# Faster PET Reconstruction with Non-Smooth Priors by Randomization and Preconditioning

Matthias J. Ehrhardt, Pawel Markiewicz, Carola-Bibiane Schönlieb

***Abstract*—Uncompressed clinical data from modern positron emission tomography (PET) scanners are very large, exceeding 300 million data points. The last decades have seen tremendous advancements in mathematical imaging tools many of which lead to non-smooth optimization problems. Most of these tools have not been translated to clinical PET data, as the algorithms for non-smooth problems do not scale well enough for large data sets. In this work, inspired by big data machine learning applications, we use advanced randomized optimization algorithms to solve the PET reconstruction problem for a very large class of non-smooth priors which includes for example total variation, total generalized variation, directional total variation and many constraints. We show on real PET data (FDG and florbetapir) from the Siemens Biograph mMR that a dozen forward (and back) projections are sufficient, thus showing that this algorithm is fast enough to bring these models into clinical practice. Moreover, the proposed algorithm is similarly fast on the unregularized problem as the clinical standard OSEM but—in contrast to OSEM—has provable convergence guarantees, robustness and stability for any subset selection.**

***Index Terms*—positron emission tomography, convex optimization, randomized optimization, non-smooth optimization, total variation, anatomical priors**

## I. Introduction

POSITRON EMISSION tomography (PET) is an important clinical imaging technique as it allows monitoring function of the human body by following a radio-active tracer. The image reconstruction process in PET is challenging as the low number of photon counts call for the Poisson noise modeling and the amount of data is excessively large on modern scanners. While most clinical systems still run non-penalized reconstructions, it has been shown that regularization improves noise control and quantification [1], [2]. In addition, recent trends suggests that non-smooth regularizers are beneficial for imaging applications as they allow smooth variations within regions without oversmoothing sharp boundaries [3], [4], [5], [6], [7], [8], [9], [10], [11]. Only a few algorithms are capable of combining non-smooth priors and the Poisson noise model, e.g. [12], [13], [14], [9].

Modern PET scanners always come with a second anatomical modality such as computed tomography (CT) or magnetic resonance imaging (MRI). The non-smooth regularizers can also be used to either incorporate anatomical knowledge from MRI or CT into the reconstruction, e.g. [15], [16], [17], [18], [19], [20], or to jointly reconstruct PET and the anatomical CT/MRI image [21], [22], [23], [24]. One of the most popular algorithms to solve the resulting non-smooth convex optimization problem is the primal-dual hybrid gradient (PDHG) algorithm[1]. While this algorithm is flexible enough to solve a variety of non-smooth optimization problems, in every iteration the linear operators that correspond to the projection and the backprojection have to be fully evaluated. Moreover, in every iteration operations on vectors that have the size of the data have to be evaluated. For modern scanners like the Siemens Biograph mMR with span-1 data format, the size of these vectors are of length 350 million and therefore limiting the applicability of this algorithm (and therefore many non-smooth models) to state-of-the-art scanners.

M. J. Ehrhardt and C.-B. Schönlieb are with the Department for Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, UK (e-mail: m.j.ehrhardt@damtp.cam.ac.uk). P. Markiewicz is with the Centre for Medical Image Computing, London WC1E 6BT, UK.

### A. Contributions

In this work we investigate a randomized variant of PDHG [25], [12] (named Stochastic PDHG or short SPDHG) where in every iteration only a subset of the data is being used and only operations related to these data need to be evaluated. The numerical examples show that this strategy leads to a significant speed-up. Moreover, we propose and evaluate the use of preconditioners for PDHG for PET image reconstruction. Both forms of acceleration can be used independently or jointly.

We show that SPDHG is competitive with OSEM on unregularized reconstruction problems but stable with respect to the choice of the subsets due to its mathematically guaranteed convergence properties. In fact SPDHG converges to the deterministic solution for any proper subset selection. In addition, SPDHG is general enough to allow for a very large class of convex (and potentially non-smooth) regularizers such as total variation, total generalized variation and the parallel level sets prior. We show on clinical data from the Siemens Biograph mMR that with this algorithm, many modern mathematical models become feasible to be used in routine clinical imaging.

A few initial findings were published in a conference paper [26]. Convergence proofs under varying smoothness assumptions can be found in [27]. The numerical experiments use the open-source packages NiftyPET [28] and ODL [29] which allow efficient CUDA accelerated computations with the convenience of Python.

### B. PET Reconstruction via Optimization

The PET reconstruction problem can be formulated as the solution to the optimization problem

$$\min_{u \geq 0} \Big\{ D(\mathbf{R}u) + \alpha R(u) \Big\} \tag{1}$$

[1]also known as the Chambolle–Pock algorithm

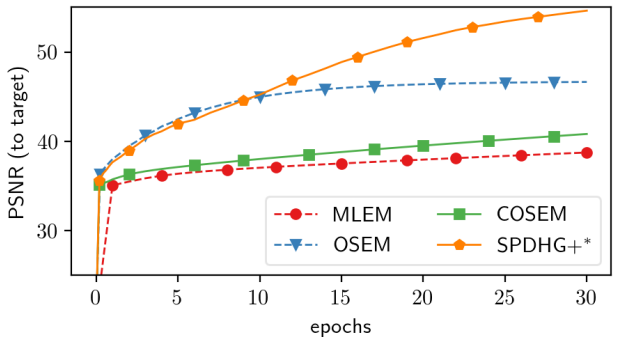


Fig. 1. **Quantitative reconstruction speed.** Comparison of MLEM, OSEM (21 subsets), COSEM (252) and the proposed SPDHG+ (252) in terms of peak signal-to-noise (PSNR) ratio with respect to an optimal solution for the `amyloid` dataset approximated by 5,000 MLEM iterations. OSEM and SPDHG+ are clearly faster than MLEM and COSEM. *proposed

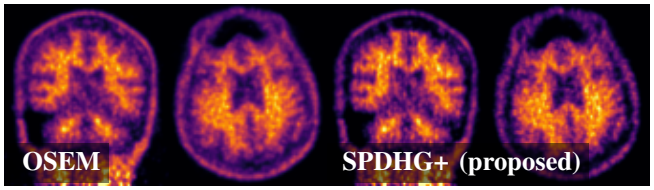


Fig. 2. **Qualitative reconstruction speed.** Visual comparison of OSEM (21 subsets) and SPDHG+ (252) after 10 epochs for maximum likelihood reconstruction. Both algorithms achieve very similar clinically relevant images with only 10 operator evaluations.

