Title

DiFiR-CT: Distance field representation to resolve motion artifacts in computed tomography

Journal

Medical Physics, 50(3)

ISSN

0094-2405

Authors

Gupta, Kunal

Colvert, Brendan

Chen, Zhennong

et al.

Publication Date

2023-03-01

DOI

10.1002/mp.16157

Peer reviewed

HHS Public Access

Author manuscript

Med Phys. Author manuscript; available in PMC 2023 November 28.

Published in final edited form as:

Med Phys. 2023 March ; 50(3): 1349–1366. doi:10.1002/mp.16157.

DiFiR-CT: Distance Field Representation to resolve motion artifacts in Computed Tomography

Kunal Gupta,
Dept. of Computer Science Engineering, University of California San Diego.

Brendan Colvert,
Dept. of Bioengineering, University of California San Diego. B.C is now with Medtronic.

Zhennong Chen,
Dept. of Bioengineering, University of California San Diego.

Francisco Contijoch
Dept. of Bioengineering, Radiology, University of California San Diego.

Abstract

Background: Motion during data acquisition leads to artifacts in computed tomography (CT) reconstructions. In cases such as cardiac imaging, not only is motion unavoidable, but evaluating the motion of the object is of clinical interest. Reducing motion artifacts has typically been achieved by developing systems with faster gantry rotation or via algorithms which measure and/or estimate the displacement. However, these approaches have had limited success due to both physical constraints as well as the challenge of estimating non-rigid, temporally varying, and patient-specific motion fields.

Purpose: To develop a novel reconstruction method which generates time-resolved, artifact-free images without estimation or explicit modeling of the motion.

Methods: We describe an analysis-by-synthesis approach which progressively regresses a solution consistent with the acquired sinogram. In our method, we focus on the movement of object boundaries. Not only are the boundaries the source of image artifacts, but object boundaries can simultaneously be used to represent both the object as well as its motion over time without need for an explicit motion model. We represent the object boundaries via a signed distance function (SDF) which can be efficiently modeled using neural networks. As a result, optimization can be performed under spatial and temporal smoothness constraints without the need for explicit motion estimation.

Results: We illustrate the utility of DiFiR-CT in three imaging scenarios with increasing motion complexity: translation of a small circle, heart-like change in an ellipse's diameter, and a complex topological deformation. Compared to filtered backprojection, DiFiR-CT provides high quality image reconstruction for all three motions without hyperparameter tuning or change to the architecture. We also evaluate DiFiR-CT's robustness to noise in the acquired sinogram and found its reconstruction to be accurate across a wide range of noise levels. Lastly, we demonstrate how

Correspondence : fcontijoch@ucsd.edu.

the approach could be used for multi-intensity scenes and illustrate the importance of the initial segmentation providing a realistic initialization. Code and supplemental movies are available at https://kunalmgupta.github.io/projects/DiFiR-CT.html.

Conclusions: Projection data can be used to accurately estimate a temporally-evolving scene without the need for explicit motion estimation using a neural implicit representation and analysis-by-synthesis approach.

I. Introduction

X-ray computed tomography (CT) can non-invasively and volumetrically evaluate patient anatomy with high spatial resolution, which has led to a widespread clinical use. Unfortunately, image quality can be reduced if motion occurs during the acquisition, despite fast acquisition of individual projections. This is particularly problematic when imaging fast moving structures (e.g., the heart) as current single-source conebeam scanners require 200–300 ms to perform a full gantry rotation. As a result, while CT is used for the evaluation of suspected coronary artery disease[1] and acute chest pain[2], even small motions (~1.2mm) can result in significant blurring of key structures, such as coronary arteries[3]. Further, motion of highly-attenuating structures, such as metallic devices, can further impair clinical assessment[4]. Besides static, single frame imaging, CT is increasingly being used to evaluate heart dynamics[5,6,7]. Various techniques have been developed to avoid, reduce, and/or correct motion artifacts. Broadly, approaches aim to either reduce the amount of data needed for image reconstruction[8,9,10] or correct for the underlying object motion[11,12]. However, as described in Section II., current methods remain limited.

In this work, we propose a new approach for time-resolved reconstruction of CT images of moving objects without estimation or explicit correction of the underlying object motion. Traditional intensity-based image reconstruction aims to represent each spatial position of the scene and therefore becomes suboptimal for scenes with motion as the problem becomes ill-posed. Our approach is based on the insight that object boundaries can simultaneously be used to represent an object's shape as well as its motion over time. As shown in the *toy example* in Fig 1, a signed distance function-based representation provides an alternative, efficient representation to intensity mapping which we leverage to solve for both the shape and motion.

Toy example:

Consider a white 2D ellipse which moves from left to right, over a black background, across Y time frames. We can represent the scene at each time frame by measuring a per-pixel quantity $f(\mathbf{x})$ where $\mathbf{x}$ refers to the coordinate of a certain pixel in the image. Traditionally, $f(\mathbf{x})$ is the image intensity. In this toy example, $f(\mathbf{x})$ would assume a discrete value (white or black). Alternatively, $f(\mathbf{x})$ can be defined as a signed distance function i.e. SDF($\mathbf{x}$) which is a continuous value equal to the smallest signed distance of $\mathbf{x}$ to the object's boundary (shown by the radii of green and blue circles). Note that SDF($\mathbf{x}$) < 0 if $\mathbf{x}$ is inside the object's boundary and SDF($\mathbf{x}$) > 0 when outside the object. While both definitions of $f(\mathbf{x})$ represent the object's state in each of the three frames, the change in $f(\mathbf{x})$ as a function of time (bottom) is different. The intensity-based representation (bottom left) leads to discontinuous

changes over time which are hard to represent by a Lipschitz continuous function. On the other hand, the signed distance function (bottom right) yields a smooth curve which can be easily fitted. This example highlights how the *signed distance may be a superior representation of an object's spatiotemporal state compared to the commonly used image intensity*.

A discrete voxel grid can be used to represent the SDF of a scene but it leads to a huge memory footprint when extended to the temporal domain. This motivates our use of neural network-based representations, commonly referred to as neural implicit representations (NIRs), as they are capable of efficiently representing complex scenes[13]. Neural representations are highly flexible, can represent arbitrarily complex scenes with only a few MB of memory[14], and can be easily optimized via gradient descent.

In this work, we use NIRs to map each spatiotemporal position to its corresponding signed distance. Each scene is represented as objects-of-interest with temporally-evolving boundaries by means of neural implicit SDF representation instead of as scalar-valued intensity image or set of images. Via optimization, we estimate the position and displacement of object boundaries over time which is most consistent with the acquired sinogram. Instead of using data-driven priors, we promote convergence towards low frequency solutions (i.e., physically plausible reconstructions) by using Fourier coefficients to represent the temporal evolution of the SDF and user-specified spatial and temporal smoothness constraints. Therefore, our approach solves for the objects position and displacement without assuming a particular type of object motion. It also does not require additional data (e.g., extra CT projections, motion field estimates, or physiologic signals such as the ECG).

Prior work has demonstrated that projections can be used to identify the position of object boundaries[15,16,17]. However, these efforts did not address the challenges encountered with a moving scene. Further, prior work has demonstrated how NIRs can be optimized directly from projection images via volumetric differentiable renderers[13,18]. Reconstruction of CT images is analogous to the inverse volumetric object rendering problem dealt with in these previous works. As part of our framework DiFiR-CT, we propose an algorithm that optimizes the spatiotemporal representation of a moving object in both the image domain and neural domain across three stages: Initialization (Section IV.A), Training (Section IV.B) and Refinement/Export (Section IV.C). We show that this framework is robust to the choice of hyperparameters and performs well despite the presence of noise in the acquired sinogram (Section V.) across several different problems (Section VI.).

Therefore, the contributions of this work are as follows:

- We pose CT image reconstruction as an object boundary-based problem and show that this enables improved representation and reconstruction of moving objects.
- We present a pipeline, DiFiR-CT, which combines neural rendering with an implicit representation to reconstruct objects undergoing fast and complex deformations.

- We demonstrate that DiFiR-CT can reconstruct movies of moving objects imaged with standard axial CT acquisitions without assuming priors related to the motion.
- We highlight that DiFiR-CT performance is robust to the choice of hyperparameters, is robust to noise in the acquired sinogram, and can be readily used to solve for a wide range of motions.

This paper is organized as follows. Section II. discusses prior works related to DiFiR-CT. A component-wise description of DiFiR-CT is covered in Section III. followed by implementation details in Section IV. Sections V. and VI. respectively detail the experiments and results.

II. Background

II.A. Avoidance, Estimation, or Reduction of Motion Artifacts

Image artifacts can be minimized by limiting motion during acquisition of CT data. While anatomy of interest can often be held still, this is challenging when imaging the heart due to its near-constant motion. Instead, artifacts due to cardiac motion can be reduced by imaging during portions of the cardiac cycle with relatively small motion (either the end-systole or diastasis periods)[19]. However, this can be of limited use in patients with high or irregular heart rates. Further, avoiding periods of object motion may limit assessment of key properties such as joint articulation or cardiac dynamics.

Faster gantry rotation can also be used to reduce motion artifacts. Increasing the rotation speed for single-source systems is mechanically challenging and requires an increase in x-ray tube power to maintain similar image quality[11]. Dual-source systems have been shown to improve the temporal resolution[20,21]. Unfortunately, the increased mechanical complexity or additional sources has limited design of systems with more than two sources[22]. While non-mechanical, electron beam CT systems, initially introduced in the 1980s, can achieve fast acquisition times (~50ms), suboptimal image quality has limited clinical adoption[23,24]. As a result, most current clinical CT systems are of the single-source design and reconstruct images with temporal footprints >100ms which can result in significant motion artifacts when imaging cardiac structures.

Computer vision approaches have proven beneficial in solving the significantly under-determined problem of jointly estimating image intensity and motion over time on a per-pixel basis[25,26]. For example, partial angle images can be used to estimate and correct for temporally-constant but spatially-varying motion[12,27,28]. One such approach (“SnapShot Freeze”, GE Healthcare) has been developed into a clinical solution and been shown to improve image quality[29] and reduce the presence of artifacts[30]. More recent work has leveraged machine learning to improve motion estimation[31]. However, the ability of these methods correct complex motions remains unclear despite advanced phantoms such as XCAT[32] being used to improve evaluation[33].

Machine learning has also been used to correct motion artifacts in reconstructed images. For example, Ko et al. developed a deep convolutional neural network (CNN) to compensate

for both rigid and nonrigid motion artifacts[34]. Similarly, Lossau et al. used three CNNs to reduce the metal artifacts created by pacemakers[35]. The classification of coronary artery plaques blurred by motion has also been improved by fine tuning of inception v3 CNN[36].

II.B. Neural Implicit Representations and Reconstruction

Neural implicit representations (NIRs) are an efficient approach for storage of high resolution information. This can be attributed to the fact that NIRs are continuously differentiable, which allows for both theoretically infinite resolution and efficient optimization using classic gradient descent techniques[37]. In contrast, conventional representations either require significantly higher memory usage (voxels[38,39]), make compromises regarding topology information (point clouds[40]) or give unrealistic surfaces after optimization (meshes[41]).

