.

After the fluorescence data acquisition is finished, a hepatobiliary contrast agent for XCT imaging, Fenestra LC (Advanced Research Technologies, Montreal, California), is injected at a dose of 15 mL/kg body weight through the tail vein. One hour after the injection, the XCT images are collected to provide structural prior information. The x-ray tube works at 45 kVp and 1 mA during the scan, and the XCT images are collected by a complementary metal oxide semiconductor flat-panel detector (C7921-02, Hamamatsu, Japan).

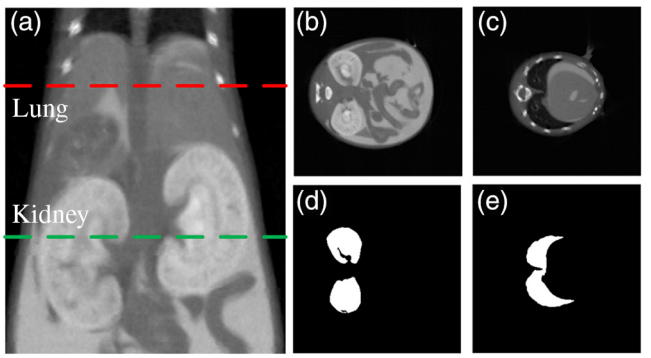

Fig. 8 XCT results of the mouse experiments. (a) Coronal XCT image in the chest region of the mouse. (b) and (c) Transversal XCT images indicated by the green and red dashed lines in (a), respectively. (d) and (e) Manually segmented organs corresponding to (b) and (c).

Table 5 Original and PCA compressed data sizes for the mouse experiments.

	Measurements M		
Voxels N	Original	PCA	Compression ratio
6062	12,968	2507	5.2

In the mouse experiments, a total of 12,968 measurements are acquired. As shown in Fig. 8(a), a chest region with a height of 2.5 cm of the mouse is used to reconstruct the parametric images, and the reconstructed region is discretized into 6062 voxels. The transversal XCT images indicated by the green and red dashed lines in Fig. 8(a) are shown in Figs. 8(b) and 8(c). The XCT images are manually segmented into four regions: liver, lungs, kidneys, and the background. The Laplacian-type[25] regularization matrix is constructed according to the segmentation results. Figures 8(d) and 8(e) depict the segmented results corresponding to Figs. 8(b) and 8(c). The structural priors are used to create a heterogeneous model by assigning different optical properties to relevant regions, as shown in Table 1. Additionally, the PCA-based acceleration method is compared with the proposed GPU method, and the CPV is set to 0.995.

3.2.2 *Experimental results*

As shown in Table 5, the number of the measurements M is reduced from 12,968 to 2507 with PCA, and the compression ratio is about five times. The time consumptions of the traditional, GPU, and PCA methods are listed in Table 6. The acceleration ratios of the GPU and PCA methods are 9.8 and 9.4, respectively. The acceleration performances of the GPU and PCA methods are very close.

The RD*s* obtained with the GPU and PCA methods are shown in Table 7. The RDs of four parameters reconstructed with the PCA method are all larger than those with the GPU method. With a CPV value of 0.995, the reconstruction quality

Table 6 Time consumptions and acceleration ratios of the GPU and PCA methods for the mouse experiments.

Time(s)			Acceleration ratio	
Traditional	GPU	PCA	GPU	PCA
32,512	3326	3466	9.8	9.4

Table 7 RDs of the results obtained with the GPU and PCA methods for the mouse experiments.