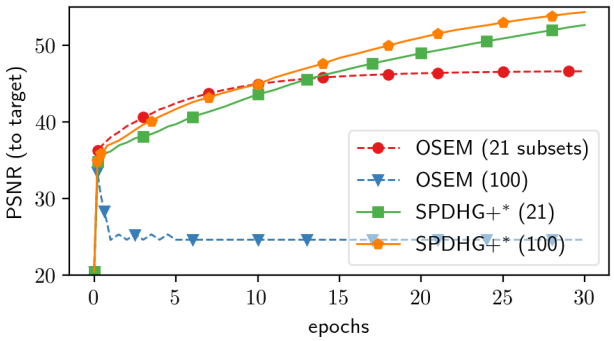


Fig. 3. **Number of subsets stability.** OSEM and SPDHG+ are compared for a varying number of subsets. While the speed of SPDHG+ increases with the number of subsets, OSEM fails to converge to the same solution for 100 subsets. *proposed

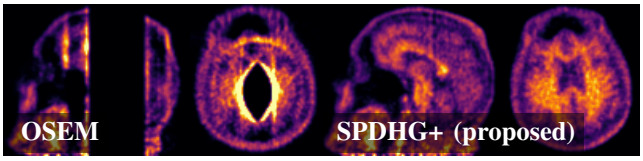


Fig. 4. **Subset selection stability.** In this example both OSEM and SPDHG+ take 21 subsets with bins equally and equidistantly divided into the 21 subsets. In contrast to OSEM, SPDHG+ is robust with respect to this subset selection and achieves a clinically relevant solution.

where $D(\mathbf{R}u)$ measures the fidelity of our estimated image $x$ with the data and the regularizer $\alpha R(u)$ penalizes features that are not desirable in the solution. In other words the regularizer can be used to avoid that our image estimate fits the data too closely which is needed due to noise. The data term $D(\mathbf{R}u)$ is (up to constants independent of $u$) the negative log-likelihood of the multi-variate Poisson distribution

$$D(y) = \sum_{i=1}^{M} y_i + r_i - b_i + b_i \log\left(\frac{b_i}{y_i + r_i}\right),$$

where the expected value is the projected image $\mathbf{R}u$ summed with the estimated background activity $r$. The latter is needed in order to model non-linear effects such as scatter and randoms. The data term measures the distance of the estimated data $\mathbf{R}u + r$ to the measured data $b$ in the sense that $D(\mathbf{R}u) \geq 0$ and $D(\mathbf{R}u) = 0$ if and only if $\mathbf{R}u + r = b$.

### C. Motivating Example: OSEM

If there is no regularizer, *i.e.* $\alpha R = 0$, the most common algorithm to solve the optimization (1) is the *maximum likelihood expectation maximization* algorithm (MLEM) [30] which is given by

$$u^{k+1} = \frac{u^k}{\mathbf{R}^T 1} \mathbf{R}^T \left(\frac{b}{\mathbf{R}u^k + r}\right) \tag{2}$$

where all operations have to be understood component-wise. The computational bottleneck in the MLEM algorithm is the evaluation of the operator $\mathbf{R}$ and its transpose $\mathbf{R}^T$ once in each iteration. To overcome this hurdle, it has been proposed to change the update and evaluate the operator and its adjoint only on a subset of the data in each iteration. If we divide our data into $m$ subsets and at every iteration we choose $j = \mathrm{mod}\,(k, m)$ and update

$$u^{k+1} = \frac{u^k}{\mathbf{R}_j^T 1} \mathbf{R}_j^T \left(\frac{b_j}{\mathbf{R}_j u^k + r_j}\right) \tag{3}$$

then this algorithm became known as *ordered subsets expectation maximization* (OSEM) [31]. While this change of the update equation reduces the computational burden by $1/m$ it is in general not guaranteed to converge and even if it converges it may not converge to the same modeled solution. This can be for instance seen in figure 1 where OSEM with 42 subsets (lilac curve) stop decreasing the objective after about 10 epochs (1 epoch = 1 full evaluation of the projection operator). Subsequently, a convergent version of OSEM, called *convergent OSEM* (COSEM), has been developed [32]. While it comes with mathematical convergence guarantees it is much slower than OSEM and therefore never became popular for the reconstruction of clinical PET data. A comparison of MLEM, OSEM and COSEM on clinical FDG data is shown in figure 1.

MLEM has been extended to include smooth [33] and certain non-smooth [9] prior information, however, conceptually both algorithms would intrinsically struggle with the ordered subset acceleration. Also other algorithms have been "accelerated" base don the ordered subset idea, e.g. [34], but inherit its intrinsic instability. See [35] for a numerical comparison and [1], [2] for a validation on clinical PET data.

For differentiable priors, a surrogate based technique allows for stable subset acceleration [36], [37], [38]. In this work we propose a algorithm that is provably convergent and thus stable and robust. It is flexible enough to be applicable to a large variety of convex and non-smooth priors and as efficient as OSEM-type algorithms.

## II. Mathematical Model

### A. Non-Smooth PET Reconstruction with Subsets

As outlined earlier, PET reconstruction can be formulated in terms of the optimization problem (1). Computationally, it is very convenient to rewrite the optimization problem (1) in terms of subsets. Therefore, let $\{S_i\}$ be a partition of $[M]$, in the sense that $\cup_{i=1}^m S_i = [M]$, where we used the notation $[M] \coloneqq \{1, \dots, M\}$. It is not necessary to assume that $S_i \cap S_j = \emptyset$ for $i \neq j$. For notational simplicity we will restrict ourselves to the latter case. Then we can write

$$D_i(y) \coloneqq \sum_{j \in S_i} \varphi_j(y_j) \tag{4}$$

where we slightly abuse the notation by indexing both $y$ and $\varphi$ by $j \in S_i$. The distance function for every data point is given by

$$\varphi(y) \coloneqq \begin{cases} y + r - b + b \log\left(\frac{b}{y+r}\right) & \text{if } y + r > 0 \\ \infty & \text{else} \end{cases} \tag{5}$$

where we omitted the index $j$ at $\varphi, y, r$ and $b$ for readability. Algorithms from convex optimization require the problem to be defined over an entire vector space which we satisfy by extending $\varphi$ to $\infty$ for non-positive estimated data $y + r$. The data and the background are photon counts and therefore have a natural non-negativity constraint. To cover the case of vanishing data, it is standard to define $0 \log 0 \coloneqq 0$.

We model the non-negativity constraint for the image $x$ with the indicator function $\imath_+$, which is defined as

$$\imath_+(u) = \begin{cases} 0 & \text{if } u \geq 0 \\ \infty & \text{else.} \end{cases} \tag{6}$$

Thus, we end up with the unconstrained optimization problem

**Problem 1** (PET Reconstruction with Subsets)**.**

$$u^\sharp \in \arg\min_{u \in \mathbb{R}^N} \left\{ \sum_{i=1}^m D_i(\mathbf{R}_i u) + \alpha R(u) + \imath_+(u) \right\} \tag{7}$$

We would like to stress that solving problem (7) is equivalent to solving the original problem (1) for any choice of subsets. In fact, the subset selection becomes a reconstruction parameter that may be varied to speed up the reconstruction procedure.

Often, our prior assumptions involve linear operators, too. One of the most prominent examples of this is the total variation which we can write as $R(u) = \mathrm{TV}(u) = \|\nabla x\|_{2,1}$. By this mean, that we take this 2-norm locally, i.e. at every voxel we take the 2-norm of the spatial gradient, and the 1-norm globally, i.e. we sum over all voxels. Similarly, we use the directional total variation $R(u) = \mathrm{dTV}(u) = \|\mathbf{D}\nabla u\|_{2,1}$ to incorporate a-priori knowledge about the solution given by some anatomical prior image, see [17], [18], [19], [39] for details.