Since classic shape reconstruction works by Park et al. and Mescheder et al.[42,43], NIRs have been used for novel view synthesis[13,44,45,46] and multi-view reconstruction[47,48,49]. Recently, NIRs have been used to improve CT[50,51,52,53,54,55] and MR[56] imaging. The use of NIRs to help solve inverse problem of CT reconstruction was first shown by Sun[50]. Such techniques have since shown to enable super resolution[51] and improve sparse CT reconstruction[52,53,54] In most cases, these approaches model the data acquisition process to optimize the NIRs for static objects. In contrast, the focus of this paper is to reconstruct objects that move during data acquisition. NIRs have been extended to dynamic scenes by warping a learned template scene with estimated motion field[55]. In this work, the scene and motion fields are modeled separately. Due to the reliance on a template, this method fails to handle complex topology changes. Our approach aims to accurately capture boundary motion and enable time-resolved reconstruction by representing the position and motion of object boundaries via SDFs, a physically motivated representation. Moreover, we demonstrate that our technique can utilize the result of conventional reconstructions (in our case, filtered backprojection) as an effective initialization. We refer readers to[14] for a survey on NIRs.

III. DiFiR-CT

Here, we describe the core components of our approach, namely the signed distance-based neural implicit representation and our differentiable renderer which allows for optimization of the representation via analysis-by-synthesis[57].

III.A. Signed distance-based NIR

A signed distance-based implicit representation seeks to store the object's shape as a zero level-set of a higher dimensional function. This can be extended to scenes with multiple objects by representing each object with their own signed distance function. In this work, we implicitly represent a scene using a vector of signed distance functions (SDF) $\mathrm{f} : \mathbb{R}^N \times \mathbb{R} \rightarrow \mathbb{R}^K$ to map the spatiotemporal coordinate $\langle \mathbf{x}, t \rangle \in \mathbb{R}^N \times \mathbb{R}$ to the boundaries of K objects in spacetime represented by their SDF values $\mathbf{f}(\mathbf{x}, t) \in \mathbb{R}^K$ where N is the number of spatial dimensions and K is the number of objects represented. As discussed in Section II.D, fitting a neural network is an efficient and differentiable alternative to defining SDFs

on a discrete set of grids. Further, a neural network compactly adapts the spatiotemporal resolution of the representation to the complexity of the underlying object.

We approximate the SDF representation $\mathbf{f}(\mathbf{x}, t)$ by a neural network $\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w})$, where $\mathbf{w}$ is the set of network weights (see Fig. 2). We do this by creating two sub-networks that jointly approximate the shape of the SDF $\hat{\mathbf{g}}_{\mathrm{E}}$ as well as its spatiotemporal displacement $\hat{\mathbf{g}}_{\mathbf{V}}$. While this was done explicitly via two sub-networks and was intended to ensure the network had sufficient expressive capacity, optimization of the MLP for specific tasks is left for future work.

EncoderNet—For each location $\mathrm{x} \in \mathbb{R}^N$, the stationary component of the object's SDF was represented using a network EncoderNet $\hat{\mathbf{g}}_{\mathbf{E}}(\mathbf{x}; \mathbf{w}_{\mathbf{E}}): \mathbb{R}^N \rightarrow \mathbb{R}^K$.

VelocityNet—For each location $\mathbf{x} \in \mathbb{R}^N$, the temporal evolution of an object's SDF was parameterized using Fourier coefficients $(\mathscr{A}_{ij}, \mathscr{B}_{ij}), i \in \{1, \cdots, M\}, j \in \{1, \cdots, K\}$ which are respectively the coefficients for sines and cosines of K objects. The coefficients were fit by a network $\hat{\mathbf{g}}_{\mathbf{F}}(\mathbf{x}, t; \mathbf{w}_{\mathbf{F}}): \mathbb{R}^N \rightarrow \mathbb{R}^{2MK}$. Then, $\hat{\mathbf{g}}_{\mathbf{V}}(\mathbf{x}, t; \mathbf{w}_{\mathbf{F}}): \mathbb{R}^N \rightarrow \mathbb{R}^K$ was computed as:

$$\hat{\mathbf{g}}_{\mathbf{V}}(\mathbf{x}, t; \mathbf{w}_{\mathbf{F}}) = \frac{1}{M} \begin{bmatrix} \mathscr{A}_{11}(\mathbf{x}) & \cdots & \mathscr{A}_{1M}(\mathbf{x}) & \mathscr{B}_{11}(\mathbf{x}) & \cdots & \mathscr{B}_{1M}(\mathbf{x}) \\ \mathscr{A}_{21(\mathbf{x})} & \cdots & \mathscr{A}_{2M}(\mathbf{x}) & \mathscr{B}_{21}(\mathbf{x}) & \cdots & \mathscr{B}_{2M}(\mathbf{x}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathscr{A}_{K1}(\mathbf{x}) & \cdots & \mathscr{A}_{KM}(\mathbf{x}) & \mathscr{B}_{K1}(\mathbf{x}) & \cdots & \mathscr{B}_{KM}(\mathbf{x}) \end{bmatrix} \begin{bmatrix} sin(2\pi\omega_1 t) \\ sin(2\pi\omega_2 t) \\ \vdots \\ sin(2\pi\omega_M t) \\ cos(2\pi\omega_1 t) \\ cos(2\pi\omega_2 t) \\ \vdots \\ cos(2\pi\omega_M t) \end{bmatrix} \quad (1)$$

Here, $\omega_i \sim \mathcal{N}(0, F_{\max})$ is a random variable sampled from a normal distribution with standard deviation F_{max} (Hz). We modeled our randomly sampling of Fourier features on work done by Rahimi and Recht[58]. However, we found that sampling from a normal distribution (instead of a uniform distribution) led to better optimization. The random sampling is performed once, at the beginning of each reconstruction. In practice, we expect physical constraints allow for bandlimiting of the Fourier representation and parameterization of the network with F_{max}. We determined F_{max} empirically with the goal of encouraging VelocityNet to represent physically-relevant displacement fields and found M=128 frequencies to be sufficient for our test cases despite earlier results utilizing M=256[59]. As shown in Equation 1, $\hat{\mathbf{g}}_{\mathbf{V}}(\mathbf{x}, t; \mathbf{w}_{\mathbf{F}})$ yields a distance value based on the Fourier representation encoded by $\hat{\mathrm{g}}_{\mathbf{F}}(\mathbf{x}, t; \mathbf{w}_{\mathbf{F}})$. In other words, $\hat{\mathbf{g}}_{\mathbf{V}}(\mathbf{x}, t; \mathbf{w}_{\mathbf{F}})$ recombines the 2MK output of $\hat{\mathbf{g}}_{\mathbf{F}}(\mathbf{x}, t; \mathbf{w}_{\mathbf{F}})$ into the desired K-dimensional SDF displacement values such that they can be summed with the result from the EncoderNet.

Implementation SIRENs[60] are an efficient framework capable of capturing high frequency information and, therefore, were used to implement both EncoderNet and VelocityNet. As a result, the overall representation is given by

$$\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w}) = \hat{\mathbf{g}}_{\mathbf{V}}(\mathbf{x}, t; \mathbf{w}_{\mathbf{F}}) + \hat{\mathbf{g}}_{\mathbf{E}}(\mathbf{x}; \mathbf{w}_{\mathbf{E}}) \quad (2)$$

where $\mathbf{w}$ is the union of EncoderNet weights $\mathbf{w}_{\mathbf{E}}$ and VelocityNet weights $\mathbf{w}_{\mathbf{F}}$. i.e. $\mathbf{w} = \mathrm{w}_{\mathbf{E}} \cup \mathrm{w}_{\mathbf{F}}$.

III.B. Differentiable Renderer

The differentiable renderer projects a spatiotemporal attenuation intensity $I(\mathbf{x}, t)$ to the sinogram domain $r(l, \theta)$ using a model of the CT system. This is done by integrating attenuation encountered by rays traveling from the source to a specific detector position l at a gantry position θ. As the gantry rotates over time t, the spatiotemporal attenuation intensity $I(\mathbf{x}, t)$ is *rendered* resulting in the sinogram value $r(l, \theta)$ as defined by the following integral:

$$r(l, \theta) = \int_u I(\mathbf{x}, t) \Gamma(\theta(t)) du \quad (3)$$

where u is the path the ray traverses through the scene and $\Gamma(\theta(t))$ is the time-varying two-dimensional rotation matrix which describes the gantry rotation by an angle θ about the center of the scene which varies as a function of time. Note that the the computation for $r(l, \theta)$ can be performed *differentiably* using appropriate sampling of the above integral. This means that the spatiotemporal attenuation intensity $I(\mathbf{x}, t)$ can be updated based on some error functional $E(r)$ defined on the rendered sinogram $r(l, \theta)$ by backpropagating its gradient $\frac{\partial E(r)}{\partial r(l, \theta)}$.

In our framework, the attenuation intensity $\hat{I}(\mathbf{x}, t)$ is encoded by means of object-wise spatiotemporal SDFs encoded in the network $\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w})$. Note, hat notation $\hat{\cdot}$ is used to distinguish quantities *predicted* by a neural network from ground truth values. $\hat{I}(\mathbf{x}, t)$ is extracted from $\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w})$ in a *differentiable* manner so that gradients of the error functional ∂E can be used to directly update the weights $\mathbf{w}$ of our implicit representation $\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w})$. As shown in Fig 3, the process begins by uniformly querying the SDF representation $\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w})$ over the N-dimensional grid of size d at some time frame t. This results in a unique, albeit discrete, SDF representation $\hat{\mathbf{f}}_{\square}(\mathbf{x}, t, k, \hat{\mathbf{g}})$ for each of the k objects as separate channels of the network's output volume. Note, the $\square$ indicates a discrete representation so $\hat{\mathbf{f}}_{\square}(\mathbf{x}, t, k, \hat{\mathbf{g}})$ is the estimate of the SDF sampled on a discrete pixel/voxel grid. This discrete representation is then converted into an occupancy via $\zeta(y)$, defined as:

$$\zeta(y) = min(1, max(0, \mu^*(\sigma(y) - 0.5))) \quad (4)$$

where μ is a scaling factor that controls the sharpness of the boundary and σ refers to the sigmoid function. The spatiotemporal intensity map $\hat{I}(\mathbf{x}, t)$ can be computed by taking the inner product of per-channel occupancy and attenuation intensity $a(k)$ (assumed to be known apriori).

$$\hat{I}(\mathbf{x}, t) = a(k)\zeta(\hat{f}_{\square}(\mathbf{x}, t, k, \hat{g})) \quad (5)$$

In order to perform accurate rendering and model the spatially-integrating detectors used for CT, the derived occupancy grid $\hat{\mathbf{f}}\left(\mathbf{x}, t; \mathbf{w}\right)$, was upsampled k_{samp} times using bilinear interpolation before performing the rendering operation. Subsequently, the projected sinogram $\hat{r}(l, \theta)$ was downsampled via average pooling. These steps describe a differentiable rendering operator $R\left(\cdot; \theta\right): \mathbb{R}^{d^N} \times \mathbb{R} \rightarrow \mathbb{R}^{d^N - 1}$ which can map an SDF representation $\hat{g}(\mathbf{x}, t; \mathbf{w})$ to the projection domain $\hat{r}(l, \theta)$ via appropriate discretization. This procedure is end-to-end differentiable and allows for backpropagating gradients of a functional E such as the L_1 distance between $r(l, \theta)$ and $\hat{r}(l, \theta)$ to the network weights $\mathbf{w}$.

IV. Reconstruction using DiFiR-CT

DiFiR-CT leverages the fact that the spatiotemporal attenuation map of a moving object can be represented both explicitly in the image domain as well as implictly in the neural domain. That is, as weights of a neural network. As outlined in Figure 4, DiFiR-CT performs optimization in both of these domains across three stages: Initialization, Training, Refinement/Export.