	A	B	α	β
GPU	0.018	0.009	0.010	0.012
PCA	0.053	0.035	0.021	0.027

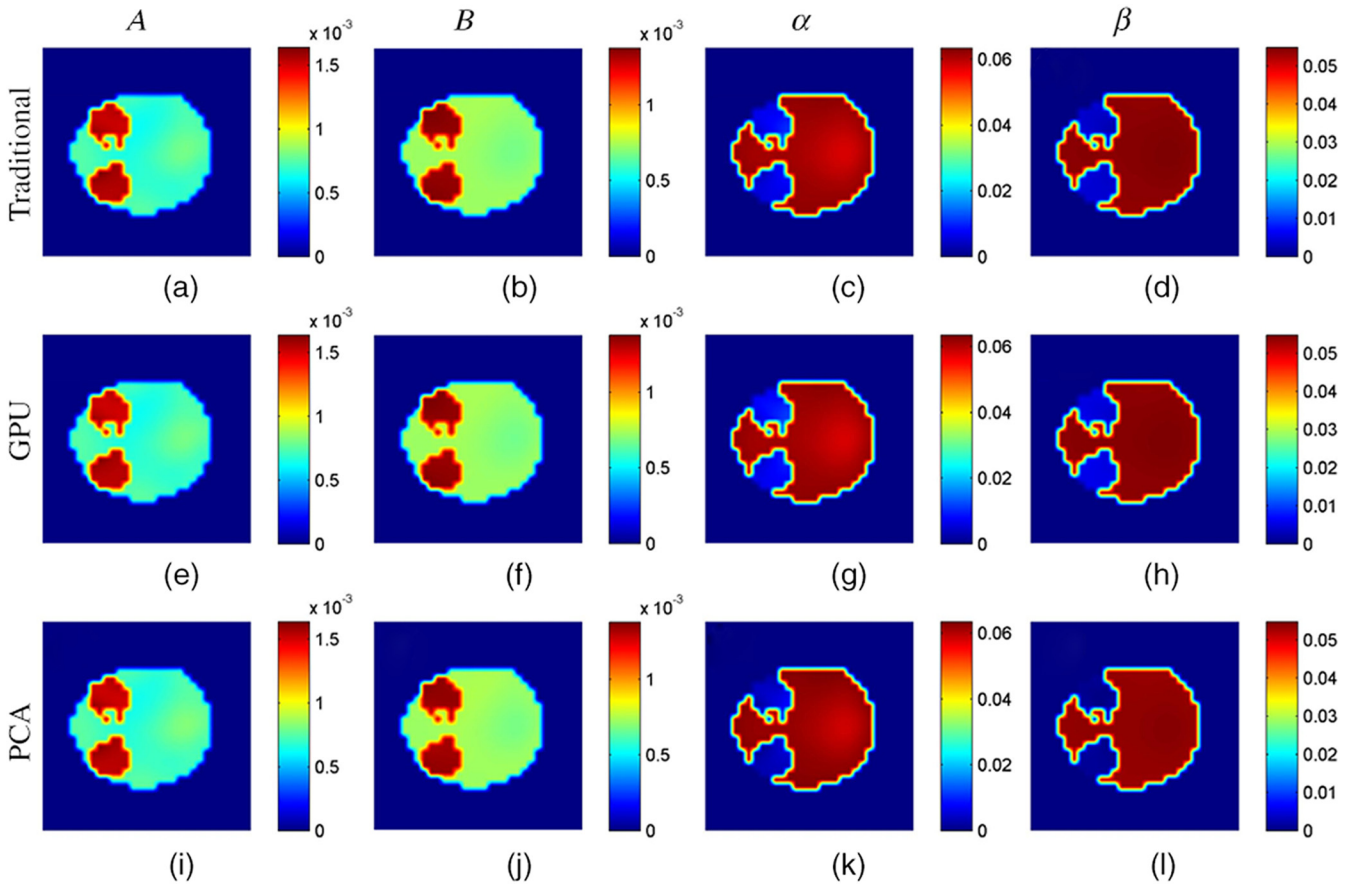

Fig. 9 Cross-sectional parametric images for the mouse experiments in the region of lung corresponding to the red dashed line in Fig. 8(a). (a)–(d) The results reconstructed with the traditional method. (e)–(h) The results reconstructed with the GPU method. (i)–(l) The results reconstructed with the PCA method.

of the GPU method is better than PCA, while the acceleration performances of these two methods are close.

The cross-sectional parametric images obtained with the traditional, GPU, and PCA methods for the mouse experiments are shown in Fig. 9, and the images are taken from the lung region of the mouse corresponding to the red dashed line in Fig. 8(a). Figure 10 shows the cross-sectional images in the kidney region of the mouse corresponding to the green dashed line in Fig. 8(a). For all the reconstruction images, no visual differences can be observed.

4 Discussions

DFMT is a promising technique that can be used in tumor detection, drug development, and metabolic research.[3] The previously proposed direct reconstruction method can improve the quality of the DFMT result.[5] However, the implementation of the direct reconstruction method is limited due to the large time consumption. Generally, several hours are needed to reconstruct the DFMT images. Therefore, it is necessary to accelerate the direct method.

In this study, an acceleration strategy using GPU for DFMT is presented. The results of the numerical simulations and *in vivo* experiments demonstrate that this strategy can efficiently reduce the time cost of the reconstruction, with nearly no quality degradation.