Solving problem (7) is challenging, even when the involved variables are small and matrix-vector products are easy to compute. The difficulty arises due to problems on its regularity. The data term $D_i$ is not finite everywhere and while it is differentiable on its effective domain $\mathrm{dom}(D_i) \coloneqq \{u \mid D_i(u) < \infty\}$, the gradient is not Lipschitz continuous. In addition, further non-smoothness comes from the constraint $\imath_+$ and the prior $R$ may be non-smooth as well. All of this being said, in PET reconstruction, the variable sizes are actually very large and matrix-vector products very expensive to compute.

To apply optimization algorithms to solve (7), we reformulate it as a generic optimization problem of the form

**Problem 2** (Generic Optimization Problem)**.**

$$x^\sharp \in \arg\min_{x \in X} \left\{ \sum_{i=1}^n f_i(\mathbf{A}_i x) + g(x) \right\}. \tag{8}$$

For instance, for unregularized reconstructions, i.e. $\alpha R = 0$, we may make the association

$$n = m, \quad g = \imath_+, \quad f_i = D_i, \quad \mathbf{A}_i = \mathbf{R}_i$$

and reconstructions regularized by the total variation, i.e. $R(u) = \|\nabla u\|_{2,1}$, can by achieved by

$$\begin{aligned} n &= m + 1, & g &= \imath_+ \\ f_i &= D_i, i \in [m], & f_n &= \alpha \|\cdot\|_{2,1} \\ \mathbf{A}_i &= \mathbf{R}_i, i \in [m], & \mathbf{A}_n &= \nabla. \end{aligned} \tag{9}$$

### B. Optimization with Saddle-Point Problems

Instead of solving problem (7) directly, it is more efficient to reformulate the minimization problem as a saddle point problem by making use of the concept of the *convex conjugate* of a functional, see e.g. [40].

**Definition 1** (Convex Conjugate)**.** *Let $f : Y \to \mathbb{R}_\infty \coloneqq \mathbb{R} \cup \{\infty\}$ be a function with extended real values. Then we define the Fenchel conjugate of $f$ as the function $f^* : Y \to \mathbb{R}_\infty$ with*

$$f^*(y) = \sup_x \{\langle y, x \rangle - f(x)\}.$$

For convex, proper and lower semi-continuous (lsc) functions $f$ we have that $f^{**} = f$, see e.g. [40], and thus $f(x) = \sup_y \{\langle x, y \rangle - f^*(y)\}$. Then, with $Y = \prod_{i=1}^n Y_i$, problem 2 is equivalent to

**Problem 3** (Generic Saddle Point Problem)**.**

$$\min_{x \in X} \sup_{y \in Y} \left\{ \sum_{i=1}^n \langle \mathbf{A}_i x, y_i \rangle - f_i^*(y_i) + g(x) \right\}. \tag{10}$$

**Example 1.** *The convex conjugate of the PET distance function* (4) *is given by $D_i^*(y) = \sum_{j \in S_i} \varphi_j^*(y_j)$ with*

$$\varphi^*(y) = \begin{cases} -yr - b \log(1 - y), & \text{if } y \leq 1 \\ \infty, & \text{else} \end{cases} \tag{11}$$

*where we omitted the index $j$ at $\varphi, y, b$ and $r$ for readability and define $-b \log(0) \coloneqq \infty$ for $b > 0$.*

As some (or all) of the $f_i$ in (8) are non-smooth we make use of the proximity operator of these. Our definition varies slightly from the usual definition as we allow the step size parameter to be matrix-valued. For a symmetric and positive definite matrix $\mathbf{S}$, we define the weighted norm $\|x\|_\mathbf{S}$ as $\|x\|_\mathbf{S}^2 \coloneqq \|\mathbf{S}^{-1/2} x\|^2 = \langle \mathbf{S}x, x \rangle$.

**Definition 2** (Proximity Operator)**.** *Let* $\mathbf{S}$ *be a symmetric and positive definite matrix. Then we define the proximal operator of* $f$ *at the proximal point* $x$ *with metric or step size* $\mathbf{S}$ *as*

$$\operatorname{prox}_f^{\mathbf{S}}(x) \coloneqq \arg\min_z \left\{ \|z - x\|_{\mathbf{S}}^2 + f(z) \right\} .$$

From here on, $\mathbf{S}$ and $\mathbf{T}$ will always be diagonal, symmetric and positive definite matrices.

**Example 2.** *The proximity operator of the non-negativity constraint* (6) *is given component-wise by*

$$\operatorname{prox}_{i_+}^{\mathbf{T}}(x) = \max(x, 0) .$$

**Example 3.** *Let* $\mathbf{S}_i = \operatorname{diag}((\sigma_j)_{j \in S_i})$*. The proximal operator of the convex conjugate of the PET distance* (11) *can be computed component-wise as* $[\operatorname{prox}_{D_i^*}^{\mathbf{S}_i}(y)]_j = \operatorname{prox}_{\varphi_j^*}^{\sigma_j}(y_j)$*. For each component, the proximal operator is given by*

$$\operatorname{prox}_{\varphi^*}^{\sigma}(y) = \frac{1}{2}\left[ w + 1 - \left( (w-1)^2 + 4\sigma b \right)^{1/2} \right]$$

*where we again omitted the indices* $j$ *for readability and denoted* $w = y + \sigma r$*.*

## III. Algorithm

An algorithm that solves (10) and therefore the PET reconstruction problem for a very large class of regularizers is the SPDHG, originally proposed and analytically studied in [27]. It is summarized in algorithm 1. It consists of very simple operations involving only basic linear algebra, matrix-vector multiplications and the evaluations of proximity operators.

### A. Convergence

SPDHG is guaranteed to converge for any $f_i$ and $g$ which are convex, proper and lsc. We know state a very general convergence result which can be easily derived from the results proven in [27]. For more details and convergence rates we refer the reader to [27].

**Theorem 1** (Convergence)**.** *Assume that the sampling is proper, i.e. the probability* $p_i$ *for an index* $i \in [n]$ *to be sampled is positive. Let the step length parameters* $\mathbf{T} = \min_{i \in [n]} \mathbf{T}_i, \mathbf{S}_i$ *be chosen such that for all* $i \in [n]$

$$\left\| \mathbf{S}_i^{1/2} \mathbf{A}_i \mathbf{T}_i^{1/2} \right\|^2 < p_i \tag{12}$$

*holds. Then for any initialization, the iterates* $(x, y)$ *of SPDHG (algorithm 1) converge to a saddle point of* (10) *(in an expected Bregman sense).*

For illustration purposes, we summarize a special case of SPDHG with uniform sampling for the simpler unregularized problem in algorithm 2.

