Scatter Correction based on GPU-accelerated Full Monte Carlo Simulation for Brain PET/MRI

Bo Ma, Michaela Gaens, Liliana Caldeira, Julian Bert, Philipp Lohmann, Lutz Tellmann, Christoph Lerche, *Member, IEEE,* JUrgen Scheins, Elena Rota Kops, Hancong Xu, Mirjam Lenz, Uwe Pietrzyk, and Nadim Jon Shah

Abstract—Accurate scatter correction is essential for qualitative and quantitative PET imaging. Up to now, scatter correction based on Monte Carlo simulation (MCS) has been recognized as the most accurate method of scatter correction for PET. However, the major disadvantage of MCS is its long computational time, which makes it unfeasible for clinical usage. Meanwhile, single scatter simulation (SSS) is the most widely used method for scatter correction. Nevertheless, SSS has the disadvantage of limited robustness for dynamic measurements and for the measurement of large objects. In this work, a newly developed implementation of MCS using graphics processing unit (GPU) acceleration is employed, allowing full MCS-based scatter correction in clinical 3D brain PET imaging. Starting from the generation of annihilation photons to their detection in the simulated PET scanner, all relevant physical interactions and transport phenomena of the photons were simulated on GPUs. This resulted in an expected distribution of scattered events, which was subsequently used to correct the measured emission data. The accuracy of the approach was validated with simulations using GATE (Geant4 Application for Tomography Emission), and its performance was compared to SSS. The comparison of the computation time between a GPU and a single-threaded CPU showed an acceleration factor of 776 for a voxelized brain phantom study. The speedup of the MCS implemented on the GPU represents a major step towards the application of the more accurate MCS-based scatter correction for PET imaging in clinical routine.

Index Terms—GPU, Monte Carlo simulation, PET, Scatter Correction, Single Scatter Simulation.

This work was financially supported by the OCPC scholarship program between China and Germany and the China Postdoctoral Science Foundation 2015M570154. *(Corresponding author: Bo Ma)*.

B. Ma, P. Lohmann, L. Tellmann, C. Lerche, J. Scheins, E. Rota Kops, H. Xu, M. Lenz, U. Pietrzyk and N. J. Shah are with the Institute of Neuroscience and Medicine-4, Forschungszentrum Juelich, 52425 Juelich, Germany (e-mail: b.ma@fz-juelich.de; p.lohmann@fz-juelich.de; l.tellmann@fz- juelich.de; c.lerche@fz-juelich.de; j.scheins@fz-juelich.de; e.rota.kops@fz- juelich.de; ha.xu@fz-juelich.de; m.lenz@fz-juelich.de).

M. Gaens was with the Institute of Neuroscience and Medicine-4, Forschungszentrum Juelich, 52425 Juelich, Germany. She is now with Heidelberg Engineering GmbH, 69115 Heidelberg, Germany (e-mail: michaela-gaens@gmail.com).

J. Bert is with the LaTIM, INSERM, UMR1101, Universite de Bretagne Occidentale, 29238 CHRU Brest, France (e-mail: julien.bert@univ-brest.fr).

L. Caldeira was with the Institute of Neuroscience and Medicine-4, Forschungszentrum Juelich, 52425 Juelich, Germany. She is now with the Department for Diagnostic and Interventional Radiology, University Hospital Cologne, 50937 Cologne, Germany.

B. Ma is also with the Institute of High Energy Physics, Chinese Academy of Sciences, 100049 Beijing, China.

L. Caldeira and U. Pietrzyk are also with the Faculty of Mathematics and Natural Science, University of Wuppertal, 42119 Wuppertal, Germany (e-mail: l. caldeira@fz-juelich.de; Uwe.Pietrzyk@web.de).

N. J. Shah is also with the Department of Neurology, Faculty of Medicine, JARA, RWTH Aachen University, 52062 Aachen, Germany (email: n.j.shah@fz-juelich.de).

I. Introduction

ACCURATE scatter correction[1] is essential for qualitative and quantitative PET imaging. Up to now, scatter correction based on Monte Carlo simulation (MCS) has been recognized as the most accurate method [1]-[3]. However, the major disadvantage of MCS is the long computation time which makes it currently unfeasible for the clinical environment. For example, a scatter simulation for a brain PET measurement can take as long as several weeks [4]. At present, the most widely used approach for scatter correction in clinical PET imaging is the single scatter simulation (SSS) [5], [6]. SSS implementations have led to a significant improvement towards accurate scatter estimation in clinical 3D PET imaging. However, SSS is still associated with certain approximations, which are sometimes unreliable [7]—[9]. More specifically, tail fitting, which is usually the last step for the estimation of the scatter distribution, is prone to fail occasionally, especially for dynamic PET data with low count statistics or objects filling almost the entire field-of-view (FOV). This may lead to significant errors during image quantification, potentially hampering applications such as the evaluation of therapy results, which are particularly dependent on image-derived quantitative parameters [7]-[9]. In order to improve overall PET quantification, there have been several proposals towards the application of physically more accurate MCS for scatter correction [4], [9]-[12]. Comparing to the unreliable tail fitting of SSS, the scaling factor in MCS is derived on the basis of all true events by taking into account the associated physical effects. This makes the MCS methods more robust and more stable than SSS methods.

In recent years, the highly parallel computing power of graphics processing units (GPUs) has come into focus for the development of PET imaging systems [13]-[15]. The simulation of the transport and interaction of photons can be treated individually and independently, which is suited for the inherent parallel computation architecture of GPUs. This has led to the pursuit of GPU-based MCS for medical applications in recent years [16]—[21]. Among these proposed implementations, the framework of the GPU accelerated Geant4 based Monte Carlo Simulation (GGEMS) reported by Bert and colleagues [19], [20] is the only one which targets both imaging and therapeutic applications. GGEMS is based on the implementation of relevant physical processes of the well-validated Geant4 libraries. However, the previous implementation of the GGEMS did not include a full simulation of the detection procedure in the scanner. An alternative approach is the hybrid CPU/GPU architecture for GATE (Geant4 Application for Tomography Emission) [20], the detection module of which still relies on standard CPU code, slowing down the overall simulation. For the other existing approaches, the physical effects within the detectors have also not been considered, and they rely exclusively on ray tracing based on a geometrical description of the detectors. However, physically accurate modeling of the detection processes is

an essential element to provide reliable simulation results.

In this work, the GGEMS based simulation, which was extended to comprise the detection module, was implemented and validated for brain PET imaging [22]. The proposed approach was presented and evaluated on a hybrid 3T MR- BrainPET developed by Siemens Healthineers [23]. It takes into account all available information, including the attenuation map for the head and the MR coils which cause additional scatter, to guarantee the accuracy of scatter estimation. In this paper, the detail of the proposed approach is firstly described, then its performance is evaluated and compared with GATE and SSS-based scatter correction using both phantoms and patient datasets.

II. Methods Implementation

A. BrainPET Scanner

The BrainPET is designed as an insert to be placed in the bore of the MR scanner. It consists of 192 detector blocks arranged in 32 copper-shielded detector cassettes which are placed on a ring. For each cassette, six detector blocks are axially aligned, each block consisting of 12 x 12 lutetium oxyorthosilicate (LSO) crystals. The size of the LSO crystals is 2.5 x 2.5 x 20.0 mm^3. The scanner has an axial FOV of 19.2 cm and a transaxial FOV of 36.5 cm in diameter. When the BrainPET is inserted into the conventional 3T MRI scanner (Siemens 3T TIM Trio with minor modifications), the standard patient bed is replaced by a vertically fixed bed and two adapted head coils. These coils are optimized for PET with respect to minimal attenuation for 511 keV photons [23].

B. GPU Implementation of the Monte Carlo Simulation

The framework of GGEMS described previously by Bert et al. [19] is used as a basis for this work. To keep this paper concise and readable, the procedures of the photon generation and tracking in the phantom, which are demonstrated explicitly in [19], are also briefly introduced. Similar to other approaches [16], [17], the strategy of one thread per particle, or in this case one thread per photon, is adopted for the GPU implementation, and is realized using Compute Unified Device Architecture (CUDA) developed by NVIDIA. This means that each individual thread simulates the entire trajectory of one annihilation photon, enabling the simulation of thousands of photons in parallel. More specifically for PET simulations, two stacks of photons are created, corresponding to the two photons originating from the same annihilation. Both stacks are realized as a CUDA C structure of arrays residing in the global device memory, allowing coalesced access by all threads. Each array of the structure corresponds to one parameter of the photons, including the position, the direction, the energy, the time-of-flight (TOF), the occurrence number of Compton scattering, and a flag indicating whether the photon is still active or not (absorbed or below the defined energy threshold).

The general workflow of the implemented GGEMS is shown in Fig. 1. After the initialization of the pseudo-random number generator (PRNG), the annihilation photon pairs are generated according to the activity distribution of the input emission image. Subsequently, the photons' trajectories are tracked through the voxelized phantom (including the phantom and/or coils), which consists of defined materials, and physical interactions between photons and the material are taken into account. Upon leaving the voxelized phantom, the photons are tracked from the phantom to the detector blocks using ray tracing. In the detectors, the interactions and transport of the photons are simulated, subjected to relevant physical effects and detector boundaries. The TOF for the photon and its position within the detector are recorded. After the detection module, data are copied back from the GPU to the CPU and the simulated photons are time-stamped and sorted into coincidence events. The final output data are similar to PET list-mode data with an additional flag indicating whether the coincidence event is an unscattered or scattered true event or a random event. Starting from the initialization of PRNG to the photon detection, all steps are realized as separate kernel functions on the GPU. Time stamping and coincidence sorting are implemented on the CPU.