As previously mentioned, the performances of the GPU and PCA methods are compared. For the GPU method, an appropriate CPV should be carefully selected to balance the calculation speed and the reconstruction accuracy.[11] A smaller CPV achieves faster computational speed, while a larger CPV obtains better reconstruction quality. In addition, the selection of CPV also depends on the specific data used. In this paper, the CPV was empirically selected to be 0.7 and 0.995 in the simulations and *in vivo* experiments, respectively. The GPU method does not need such a data-dependent parameter, and it may be convenient to use once it is implemented. On the other hand, information is lost to some extent in the PCA method, which could reduce the quality of the reconstruction results, especially when the CPV is set to be too small. On the contrary, all the measurement information can be retained when using the GPU method.

Furthermore, the PCA- and GPU-based methods are two different approaches to acceleration of FMT reconstruction and can be combined together to further increase the computational speed. Simulation case 1 is used to study the performance of the combination of the PCA and GPU methods, and the CPV is set to be 0.7. The time consumption of the combination method is 452 s, i.e., the acceleration ratio is increased to 48.9. The maximum RD is 0.043, which is slightly larger than that of the PCA method (0.038). The results demonstrate that, by combining the PCA and GPU methods, the computational speed can be

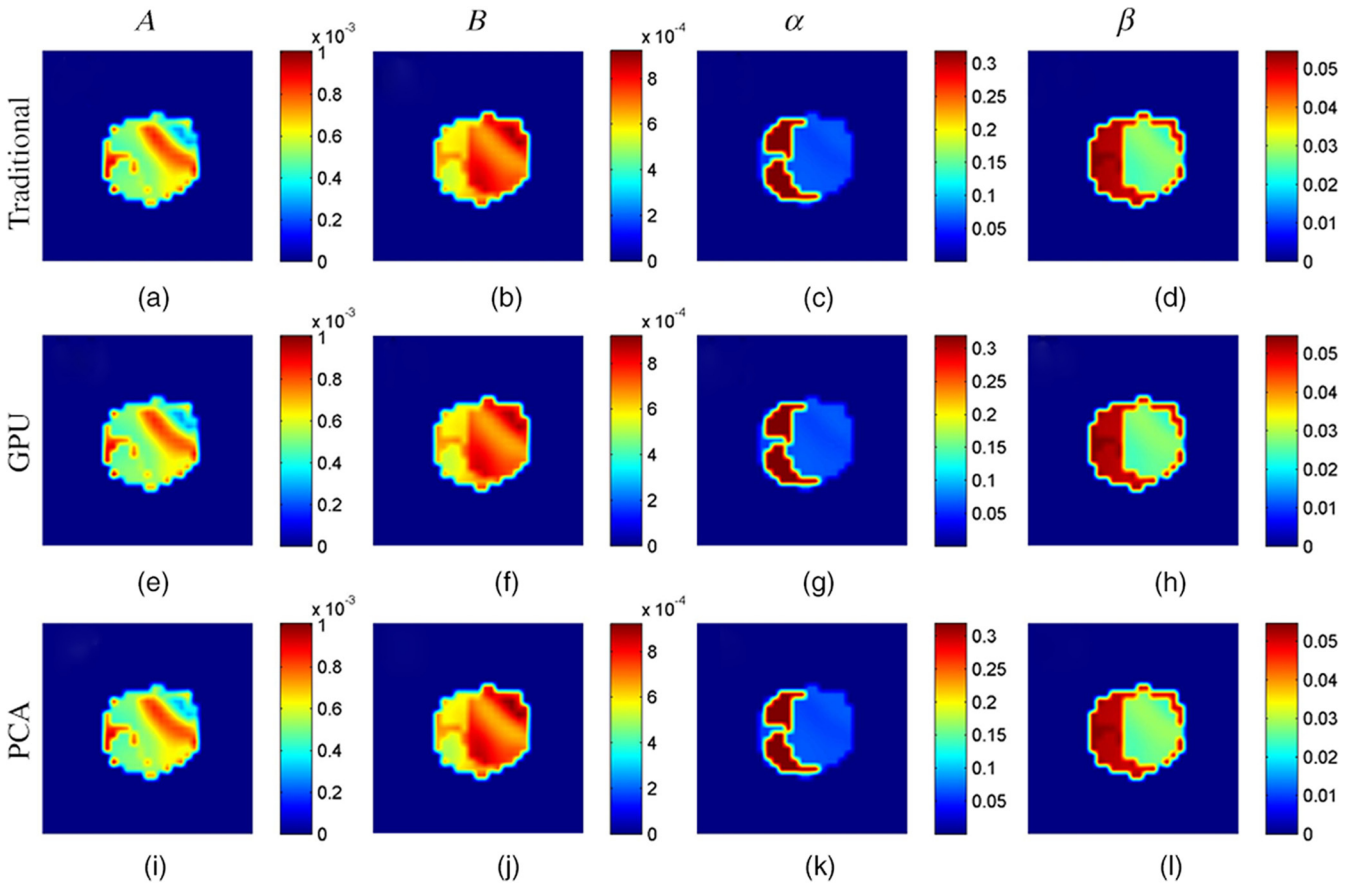

Fig. 10 Cross-sectional parametric images for the mouse experiments in the region of kidneys corresponding to the green dashed line in Fig. 8(a). (a)–(d) The results reconstructed with the traditional method. (e)–(h) The results reconstructed with the GPU method. (i)–(l) The results reconstructed with the PCA method.

significantly improved, and the loss of the image quality is acceptable.

5 Conclusion

In conclusion, the direct DFMT reconstruction algorithm is accelerated using a GPU-based strategy. The feasibility of this method is confirmed by numerical simulations and *in vivo* experiments. According to the results, the time consumptions are reduced to ~10% of the traditional method. The average RD of simulations and *in vivo* experiments is less than 2%, which means the errors between the acceleration method and the traditional method are negligible.