**Remark 1** (Computational Efficiency)**.** *Each iteration of the algorithm 2 is computationally efficient as only projections and backprojections corresponding to the randomly selected subset* $i$ *of the data are required. However, the algorithm maintains the whole backprojected dual variable* $z = \mathbf{R}^T y = \sum_{i=1}^m \mathbf{R}_i^T y_i$ *and in each iteration updates the primal variable with it.*

**Algorithm 1** Stochastic Primal-Dual Hybrid Gradient (SPDHG) to solve (10). Default values given in brackets.

**Input:** iterates $x(=0)$, $y(=0)$, step parameters $\mathbf{S} = \{\mathbf{S}_i\}$, $\mathbf{T}$
**Output:** reconstructed image $x$

$\overline{z} = z = \mathbf{A}^T y \ (= 0)$
**for** $k = 1, \dots$ **do**
  $x = \operatorname{prox}_g^{\mathbf{T}}(x - \mathbf{T}\overline{z})$
  Select $i \in [n]$ at random with probability $p_i$
  $y_i^+ = \operatorname{prox}_{f_i^*}^{\mathbf{S}_i}(y_i + \mathbf{S}_i \mathbf{A}_i x)$
  $\Delta z = \mathbf{A}_j^T \left( y_j^+ - y_j \right)$
  $y_i = y_i^+$
  $\overline{z} = z + \left(1 + \frac{1}{p_i}\right) \Delta z$
  $z = z + \Delta z$

**Algorithm 2** SPDHG+ for the unregularized problem to compute the results in figure 2.

**Output:** reconstructed image $x$

$x = 1$, $y = 0$
$\overline{z} = z = 0$
$\mathbf{S}_i = \operatorname{diag}(1/\mathbf{R}_i 1), 1 \le i \le m$
$\mathbf{T} = \min_{i \in [m]} \operatorname{diag}(1/(m\mathbf{R}_i^T 1))$
**for** $k = 1, \dots$ **do**
  $x = \max(x - \mathbf{T}\overline{z}, 0)$
  Select $i \in [m]$ uniformly at random
  $w = y_i + \mathbf{S}_i(\mathbf{R}_i x + r_i)$
  $y_i^+ = \frac{1}{2}\left[ w + 1 - \left( (w-1)^2 + 4\mathbf{S}_i b_i \right)^{1/2} \right]$
  $\Delta z = \mathbf{R}_i^T (y_i^+ - y_i)$
  $y_i = y_i^+$
  $\overline{z} = z + (m+1)\Delta z$
  $z = z + \Delta z$

**Remark 2** (Memory Requirements)**.** *The memory requirement of the algorithm is higher compared to OSEM or gradient descent but still reasonably low. It requires memory equivalent to two images* $(z, \overline{z})$ *and up to twice the binned sinogram data* $(y, y^+)$ *in addition to the necessary memory consumption (output image, sinogram data, background).*

**Remark 3** (Sampling)**.** *The algorithm allows any kind of random selection as long as the draws are independent and the probability that block* $i$ *is being selected with positive probability* $p_i > 0$*. We will investigate two choices of sampling in the numerical section of this paper. A more thorough numerical and theoretical investigation will be subject of future work.*

### B. Step Sizes and Preconditioning

We will now discuss two different choices of step sizes under which the algorithm 1 is guaranteed to converge.

**Theorem 2** (Step Size Parameters)**.** *Let* $\rho < 1$*. Then, condition* (12) *of Theorem 1 is satisfied by*

$$\mathbf{S}_i = \frac{\rho}{\|\mathbf{A}_i\|}\mathbf{I} , \quad \mathbf{T}_i = \frac{\rho p_i}{\|\mathbf{A}_i\|}\mathbf{I} . \tag{13}$$

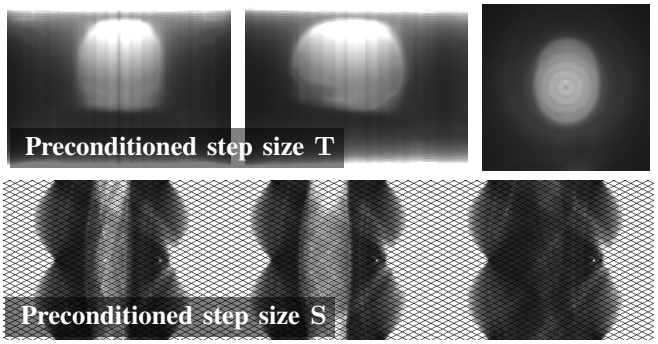


Fig. 5. Preconditioned parameters $\mathbf{T}$ (top) and $\mathbf{S}$ (bottom) (14) for the data set FDG. Besides the boundary the step sizes are large in interesting regions clearly showing the head of the patient.

*Moreover, if $\mathbf{A}_i$ has only non-negative elements, then condition* (12) *is also satisfied by*

$$\mathbf{S}_i = \operatorname{diag}\left(\frac{\rho}{\mathbf{A}_i 1}\right), \quad \mathbf{T}_i = \operatorname{diag}\left(\frac{\rho p_i}{\mathbf{A}_i^T 1}\right). \tag{14}$$

**Remark 4.** *If $n = 1$ and $p_i = 1$, then the step sizes* (13) *can be identified with the scalar step sizes $\sigma_i = \rho/\|\mathbf{A}_i\|$ and $\tau = \rho/\|\mathbf{A}_i\|$ which are commonly chosen for PDHG.*

**Remark 5.** *Note that the non-negativity condition holds for the PET forward operator (and any other ray tracing based operator). Moreover, the step size $\mathbf{T}$ in* (14) *resembles the sensitivities used in the update of MLEM* (2) *and OSEM* (3)*. In addition, a similar preconditioning is performed for the dual variable in the data space.*

An example of step sizes with preconditioning (14) is shown in figure 5.

## IV. Numerical Results

### A. Data

We validate the numerical performance of the proposed algorithm on two data sets. The two separate PET brain datasets each use a distinct radiotracer: [$^{18}$F]FDG in epilepsy and [$^{18}$F]florbetapir in a neuroscience sub-study (Insight'46) of the Medical Research Council National Survey of Health and Development [41]. The epileptic patient was injected with 250 Mbq of FDG, one hour before the 15-minute PET acquisition. The neuroscience volunteer was injected with 370 MBq of florbetapir and scanned dynamically for one hour, starting at the injection time. The last ten minutes were used as a measurement of amyloid deposition, which for the participant was negative.

### B. Results for Total Variation

In this section we analyze the impact of various choices within SPDHG on its performance, from randomness over sampling to preconditioning. The test case is total variation regularization, thus we consider the setting (9).

*1) Randomness:* Figure 6 shows the effect of randomness where we compare the deterministic PDHG to SPDHG with uniform sampling and scalar step sizes (13) with two different number of subsets. The horizontal axis reflects the number of projections in each algorithm, we call one full projection for the whole data one epoch. We can easily see that both random variants are faster than then deterministic PDHG. Moreover, the randomized SPDHG becomes faster by choosing a larger number of subsets.

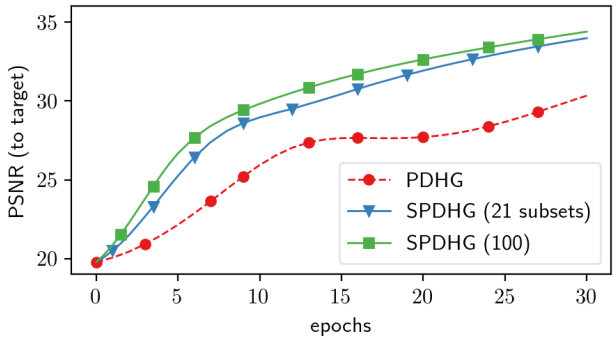


Fig. 6. **Deterministic v randomized.** The results for FDG with TV regularization show that the randomized algorithms are much faster than its deterministic counterpart. Moreover, more subsets leads to a faster algorithm.