C. Random Number Generator

Following Bert et al. [19], the Brent-XOR256 PRNG is employed in this implementation. The choice of this PRNG is based on its long period, fast computation of random numbers and ease of use for GPU architectures with the merit of generating one random number at a time. This allows each thread to handle its own PRNG independently. Different random seeds are used to initialize the individual PRNGs, generating uncorrelated series of random numbers. Unlike [19], in which the initial random seeds are created on the CPU, in this work the random seeds are created on the GPU by the random number generator provided by the cuRAND library [24].

D. Photon Tracking

During its simulated lifetime, a photon is processed by three sequential GPU kernels. In the first kernel, annihilation photon pairs are generated according to the activity distribution of the input emission image. The second kernel simulates the photon's propagation through the voxelized phantom which contains information about materials and their physical properties for each voxel. In the third kernel, the propagation and interactions of photons in the detector blocks are simulated. Materials for the phantom and detectors are defined in a text file using the relative weights for each compound.

1) Physical Effects: The physical effects of the GGEMS are realized on GPUs based on adapting the electromagnetic standard processes of Geant4 [25]. Only the processes relevant for PET, i.e.

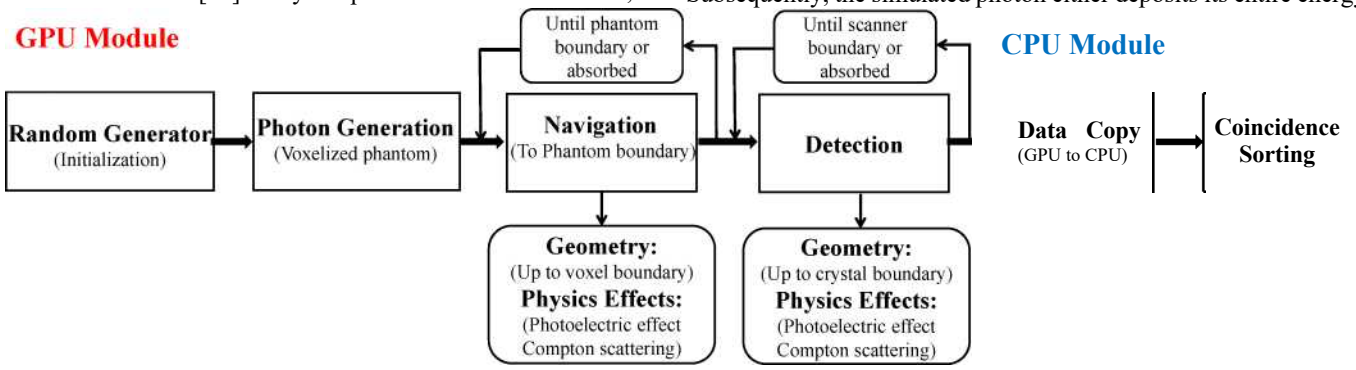


Fig. 1. Flowchart of the implementation of GPU based Monte Carlo simulation for the 3T MR-BrainPET.

photoelectric effects and Compton effects, are included. For the computation of total photoelectric cross sections, the parameterized table proposed by Biggs and Lighthill [26] is applied. Specific cross sections are calculated by loglog interpolation from the parameterized table, which is stored in the constant memory of the GPU for fast access. The Compton cross-section is analytically calculated according to the Klein-Nishina formula on the GPU with no need for externally located data. The generation of secondary electrons (ejected photoelectrons and Compton recoil electrons) is omitted, since, due to their short ranges compared to the dimensions of the crystals, their energy deposition can be treated as pointlike [27]. The transport of optical photons generated by the scintillation process is not explicitly simulated.

2) Photon Propagation and Interaction: For each step of the photon navigation inside the voxelized phantom, the total cross sections of the photoelectric effect and Compton scattering are determined according to the photon energy and material. The transport free paths for these two effects are calculated with the combination of the total cross-section and a random number. Besides, the distance between the annihilation photon and the voxel boundary is also computed. Depending on which distance is the shortest, the annihilation photon either propagates to the next voxel in its path or undergoes physical interactions of photoelectric effect or Compton scattering. If the photon proceeds to the next voxel, the stopping position will be at the boundary of the voxel. A new calculation will be started subsequently for the free path of these three procedures. For photoelectric interactions, the energy of the current annihilation photon is registered as energy deposition, and simulation of the annihilation photon transport is set to an inactive state. For the Compton interactions, the annihilation photon deposits part of its energy according to a randomly sampled scattering angle and the program continues to simulate the propagation of the scattered photon with the residual energy.

3) Annihilation Photon Detection: After leaving the vox- elized phantom, the annihilation photons are tracked from the phantom to the detector blocks using ray tracing. Comparing to the work of Bert et al. [19], this step is also implemented on the GPU for further acceleration of the simulation. The photon propagating processes inside the detector blocks are similar to those in the voxelized phantom. Total cross sections for the photoelectric effect and Compton scattering are calculated according to the material of the crystals and the energy of the annihilation photon. Besides, the distance between the photon and the boundary of the detector block on the annihilation photon's trajectory is also determined. Subsequently, the simulated photon either deposits its entire energy, part of its energy or propagates to an adjacent detector block, according to the respective transport free path. Similar to the GATE implementation, the registered energy of single events (singles) is computed from all physical interactions (hits) of the corresponding photon inside a specific detector block. In this implementation, the centroid of all hits is weighted by the deposited energy of each simulated interaction and updated successively at each interaction to reduce the number of parameters that needed to be stored. For each photon, only one single is stored. If the photon interacts in more than one detector block, the single corresponding to the block with the highest energy deposition is stored.

4) Coincidence Sorting: After the detection kernel, the two photon stacks which correspond to the back-to-back annihilation photon pair are copied from the GPU to the CPU for coincidence processing. The time stamp (T_i) for each photon pair (i) is built from three parts:

$$T_i = t_1 + t_2 + t_3 \quad (1)$$

where t_1 is the time stamp of the previous photon pair (T_{i-1}), t_2 is the time interval between two positron emissions which is generated according to the total activity of the emission image, and t_3 is the TOF for each individual photon of the two back-to-back photons. These time stamps are used to sort singles into coincidences according to a given coincidence timing window, which is 12 ns for the BrainPET insert in this study. In the current implementation, multiple coincidences are discarded. Coincidence events are subsequently stored in a binary file analogous to PET list mode data, with the information of the two crystal IDs and a flag indicating whether the coincidence is a true, scattered, or random event.

E. Construction of the Attenuation Volume

To track the annihilation photons, a volumetric image dataset defining the attenuation properties of the phantom, patient and additional equipment (e.g., an MR coil) is needed. For the combined MR-BrainPET, it is necessary to include attenuation maps of the patients' head or the phantom, the MR head coils, and the head holder (or bed). For the phantom measurements used in this study, the attenuation maps were obtained through a transmission scan (using 511 keV photons) performed on a Siemens ECAT HR+ PET scanner. For the individual patient measurements, the MR-derived attenuation maps were obtained using the template-based approach

[28]. Two attenuation maps were prepared for the MR head coils. One was obtained through a transmission scan on the Siemens ECAT HR+ PET scanner, which was used for the image reconstruction and SSS. The other one was obtained by a CT measurement. Because GGEMS requires materials types instead of attenuation coefficients, a CT image is more appropriate to separate the different components of the coil. According to the Hounsfield units in the CT image, the MR coils were segmented into compartments of plastic housing and copper wires. For the GGEMS of a hybrid PET/MRI measurement, this coil image was merged with the patient attenuation map in order to include all potential scatter sources. Both the attenuation maps have the same dimension and voxel size as the emission image.

F. Scatter Correction Procedure

A simplified flowchart of the procedure for the estimation of the scatter distribution is shown in Fig. 2. Input data, including an initial reconstructed emission image (with all necessary corrections except scatter correction), an MR- derived attenuation map of the patient's head [28], and a CT- derived attenuation template of the MR coils, were used for the simulation. Based on the attenuation map, the emission image was masked to allow photons to be generated only inside the actual object and to avoid erroneous photon generation caused by spurious reconstructed activity outside the object. After the simulation, two datasets, including the unscattered and the scattered true events, were generated. Both datasets were normalized for geometrical crystal efficiencies using a normalization file derived from the simulation of an air-filled cylinder covering the whole FOV. For the measured data, normalization and random corrections were carried out using the standard procedures based on Badawi and Marsden [29] and Byars et al. [30], respectively. Subsequently, the total counts of the simulated data (true and scattered coincidences) and the measured data in the projection space were compared to obtain a global scaling factor. This scaling factor was used to scale the simulated scatter distribution in the projection space, which could be directly subtracted from the emission data or incorporated into the iterative reconstruction algorithm such as 3D OP-OSEM (Ordinary Poisson-Ordered Subsets Expectation Maximization) [31].