Briefly, Initialization is intended to obtain well-behaved gradients. One way we do this is by leveraging the fact that the norm of the gradient of the signed distance field is always unity (i.e., the Eikonal constraint[60]). Further, instead of using a random initialization for the neural SDF representation, we utilize the filtered back projection result. Algorithm 1 identifies objects of interest in an image reconstructed via filtered back projection (FBP) using a simple intensity-based segmentation. Algorithm 2 encodes the resulting segmentation (binary background-foreground) image as signed distance functions and Algorithm 3 performs the explicit-to-implicit representation conversion.

The training portion begins with Algorithm 4 which is responsible for updating (a.k.a. training) the neural SDF representation to match the sinogram data. Subsequently, **Algorithm 5** performs the implicit-to-explicit conversion and **Algorithm 6** creates the occupancy image for the discretized SDF map.

Results improved when DiFiR-CT was re-initialized using the result of it's first prediction. Therefore, Refinement/Export consists of applying Algorithms 2 – 6 on the initial result of DiFiR-CT.

IV.A. Initialization

The SDF of K foreground objects of interest is initialized using the filtered back projection (FBP) reconstruction images $I_{FBP}(\mathbf{x}, t)$. Here, the number of objects of interest K is defined apriori. Defining both a background class as well as several additional classes κ such that $K' = K + \kappa + 1$ improved initialization.

Algorithm 1 performs intensity-based segmentation using a Gaussian Mixture Model[61] $GMM(I_{FBP}(\mathbf{x}, t), k')$ to create segmentation images $G(\mathbf{x}, t)$. This results in binary classification images $C(\mathbf{x}, t, k')$. From these binary images, the mass $\mathscr{M}(k')$ of each foreground class was

calculated. After removal of the background $c_{background}$, defined as the class with the largest mass, the binary images $C(\mathbf{x}, t, k)$ for the K largest classes were kept.

Algorithm 1: Object Segmentation

Input: FBP images $I_{FBP}(\mathbf{x}, t) \forall t \in T$
number of objects K
buffer classes κ
Output: Binary classification images $C(\mathbf{x}, t, k) \forall t \in T$

Randomly, select $\tau \subset T$
$Segmentor \leftarrow \text{GMM}(I_{FBP}(\mathbf{x}, t), k')\ t \in \tau$
$G(\mathbf{x}, t, k') \leftarrow Segmentor.\text{fit}(I_{FBP}(\mathbf{x}, t) \forall t \in T)$
$\mathcal{M} \leftarrow \{\cdots, \sum_{x \in \Omega} G(\mathbf{x}, t, c_i), \cdots\}$
$c_{background} \leftarrow \underset{c}{argmax}\ \mathcal{M}$
$\mathcal{M} \leftarrow \mathcal{M}.\text{delete}(c_{background})$
$\{c_i\} \leftarrow \text{TopKIndices}(\mathcal{M})$
$C(\mathbf{x}, t, k) \leftarrow G(\mathbf{x}, t, \{c_i\})$
return $C(\mathbf{x}, t, k)$;

Algorithm 2: Convert Binary images to SDF

Input : Binary class images $C(\mathbf{x}, t, k) \forall t \in T$
Output: SDF images $\hat{f}(\mathbf{x}, t, k) \forall t \in T$

$\tilde{B}(\mathbf{x}, t, k) \leftarrow \text{TotalVariationMinimizer}(C(\mathbf{x}, t, k))$
$\tilde{D}(\mathbf{x}, t, k) \leftarrow \text{DistanceTransform}(\tilde{B}(\mathbf{x}, t, k))$
$\hat{f}(\mathbf{x}, t, k) \leftarrow \text{TotalVariationMinimizer}(\tilde{D}(\mathbf{x}, t, k))$
return $\hat{f}(\mathbf{x}, t, k)$;

Binary class images $C(\mathbf{x}, t, k)$ representing pixels of each label across time were then converted into SDF images $\hat{f}(\mathbf{x}, t, k)$ using Algorithm 2. Our approach uses total variation minimization[62] to smooth the images and then performs the distance transform over the entire domain to obtain SDF images $\hat{f}(\mathbf{x}, t, k)$.

To create neural implicit representations from the explicit SDF images $\hat{f}(\mathbf{x}, t, k)$, we performed optimization as described in Algorithm 3. A randomly initialized neural network $\mathbf{g}(\mathbf{x}, t; \mathbf{w})$ is optimized into a network $\hat{\mathbf{g}}_o(\mathbf{x}, t; \mathbf{w})$ which best approximates the explicit SDF image $\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w}) \approx \hat{f}(\mathbf{x}, t, k)$. The optimization is directly supervised and aims to minimize the SDF differences $\mathscr{L}_{SDF}$ and satisfy the Eikonal constraint $\mathscr{L}_{Eikonal}$. The optimization ends when the maximum number of iterations *maxIterations* have been reached.

IV.B. Training

At this point, the SDF representation $\hat{\mathbf{g}}_o(\mathbf{x}, t; \mathbf{w})$ includes motion artifacts present in I_{FBP}. Algorithm 4 optimizes the neural network SDF representation $\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w})$ to best explain the acquired sinogram $r(\mathbf{l}, \theta)$ with the minimum total variation in space $\| \nabla_x \hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w}) \|_1$ and time $\| \nabla_t \hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w}) \|_1$.

The current neural implicit SDF prediction $\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w})$ can be converted to a spatiotemporal intensity map $\hat{I}(\mathbf{x}, t)$ and projected to the sinogram domain via previously described renderer (Section III.) $R(\hat{I}(\mathbf{x}, t), \theta)$. This results in a sinogram estimate $\hat{r}(\mathbf{l}, \theta)$ used to calculate the sinogram loss $\mathscr{L}_{Sinogram}$ — the difference between the current estimate and the acquired sinogram $r(\mathbf{l}, \theta)$. This loss is combined with the Eikonal constraint $\mathscr{L}_{Eikonal}$ and spatial $\mathscr{L}_{TVS}$ and temporal TV $\mathscr{L}_{TVT}$ loss terms to improve optimization. The optimization was performed until the projection loss $\mathscr{L}_{Sinogram}$ decreased below a certain threshold *minLoss* or maximum number of iterations *maxIterations* was reached.

Algorithm 3: Initializing Neural SDF Representation

Input : Neural SDF $\mathbf{g}(\mathbf{x}, t; \mathbf{w})$, SDF images $f(\mathbf{x}, t, k)$

Output: Initialized neural SDF $\hat{\mathbf{g}}_o(\mathbf{x}, t; \mathbf{w}) \forall t \in T$

$\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w}) = \mathbf{g}(\mathbf{x}, t; \text{rand}(\mathbf{w}))$

for *iteration* $<$ *maxIterations* **do**

  Sample $t \sim T$

  $\mathcal{L}_{SDF} \leftarrow \frac{1}{|\Omega|} \sum_{x \in \Omega} \left| \hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w}) - \hat{f}(\mathbf{x}, t) \right|$

  $\mathcal{L}_{Eikonal} \leftarrow \frac{1}{|\Omega|} \sum_{x \in \Omega} |\|\nabla_x \hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w})\|_2 - 1|$

  $\mathcal{L} \leftarrow \mathcal{L}_{SDF} + \lambda \mathcal{L}_{Eikonal}$

  $\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla \mathcal{L}$

$\hat{\mathbf{g}}_o(\mathbf{x}, t; \mathbf{w}) \leftarrow \hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w})$

return $\hat{\mathbf{g}}_o(\mathbf{x}, t; \mathbf{w})$;

IV.C. Refinement/Exporting

To generate spatiotemporal intensity images from the neural SDF representation, we first convert the neural SDF into a set of explicit SDF images $\hat{f}_{\square}(\mathbf{x}, t, k)$. This is achieved by sampling the neural SDF $\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w})$ over a n-dimensional grid at a desired spatial resolution (Algorithm 5). The resulting SDF image is then binarized $\hat{B}(\mathbf{x}, t, k) = \zeta(\hat{f}_{\square}(\mathbf{x}, t, k))$ (**Algorithm 6**). Binarized images $\hat{B}(x, t, k; w)$ are then used for Refinement, a second pass through the algorithm (starting with Algorithm 2). After Refinement, images were exported and analyzed.

Algorithm 4: Optimizing Neural SDF Representation

Input : Initialized SDF $\hat{\mathbf{g}}_o(\mathbf{x}, t; \mathbf{w}) \forall t \in T$, acquired sinogram $r(\mathbf{l}, \boldsymbol{\theta})$

Output: Optimized SDF $\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w}) \forall t \in T$

for *iteration* $<$ *maxIterations* **do**

  Sample $t \sim T$;

  Compute projection $\hat{r}(\mathbf{l}, \boldsymbol{\theta}) \leftarrow R(\hat{I}(\mathbf{x}, t), \theta(t))$;

  $\mathcal{L}_{Sinogram} \leftarrow \frac{1}{|\Omega|} \sum_{x \in \Omega} |r(\mathbf{l}, \boldsymbol{\theta}) - \hat{r}(\mathbf{l}, \boldsymbol{\theta})|$

  $\mathcal{L}_{Eikonal} \leftarrow \frac{1}{|\Omega|} \sum_{x \in \Omega} |\|\nabla_x \hat{\mathbf{g}}_o(\mathbf{x}, t; \mathbf{w})\|_2 - 1|$

  $\mathcal{L}_{TVS} \leftarrow \frac{1}{|\Omega|} \sum_{x \in \Omega} \|\nabla_s \hat{\mathbf{g}}_o(\mathbf{x}, t; \mathbf{w})\|_1$

  $\mathcal{L}_{TVT} \leftarrow \frac{1}{|\Omega|} \sum_{x \in \Omega} \|\nabla_t \hat{\mathbf{g}}_o(\mathbf{x}, t; \mathbf{w})\|_1$

  $\mathcal{L} \leftarrow \mathcal{L}_{Sinogram} + \lambda_1 \mathcal{L}_{Eikonal} + \lambda_2 \mathcal{L}_{TVS} + \lambda_3 \mathcal{L}_{TVT}$

  $\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla \mathcal{L}$

  if $\mathcal{L}_{Sinogram} < minLoss$ **then**

    break

$\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w}) \leftarrow \hat{\mathbf{g}}_o(\mathbf{x}, t; \mathbf{w})$;

return $\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w})$;

V. Experiments

V.A. Implementation Details

The image resolution used during training was set to $n = 128$ with upsampling factor $k_{samp} = 2$. For optimal performance, we used $F_{max} = 3.0$ and $\mu = 50$. In Algorithm 1, we sampled τ such that $|\tau| = 0.02 * |T|$ and set $\kappa = 3$. For Algorithms 3 & 4 the learning rate was $\alpha = 1 \times 10^{-5}$ and decayed by a factor of 0.95 every 200 steps. The optimization procedures was run with *maxIterations* = 5000, *minLoss* = 0.08, and $\lambda = 0.1$ and performed 5 times, each with a different random seed for the neural network. For best results, we trained with minibatch $|t| = 20$. As described below, we observed optimal results with $\lambda_1 = 0.1$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.5$. All experiments were performed using a 2D parallel beam CT geometry of a single foreground class (i.e. $K = 1$) except for the multi-object study. Optimizing DiFiR-CT takes ~ 30 mins on a single NVIDIA 1080Ti GPU.

V.B. Evaluation of DiFiR-CT Parameters

First, we simulated a simple cause of motion corruption: translation of a circle during imaging. This was intended to mimic motion of a coronary artery[3]. A circle with a diameter

25% of the size of the image was placed that same distance (25% of the image) from the center and angularly translated during the acquisition. We used this toy problem to explore the impact of λ_1, λ_2, and F_{max} on the ability of DiFiR-CT to faithfully reconstruct a moving object. λ_1 and λ_2 are the weight of spatial and temporal total variation in the cost function used to supervise learning and F_{max} is the maximum Fourier coefficient available for representation of the temporal changes.