*2) Sampling:* The effect of different choices of sampling is shown in figure 7. We compare two different samplings: uniform sampling and balanced sampling. The uniform sampling chooses all indices $i \in [n]$ with equal probability $p_i = 1/n$. In contrast, for balanced sampling we choose with uniform probability either data or regularization. If we choose data, then we select a subset again randomly with uniform probability. Thus, the probability for each subset of the data to be selected is $p_i = 1/(2m)$ and for the regularization to be selected $p_n = 1/2$. We make two observations. First, balanced sampling is always faster than uniform sampling. This shows the importance of updating the dual variable associated to the regularizer. Second, for either sampling choosing a larger number of subsets again improves the performance.

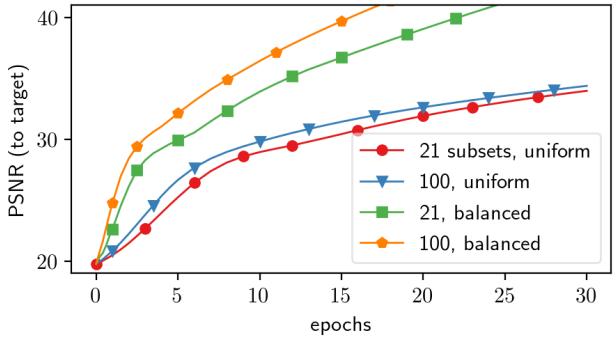


Fig. 7. **Uniform v balanced sampling.** Setting as in figure 6. In addition to increasing the number of subsets, the sampling is also very important for the speed of the algorithm: 21 subsets with balanced sampling is faster than 100 subsets with uniform sampling.

*3) Preconditioning:* As shown in theorem 2, the step size parameters $\mathbf{T}$ and $\mathbf{S}_i$ can be chosen either as scalars (13) or as vectors (14), the latter can be seen as a form of preconditioning. The results are shown in figure 8 where SPDHG

refers to the variant with scalar step sizes and SPDHG+ to its preconditioned version. Similar to previous observations, we see that preconditioning may accelerate the convergence of either the deterministic PDHG or the randomized SPDHG. Moreover, combining these yields an even faster algorithm.

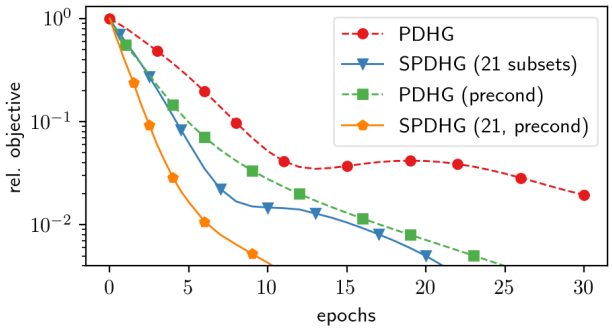


Fig. 8. **Preconditioning** can be used with and without randomization. For the setting as in figure 6, the preconditioned algorithms are much faster than without preconditioning.

*4) Performance of Proposed Algorithm:* Based on the previous three examples, we propose to combine randomization, balanced sampling and preconditioning to get the fastest algorithm. Figure 9 shows the visual performance of PDHG and SPDHG+. In contrast to the deterministic PDHG, the proposed SPDHG+ yields a clinically relevant image after only 10 epochs.

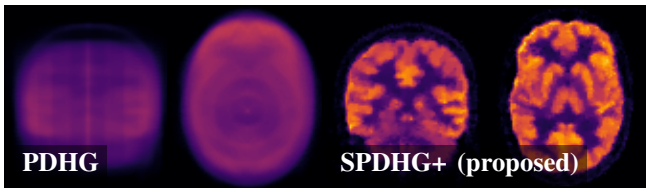


Fig. 9. The qualitative results for the setting as in figure 6 show that in contrast to the deterministic PDHG), the proposed SPDHG+ yields a clinically relevant image after only 10 epochs.

### C. Results for anisotropic Total Variation

Anisotropic total variation decouples the penalization of the derivatives. The mathematical model is similar to the isotropic TV model (9), the only difference being the norm how the total variation is measured: $f_n = \alpha\|\cdot\|_{1,1}$. It can be seen in figure 10 for `amyloid` that with randomization and preconditioning only a few epochs are needed to reconstruct clinically relevant images.

### D. Results for Directional Total Variation

Anatomical information from a co-registered MRI is available on combined PET-MR scanners. The structural information of the anatomy can be utilized by the directional total variation regularization, see [17], [18], [19], [39] for details. The mathematical model is very similar to the total variation model (9), except for an additional matrix $\mathbf{D}$. Thus, the only difference is $\mathbf{A}_n = \mathbf{D}\nabla$. It can be seen in figure 11 for `amyloid` that with randomization and preconditioning only a few epochs are needed to reconstruct clinically relevant images.

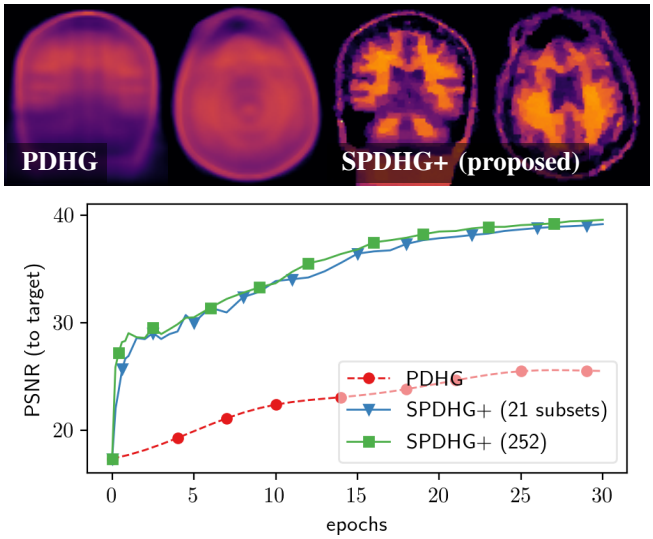


Fig. 10. **Top:** PDHG and SPDHG+ reconstructions from `FDG` data with anisotropic TV regularization after 10 epochs. Only a few epochs are needed to reconstruct clinically relevant images if randomization and preconditioning are used. **Bottom:** The quantitative results confirm these observations and show a significant speed-up with both randomization and preconditioning. Increasing the number of subsets from 21 to 252 did not increase the performance much.