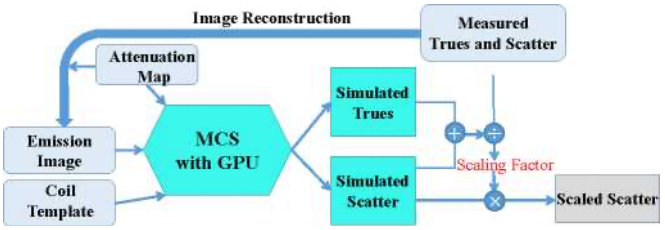


Fig. 2. Simplified flowchart of GGEMS-based scatter correction. Specific input data from measurements are shaded in blue and the output of the scatter simulation is shaded in grey.

III. Methods Evaluat ion

A. Accuracy Validation of GGEMS

In order to validate the complete GGEMS proposed in this work, several comparisons were performed with the CPUbased GATE (V7.2) simulation. The compiler (CUDA version: 9.0, gcc version: 5.4.0) and the operating system (Ubuntu 16.04) for GGEMS and GATE are the same throughout this paper. GATE has already been extensively validated with measurements and was therefore used as a reference for the performance evaluation of the proposed implementation [32]. A homogeneous water phantom placed in the center of the BrainPET insert was simulated. The phantom was realized as a 3D image volume of 100 x 100 x 100 voxels with a voxel size of 1.25 x 1.25 x 1.25mm^3. The source activity was set to 1 MBq and distributed homogeneously in the phantom. In the GATE simulation, back-to-back annihilation photons were simulated. Physical processes based on Geant4's electromagnetic standard models, including photoelectric effect, Compton effect and electron ionization, were applied. The energy window was set to 420 keV - 600 keV, which corresponded to the energy threshold of the BrainPET insert. To ensure secondary electrons were not tracked within the phantom, large range-cuts were used in GATE. However, to obtain more accurately detected events for the photon detection procedure, secondary electrons were considered for the GATE digitizer module. Ten simulation runs were carried out for both GATE simulation and GGEMS, with one million annihilation photon pairs simulated for each run. Since the GGEMS model had already been validated for photon tracking inside the phantom [19], the validation in this work focused on the newly added functionalities, including the physical interactions inside the detector blocks, the recording of single events, and the sorting of coincidence events.

B. Comparison of Computation Time between the GPU and the CPU

To evaluate the computational efficiency, simulation time was compared between GGEMS running on a single GPU of a dual GPU card (Geforce GTX 690 with 1536 CUDA cores for each GPU) with the standard GATE (V7.2) simulation running on a single CPU (Intel Core i7-3770, 3.4 GHz). The number of threads per block was defined by using the CUDA occupancy calculator provided by NVIDIA (256 threads per block in this work). Since modern CPUs normally have a varying number of cores, and the validated standard GATE

implementation usually uses one single core, we also chose one core of the CPU for the GATE simulation. A voxelized human brain phantom based on a dataset from the BrainWeb database [33] was used for the simulation. It was derived from an MR image that was segmented into different tissue classes. For the PET simulation, different activity concentrations were assigned to the grey and white matter compartments with a ratio of 4:1. In the attenuation volume, four tissue classes (air, brain (soft tissue including gray matter, white matter, skin etc.), water (cerebrospinal fluid) and skull) were taken into account. Both the activity and attenuation image datasets consisted of 210 x 210 x 153 voxels with a voxel size of
1.25 1.25 1.25 mm^3.

C. Effect of Detector Modeling in the Scatter Estimation

The quantitative effect of detector modeling of GGEMS for scatter estimation was evaluated using a cylindrical phantom with different sized spheres. The NEMA NU 2-2012 protocol [34] employs a thorax-shaped phantom which includes a cylindrical insert approximating the lung and a cylinder with hot and cold spheres of different sizes to simulate lesions. Since this phantom is too large to be used in the BrainPET scanner, it is replaced by a

smaller, custom-made cylinder containing only hot and cold spheres (hereafter referred to as "sphere phantom"). Compared to the NEMA standard, this phantom has 25% smaller dimensions, resulting in a cylinder of 150 mm in diameter and spheres of 27 mm, 22 mm, 17 mm, 13 mm, 10 mm and 8 mm in diameter. Two validations were carried out to test the impact of the detector modeling on the scatter estimation.

1) Detector Modeling Validation using Simulations: According to the NEMA protocol, the two largest spheres were filled with non-active water, while the activity ratios between the four smallest spheres and the background were set to 8:1. Three simulations were carried out using GGEMS. The first simulation (S_1) included the detector modeling, while the second one (S_2) did not, both of which with the simulation number of 4.8×10^{10}. The third simulation (S_3) which was taken as the ground truth, included the detector modeling, with simulated photons of 2.4×10^{11}. Two image reconstructions were carried out using the vendor provided OP-OSEM scheme with 2 subsets and 64 iterations for each subset, in which the prompt data (without random events) were from S_3, while the scatter estimations were from S_1 and S_2, respectively. Both the scatter distributions from S_1 and S_2 were scaled to S_3 according to their total counts of the detected true events (scattered and unscattered events).

For data analysis, according to the NEMA protocol, circular ROIs of the known size of the spheres were drawn on the transaxial image around the sphere centers. For each sphere, three ROIs with the same size were drawn in the background of the same image plane (n) and in image planes n±1 and n±2, resulting in 15 background ROIs for each sphere size (see Fig. 3). Contrast recovery coefficients (CRCs) were calculated for the cold (CRC_{cold}) and hot (CRC_{hot}) spheres according to (2) and (3)

$$CRChot = \frac{\left(\frac{C_{hot}}{C_{background}}\right) - 1}{\left(\frac{A_{hot}}{A_{background}}\right) 1} \quad (2)$$

$$CRC_{coid} = 1 - \frac{C\,C^{cold}}{C_{background}} \quad (3)$$

respectively. C_{hot}, C_{cold} and $C_{background}$ are the activity concentrations recovered in the reconstructed image for hot spheres, cold spheres and the background, respectively. A_{hot} and $A_{background}$ represent the activity concentrations filled in the hot spheres and the background.

2) Detector Modeling Validation using Measurements: A phantom measurement which was accordant with the simulation mentioned above was implemented. In the measurement, ^{18}F solution was filled into the four hot spheres and the background with the activity concentration of 74.1 *kB q/ml* and 8.6 *kBq/ml*, respectively (ratio: 8.6:1). The phantom was placed inside the MR coil and both MRI and PET data were acquired simultaneously. Acquisition time was 30 minutes with 5.6×10^8 prompt events. Since the diameter of the tubule connecting to the sphere was very small, bubbles were shown to be present in the spheres despite repeated refill attempts. Due to radioprotection consideration for the technical staff, we decided to use the measurements with bubbles. For this, we manually delineated the regions in the spheres to exclude the bubbles (effective region for the rest of the manuscript) on the MR image and applied these regions to the PET image, which was previously co-registered to the MR image (Fig. 3). Images were reconstructed using all corrections in the OP- OSEM scheme with 2 subsets and 64 iterations for each subset. Two scatter estimates (with and without detector modeling) of GGEMS were used in the image reconstruction, both with the simulation of 4.8×10^{10} annihilation photon pairs. As for the attenuation map performed on the ECAT HR+ PET scanner, due to the existence of the air bubbles, the position of the phantom for the transmission scan was the same as in the emission scan. The reconstructed attenuation map was then aligned to the reconstructed emission image.

CRCs for both the real regions which included the air bubbles and the effective regions which excluded the bubbles were computed. The diameters for the circular background ROIs corresponding to the effective regions of the spheres were calculated according to the pixel number of the effective regions.

D. GGEMS Validation against SSS in Phantom Studies

Phantom measurements were carried out to evaluate and validate the new correction procedure using the MR-BrainPET. In III.D.1 and III.D.2, reconstructed images using GGEMS and SSS based scatter corrections were compared employing a cylindrical phantom with three cold cylinders and the sphere phantom. In III.D.3, the sphere phantom was additionally applied to evaluate the simulation time required to obtain stable results. In order to accurately mimic the clinical situation for PET/MRI measurements, the phantoms were placed inside the MR head coils for all measurements.