The impact of these parameters on DiFiR-CT was evaluated via parameter sweeps. First, λ_1 and λ_2 were varied (0, 0.05, 0.1, 0.15, 0.3, 0.5, 0.9, 1.5, 3.0, 5.0, 10.0) with F_{max} set to 3.0. Then, F_{max} was varied from 0 to 10.0 with $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$,

We evaluated a range of angular displacements per gantry rotation (0, 1, 5, 10, 20, 40, 70, 100, 120, 150, and 200 degrees per gantry rotation). DiFiR-CT results were compared to both the ground-truth vessel image as well as the FBP result using the metrics described below.

V.C. Evaluation of Noise on DiFiR-CT

The impact of noise in the acquired sinogram on the quality of DiFiR-CT reconstruction was evaluated using imaging data obtained of a circle undergoing an angular displacement of 100 degrees. Poisson-distributed quanta-counting noise[63] was added to the acquired sinogram and the incident x-ray beam intensity was modulated to create images with a range of contrast-to-noise ratios (from 3 to 40). Five different noise realizations were created and DiFiR-CT and FBP results were compared to the ground-truth using the metrics described below.

V.D. Multi-object DiFiR-CT and Impact of Initial Segmentation

As outlined above, a key step in the DiFiR-CT framework is the initialization of the SDF map of scene. Above, we described a Gaussian Mixture Model (GMM) approach that is solely based on the intensity histogram in $I_{FBP}(\mathbf{x}, t)$. However, when applied to a scene with multiple moving objects, each with different attenuations, the GMM may fail to differentiate objects based solely on the intensity distribution. Figure 5 shows a failure of the GMM approach when analyzing the FBP reconstruction of two moving dots with two different attenuations (top = 0.7, bottom = 0.2). GMM identifies two intensity values of interest by evaluating the distribution of pixel intensities. However, the second object identified is the motion artifact of the brighter dot. Therefore, we refined the segmentation by providing some spatial information. A second approach, SegSI, uses a Region-Of-Interest (ROI) to guide thresholding-based segmentation. A bounding box was defined to only contain one moving dot such that we assigned one individual class to each box. In the box, we defined an intensity threshold = $\gamma \times I_{max}$ where I_{max} is the maximum intensity in the scene in each box to capture the real object. $\gamma = 0.7$ was set empirically. In all other respects, DiFiR-CT remained the same such that we could illustrate the improvement associated with an improved initialization.

V.E. Imaging of Nonrigid, Heart-like Motion

To evaluate the utility of DiFiR-CT during non-rigid motion, we modeled heart-like motion using an ellipse with time-varying axes. In addition to static imaging, cine CT imaging can be used to evaluate heart size and dynamics. Therefore, we evaluated two imaging approaches: 1) Single-gantry rotation imaging where a sinogram spanning 360 degrees is acquired during which the ellipse diameter changes. 2) Full heart cycle imaging where multiple (typically 4–5) gantry rotations were obtained spanning the entire period of the heart-like motion. Of note, DiFiR-CT readily incorporate multiple rotation data into the reconstruction framework without modification. In this scenario, DiFiR-CT results were compared to FBP reconstructions centered at the same temporal position.

V.F. Imaging of Complex Deformations

Lastly, we demonstrate the ability of DiFiR-CT to reconstruct scenes with complex topological change. To do so, we created a complex scene where the letter “A” transforms into “B”, then into “C’, then back to “B’ and “A”. Without changes to any DiFiR-CT parameters (spatial/temporal weighting and Fourier coefficients), DiFiR-CT results were compared to FBP imaging.

V.G. Metrics to Evaluate Image Quality

DiFiR-CT was compared to conventional FBP images using mean-square error (MSE) and the foreground’s Dice coefficient where FBP images were thresholded using half of the foreground intensity. We did not compare DiFiR-CT to motion correction methods. This was motivated by the fact that the improvement obtained via correction depends on the suitability of the motion model in the correction scheme to the motion present in the acquired data. Given that current correction methods have been designed for clinical scenarios, our motion scenarios are not expected to represent realistic use cases. For DiFiR-CT, we report the median accuracy of 5 independent optimizations, where each optimization is initialized with a different random seed.

VI. Results

VI.A. Moving Coronary Vessel Imaging

As shown by the images and metrics in Figure 6, DiFiR-CT accurately reconstructed images of a small circle while FBP images showed significant artifacts ($\lambda_1 = 0.5$, $\lambda_2 = 0.5$, and $F_{max} = 3.0$). Low MSE and high (>0.9) Dice was maintained for angular displacements of up to 150° during data acquisition. Figure 6 also illustrates how the DiFiR-CT reconstruction is impacted by displacements >150°. While the shape of the circle is preserved, increased MSE and decreased Dice occur primarily due to reconstruction of the circle at the incorrect temporal position.

VI.B. Evaluation of DiFiR-CT Parameters

DiFiR-CT reconstruction of a circle with translation = 100° per gantry rotation was used to evaluated the robustness of DiFiR-CT to changes in λ_1, λ_2, and F_{max}. Figure 7 (top) illustrates the minor impact of changing the strength of spatial (λ_1) and temporal (λ_2) regularization

while maintaining $F_{max} = 3.0$. Accurate reconstruction was achieved over a wide range of regularization strength. Figure 7 (bottom) illustrates the performance of DiFiR-CT with $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$ when varying F_{max}. Decreasing F_{max} limits the temporal evolution that can be represented by the neural implicit approach. However, we observed robust DiFiR-CT reconstruction performance even at low F_{max} values. Reconstruction accuracy decreased at high F_{max}. This suggests that introduction of high frequency coefficients can result in overfitting.

VI.C. DiFiR-CT Reconstruction of Motion and Noise

Without modification of the parameters identified above ($\lambda_1 = 0.5$, $\lambda_2 = 0.5$, and $F_{max} = 3.0$), DiFiR-CT successfully reconstructed images across a range of contrast-to-noise values in the acquired sinogram. Figure 8 illustrates the high image quality of DiFiR-CT as well as the loss of image quality using the FBP approach. Despite significant object motion and low (≤ 5) CNR, DiFiR-CT maintained MSE ≤ 0.0031 and Dice ≥ 0.91 while FBP results were significantly worse in this scenario (MSE ≥ 0.02, Dice ≤ 0.05).

VI.D. Multi-object DiFiR-CT and Impact of Initial Segmentation

Reconstruction with DiFiR-CT-SegGMM was limited when $\Delta\theta > 60$ degrees (Figure 9). In contrast, DiFiR-CT-SegSI maintained high-quality motion-corrected reconstructions for all $\Delta\theta$ and achieved low RSME (<0.028) and high (>0.89) DICE for $\Delta\theta$ up to 160 degrees. This illustrates several key features of the DiFiR-CT framework. First, solving for two intensities illustrates the ability of DiFiR-CT to be extended to multi-intensity scenes. Second, a spatially-naive segmentation can lead to limited performance. However, simple modifications such as the use of a ROI can significantly improve the result. For example, one potential use of DiFiR-CT could be to correct motion artifacts of a particular object that the user selects via an ROI.

VI.E. Imaging Nonrigid, Heart-like Motion with DiFiR-CT

Without modification of the parameters identified above ($\lambda_1 = 0.5$, $\lambda_2 = 0.5$, and $F_{max} = 3.0$), DiFiR-CT successfully reconstructed images of the heart-like motion for both single gantry and full heart cycle imaging. Relative to FBP, DiFiR-CT improved imaging metrics when imaging an ellipse with changing dimensions. Figure 10 illustrates these differences.

The change in axes dimension is shown by the left two columns (temporal reformats along the x- and y-direction). Metrics of image quality (MSE and Dice) are shown on the right. For full heart cycle imaging with beating velocity = 3.0, DiFiR-CT decreased MSE (FBP: median 0.005, IQR 0.003–0.009, DiFiR-CT: median 0.001, IQR 0.001–0.002). Further, DiFiR-CT increased the percentage of frames with MSE < 0.005 from 46% (FBP) to 100%. DiFiR-CT also improved Dice (FBP: median 0.84, IQR 0.73–0.88 to DiFiR-CT: median 0.96, IQR 0.95–0.97). Supplemental Movie 1 illustrates the ground truth motion and improvement of DiFiR-CT reconstruction, relative to FBP.

VI.F. Imaging Complex Topology Changes with DiFiR-CT

Without modification of the DiFiR-CT framework or parameters ($\lambda_1 = 0.5$, $\lambda_2 = 0.5$, and $F_{max} = 3.0$), DiFiR-CT successfully reconstructed data acquired during a complex letter warping scene. Figure 11 shows five frames of the deformation with the groundtruth (top) as well as FBP (middle) and DiFiR-CT (bottom) reconstructions. DiFiR-CT significantly reduced the severity of artifacts created when using FBP. The plot of MSE and Dice scores as a function of time further illustrate the improvement. DiFiR-CT decreased error as measured via MSE (FBP: median 0.011, IR 0.008–0.019 to DiFiR-CT: median 0.007, IQR 0.004–0.013). Further, DiFiR-CT increased the percentage of the frames with MSE< 0.005 from 15.2% to 34%. DICE scores also improved with DiFiR-CT (FBP: median 0.78, IQR 0.66–0.85, DiFiR-CT: median 0.86, (IQR 0.80–0.91). The percentage of frames with Dice > 0.85 increased from 25.8% for FBP to 58.8% with DiFiR-CT. Supplemental Movie 2 illustrates the ground truth and FBP and DiFiR-CT reconstruction of the complex scene.

VII. Discussion

DiFiR-CT combines implicit representation of an object by a signed distance function (SDF) with differentiable rendering to successfully enable time-resolved imaging free from motion artifacts despite data acquisition occurring during motion. DiFiR-CT takes advantage of several important features of SDFs – namely, that they represent movement of a boundary as a spatially and temporally smooth evolution. DiFiR-CT represents the scene as SDFs which evolve over time using an implicit representation; in this case, a neural network. Differentiable rendering is used to improve the estimate of the scene by comparing the observed CT data with the SDF-based representation in the sinogram domain. The framework also enables additional regularization such as penalizing deviations from the Eikonal constraint and minimizing spatial and temporal variations. We demonstrated the utility of DiFiR-CT in three different imaging scenarios without changes to the architecture. Specifically, DiFiR-CT was readily applied to objects with different motions as well as data spanning one or more than one gantry rotation. These cases highlight how DiFiR-CT can be used to accurately reconstruct objects undergoing 1) translation, 2) heartbeat-like affine changes in diameter, and 3) complex topological changes (warping of letters). This flexibility was facilitated by the fact that DiFiR-CT does not utilize an explicit motion model.

In our evaluation of DiFiR-CT, we observed that accurate reconstructions were possible even with a fairly limited set of Fourier coefficients. This is likely due to the beneficial fact that SDFs used to parameterize the reconstruction problem evolve smoothly over space and time. We also observed reconstruction performance decreased if F_{max} increased. This highlights the benefit of limiting the complexity of the solution via F_{max}. An added advantage is that this hyperparameter has a simple physical interpretation, the bandwidth of the motion, which could facilitate its selection when applying DiFiR-CT to novel applications. It is likely that additional constraints such as preservation of mass could improve imaging in certain scenarios but this is left for future work targeting specific clinical applications.