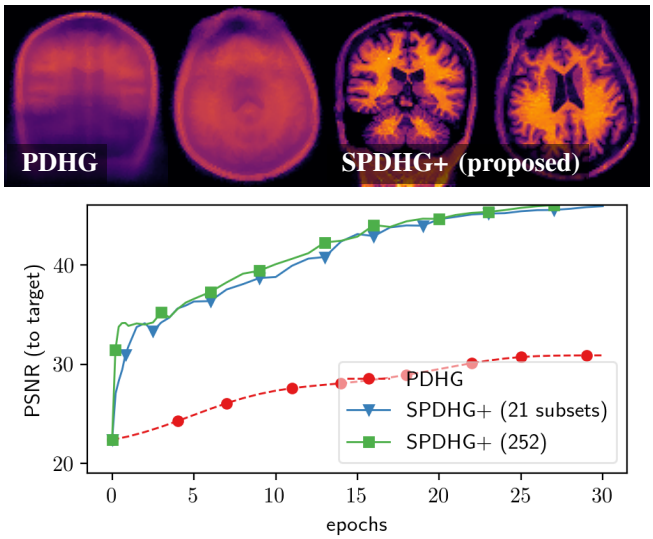


Fig. 11. Directional TV regularizations (which uses MRI information) for `amyloid` data. The qualitative **top** and quantitative **bottom** results show that only a few epochs are needed to reconstruct clinically relevant images if randomization and preconditioning are used.

### E. Results for Total Generalized Variation

More sophisticated regularization involves the total generalized variation (TGV) [6], [10], [11]

$$\mathrm{TGV}_{\alpha_0,\alpha_1}(u) = \inf_w \{\alpha_0\|\nabla u - w\|_{2,1} + \alpha_1\|\mathcal{E}w\|_{2,1}\} \quad (15)$$

which can balance first and second order regularization, i.e. it allows for edge-preserved reconstruction while avoiding the stair-casing artifact. We can solve the TGV regularized PET reconstruction problem by solving (8) with the assignment $x =$

$(u, w)$ and

$$\begin{aligned} n &= m + 2, & g(x) &= \imath_+(u) \\ f_i &= D_i, i \in [m], & \mathbf{A}_i &= (\mathbf{R}_i, 0), i \in [m] \\ f_{n-1} &= \alpha_0 \|\cdot\|_{2,1}, & f_n &= \alpha_1 \|\cdot\|_{2,1} \\ \mathbf{A}_{n-1} &= (\nabla, -\mathbf{I}), & \mathbf{A}_n &= (0, \mathcal{E})\,. \end{aligned} \tag{16}$$

The numerical results shown in figure 12 are in line with the previous findings indicating that randomization and preconditioning can significantly speed up the reconstruction.

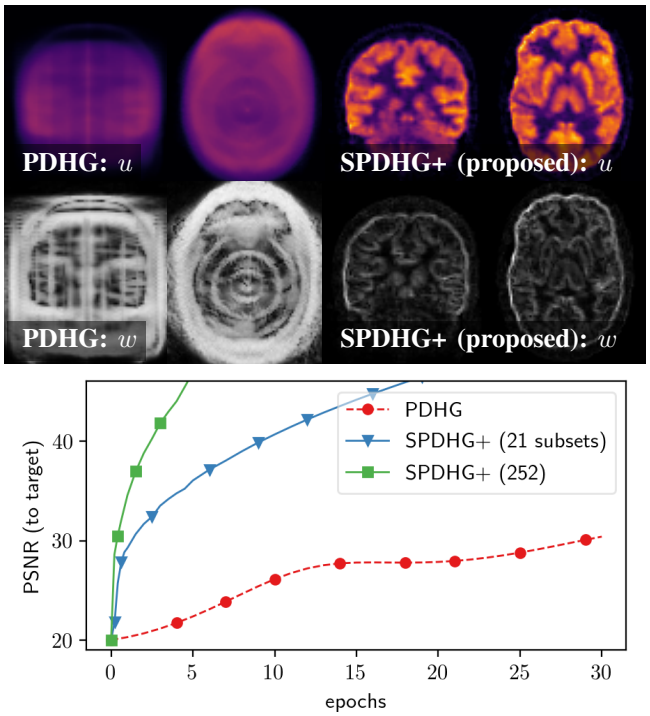


Fig. 12. TGV regularized reconstruction for the `FDG` data. The qualitative **top** and quantitative **bottom** results show that only a few epochs are needed to reconstruct clinically relevant images if randomization and preconditioning are used. This is visible for both the actual images $u$ and for the reconstructed vector field $w$.

### F. Comparison of Mathematical Models

We conclude this section by a comparison of various methods on both data sets in figures 13 and 14. While we leave the detailed visual comparisons to the reader, we would like to note that all these images use the same number of forward operator evaluations so have basically the same computational cost.

## V. Discussion

The extensive numerical experiments all consistently confirm that randomization and preconditioning both speed up the reconstruction. These trends were irrespective of the data set and the regularization model used. The convergence speed in our work was abstractly defined by a solution of the underlying mathematical optimization model approximated with way too many iterations than would be feasible in routine clinical practice. This strategy was chosen intentionally as we did not want to target a specific clinical use case. After these

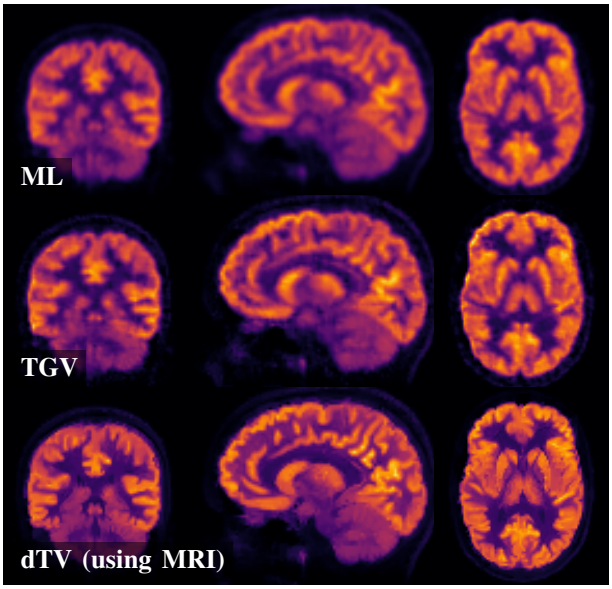


Fig. 13. Comparison of several reconstruction approaches for the `FDG` data. All approaches have about the same computational cost.

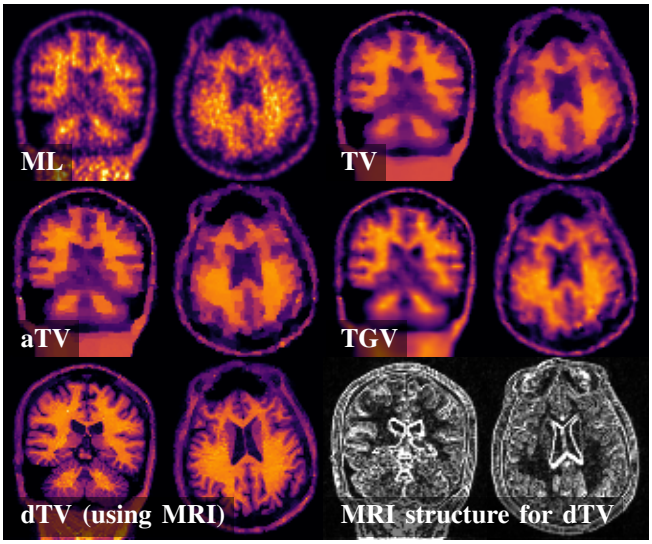


Fig. 14. Comparison of several reconstruction approaches for the `amyloid` data. All approaches have about the same computational cost.

successful initial trials, in the future we will collaborate with medical researchers and clinicians to focus on specific use cases where each use case defines its own metric of what image we wish to reconstruct.