1) Cylindrical Phantom with Cold Cylindrical Inserts: For the quality assessment of the scatter correction, a measurement using a

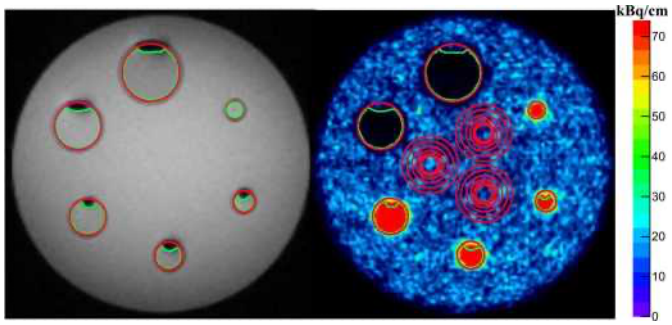


Fig. 3. Transaxial image of the co-registered T1-weighted MR image (left) and the PET image (right) of the cylindrical phantom with spheres. Effective regions (green) were manually delineated within the spheres (red) to exclude air bubbles. ROIs in the background have the same number of pixels as the corresponding spheres.

cylindrical phantom with three cold cylindrical inserts was carried out. The phantom was 15.0 cm in diameter and 25.5 cm in length. The three cold cylindrical inserts consisting of air, water and polytetrafluoroethylene (PTFE), have a diameter of 4 . 0 cm and a length of 19 . 0 cm (Fig. 4 (a)). The wall of the phantom and the inserts was made of poly methyl methacrylate (PMMA) with a thickness of 5.0 mm and 2.0 mm, respectively. The phantom was filled with 85 MBq of ^{18}F solution, and the acquisition time was 15 minutes with 9.3 x 10^8 total coincidence events. Images were reconstructed using the OP-OSEM algorithm with 2 subsets and 64 iterations for each subset, including all standard corrections. For the GGEMS-based scatter estimation, 6.0x 10^{10} annihilation photon pairs were simulated. Twelve circular regions of interest (ROIs) of 22.5 mm in diameter were drawn in the background of the phantom and at the center of the inserts for 90 transaxial planes at the center of the reconstructed image. Since the insert regions did not contain radioactivity, the reconstructed residual activity within these regions originated from the inaccuracies of the scatter corrections. The inaccuracy can be assessed by relating the activity concentration of the insert regions C_{insert} to that of the background regions $C_{background}$ in the image (Fig. 4 (c)) according to (4)

where RE (Residual Error) indicates the relative erroneous activity concentration in the cold insert regions caused by the misplacement of scattered events.

Besides, radial distributions of scatter events in the projection

$$RE = \frac{C_{insert}}{C_{background}} \quad (4)$$

space from GGEMS (including total, single and multiple scatter events) and SSS are also demonstrated.

2) Cylindrical Phantom with Hot and Cold Spheres: The sphere phantom was also applied to compare the SSS and GGEMS based scatter estimations. According to the NEMA protocol, the two largest spheres were filled with non-active water, while the activity ratios between the four smallest spheres and the background were resulted to be 8.6:1 (M_1) and 4.4:1 (M_2) due to the manual procedure. The detail of M_1

is introduced in III.C.2. For M_2, the activity concentrations of the four hot spheres and the background were 45.6 kBq/ml and 10.3 kBq/ml, respectively. The acquisition time was 30 minutes with 5.6 x 10^8 prompt events. Scatter estimates of SSS were calculated from the vendor provided SSS algorithm, and the GGEMS-based scatter correction was based on the simulation of 4.8 x 10^{10} annihilation

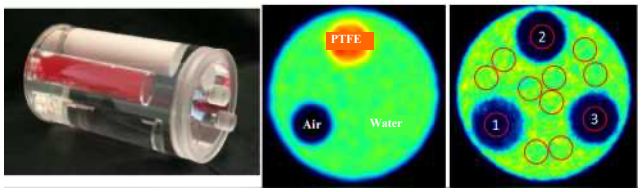


Fig. 4. Cylindrical phantom with three cold inserts. (a) Phantom picture, the water insert was dyed to improve its visibility; (b) Attenuation map; (c) ROIs used for analysis on one transaxial plane. ROIs 1, 2, and 3 correspond to air, PTFE and water inserts, respectively. The other nine ROIs are background.

photon pairs. The CRCs of the reconstructed images based on SSS and GGEMS scatter corrections were compared using the method mentioned in III.C.

3) Adequate Number of Simulated Photons: The sphere phantom was additionally employed to evaluate the simulation time required to obtain stable results of CRCs for the reconstructed images. There are two adjustable parameters which directly influence the simulation time: (i) the size of the image matrix used for photon tracking and (ii) the total number of simulated photons [4]. Therefore, CRCs were calculated with different numbers of simulated photons, using the sphere phantom with an activity concentration ratio of 8.6:1 between the hot spheres and the background. Two image matrices were tested as input data for GGEMS: one consisted of 256 x 256 x 153 voxels with a voxel size of 1.25x 1.25x 1.25 mm^3; the other consisted of 128x 128x 77 voxels with a voxel size of 2.50 x 2.50 x 2.50 mm^3. Both scatter estimates were subsequently employed as input scatter sinograms for the image reconstruction using the vendor provided OP-OSEM scheme with 2 subsets and 64 iterations for each subset. The dimensions and voxel size for all the reconstructed images were the same, i.e., 256x 256x 153 and
1.25 x 1.25 x 1 .25 mm^3, respectively.

E. GGEMS Application for Human Studies

The proposed GGEMS-based scatter correction was applied to two human datasets acquired with different radiotracers, namely ^{18}F-FDG (fluorodeoxyglucose) and ^{18}F-FET (fluoroethyl-L-tyrosine). In the first case, 183 MBq of ^{18}F- FDG was administered to the patient in a single bolus injection and data acquisition (60 min) was started simultaneously with the injection, leading to a total prompt event count of 1.3x 10^9. In the second case, 241 MBq of ^{18}F -FET was administered to a brain tumor patient in a single bolus injection. Again, the data acquisition started simultaneously with the injection and lasted for 50 min with 5.5 x 10^8 measured prompt events acquired. The standard clinical workflow with OP- OSEM reconstruction (2 subsets with 32 iterations for each subset) was used for image reconstruction. Data corrections, including dead time, random, attenuation and scatter correction were applied. Scatter estimates from the vendor-provided SSS and our GGEMS-based implementation (4.8 x 10^{10} simulated photon pairs) were compared using both the sinogram data and the reconstructed images.

Since the ground truth is not known, comparisons between both scatter correction methods, namely SSS and GGEMS, were performed for the qualitative assessment. For the ^{18}F- FET measurement, tumor-to-brain ratios (TBRs) are commonly used in the diagnosis of brain tumors, so differences between GGEMS and

SSS on TBRs were evaluated. The mean TBR (TBR_{mean}) was generated by dividing the mean standardized uptake value (SUV) of ^{18}F—FET in the tumor area (T_{mean}) by the mean value of the normal brain tissue (B_{mean}) for the time frame of 20-40 min after injection. B_{mean} was calculated with the mean value of a large oval volume-of-interest (VOI) placed in healthy brain tissue on the contralateral hemisphere including white and grey matter. For T_{mean}, a 3D auto-contouring process using a TBR of 1.6 was applied (Fig. 5) [35].

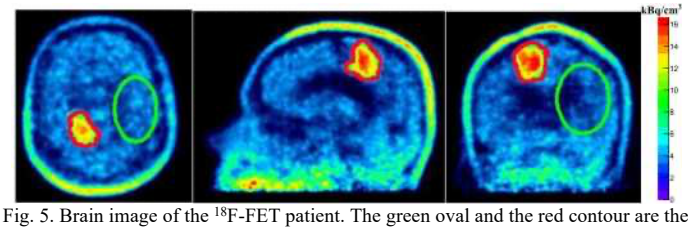


Fig. 5. Brain image of the ^{18}F-FET patient. The green oval and the red contour are the VOIs of background and the tumor, respectively.

IV. Results

A. Accuracy Validation of GGEMS

As shown in Table I, there is a very good agreement between GGEMS and GATE for both photoelectric and Compton interactions regarding the average values and the standard deviations. The small observed differences of 0.12% for photoelectric interactions and 0.06% for Compton interactions are well within the expected statistical variations. This validates the accuracy of the physical processes of the new GPU implementation within the detector module. A small discrepancy of 0.86% can be observed for the detected single events. This translates into a difference of 1.79% for true coincidence events. Table I also lists the number of events undergoing Compton scattering within the water phantom prior to their detection. The scatter fractions provided by GATE and GGEMS were also found to be in good agreement (11.69% vs. 11.63% for single events and 23.01% vs. 22.96% for coincidence events). Similarly, there is a good agreement for the random fraction, which was very low in this example (2.17% vs. 2.19%).

B. Comparison of Computation Time between the GPU and the CPU

The simulation time of one million annihilation photon pairs for the GGEMS on a single GPU was 1.5 seconds, while for the GATE simulation on one CPU core, the simulation time was 1165 seconds, resulting into a speedup factor of 776.

TABLE I
AVERAGE (AVE) AND STANDARD DEVIATION (STD) OF INTERACTIONS INSIDE THE DETECTOR, DETECTED SINGLES AND COINCIDENCES OBTAINED FROM 10 SIMULATIONS. (*RELATIVE TO GATE)

		GGEMS	GATE	Diff.
PE Interactions	ave	530436	531070	-634 (-0.12 %)*
	std	604	880	
Compton Interactions	ave	422147	422386	-239 (-0.06 %)*
	std	812	980	
Single events	ave	299344	296788	+2556 (+0.86 %)*
	std	388	472	
Scattered singles	ave	34828	34684	+144 (+0.41 %)*
	std	230	184	
True coincidence	ave	30988	30442	+546 (+1.79 %)*
	std	188	126	
Scatter coincidence	ave	7116	7006	+110 (+1.58 %)*
	std	94	83	
Random coincidence	ave	679	660	+19 (+2.86 %)*
	std	37	27	

C. Effect of Detector Modeling in the Scatter Estimation

1) Detector Modeling Validation using Simulations: The CRCs of the reconstructed images using scatter estimates with and without detector modeling are summarized in Table II. For the GGEMS with detector modeling, the CRCs have an obvious improvement of between 5.7% and 16.1% for all the spheres comparing to those without detector modeling.