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant Nos. 81227901, 81271617, 61322101, and 61361160418; and the National Major Scientific Instrument and Equipment Development Project under Grant No. 2011YQ030114. All applicable international, national, and/or institutional guidelines for the care and use of animals were followed. All procedures performed in studies involving animals were in accordance with the ethical standards of the Ethics Committee of Tsinghua University. This article does not contain any studies with human participants performed by any of the authors.

References

1. M. Gurfinkel et al., "Pharmacokinetics of ICG and HPPH-car for the detection of normal and tumor tissue using fluorescence, near-infrared reflectance imaging: a case study," *Photochem. Photobiol.* **72**(1), 94–102 (2000).
2. E. M. C. Hillman and A. Moore, "All-optical anatomical co-registration for molecular imaging of small animals using dynamic contrast," *Nat. Photonics* **1**(9), 526–530 (2007).
3. G. Zhang et al., "Imaging of pharmacokinetic rates of indocyanine green in mouse liver with a hybrid fluorescence molecular tomography/x-ray computed tomography system," *J. Biomed. Opt.* **18**(4), 040505 (2013).
4. G. L. Zhang et al., "A direct method with structural priors for imaging pharmacokinetic parameters in dynamic fluorescence molecular tomography," *IEEE Trans. Biomed. Eng.* **61**(3), 986–990 (2014).
5. G. L. Zhang et al., "Full-direct method for imaging pharmacokinetic parameters in dynamic fluorescence molecular tomography," *Appl. Phys. Lett.* **106**(8), 081110 (2015).
6. A. B. Milstein, K. J. Webb, and C. A. Bouman, "Estimation of kinetic model parameters in fluorescence optical diffusion tomography," *J. Opt. Soc. Am. A.* **22**(7), 1357–1368 (2005).
7. A. Alacam and B. Yazici, "Direct reconstruction of pharmacokinetic-rate images of optical fluorophores from NIR measurements," *IEEE Trans. Med. Imaging* **28**(9), 1337–1353 (2009).
8. X. Cao et al., "An adaptive Tikhonov regularization method for fluorescence molecular tomography," *Med. Biol. Eng. Comput.* **51**(8), 849–858 (2013).
9. J. Dutta et al., "Joint L-1 and total variation regularization for fluorescence molecular tomography," *Phys. Med. Biol.* **57**(6), 1459–1476 (2012).