Further, in our examples, we show how this approach can be used to image one or two foreground objects with an empty background. However, in practice, the background class could be used to represent the position and intensity of static objects. Future work aims to extend this framework to resolve scenes with more complex background intensity.

We illustrate DiFiR-CT results without modification of parameters λ_1, λ_2, and F_{max}. However, the optimal choice of these parameters is expected to vary depending on the specific imaging scenarios. Further, the rotation speed of CT systems can depend on both the design of the system and the clinical protocol. Evaluating the optimal combination of parameters is left for future work where specific scenarios are evaluated.

We illustrate how DiFiR-CT can seamlessly reconstruct data acquired using two common CT acquisition approaches (single and multiple gantry rotation imaging). Specifically, DiFiR-CT reconstructed imaging data spanning multiple gantry rotations simultaneously without explicit labeling of timepoints or the need for specific, complementary images. This is significantly different than approaches such as FBP which reconstruct each frame independently. As a result, in addition to improving single gantry reconstruction, DiFiR-CT also has the potential to improve reconstruction quality of each frame acquired during a multiple gantry acquisition by leveraging additional temporal information. While acquiring additional temporal information (beyond the typical half- or single gantry rotation) increases the dose delivered to the patient, it may enable DiFiR-CT to resolve the dynamics of objects that have been significantly hampered by imaging artifacts.

We believe our findings are noteworthy and novel but our study had several limitations. First, we quantified DiFiR-CT image quality relative to the groundtruth and conventional FBP reconstruction instead of direct comparison to previously-reported motion-correction approaches. We chose not to compare DiFiR-CT to techniques specially crafted for certain imaging scenarios such as translation of a coronary artery. We did so because we expect performance of motion correction approaches to depend significantly on the specifics of the application. For example, we expect motion-correction approaches to accurately reconstruct scenes when the object motion agrees with the method's motion model. However, limited performance may occur if the object motion is different from the motion model. While our approach does not require an *a priori* motion model, it is difficult to ensure that our examples adhere to the constraints incorporated into current approaches. Comparison to current motion-correction algorithms is planned for future work in specific clinical scenarios.

Second, we demonstrated the use of DiFiR-CT using a single-source 2D parallel beam geometry. This was done to simplify interpretability of our findings. However, the 4D problem is a major motivation behind our use of a neural representation for the distance field. By storing the representation as a neural network, our approach does not need to represent the number of pixels/voxels in the final reconstructed image. For example, in our 2D examples, we reconstruct 720 128×128 images (11.8 million pixels) using neural representations with ~200k weights. Practically speaking, our approach has the advantage that encoding a 3D field simply involves adding one more channel/input dimension to our neural representation. Future work is needed to determine the exact architecture needed to

accurately represent a 3D distance field but we expect our method will not suffer from the same scalability challenges faced by voxel-based representations.

Third, while our algorithm has been designed to solve for multiple foreground classes, it currently assumes an object has a uniformity intensity which does not change over time. While non-uniform intensities could be incorporated into our framework, how such a system is optimized is left for future work.

DiFiR-CT could complement previously described motion correction methods that operate on intensity values by providing accurate boundary position information. When solving for multiple moving objects, we observed that accurate optimization begins to depend, to a greater extent, on the initial segmentation of the filtered backprojection image and initialization of the algorithm. The use of a bounding box may be limited in certain clinical applications and was selected as a simple refinement to our process. We expect that more advanced initializations, such as semantic segmentations provided by experts or automated methods, could serve as robust initializations. However, the availability of such segmentations will depend on the clinical application. Therefore, while we demonstrate the possibility of imaging multiple objects, further work is needed to further optimize the initialization, particularly for more complex scenes.

Fourth, we utilized SIRENs as the neural network representation for the spatiotemporal SDF. However, certain networks may provide a better representation, especially when used for specific imaging applications/motion fields. This is left for future work.

Finally, dual-source systems and fan- or cone-beam geometries are expected to change the relationship between object motion and acquired CT data. While we expect DiFiR-CT to improve image quality in these scenarios, we leave these extensions as future work.

VIII. Conclusions

A novel reconstruction framework, DiFiR-CT, can be used to reconstruct CT imaging data acquired during object motion in a time-resolved manner free from motion artifacts. Our approach leverages a neural implicit scheme and does not require a prior motion models or explicit motion estimation. Representing moving boundaries using a signed distance metric and neural implicit framework enables ‘analysis-by-synthesis’ to identify a solution consistent with the observed sinogram as well as spatial and temporal consistency constraints.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Neglia D et al. , Detection of significant coronary artery disease by noninvasive anatomical and functional imaging, Circulation: Cardiovascular Imaging 8, e002179 (2015). [PubMed: 25711274]
2. Hoffmann U et al. , Coronary CT angiography versus standard evaluation in acute chest pain, New England Journal of Medicine 367, 299–308 (2012). [PubMed: 22830462]

3. Contijoch F, Stayman JW, and McVeigh ER, The impact of small motion on the visualization of coronary vessels and lesions in cardiac CT: A simulation study, Medical Physics 44, 3512–3524 (2017). [PubMed: 28432820]
4. De Man B, Nuyts J, Dupont P, Marchal G, and Suetens P, Metal streak artifacts in X-ray computed tomography: a simulation study, in 1998 IEEE Nuclear Science Symposium Conference Record. 1998 IEEE Nuclear Science Symposium and Medical Imaging Conference (Cat. No. 98CH36255), volume 3, pages 1860–1865, IEEE, 1998.
5. Maffei E et al. , Left and right ventricle assessment with Cardiac CT: validation study vs. Cardiac MR, European radiology 22, 1041–1049 (2012). [PubMed: 22270140]
6. Pourmorteza A, Schuleri KH, Herzka DA, Lardo AC, and McVeigh ER, A new method for cardiac computed tomography regional function assessment: stretch quantifier for endocardial engraved zones (SQUEEZ), Circulation: Cardiovascular Imaging 5, 243–250 (2012). [PubMed: 22342945]
7. Pourmorteza A, Chen MY, van der Pals J, Arai AE, and McVeigh ER, Correlation of CT-based regional cardiac function (SQUEEZ) with myocardial strain calculated from tagged MRI: an experimental study, The international journal of cardiovascular imaging 32, 817–823 (2016). [PubMed: 26706935]
8. He J, Yang Y, Wang Y, Zeng D, Bian Z, Zhang H, Sun J, Xu Z, and Ma J, Optimizing a Parameterized Plug-and-Play ADMM for Iterative Low-Dose CT Reconstruction, IEEE Transactions on Medical Imaging 38, 371–382 (2019). [PubMed: 30106717]
9. Zhang Y, Hu D, Zhao Q, Quan G, Liu J, Liu Q, Zhang Y, Coatrieux G, Chen Y, and Yu H, CLEAR: Comprehensive Learning Enabled Adversarial Reconstruction for Subtle Structure Enhanced Low-Dose CT Imaging, IEEE Transactions on Medical Imaging 40, 3089–3101 (2021). [PubMed: 34270418]
10. Wu W, Hu D, Niu C, Yu H, Vardhanabhuti V, and Wang G, DRONE: Dual-Domain Residual-based Optimization NEtwork for Sparse-View CT Reconstruction, IEEE Transactions on Medical Imaging 40, 3002–3014 (2021). [PubMed: 33956627]
11. Flohr TG, Allmendinger T, Bruder H, Schwemmer C, Kappler S, and Schmidt B, Future Technological Advances in Cardiac CT, pages 873–892, Humana Press, Totowa, NJ, 2019.
12. Pack JD, Manohar A, Ramani S, Claus B, Yin Z, Contijoch FJ, Schluchter AJ, and McVeigh ER, Four-Dimensional computed tomography of the left ventricle, Part I: motion artifact reduction, Medical Physics (2021).
13. Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, and Ng R, Nerf: Representing scenes as neural radiance fields for view synthesis, in European conference on computer vision, pages 405–421, Springer, 2020.
14. Xie Y, Takikawa T, Saito S, Litany O, Yan S, Khan N, Tombari F, Tompkin J, Sitzmann V, and Sridhar S, Neural Fields in Visual Computing and Beyond, Computer Graphics Forum (2022).
15. Thirion J-P, Segmentation of tomographic data without image reconstruction, IEEE transactions on medical imaging 11, 102–110 (1992). [PubMed: 18218362]
16. Thirion J-P, Direct extraction of boundaries from computed tomography scans, IEEE transactions on medical imaging 13, 322–328 (1994). [PubMed: 18218508]
17. Dahl VA, Dahl AB, and Hansen PC, Computing segmentations directly from x-ray projection data via parametric deformable curves, Measurement Science and Technology 29, 014003 (2017).
18. Wang P, Liu L, Liu Y, Theobalt C, Komura T, and Wang W, NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction, arXiv preprint arXiv:2106.10689 (2021).
19. Lin E and Alessio A, What are the basic concepts of temporal, contrast, and spatial resolution in cardiac CT?, Journal of cardiovascular computed tomography 3, 403–408 (2009). [PubMed: 19717355]
20. Flohr TG et al. , First performance evaluation of a dual-source CT (DSCT) system, European radiology 16, 256–268 (2006). [PubMed: 16341833]
21. McCollough CH, Schmidt B, Yu L, Primak A, Ulzheimer S, Bruder H, and Flohr TG, Measurement of temporal resolution in dual source CT, Medical physics 35, 764–768 (2008). [PubMed: 18383698]