The focus of this contribution was to improve the reconstruction for non-smooth priors to allow the use of regularization functionals like total variation and its descendants like total generalized variation and directional total variation. However, as long as the proximal operators are simple to evaluate, the proposed randomized and preconditioned algorithm can be applied to any other model, too. It would be of huge interest to compare this algorithm to convergent subset accelerated algorithms like BSREM [36], [37], TRIOT [38] and OS-SPS [37].

We highlighted the improvements from choosing different

distributions by "comparing uniform" and "balance sampling". Further improvements are expected by optimizing the probability selection of this algorithm. This can either be an optimal distribution that is constant along the iterations or even developing over the course of the iterations.

Moreover, the algorithm does not exploit any special structure of our optimization like smoothness nor strong convexity. It is likely that exploiting these properties will lead to additional speed-up. However, as these properties for the PET data term depend on the acquired data, it is unlikely that a straight forward approach will be sufficient.

## VI. Conclusion

We have shown that using randomization and preconditioning can dramatically speed up the convergence of an iterative reconstruction algorithm. The speed up was so large that many mathematical imaging models are now feasible to be used in daily clinical routine. While all observations in this contribution where consistent among two data sets with different tracers, more studies are needed to confirm the consistent benefit of this reconstruction strategy. In any case, this algorithmic advancement has the potential to hugely impact the PET reconstruction landscape as advanced mathematical models can now be combined with efficient and convergent subset acceleration.

## Acknowledgement

M.J.E. and C.-B.S. acknowledge support from Leverhulme Trust project "Breaking the non-convexity barrier", EPSRC grant "EP/M00483X/1", EPSRC centre "EP/N014588/1", the Cantab Capital Institute for the Mathematics of Information, and from CHiPS and NoMADS (Horizon 2020 RISE project grants). In addition, M.J.E. acknowledges support from the EPSRC platform grant "EP/M020533/1". Moreover, C.-B.S. is thankful for support by the Alan Turing Institute. In addition, all authors gratefully acknowledge the hardware donation by the NVIDIA Corporation.

## References

[1] E. J. Teoh, D. R. McGowan, R. E. Macpherson, K. M. Bradley, and F. V. Gleeson, "Phantom and Clinical Evaluation of the Bayesian Penalized Likelihood Reconstruction Algorithm Q.Clear on an LYSO PET/CT System," *Journal of Nuclear Medicine*, vol. 56, no. 9, pp. 1447–1452, 2015. [Online]. Available: http://jnm.snmjournals.org/cgi/doi/10.2967/jnumed.115.159301

[2] S. Ahn, S. G. Ross, E. Asma, J. Miao, X. Jin, L. Cheng, S. D. Wollenweber, and R. M. Manjeshwar, "Quantitative comparison of OSEM and penalized likelihood image reconstruction using relative difference penalties for clinical PET," *Physics in Medicine and Biology*, vol. 60, no. 15, pp. 5733–5751, 2015.

[3] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear Total Variation based Noise Removal Algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.

[4] S. Setzer, G. Steidl, and T. Teuber, "Deblurring Poissonian images by split Bregman techniques," *Journal of Visual Communication and Image Representation*, vol. 21, no. 3, pp. 193–199, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.jvcir.2009.10.006

[5] M. Benning, P. Heins, and M. Burger, "A Solver for Dynamic PET Reconstructions based on Forward-Backward-Splitting," Tech. Rep., 2010.

[6] K. Bredies, K. Kunisch, and T. Pock, "Total Generalized Variation," *SIAM Journal on Imaging Sciences*, vol. 3, no. 3, pp. 492–526, 2010.

[7] S. Anthoine, J.-F. Aujol, Y. Boursier, and C. Melot, "Some Proximal Methods for CBCT and PET tomography," in *Wavelets and Sparsity XIV*, M. Papadakis, D. Van De Ville, and V. K. Goyal, Eds., vol. 8138, sep 2011. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1267470

[8] M. Burger and S. Osher, "A Guide to the TV Zoo," in *Level Set and PDE Based Reconstruction Methods in Imaging*, ser. Lecture Notes in Mathematics. Cham: Springer International Publishing, 2013, vol. 2090, pp. 1–70. [Online]. Available: http://link.springer.com/10.1007/978-3-319-01712-9

[9] A. Sawatzky, C. Brune, T. Koesters, F. Wübbeling, and M. Burger, "EM-TV Methods for Inverse Problems with Poisson Noise," in *Level Set and PDE Based Reconstruction Methods in Imaging*, ser. Lecture Notes in Mathematics. Springer International Publishing Switzerland, 2013, vol. 2090. [Online]. Available: http://link.springer.com/10.1007/978-3-319-01712-9

[10] K. Bredies and M. Holler, "A TGV-Based Framework for Variational Image Decompression, Zooming, and Reconstruction. Part II: Numerics," *SIAM Journal on Imaging Sciences*, vol. 8, no. 4, pp. 2851–2886, 2015. [Online]. Available: http://epubs.siam.org/doi/10.1137/15M1023865

[11] ——, "Regularization of linear inverse problems with total generalized variation," *Journal of Inverse and Ill-Posed Problems*, vol. 22, no. 6, pp. 871–913, 2014.

[12] A. Chambolle and T. Pock, "A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, dec 2011. [Online]. Available: http://link.springer.com/10.1007/s10851-010-0251-1

[13] F.-X. Dupe, M. J. Fadili, and J.-L. Starck, "Inverse Problems with Poisson Noise: Primal and Primal-Dual Splitting," in *IEEE International Conference on Image Processing*, 2011, pp. 1901–1904.

[14] M. A. T. Figueiredo and J. M. Bioucas-Dias, "Frame-based deconvolution of poissonian images using alternating direction optimization," in *International Conference on Image Processing, ICIP*, vol. 19, no. 12, 2010, pp. 3549–3552.

[15] J. E. Bowsher, H. Yuan, L. W. Hedlund, T. G. Turkington, G. Akabani, A. Badea, W. C. Kurylo, C. T. Wheeler, G. P. Cofer, M. W. Dewhirst, and G. A. Johnson, "Utilizing MRI Information to Estimate F18-FDG Distributions in Rat Flank Tumors," in *IEEE Nuclear Science Symposium and Medical Imaging Conference*, 2004, pp. 2488–2492.

[16] M. Hintermüller, M. Holler, and K. Papafitsoros, "A function space framework for structural total variation regularization with applications in inverse problems," 2017. [Online]. Available: http://arxiv.org/abs/1710.01527

[17] M. J. Ehrhardt, P. Markiewicz, M. Liljeroth, A. Barnes, V. Kolehmainen, J. Duncan, L. Pizarro, D. Atkinson, B. F. Hutton, S. Ourselin, K. Thielemans, and S. R. Arridge, "PET Reconstruction with an Anatomical MRI Prior using Parallel Level Sets," *IEEE Transactions on Medical Imaging*, vol. 35, no. 9, pp. 2189–2199, 2016. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7452643

[18] M. J. Ehrhardt and M. M. Betcke, "Multi-Contrast MRI Reconstruction with Structure-Guided Total Variation," *SIAM Journal on Imaging Sciences*, vol. 9, no. 3, pp. 1084–1106, 2016. [Online]. Available: http://arxiv.org/abs/1511.06631http://epubs.siam.org/doi/10.1137/15M1047325

[19] G. Schramm, M. Holler, A. Rezaei, K. Vunckx, F. Knoll, K. Bredies, F. Boada, and J. Nuyts, "Evaluation of Parallel Level Sets and Bowsher's Method as Segmentation-Free Anatomical Priors for Time-of-Flight PET Reconstruction," *IEEE Transactions on Medical Imaging*, vol. 0062, no. c, pp. 1–1, 2017.