TABLE II
CRC COMPARISON OF SIMULATION RESULTS FOR GGEMS WITH (CRC_D) AND WITHOUT (CRC_{ND}) DETECTOR MODELING.

Spheres	CRC_d (%)	CRC_{nd} (%)	Diff. (p.p.)
27 mm cold	83.2 ± 0.6	75.2 ± 1.1	8.0
22 mm cold	79.6 ± 1.0	73.9 ± 1.1	5.7
17 mm hot	91.2 ± 5.6	75.1 ± 3.9	16.1
13 mm hot	82.7 ± 7.0	66.6 ± 5.2	16.1
10 mm hot	77.0 ± 8.5	62.6 ± 6.7	14.4
8 mm hot	69.1 ± 7.9	55.9 ± 6.4	13.2

2) Detector Modeling Validation using Measurements: For the phantom measurement, the CRCs of the reconstructed images with detector modeling also outperform those without detector modeling (5.7%-18.3%), especially for the hot spheres as shown in Table III.

D. GGEMS Validation against SSS in Phantom Studies

1) Cylindrical Phantom with Cold Cylindrical Inserts:

The result of the comparison between GGEMS and SSS based scatter corrections for this experiment is summarized in Table IV. Quantitatively, the residual error of the GGEMS- based method was smaller than that of the SSS approach for all three inserts, although the difference is rather small. Besides, for both methods, the residual intensity of the air insert was much higher than that in the other two inserts. In Fig. 6, the radial distributions of the scatter events from GGEMS (including total, single and multiple scatter events) and SSS are shown. Comparing to SSS, both single scatter events and multiple scatter events are taken into account in GGEMS.

Multiple scatter is more likely near the center of the phantom, and this results in a larger discrepancy between GGEMS and SSS in the center. However, because the phantom diameter is small and its structure is simple, the scatter estimation from SSS is in good consensus to GGEMS, with a difference of only 1.4% for the total scatter counts.

TABLE III
CRC COMPARISON OF MEASUREMENT RESULTS FOR GGEMS WITH (CRC_D) AND WITHOUT (CRC_{ND}) DETECTOR MODELING. BOTH THE RESULTS FOR THE REAL REGIONS WITH BUBBLES INSIDE THE SPHERES AND THE EFFECTIVE REGIONS WITHOUT BUBBLES ARE GIVEN.

Measurements	Spheres	Real Region			Effective Region		
		CRC_d (%)	CRC_{nd} (%)	Diff. (p.p.)	CRC_d(%)	CRC_{nd} (%)	Diff. (p.p.)
M(8.6:1)	27 mm cold	84.8 ± 0.5	79.1 ± 2.2	5.7	88.5 ± 0.4	82.5 ± 2.0	6.0
	22 mm cold	81.7 ± 0.8	74.9 ± 2.7	6.8	85.9 ± 0.6	78.1 ± 2.5	7.9
	17 mm hot	89.9 ± &0	71.6 ± 9.9	1&3	93.2 ± 7.7	76.3 ± 11.5	16.9
	13 mm hot	77.6 ± 10.9	61.8 ± 10.5	15.8	86.6 ± 11.0	71.4 ± 13.6	15.2
	10 mm hot	6&3 ± 13.2	52.8 ± 10.6	15.5	82.4 ± 12.5	66.3 ± 15.1	16.1
	8 mm hot	75.8 ± 16.7	59.7 ± 12.3	16.1	77.5 ± 13.6	63.9 ± 15.3	13.5

TABLE IV
COMPARISON OF GGEMS AND SSS . GIVEN VALUES OF RE INDICATE THE MAGNITUDE OF THE RELATIVE ERRONEOUS ACTIVITY CONCENTRATION IN THE COLD REGIONS (IDEALLY ZERO).

	RE_{GGEMS} (%)	RE_{SSS} (%)
Air	10.4 ± 0.8	11.4 ± 0.7
Water	4.1 ± 0.7	4.3 ± 0.5
PTFE	3.0 ± 0.6	3.1 ± 0.4

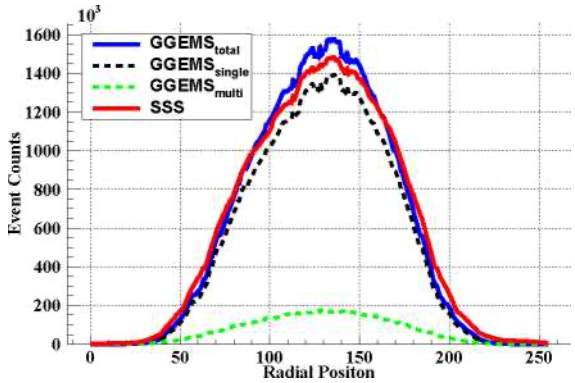


Fig. 6. Radial distribution of scatter sinograms estimated by GGEMS and SSS

2) Cylindrical Phantom with Hot and Cold Spheres: The CRCs of the reconstructed images using GGEMS and SSS based scatter corrections are summarized in Table V, where both the real regions impaired by air bubbles and the effective regions excluding air bubbles are shown. For the cold spheres, the CRCs are similar for both methods. In contrast, for the hot spheres, an improvement of between 4.7% and 11.7% is found for the GGEMS-based scatter correction. Although the bubbles in the spheres affect the CRC values, the improvement in the CRCs for the hot spheres when using GGEMS is consistent in both cases, with and without correction for bubbles. The results for the smaller spheres (8 mm and 10 mm) should be interpreted carefully because of the small number of pixels in the corresponding ROIs.

3) Adequate Number of Simulated Photons: Fig. 7 shows the dependency of the CRCs on the simulated photon numbers. The CRC values of the cold spheres exhibit faster convergence and stabilize at around 1.2 x 10^{10} simulated photon pairs, while the CRC values of the hot spheres stabilize at around 3.0 x 10^{10} simulated photon pairs[2]. An overall comparison, including the effect of the image matrix size, is given in Fig. 8 for the hot spheres. It is shown that the CRCs have an improvement for GGEMS over SSS starting from 1.2 x 10^{10} simulated photon pairs. The reduced image matrix and the full matrix show similar results with the largest discrepancy in the smallest sphere (8 mm), which is not as clearly resolved in the case of 2.5 x 2.5 x 2.5 mm^3 voxels compared to the case of 1.25 x 1.25 x 1.25 mm^3 voxels. The average simulation time on a single GPU is 1.1 s and 0.7 s per million photon pairs for the full image matrix and the reduced matrix, respectively.

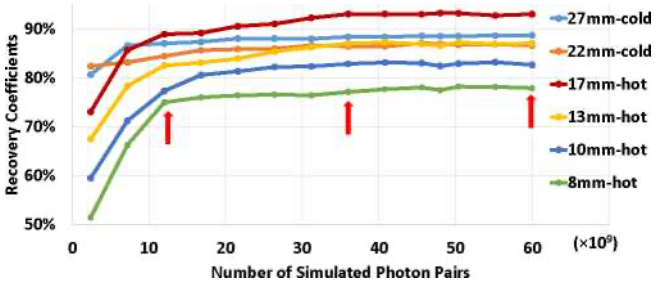


Fig. 7. Dependence of CRC values on simulation statistics for the sphere phantom.

E. GGEMS Application for Human Studies

For the ^{18}F-FDG PET human measurement, the axial and radial distributions of the scatter sinograms generated by GGEMS and SSS as well as the prompt data (after random and normalization correction) are shown in Fig. 9. The event counts for both the axial and radial distributions are the summation of counts from all the planes (cross and direct planes) of the sinogram. The total number of scatter events for GGEMS is 5.5% higher than that obtained with SSS. In the central regions (between the blue dashed lines), the difference is as high as 13.2%. The differences of the scatter

[2]Transverse view of the scatter sinograms with different simulation statistics can be found in the supplementary files. Readers can get an impression on the noise for different statistics

TABLE V

COMPARISON OF CRCS FOR BOTH SCATTER CORRECTION APPROACHES. RESULTS OF TWO MEASUREMENTS (M_1 AND M_2) WITH DIFFERENT SPHERE-TO-BACKGROUND ACTIVITY CONCENTRATION RATIOS ARE GIVEN. FOR EACH MEASUREMENT, BOTH THE RESULTS FOR THE REAL REGIONS WITH BUBBLES INSIDE THE SPHERES AND THE EFFECTIVE REGIONS WITHOUT BUBBLES ARE GIVEN.