22. Besson G, New CT system architectures for high temporal resolution with applications to improved geometric dose efficiency and cardiac imaging, Medical physics 42, 2668–2678 (2015). [PubMed: 25979065]
23. Agatston AS, Janowitz WR, Hildner FJ, Zusmer NR, Viamonte M, and Detrano R, Quantification of coronary artery calcium using ultrafast computed tomography, Journal of the American College of Cardiology 15, 827–832 (1990). [PubMed: 2407762]
24. Budoff MJ et al. , Ultrafast computed tomography as a diagnostic modality in the detection of coronary artery disease: a multicenter study, Circulation 93, 898–904 (1996). [PubMed: 8598080]
25. Isola A, Ziegler A, Koehler T, Niessen W, and Grass M, Motion-compensated iterative cone-beam CT image reconstruction with adapted blobs as basis functions, Physics in Medicine & Biology 53, 6777 (2008). [PubMed: 18997267]
26. Tang Q, Cammin J, Srivastava S, and Taguchi K, A fully four-dimensional, iterative motion estimation and compensation method for cardiac CT, Medical physics 39, 4291–4305 (2012). [PubMed: 22830763]
27. Kim S, Chang Y, and Ra JB, Cardiac motion correction based on partial angle reconstructed images in X-ray CT, Medical Physics 42, 2560–2571 (2015). [PubMed: 25979048]
28. Hahn J, Bruder H, Allmendinger T, Stierstorfer K, Flohr T, and Kachelriess M, Reduction of motion artifacts in cardiac CT based on partial angle reconstructions from short scan data, in Medical Imaging 2016: Physics of Medical Imaging, volume 9783, page 97831A, International Society for Optics and Photonics, 2016.
29. Sheta HM, Egstrup K, Husic M, Heinsen LJ, and Lambrechtsen J, Impact of a motion correction algorithm on quality and diagnostic utility in unselected patients undergoing coronary CT angiography, Clinical imaging 40, 217–221 (2016). [PubMed: 26995573]
30. Sheta HM, Egstrup K, Husic M, Heinsen LJ, Nieman K, and Lambrechtsen J, Impact of a motion correction algorithm on image quality in patients undergoing CT angiography: A randomized controlled trial, Clinical imaging 42, 1–6 (2017). [PubMed: 27838576]
31. Maier J, Lebedev S, Erath J, Eulig E, Sawall S, Fournié E, Stierstorfer K, Lell M, and Kachelrieß M, Deep learning-based coronary artery motion estimation and compensation for short-scan cardiac CT, Medical Physics (2021).
32. Segars WP, Sturgeon G, Mendonca S, Grimes J, and Tsui BM, 4D XCAT phantom for multimodality imaging research, Medical physics 37, 4902–4915 (2010). [PubMed: 20964209]
33. Lee D, Choi J, Kim H, Cho M, and Lee K-Y, Validation of a novel cardiac motion correction algorithm for x-ray computed tomography: From phantom experiments to initial clinical experience, PLOS ONE 15, 1–18 (2020).
34. Ko Y, Moon S, Baek J, and Shim H, Rigid and non-rigid motion artifact reduction in X-ray CT using attention module, Medical Image Analysis 67, 101883 (2021). [PubMed: 33166775]
35. Lossau T, Nickisch H, Wissel T, Morlock M, and Grass M, Learning metal artifact reduction in cardiac CT images with moving pacemakers, Medical image analysis 61, 101655 (2020). [PubMed: 32092679]
36. Zhang Y, van der Werf NR, Jiang B, van Hamersvelt R, Greuter MJ, and Xie X, Motion-corrected coronary calcium scores by a convolutional neural network: a robotic simulating study, European radiology 30, 1285–1294 (2020). [PubMed: 31630233]
37. Kingma DP and Ba J, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
38. Riegler G, Osman Ulusoy A, and Geiger A, Octnet: Learning deep 3d representations at high resolutions, in Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3577–3586, 2017.
39. Gupta K, Sekhar N, Vigneault DM, Scott AR, Colvert B, Craine A, Raghavan A, and Contijoch FJ, Octree Representation Improves Data Fidelity of Cardiac CT Images and Convolutional Neural Network Semantic Segmentation of Left Atrial and Ventricular Chambers, Radiology: Artificial Intelligence 3, e210036 (2021). [PubMed: 34870221]
40. Qi CR, Su H, Mo K, and Guibas LJ, Pointnet: Deep learning on point sets for 3d classification and segmentation, in Proceedings of the IEEE conference on computer vision and pattern recognition, pages 652–660, 2017.

41. Gupta K and Chandraker M, Neural Mesh Flow: 3D Manifold Mesh Generation via Diffeomorphic Flows, Advances in Neural Information Processing Systems 33 (2020).
42. Park JJ, Florence P, Straub J, Newcombe R, and Lovegrove S, Deepsdf: Learning continuous signed distance functions for shape representation, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 165–174, 2019.
43. Mescheder L, Oechsle M, Niemeyer M, Nowozin S, and Geiger A, Occupancy networks: Learning 3d reconstruction in function space, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4460–4470, 2019.
44. Sitzmann V, Zollhöfer M, and Wetzstein G, Scene representation networks: Continuous 3d-structure-aware neural scene representations, arXiv preprint arXiv:1906.01618 (2019).
45. Liu L, Gu J, Lin KZ, Chua T-S, and Theobalt C, Neural sparse voxel fields, arXiv preprint arXiv:2007.11571 (2020).
46. Lombardi S, Simon T, Saragih J, Schwartz G, Lehrmann A, and Sheikh Y, Neural volumes: Learning dynamic renderable volumes from images, arXiv preprint arXiv:1906.07751 (2019).
47. Jiang Y, Ji D, Han Z, and Zwicker M, Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1251–1261, 2020.
48. Saito S, Huang Z, Natsume R, Morishima S, Kanazawa A, and Li H, Pifu: Pixelaligned implicit function for high-resolution clothed human digitization, in Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2304–2314, 2019.
49. Saito S, Simon T, Saragih J, and Joo H, Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 84–93, 2020.
50. Sun Y, Liu J, Xie M, Wohlberg B, and Kamilov US, CoIL: Coordinate-Based Internal Learning for Tomographic Imaging, IEEE Transactions on Computational Imaging 7, 1400–1412 (2021).
51. Wu Q, Li Y, Sun Y, Zhou Y, Wei H, Yu J, and Zhang Y, An Arbitrary Scale Super-Resolution Approach for 3-Dimensional Magnetic Resonance Image using Implicit Neural Representation, arXiv preprint arXiv:2110.14476 (2021).
52. Tancik M, Mildenhall B, Wang T, Schmidt D, Srinivasan PP, Barron JT, and Ng R, Learned initializations for optimizing coordinate-based neural representations, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2846–2855, 2021.
53. Lindell DB, Martel JN, and Wetzstein G, Autoint: Automatic integration for fast neural volume rendering, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14556–14565, 2021.
54. Shen L, Pauly J, and Xing L, NeRP: Implicit Neural Representation Learning with Prior Embedding for Sparsely Sampled Image Reconstruction, arXiv preprint arXiv:2108.10991 (2021).
55. Reed AW, Kim H, Anirudh R, Mohan KA, Champley K, Kang J, and Jayasuriya S, Dynamic CT Reconstruction from Limited Views with Implicit Neural Representations and Parametric Motion Fields, arXiv preprint arXiv:2104.11745 (2021).
56. Wu Q, Li Y, Xu L, Feng R, Wei H, Yang Q, Yu B, Liu X, Yu J, and Zhang Y, IREM: High-Resolution Magnetic Resonance Image Reconstruction via Implicit Neural Representation, in International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 65–74, Springer, 2021.
57. Li SZ and Jain A, editors, Analysis-by-Synthesis, pages 35–36, Springer US, Boston, MA, 2009.
58. Rahimi A and Recht B, Random Features for Large-Scale Kernel Machines, in Advances in Neural Information Processing Systems, edited by Platt J, Koller D, Singer Y, and Roweis S, volume 20, Curran Associates, Inc., 2007.
59. Tancik M, Srinivasan P, Mildenhall B, Fridovich-Keil S, Raghavan N, Singhal U, Ramamoorthi R, Barron J, and Ng R, Fourier features let networks learn high frequency functions in low dimensional domains, Advances in Neural Information Processing Systems 33, 7537–7547 (2020).
60. Sitzmann V, Martel J, Bergman A, Lindell D, and Wetzstein G, Implicit neural representations with periodic activation functions, Advances in Neural Information Processing Systems 33 (2020).
61. Reynolds DA, Gaussian mixture models., Encyclopedia of biometrics 741, 659–663 (2009).

62. Osher S, Burger M, Goldfarb D, Xu J, and Yin W, An iterative regularization method for total variation-based image restoration, Multiscale Modeling & Simulation 4, 460–489 (2005).

63. Whiting BR, Signal statistics in x-ray computed tomography, in Medical Imaging 2002: Physics of Medical Imaging, volume 4682, pages 53–60, SPIE, 2002.

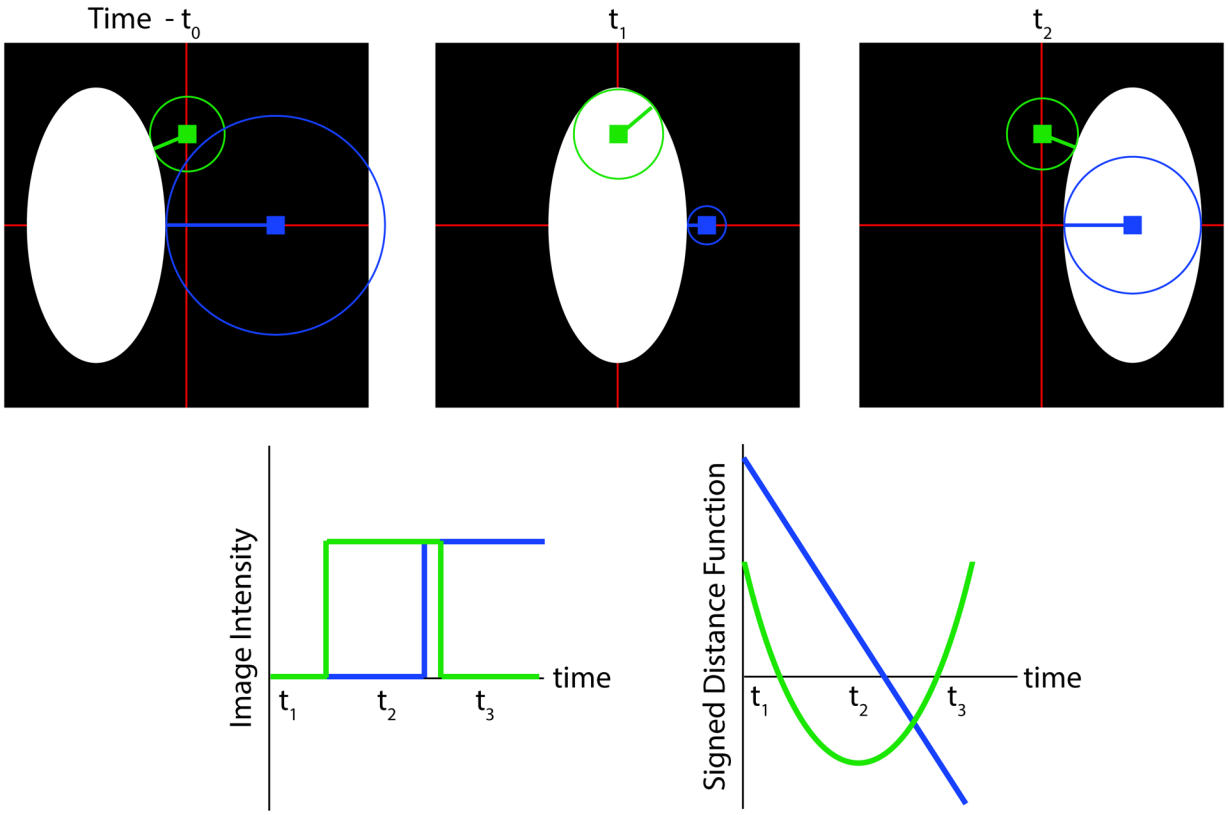

Figure 1: Illustration of the Benefit of a Signed Distance Function (SDF)-based Object Representation.

The SDF of a moving object boundary is a smooth function of time. In this scene, a white ellipse moves from left to right over a black background, across 3 frames. The per-pixel image intensity (observed at the blue and green locations) changes over time in a step-like fashion. However, the signed distance value changes smoothly which makes it amenable for fitting a Lipschitz continuous function.