[20] A. Mehranian, M. A. Belzunce, F. Niccolini, M. Politis, C. Prieto, F. Turkheimer, A. Hammers, and A. J. Reader, "PET Image Reconstruction using Multi-Parametric Anato-Functional," *Physics in Medicine and Biology*, 2017.

[21] M. J. Ehrhardt, K. Thielemans, L. Pizarro, D. Atkinson, S. Ourselin, B. F. Hutton, and S. R. Arridge, "Joint Reconstruction of PET-MRI by exploiting Structural Similarity," *Inverse Problems*, vol. 31, p. 015001, jan 2015.

[22] F. Knoll, M. Holler, T. Koesters, R. Otazo, K. Bredies, and D. K. Sodickson, "Joint MR-PET reconstruction using a multi-channel image regularizer," 2016.

[23] J. Rasch, E.-M. Brinkmann, and M. Burger, "Joint Reconstruction via Coupled Bregman Iterations with Applications to PET-MR Imaging," 2017. [Online]. Available: http://arxiv.org/abs/1704.06073

[24] A. Mehranian, M. Belzunce, C. Prieto, A. Hammers, and A. J. Reader, "Synergistic PET and SENSE MR image reconstruction using joint sparsity regularization," *IEEE Transactions on Medical Imaging*, no. c, pp. 1–1, 2017.

[25] E. Esser, X. Zhang, and T. F. Chan, "A General Framework for a Class of First Order Primal-Dual Algorithms for Convex Optimization in Imaging Science," *SIAM Journal on Imaging Sciences*, vol. 3, no. 4, pp. 1015–1046, 2010.

[26] M. J. Ehrhardt, P. J. Markiewicz, P. Richtárik, J. Schott, A. Chambolle, and C.-B. Schönlieb, "Faster PET Reconstruction with a Stochastic Primal-Dual Hybrid Gradient Method," in *SPIE Optics+Photonics:Wavelets and Sparsity XVII*, San Diego, 2017. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10394/2272946/Faster-PET-reconstruction-with-a-stochastic-primal-dual-hybrid-gradient/10.1117/12.2272946.full

[27] A. Chambolle, M. J. Ehrhardt, P. Richtárik, and C.-B. Schönlieb, "Stochastic Primal-Dual Hybrid Gradient Algorithm with Arbitrary Sampling and Imaging Applications," Tech. Rep., jun 2017. [Online]. Available: http://arxiv.org/abs/1706.04957

[28] P. J. Markiewicz, M. J. Ehrhardt, K. Erlandsson, A. Barns, J. M. Schott, D. Atkinson, S. R. Arridge, B. F. Hutton, and S. Ourselin, "NiftyPET: A high-throughput software platform for high quantitative accuracy and precision PET imaging and analysis," *submitted*, 2017.

[29] J. Adler, H. Kohr, and O. Öktem, "Operator Discretization Library (ODL)," Jan. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.249479

[30] L. A. Shepp and Y. Vardi, "Maximum Likelihood Reconstruction for Emission Tomography," *IEEE Transactions on Medical Imaging*, vol. 1, no. 2, pp. 113–22, 1982.

[31] H. M. Hudson and R. S. Larkin, "Accelerated Image Reconstruction Using Ordered Subsets of Projection Data," *IEEE Transactions on Medical Imaging*, vol. 13, no. 4, pp. 601–609, 1994.

[32] I.-t. Hsiao, A. Rangarajan, and G. Gindi, "A Provably Convergent OS-EM Like Reconstruction Algorithm for Emission Tomography," in *Proceedings of SPIE*, vol. 4684, 2002, pp. 10–19.

[33] P. J. Green, "Bayesian Reconstructions from Emission Tomography Data using a Modified EM Algorithm," *IEEE Transactions on Medical Imaging*, vol. 9, no. 893, 1990.

[34] M. G. McGaffin and J. A. Fessler, "Alternating Dual Updates Algorithm for X-ray CT Reconstruction on the GPU," *IEEE Transactions on Computational Imaging*, vol. 1, no. 3, pp. 186–199, 2015.

[35] L. Cheng, E. Asma, S. Ahn, and R. M. Manjeshwar, "Comparison of Numerical Convergence Speeds of Convergent and Accelerated Algorithms for Penalized Likelihood PET Image," in *IEEE Nuclear Science Symposium and Medical Imaging Conference*, 2013, pp. 3–6.

[36] A. R. De Pierro and M. E. B. Yamagishi, "Fast EM-like methods for maximum "a posteriori" estimates in emission tomography," *IEEE Transactions on Medical Imaging*, vol. 20, no. 4, pp. 280–288, 2001.

[37] S. Ahn and J. A. Fessler, "Globally convergent image reconstruction for emission tomography using relaxed ordered subsets algorithms," *IEEE Transactions on Medical Imaging*, vol. 22, no. 5, pp. 613–626, 2003.

[38] S. Ahn, J. A. Fessler, D. Blatt, and A. O. Hero, "Convergent Incremental Optimization Transfer Algorithms: Application to Tomography," *IEEE Transactions on Medical Imaging*, vol. 25, no. 3, pp. 283–296, 2006.

[39] L. Bungert, D. A. Coomes, M. J. Ehrhardt, J. Rasch, R. Reisenhofer, and C.-B. Schönlieb, "Blind Image Fusion for Hyperspectral Imaging with the Directional Total Variation," *Inverse Problems*, vol. 34, no. 4, p. 044003, 2018. [Online]. Available: http://arxiv.org/abs/1710.05705

[40] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2011.

[41] C. A. Lane, T. D. Parker, D. M. Cash, K. Macpherson, E. Donnachie, H. Murray-Smith, A. Barnes, S. Barker, D. G. Beasley, J. Bras, D. Brown, N. Burgos, M. Byford, M. Jorge Cardoso, A. Carvalho, J. Collins, E. De Vita, J. C. Dickson, N. Epie, M. Espak, S. M. Henley, C. Hoskote, M. Hutel, J. Klimova, I. B. Malone, P. Markiewicz, A. Melbourne, M. Modat, A. Schrag, S. Shah, N. Sharma, C. H. Sudre, D. L. Thomas, A. Wong, H. Zhang, J. Hardy, H. Zetterberg, S. Ourselin, S. J. Crutch, D. Kuh, M. Richards, N. C. Fox, and J. M. Schott, "Study protocol: Insight 46 - a neuroscience sub-study of the MRC National Survey of Health and Development," *BMC Neurology*, vol. 17, no. 1, pp. 1–25, 2017.