10. J. J. Yu et al., "Sparse reconstruction for fluorescence molecular tomography via a fast iterative algorithm," *J. Innov. Opt. Health Sci.* **7**(3), 1450008 (2014).
11. X. Cao et al., "Accelerated image reconstruction in fluorescence molecular tomography using dimension reduction," *Biomed. Opt. Express* **4**(1), 1–14 (2013).
12. T. J. Rudge, V. Y. Soloviev, and S. R. Arridge, "Fast image reconstruction in fluoresence optical tomography using data compression," *Opt. Lett.* **35**(5), 763–765 (2010).
13. T. Correia et al., "Wavelet-based data and solution compression for efficient image reconstruction in fluorescence diffuse optical tomography," *J. Biomed. Opt.* **18**(8), 086008 (2013).
14. J. Ripoll, "Hybrid Fourier-real space method for diffuse optical tomography," *Opt. Lett.* **35**(5), 688–690 (2010).
15. Q. Q. Fang and D. A. Boas, "Monte Carlo simulation of photon migration in 3D Turbid media accelerated by graphics processing units," *Opt. Express* **17**(22), 20178–20190 (2009).
16. B. Zhang et al., "The CUBLAS and CULA based GPU acceleration of adaptive finite element framework for bioluminescence tomography," *Opt. Express* **18**(19), 20201–20214 (2010).
17. X. Wang et al., "Acceleration of early-photon fluorescence molecular tomography with graphics processing units," *Comput. Math. Methods Med.* **2013**, 297291 (2013).
18. X. L. Guo et al., "A combined fluorescence and microcomputed tomography system for small animal imaging," *IEEE Trans. Biomed. Eng.* **57**(12), 2876–2883 (2010).
19. A. B. Milstein et al., "Fluorescence optical diffusion tomography," *Appl. Opt.* **42**(16), 3081–3094 (2003).
20. A. Joshi, W. Bangerth, and E. M. Sevick-Muraca, "Adaptive finite element based tomography for fluorescence optical imaging in tissue," *Opt. Express* **12**(22), 5402–5417 (2004).
21. M. A. O'Leary et al., "Fluorescence lifetime imaging in turbid media," *Opt. Lett.* **21**(2), 158–160 (1996).
22. D. J. Cuccia et al., "In vivo quantification of optical contrast agent dynamics in rat tumors by use of diffuse optical spectroscopy with magnetic resonance imaging coregistration," *Appl. Opt.* **42**(16), 2940–2950 (2003).
23. H. Shinohara et al., "Direct measurement of hepatic indocyanine green clearance with near-infrared spectroscopy: separate evaluation of uptake and removal," *Hepatology* **23**(1), 137–144 (1996).
24. D. Hyde et al., "A statistical approach to inverting the born ratio," *IEEE Trans. Med. Imaging* **26**(7), 893–905 (2007).
25. S. C. Davis et al., "Image-guided diffuse optical fluorescence tomography implemented with Laplacian-type regularization," *Opt. Express* **15**(7), 4066–4082 (2007).
26. S. R. Arridge and M. Schweiger, "A gradient-based optimisation scheme for optical tomography," *Opt. Express* **2**(6), 213–226 (1998).
27. A. Brooksby et al., "Near-infrared (NIR) tomography breast image reconstruction with a priori structural information from MRI: algorithm development for reconstructing heterogeneities," *IEEE J. Sel. Top. Quantum* **9**(2), 199–209 (2003).
28. B. Dogdas et al., "Digimouse: a 3D whole body mouse atlas from CT and cryosection data," *Phys. Med. Biol.* **52**(3), 577–587 (2007).
29. D. Hyde et al., "Performance dependence of hybrid x-ray computed tomography/fluorescence molecular tomography on the optical forward problem," *J. Opt. Soc. Am. A* **26**(4), 919–923 (2009).

Maomao Chen received his bachelor's and master's degrees in biomedical engineering from Chongqing University in 2006 and 2009, respectively. Currently, he is a PhD candidate in the Department of Biomedical Engineering, Tsinghua University. His research interest is fluorescence molecular tomography for small animal imaging.

Jiulou Zhang is a PhD candidate in the Department of Biomedical Engineering, Tsinghua University, Beijing, China. His research interest is fluorescence molecular tomography for small animal imaging.

Chuangjian Cai received his bachelor's degree in biomedical engineering, Tsinghua University, Beijing, China, in 2015. Currently, he is a PhD candidate in the Department of Biomedical Engineering, Tsinghua University. His research interest is fluorescence molecular tomography for small animal imaging.

Yang Gao received his bachelor's degree in pharmaceutical sciences from Tsinghua University, Beijing, China, in 2014. Currently, he is a PhD candidate in the Department of Biomedical Engineering, Tsinghua University. His research interest is fluorescence molecular imaging.

Jianwen Luo is a professor at Tsinghua University. He was enrolled in the Thousand Young Talents Program in 2012 and received the Excellent Young Scientists Fund from the National Natural Science Foundation of China in 2013. He serves as an advisory editorial board member of the *Journal of Ultrasound in Medicine*, associate editor of *Medical Physics*, and faculty member of Faculty of 1000. His research interest is biomedical imaging, including ultrasound imaging and fluorescence molecular imaging.