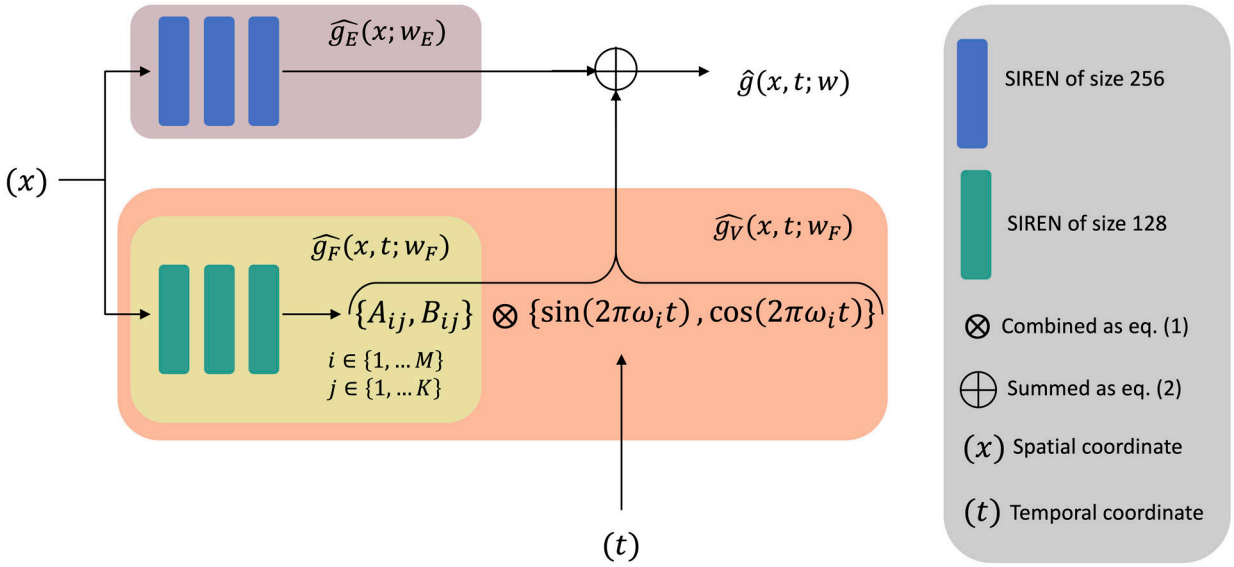

Figure 2: Neural implicit representation of the SDF.

The SDF is approximated using a neural network $\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w})$ comprising two sub-networks which estimate the shape of the SDF $\hat{g}_E$ as well as its spatiotemporal displacement $\hat{g}_V$. Neural networks are implement as SIREN[60] MLPs.

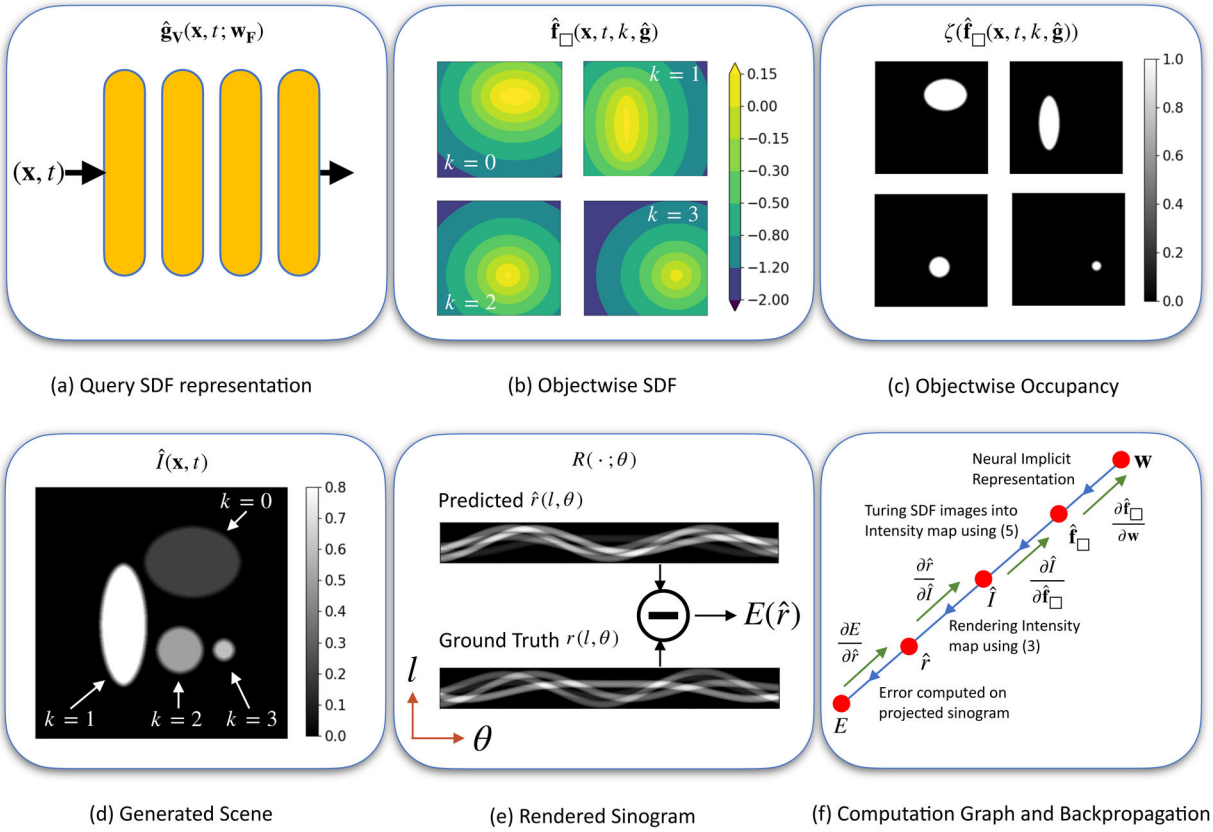

Figure 3: Differentiable rendering of signed distance based NIR.

Panel A: Given the object-wise spatiotemporal SDF representation $\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w})$, we query the network at each point $\mathbf{x}$ on the N-dimensional grid of size d at some time frame t. **Panel B:** This results in an object-wise discrete signed distance function for each of the k objects. **Panel C and D:** SDFs are then converted into object-wise occupancy values and subsequently combined to form the final scene by taking the dot product with the attenuation intensity $a(k)$. In this illustration, $a(0) = 0.2, a(1) = 0.8, a(2) = 0.5$ and $a(3) = 0.6$. **Panel E:** This scene is rendered using R that results in a sinogram $\hat{r}$ which can then be compared with ground truth sinogram r giving rise to an error functional E. **Panel F:** The entire computation graph is shown in which highlights that gradients from the functional E can be backpropagated through the renderer to the network weights $\mathbf{w}$. This is crucial for optimizing the SDF representation $\hat{\mathbf{g}}(\mathbf{x}, t; \mathbf{w})$ using only the projected sinogram data.

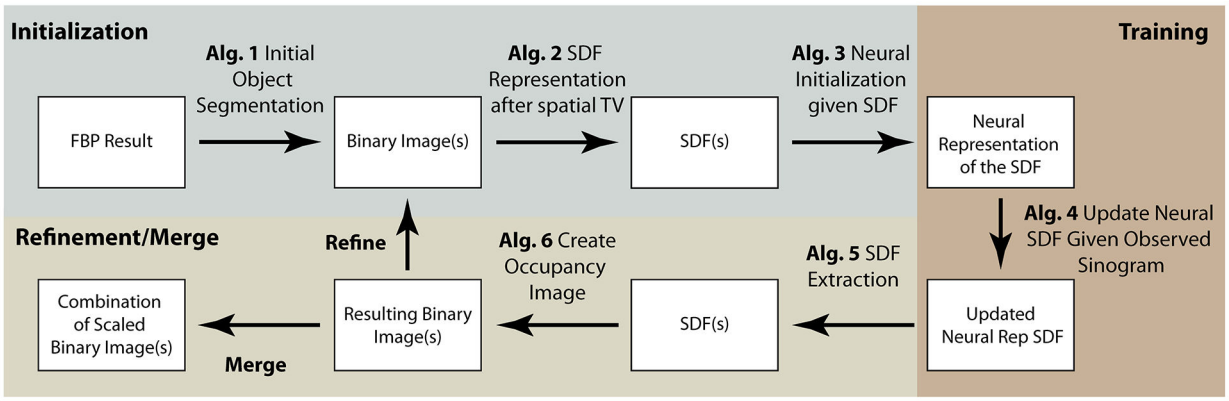

Figure 4: DiFiR-CT algorithm overview.

DiFiR-CT represents the spatiotemporal attenuation map of a moving object both explicitly in the image domain as well as implictly in the neural domain and performs optimization in both of these domains across three stages: Initialization, Training, Refinement/Export. Initialization obtains well-behaved SDF representations based on filtered back projection (FBP) reconstructed images using an intensity-based segmentation, encodes the resulting segmentation as signed distance functions, and performs the explicit-to-implicit representation conversion. Training aims to update neural SDF representation to match the sinogram data. Refinement/Merge includes the implicit-to-explicit conversion, creation of occupancy images, and scaling and combination of binary images. DiFiR-CT results improved when it was repeated using the results of the first prediction.

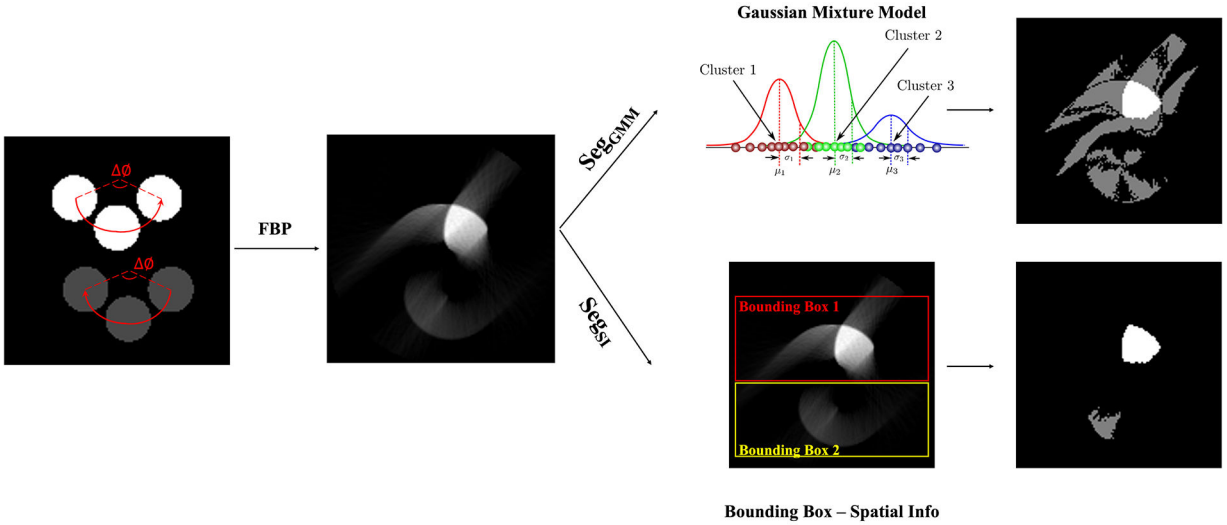

Figure 5: Intensity-only and spatially-aware object segmentation approaches used in DiFiR-CT.
The first image shows the ground truth motion of two dots (top intensity = 0.7, moving from left to right, bottom intensity = 0.2 , moving from right to left). $\Delta\theta$ is the angular displacement per gantry rotation. SegGMM: Gaussian mixture model incorrectly assigned the motion artifacts and the bottom dot as the same class. SegSI: Spatially aware segmentation utilized both spatial info (by setting bounding box in this example) and intensity info (thresholding) and led to correct detection of both top and bottom dots.