Measurements	Spheres	Real Region			Effective Region		
		CRC_{GGEMS} (%)	CRC_{SSS} (%)	Diff. (p.p.)	CRC_{GGEMS} (%)	CRC_{SSS} (%)	Diff. (p.p.)
M_1(&6:1)	27 mm cold	84.8 ± 0.5	85.0 ± 0.5	-0.2	88.5 ± 0.4	88.0 ± 0.4	0.5
	22 mm cold	81.7 ± 0.8	81.8 ± 0.7	-0.1	85.9 ± 0.6	85.0 ± 0.9	0.9
	17 mm hot	89.9 ± &0	80.4 ± 6.2	9.5	93.2 ± 7.7	83.2 ± 7.2	10
	13 mm hot	77.6 ± 10.9	69.3 ± 8.5	8.3	86.6 ± 11.0	77.0 ± 12.1	9.6
	10 mm hot	6&3 ± 13.2	61.9 ± 9.2	6.4	82.4 ± 12.5	74.9 ± 13.2	7.5
	8 mm hot	75.8 ± 16.7	66.9 ± 11.0	8.9	77.5 ± 13.6	69.2 ± 12.6	8.3
M_2 (4.4:1)	27 mm cold	85.7 ± 0.4	85.2 ± 0.4	0.5	89.3 ± 0.4	88.5 ± 0.4	0.8
	22 mm cold	80.2 ± 1.0	80.9 ± 0.8	-0.7	84.0 ± 0.9	84.2 ± 0.7	-0.2
	17 mm hot	79.6 ± 7.1	70.5 ± 5.6	9.1	88.8 ± 8.6	80.9 ± 7.5	7.9
	13 mm hot	72.4 ± &1	62.2 ± 7.0	10.2	84.4 ± 9.7	75.7 ± &9	8.7
	10 mm hot	64.0 ± &6	56.6 ± 7.5	7.4	80.0 ± 10.1	68.3 ± &6	11.7
	8 mm hot	67.5 ± 11.2	64.7 ± 9.2	2.8	76.3 ± 12.2	71.6 ± 9.6	4.7

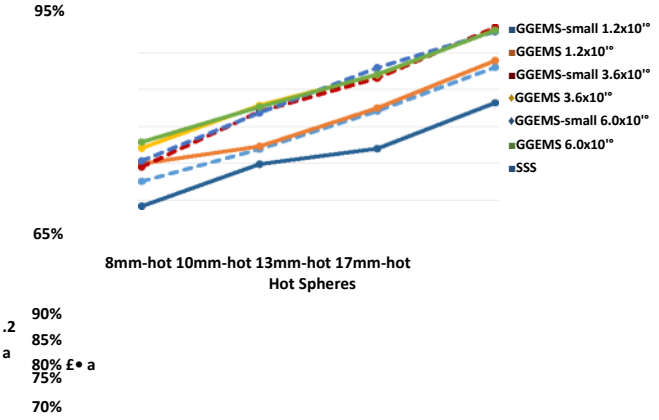


Fig. 8. Dependence of CRC values on simulated photon numbers(red arrows in Fig. 7) and image matrix size for the hot spheres.

estimation give rise to the discrepancy in the reconstructed images. Fig. 10 shows a pair of transaxial slices of the reconstructed images corrected with GGEMS and SSS based scatter corrections. The activity concentration of the image with the GGEMS-based scatter correction is 6.2% lower than that of SSS. The profile plots of a rectangular region (red region on Fig. 10 (a) and Fig. 10 (b)) are demonstrated in Fig. 10 (c), showing a higher contrast for the GGEMS-based scatter correction. The discrepancy of the activity concentration in this region is 9.8%.

Fig. 11 shows a pair of transversal slices of the reconstructed image with GGEMS and SSS based scatter correction for the ^{18}F-FET PET study. The mean activity concentration of the reconstructed image with the GGEMS-based scatter correction is 7.8% lower than that obtained with the SSS-based correction. The profile plots of a rectangular region (red region on Fig. 11 (a) and Fig. 11 (b)) are visualized in Fig. 11 (c) with an activity difference of 10.9%. The TBR_{mean} for the GGEMS-based scatter correction was 3.1 with the tumor volume of 18.1 ml, while for the SSS the TBR_{mean} was 2.8 with the tumor volume of 17.4 ml.

V. DISCUSSION

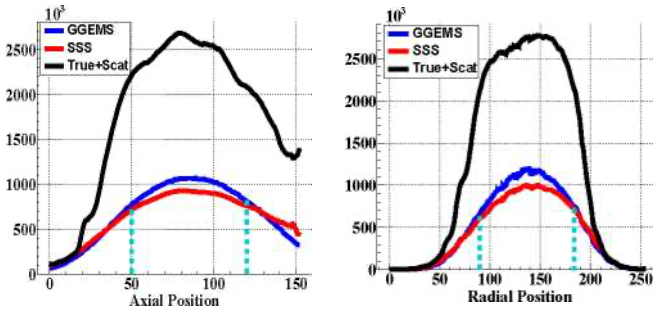
The main challenge of the Monte Carlo simulation for medical applications is the long computational time [4].

(a) (b)

Fig. 9. Axial(a) and radial(b) distribution of the coincidence events in the scatter and prompt sinograms generated by GGEMS and SSS for the ^{18}F-FDG study.

Fortunately, the simulation of ionizing radiation is perfectly suited for the parallel computation of GPUs. This is why implementations of GPU-based Monte Carlo simulation for medical applications have been pursued in recent years [16][21], [36]-[39]. Amongst all of these works, the GGEMS framework [19] is the only one which targets both imaging and therapeutic applications based on the well-validated Geant4 libraries. However, previous implementations of the GGEMS did not include a full simulation of the detection procedure in the scanner. In this work, the full simulation based on the framework of GGEMS, including the detection module and the associated coincidence-sorting algorithm, was implemented and validated for PET imaging.

There was a very good agreement between GATE and GGEMS simulation for both photoelectric and Compton interactions in the detection module, validating the accuracy of the detection procedure implemented in this work. Because of the neglect of the secondary electrons in the GPU implementation, there was a small discrepancy in the count of detected events. However, this does not affect the event positioning or the ratios of different types of coincidence events. This approximation only leads to a small global increase (0.86%) in the detected



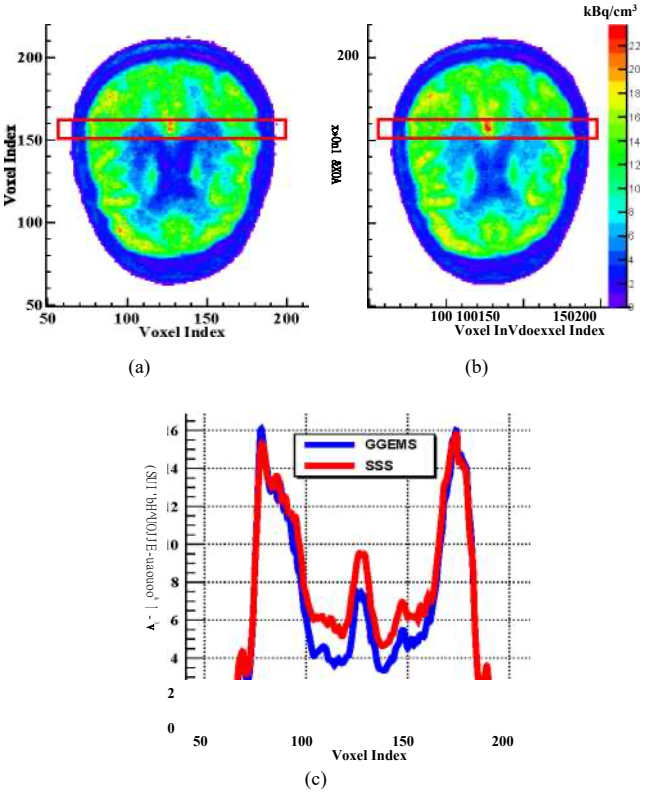


Fig. 10. A pair of transversal slices of the ^{18}F-FDG brain images corrected with (a) GGEMS -based and (b) SSS-based scatter estimates. (c) The profile plots of the marked region (red rectangle).

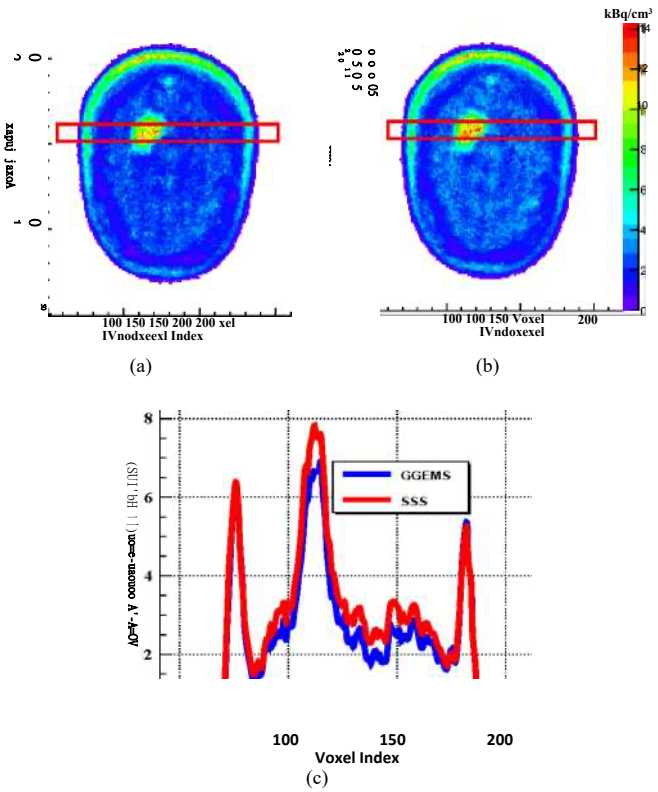


Fig. 11. A pair of transaxial slices of the ^{18}F-FET brain images corrected with (a) GGEMS -based and (b) SSS-based scatter estimates. (c) The profile plots of the marked region (red rectangle).

events, which is not relevant for most imaging applications and can therefore be considered a reasonable approximation for the application of scatter correction targeted in this work.

In terms of computational efficiency, the simulation time on a single GPU was compared with a single CPU core, resulting in a speedup factor of 776. Modern CPUs usually have several CPU cores for each CPU socket. For example, the CPU used in this work has 4 CPU cores. If all the CPU cores are used for the GATE simulation, the speed will be improved by about 3 times. In order to produce comparable output data, it was necessary to consider secondary electrons in the tracking part of the detection module for the GATE simulation. This makes the direct time comparison difficult, as the secondary electrons were neglected in the GGEMS. However, the tracking time of secondary electrons can be evaluated by GATE using only photoelectric and Compton effects in the phantom. For PET imaging, the tracking time of the secondary electrons only^{V} takes a very small fraction (< 2%) of the total simulation time. Since the proposed implementation is fully parallel, a linear reduction in simulation time can be obtained by distributing the task to more GPUs. The Geforce GTX 690 used in this work already contains two GPUs on one card, which may be used in parallel, and standard desktop computers can already be fitted with 4-6 GPUs. Furthermore, more recent generations of GPU can also provide significant improvements in speedup.

Most existing scatter estimations based on GPU-accelerated MCS do not take into account the physical effects within the detectors, using exclusively the ray tracing based on the geometrical description of the scanner. Through the comparison of the scatter estimation with and without detector modeling, both the simulation and measurement results show that the detector modeling has an obvious improvement for the image quantification. So the physically accurate modeling of the detection processes is essential to the MCS based scatter estimation in PET systems.

For the measurement of a cylindrical phantom with three cold inserts, the improvement of the RE of the cold inserts for GGEMS-based scatter correction is not obvious when compared to the SSS. Because the structure of this phantom is simple and the diameter is small, it is possible for SSS to get an accurate scatter estimation. This can also be observed dorectly in the profile of the scatter sinograms from GGEMS and SSS, which shows negligible discrepancies. For both approaches, the RE for the air insert is much larger than that for the water and PTFE inserts. This may be caused by three factors: first, the inaccuracy of the scatter correction; second, positrons which pass through the insert wall (PMMA, 2 mm thickness) annihilating with electrons inside the air compartment; third, the inaccuracy of the attenuation map acquired by the transmission scan. Since the attenuation map acquired on the ECAT HR+ PET scanner suffers from the partial volume effect, the air insert presents higher attenuation coefficients on the edge,

resulting in an overcorrection of the attenuation at its outer rim. On the other hand, the underestimation of the size of the PTFE insert in the attenuation map leads to a slight underestimation of attenuation in this region, since the attenuation coefficient of PTFE is larger than that of the surrounding water. This explains why the RE for PTFE are smaller than those for water. Furthermore, preliminary research has been carried out for the evaluation of these factors. To test the impact of the inaccurate attenuation map, the air insert was manually set with its actual size and attenuation coefficients, and was applied in SSS. In this situation, the RE reduced from 11.4% to 8.6%. To test the effect of the inaccurate scatter estimation, we simulated this measurement with MCS and reconstructed the results with only the true coincidences, resulting in a RE of 7.0%. Further studies are required for testing effects of other factors, such as the positron range.

In the measurement of the sphere phantom, the CRCs of the two cold spheres were shown to be similar between the GGEMS and SSS based scatter corrections. For the hot spheres, however, GGEMS shows superior CRCs with improvements of up to 11.7%. This indicates a potential improvement for lesion detectability compared to SSS-based scatter correction. The CRCs provided by an accurate scatter correction are especially important for truly exploiting the capabilities of high-resolution scanners, such as the BrainPET.

The dependence of the correction accuracy on simulated photon number and voxel sizes was also investigated using the sphere phantom. With a simulation time of 220 minutes on a single GPU, the performance of the GGEMS had already been better than that of SSS. When increasing the voxel size from 1.25 mm to 2.50 mm, the CRCs for the 8 mm hot sphere showed degradation. Consequently, down-sampling of the input images, as often applied in MCS, should be carefully considered [4], [11]. Using a workstation equipped with two dual GPUs, the proposed method can be applied with a simulation time of less than one hour, making it feasible for use in research-oriented clinical usage. With the rapidly increasing computing power of modern GPUs, we believe that this approach is not far from being implemented in routine clinical practice.

In terms of human studies, differences between reconstructed images with the GGEMS and SSS based scatter correction were also observed, both for the activity concentration and image contrast. For the ^{18}F-FDG patient, there was a larger difference in the central part of the brain both in the projection space and in the image space, resulting in a higher contrast for the GGEMS-based scatter correction. For the tumor patient administered with ^{18}F-FET, different values of the TBR_{mean} and tumor volumes were obtained for the GGEMS and SSS based scatter corrections. The mere existence of differences demonstrates the clinical relevance of an accurate scatter correction method. Because the ground truth is not known for patients, it is not clear which method provides more accurate estimates for these values. However, GGEMS-based scatter correction is expected to be more accurate, since this was also the case for the recovery coefficients in the phantom study. This requires further evaluation in a future study. In this paper, only one ^{18}F-FDG patient and one ^{18}F-FET tumor patient were presented. However, in the work as a whole, eight ^{18}F-FDG patients and four ^{18}F-FET tumor patients were compared in total, the results from which all showed a similar trend.

Currently, there are still some remaining limitations of the GGEMS implementation. These include the lack of explicit source characteristics which only allows "back-to-back" photon simulations, and do not incorporate more elaborate digitizer modules which would better reflect the electronics characteristics of certain PET scanners. However, for many cases, the basic configuration provided by this implementation is sufficient and additional modules may be added based on this work. Furthermore, the scaling strategies for our proposed approach and the vendor-provided SSS were different. For our approach, one scaling factor for the whole dataset was used, resulting into a more robust process, especially for the dynamic measurement. In contrast, the SSS used a series of scaling factors for each single plane of the sinogram which was more accurate for the correction of the out-field-of-view scattering events. For future work, different scaling strategies will be studied in order to gain the optimal scattering estimation, and the presented approach will be considered for integration into a GPU based image reconstruction framework for PET imaging. Besides, there is still some space for further speedup of GGEMS, such as the proper usage of the shared memory and the algorithm optimization for the detector modeling.

The proposed method has currently been applied to brain imaging from PET/MRI studies obtained with the BrainPET. However, there are no obvious limitations to applying the proposed approach for whole body PET imaging. It is also not restricted to PET/MRI imaging and can equally be applied to PET/CT imaging. In addition, with the development of TOF PET imaging, it is also important to estimate the scatter distribution in the TOF direction. For GGEMS, it can also acquire the information of the TOF and can hence be used in TOF PET systems.

VI. Conclusion

The aim of this work was to develop a full Monte Carlo simulation based on GPU-acceleration for PET applications. This allows physically accurate scatter estimations within a reasonable computational time, compatible with clinical requirements. The presented results confirm the potential of this approach. The new proposed method outperforms the standard SSS method with respect to recovery coefficient and correction accuracy in phantom measurements. Finally, the representative patient examples demonstrate the applicability of the method in the image reconstruction of real patient data. The differences in TBRs and tumor volumes demonstrate the clinical relevance of applying accurate scatter corrections.

References

[1] H. Zaidi, "Comparative evaluation of scatter correction techniques in 3D positron emission tomography," Eur. J. Nucl. Med., vol. 27, no. 12, pp. 1813—1826, 2000.

[2] C. H. Holdsworth, R. D. Badawi, P. Santos, A. D. Van den Abbeele, E. J. Hoffman, and G. El Fakhri, "Evaluation of a Monte Carlo scatter correction in clinical 3D PET," *IEEENucl Sci. Conf. R.,* pp. 2540—2544, 2004.

[3] I. Castiglioni, O. Cremonesi, M. C. Gilardi, V. Bettinardi, G. Rizzo, A. Savi, E. Bellotti, and F. Fazio, "Scatter correction techniques in 3D PET: A Monte Carlo evaluation," IEEE Trans. Nucl. Sci., vol. 46, no. 6, pp. 2053—205& 1999.

[4] C. S. Levin, M. Dahlbom, and E. J. Hoffman, "A Monte-Carlo correction for the effect of Compton-scattering in 3-D PET brain imaging," IEEE *Trans. Nucl. Sci.,* vol. 42, no. 4, pp. 1181-1185, 1995.

[5] S. Mansor, R. Boellaard, M. C. Huisman, B. van Berckel, R. C. Schuit, A. D. Windhorst, A. A. Lammertsma, and F. van Velden, "Impact of new scatter correction strategies on high-resolution research tomograph brain PET studies," *Mol. Imaging Biol.*, vol. 18, no. 4, pp. 627-635, 2016.

[6] J. M. Ollinger, "Model-based scatter correction for fully 3D PET," *Phys. Med. Biol.*, vol. 41, no. 1, pp. 153-176, 1996.

[7] K. Magota, T. Shiga, Y. Asano, D. Shinyama, J. H. Ye, A. E. Perkins, P. J. Maniawski, T. Toyonaga, K. Kobayashi *et al.*, "Scatter correction with combined single-scatter simulation and Monte Carlo simulation scaling improved the visual artifacts and quantification in 3-Dimensional brain PET/CT imaging with O-15-gas inhalation," *J. Nucl. Med.*, vol. 58, no. 12, pp. 2020-2025, 2017.

[8] J. H. Ye, X. Y. Song, and Z. Q. Hu, "Scatter correction with combined single-scatter simulation and Monte Carlo simulation for 3D PET," *IEEE Nucl. Sci. Conf. R.*, 2014.

[9] J. Teuho, V. Saunavaara, T. Tolvanen, T. Tuokkola, A. Karlsson, J. Tuisku, and M. Teras, "Quantitative evaluation of 2 scatter-correction techniques for F-18-FDG brain PET/MRI in regard to MR-based attenuation correction," *J. Nucl. Med.*, vol. 58, no. 10, pp. 1691-1698, 2017.

[10] C. H. Holdsworth, C. S. Levin, T. H. Farquhar, M. Dahlbom, and E. J. Hoffman, "Investigation of accelerated Monte Carlo techniques for PET simulation and 3D PET scatter correction," *IEEE Trans. Nucl. Sci.*, vol. 48, no. 1, pp. 74-81, 2001.

[11] C. H. Holdsworth, C. S. Levin, M. Janecek, M. Dahlbom, and E. J. Hoffman, "Performance analysis of an improved 3-D PET Monte Carlo simulation and scatter correction," *IEEE Trans. Nucl. Sci.*, vol. 49, no. 1, pp. 83-89, 2002.

[12] O. Barret, T. A. Carpenter, J. C. Clark, R. E. Ansorge, and T. D. Fryer, "Monte Carlo simulation and scatter correction of the GE advance PET scanner with SimSET and Geant4," *Phys. Med. Biol.*, vol. 50, no. 20, pp. 4823-4840, 2005.

[13] J. Bert and D. Visvikis, "A fast CPU/GPU ray projector for fully 3D listmode PET reconstruction," *IEEE Nucl. Sci. Conf. R.*, pp. 4126-4130, 2011.

[14] P. Despres and X. Jia, "A review of GPU-based medical image reconstruction," *Phys. Medica.*, vol. 42, pp. 76-92, 2017.

[15] A. Autret, M. Moreau, T. Carlier, J. Bert, O. Strauss, and D. Visvikis, "Detector modeling in PET list-mode reconstruction: comparison between pre-calculated and on-the-fly computed system matrix," *IEEE Nucl. Sci. Conf. R.*, 2015.

[16] J. Lippuner and I. A. Elbakri, "A GPU implementation of EGSnrc's Monte Carlo photon transport for imaging applications," *Phys. Med. Biol.*, vol. 56, no. 22, pp. 7145-7162, 2011.

[17] L. Jahnke, J. Fleckenstein, F. Wenz, and J. Hesser, "GMC: a GPU implementation of a Monte Carlo dose calculation based on Geant4," *Phys. Med. Biol.*, vol. 57, no. 5, pp. 1217-1229, 2012.

[18] K. S. Kim, Y. D. Son, Z. H. Cho, J. B. Ra, and J. C. Ye, "Ultra-fast hybrid CPU-GPU multiple scatter simulation for 3-D PET," *IEEE J. Biomed. Health*, vol. 18, no. 1, pp. 148-156, 2014.

[19] J. Bert, H. Perez-Ponce, Z. El Bitar, S. Jan, Y. Boursier, D. Vintache, A. Bonissent, C. Morel, D. Brasse, and D. Visvikis, "Geant4-based Monte Carlo simulations on GPU for medical applications," *Phys. Med. Biol.*, vol. 58, no. 16, pp. 5593-5611, 2013.

[20] J. Bert, H. Perez-Ponce, S. Jan, Z. El Bitar, P. Gueth, V. CupJov, H. Chekatt, D. Benoit, D. Sarrut *et al.*, "Hybrid GATE: A GPU/CPU implementation for imaging and therapy applications," *IEEE Nucl. Sci. Conf. R.*, pp. 2247-2250, 2012.

[21] M. Gaens, J. Bert, U. Pietrzyk, N. J. Shah, and D. Visvikis, "GPU- accelerated Monte Carlo based scatter correction in brain PET/MR," *IEEE Nucl. Sci. Conf. R.*, 2013.

[22] M. E. Gaens, "Monte Carlo simulation for scanner design and correction methods in PET and PET/MRI," PhD Thesis, University of Wuppertal, 2014. [Online]. Available: https://www.digibib.net/metasearch

[23] H. Herzog, K. J. Langen, C. Weirich, E. Rota Kops, J. Kaffanke, L. Tellmann, J. Scheins, I. Neuner, G. Stoffels *et al.*, "High resolution BrainPET combined with simultaneous MRI," *Nuklearmed-Nucl. Med.*, vol. 50, no. 2, pp. 74-82, 2011.

[24] "CURAND library programming guide," 2017.07. [Online]. Available: https://docs.nvidia.com/cuda/curand/index.html

[25] S. Agostinelli, J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. Arce, M. Asai, D. Axen, S. Banerjee *et al.*, "Geant4-a simulation toolkit," *Nucl. Instrum. Meth. A*, vol. 506, no. 3, pp. 250-303, 2003.

[26] F. Biggs. and R. Lighthill., "Analytical approximations for X-ray cross sections III," 1988. [Online]. Available: https://www.osti.gov/servlets/ purl/7124946

[27] S. Park, W. L. Rogers, and N. H. Clinthorne, "Effect of recoil electron range on efficiency and on spatial resolution of very high resolution animal PET," *IEEE Nucl. Sci. Conf. R.*, pp. 1756-1759, 2003.

[28] E. Rota Kops, H. Hautzel, H. Herzog, G. Antoch, and N. J. Shah, "Comparison of template-based versus CT-based attenuation correction for hybrid MR/PET scanners," *IEEE Trans. Nucl. Sci.*, vol. 62, no. 5, pp. 2115-2121, 2015.

[29] R. D. Badawi and P. K. Marsden, "Developments in component-based normalization for 3D PET," *Phys. Med. Biol.*, vol. 44, no. 2, pp. 571594, 1999.

[30] L. G. Byars, M. Sibomana, Z. Burbar, J. Jones, V. Panin, W. C. Barker, J. S. Liow, R. E. Carson, and C. Michel, "Variance reduction on randoms from delayed coincidence histograms for the HRRT," *IEEE Nucl. Sci. Conf. R.*, pp. 2622-2626, 2005.

[31] C. Michel., M. Sibomana., A. Boi., X. Bernard., M. Lonneux., M. De- frise., C. Comtat., P. E. Kinahan., and D. W. Townsend., "Preserving poisson characteristics of PET data with weighted OSEM reconstruction," *IEEE Nucl. Sci. Conf. R.*, p. 1323-9, 1998.

[32] G. Santin, D. Strul, D. Lazaro, L. Simon, M. Krieguer, M. V. Martins, V. Breton, and C. Morel, "GATE: A Geant4-based simulation platform for PET and SPECT integrating movement and time management," *IEEE Trans. Nucl. Sci.*, vol. 50, no. 5, pp. 1516-1521, 2003.

[33] "Brainweb: Simulated brain database." [Online]. Available: http: //brainweb.bic.mni.mcgill.ca/brainweb/

[34] *Performance Measurements of Positron Emission Tomographs (PETs), National Electrical Manufacturers Association. NEMA Stand Publ, 2012, NU 2*.

[35] D. Pauleit, F. Floeth, K. Hamacher, M. J. Riemenschneider, G. Reifen- berger, H. W. Muller, K. Zilles, H. H. Coenen, and K. J. Langen, "O- (2-[F-18]fluoroethyl)-L-tyrosine PET combined with MRI improves the diagnostic assessment of cerebral gliomas," *Brain*, vol. 128, pp. 678687, 2005.

[36] X. Jia, J. Schumann, H. Paganetti, and S. B. Jiang, "GPU-based fast Monte Carlo dose calculation for proton therapy," *Phys. Med. Biol.*, vol. 57, no. 23, pp. 7783-7797, 2012.

[37] E. Alerstam, T. Svensson, and S. Andersson-Engels, "Parallel computing with graphics processing units for high-speed Monte Carlo simulation of photon migration," *J. Biomed. Opt.*, vol. 13, no. 6, 2008.

[38] A. Badal and A. Badano, "AcceleratingMonte Carlo simulations of photon transport in a voxelized geometry using a massively parallel graphics processing unit," *Med. Phys.*, vol. 36, no. 11, pp. 4878-4880, 2009.

[39] Y. Lemarechal, J. Bert, N. Boussion, E. Le Fur, and D. Visvikis, "Monte Carlo simulations on GPU for brachytherapy applications," *IEEE Nucl. Sci. Conf. R.*, 2013.