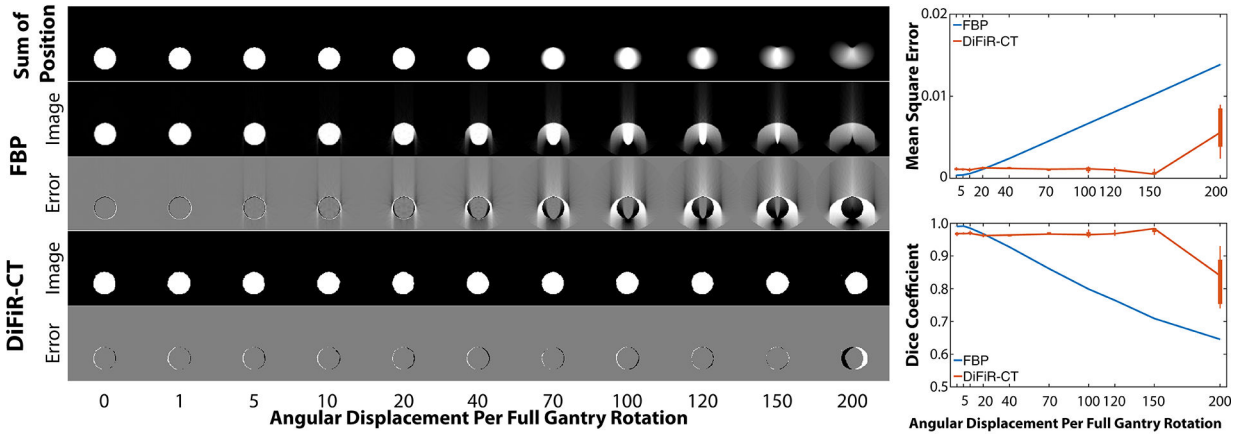

Figure 6: DiFiR-CT accurately depicts the position of a circle, despite high motion during acquisition.

FBP images are degraded with increasing motion during data acquisition. DiFiR-CT improves delineation of the translating circle. Error bars represent the standard deviation observed from 5 DiFiR-CT results using different random initialization of the networks. All images are reconstructions using 360-degrees of projections. Displacement was imparted as an angular translation (in degrees) as indicated by the legend.

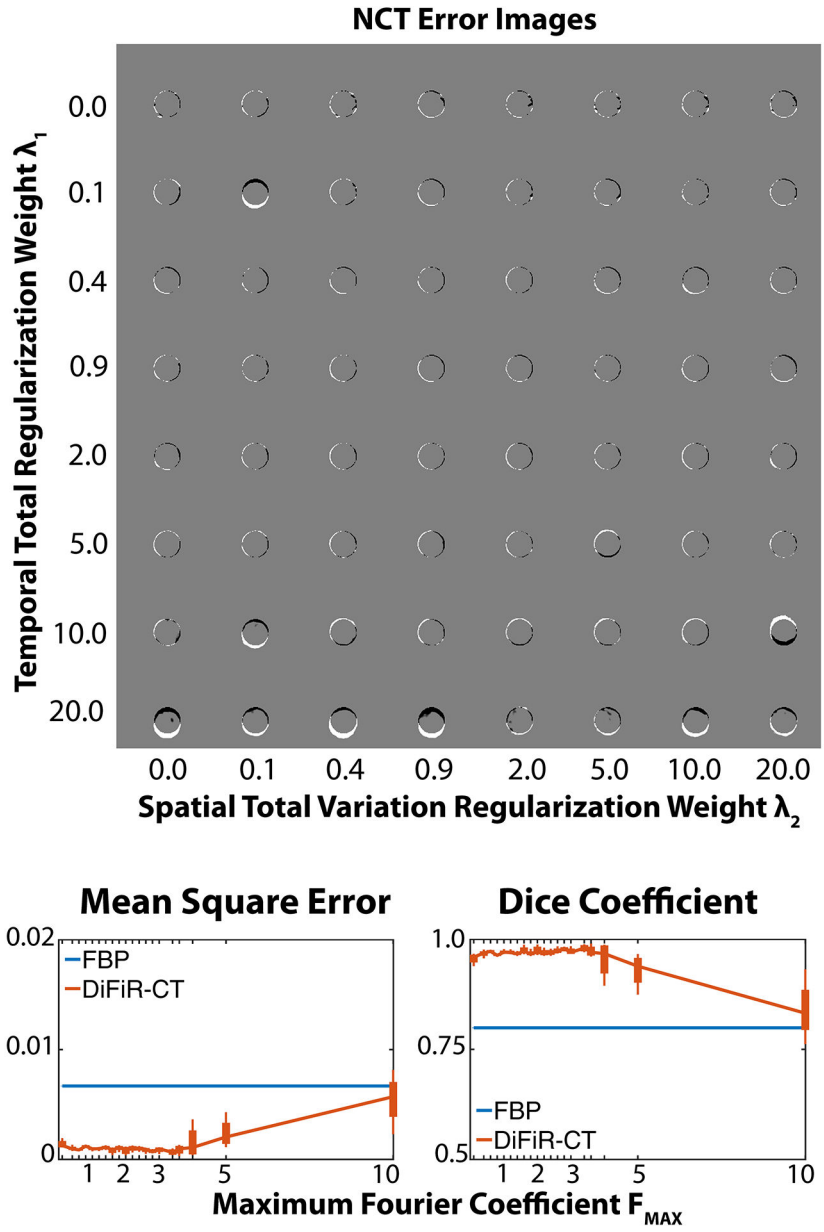

Figure 7: DiFiR-CT reconstruction of an object with angular displacement = 100° with varying spatial and temporal regularization (top) and F_{max} (bottom).

DiFiR-CT yields accurate image reconstruction over a range of spatial TV regularization. However, DiFiR-CT reconstruct can become inaccurate if temporal regularization is too highly penalized (e.g., $\lambda_1 = 20$). This corresponds to making the solution more stationary. DiFiR-CT results are also robust for a wide-range of F_{max}. At $F_{max} > 10$, the increased parameterization can lead to overfitting and robust results can be achieved with lower F_{max}.

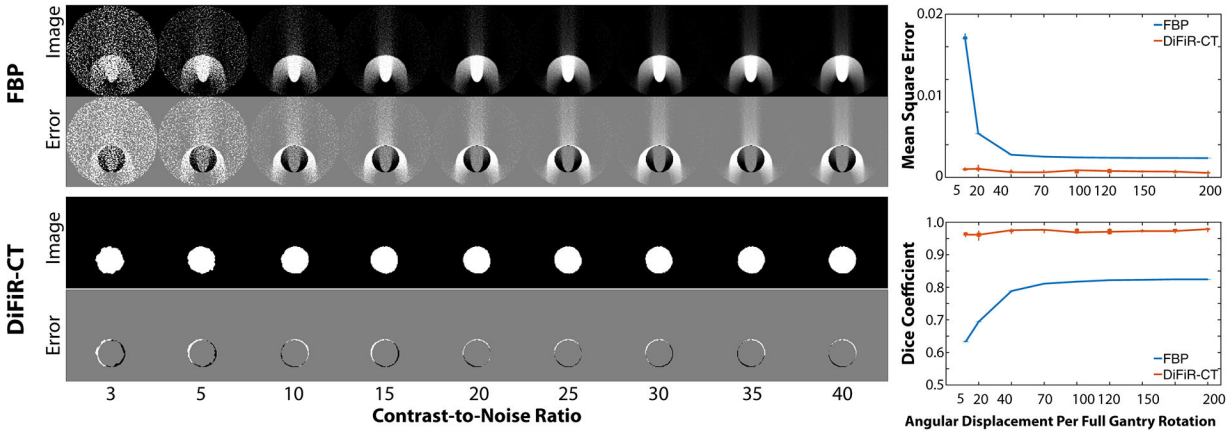

Figure 8: DiFiR-CT reconstruction with varying noise.

DiFiR-CT yields accurate image reconstruction over a range of contrast-to-noise ratios for a circle with angular displacement = 100°. However, FBP shows a decrease in image reconstruction accuracy as the noise increase (CNR decreases). Median values of five noise realizations are shown for both FBP and DiFiR-CT results.

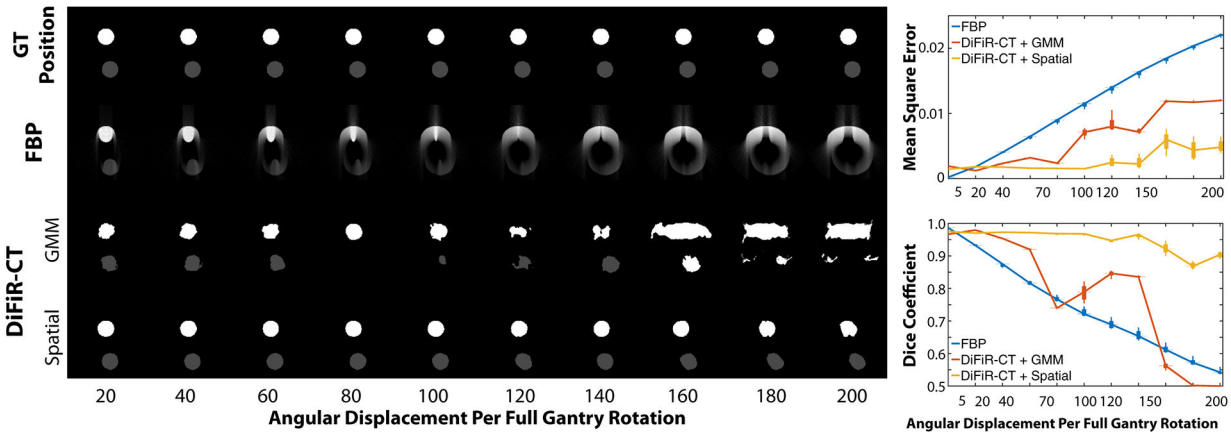

Figure 9: DiFiR-CT-SegSI accurately depicts the moving dots with two attenuations and high angular displacements.

DiFiR-CT-SegGMM failed to accurately reconstruct two circles when $\Delta\theta > 60$. DiFiR-CT-SegSI maintained high-quality motion-corrected reconstruction for all $\Delta\theta$ with higher DICE and lower MSE when compared with FBP and DiFiR-CT-GMM. $\Delta\theta$ = angular displacement per gantry rotation. FBP suffers from motion artifacts for all $\Delta\theta$.

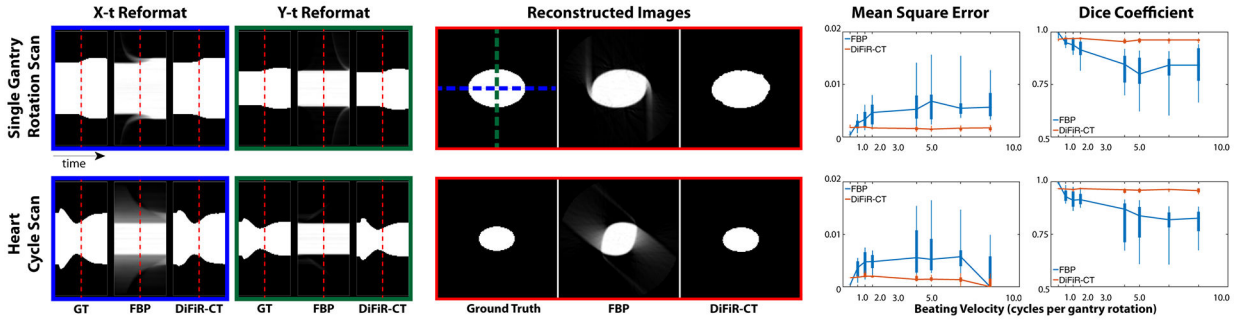

Figure 10: DiFiR-CT improves single (top) and multiple (bottom) gantry rotation imaging of a ”beating” circle.

Top: DiFiR-CT improves image quality, relative to FBP for both cases. The change in diameter of the circle along the x (blue) and y (green) axes are shown to illustrate how FBP blurring is improved by DiFiR-CT. Images reconstructed at the middle of the data acquisition (red line) are shown in the middle panel. DiFiR-CT reduces the noticeable motion artifacts along the edges. **Bottom:** DiFiR-CT can readily leverage additional temporal information available when multiple gantry rotations are acquired and improves image quality, over FBP, without modification.

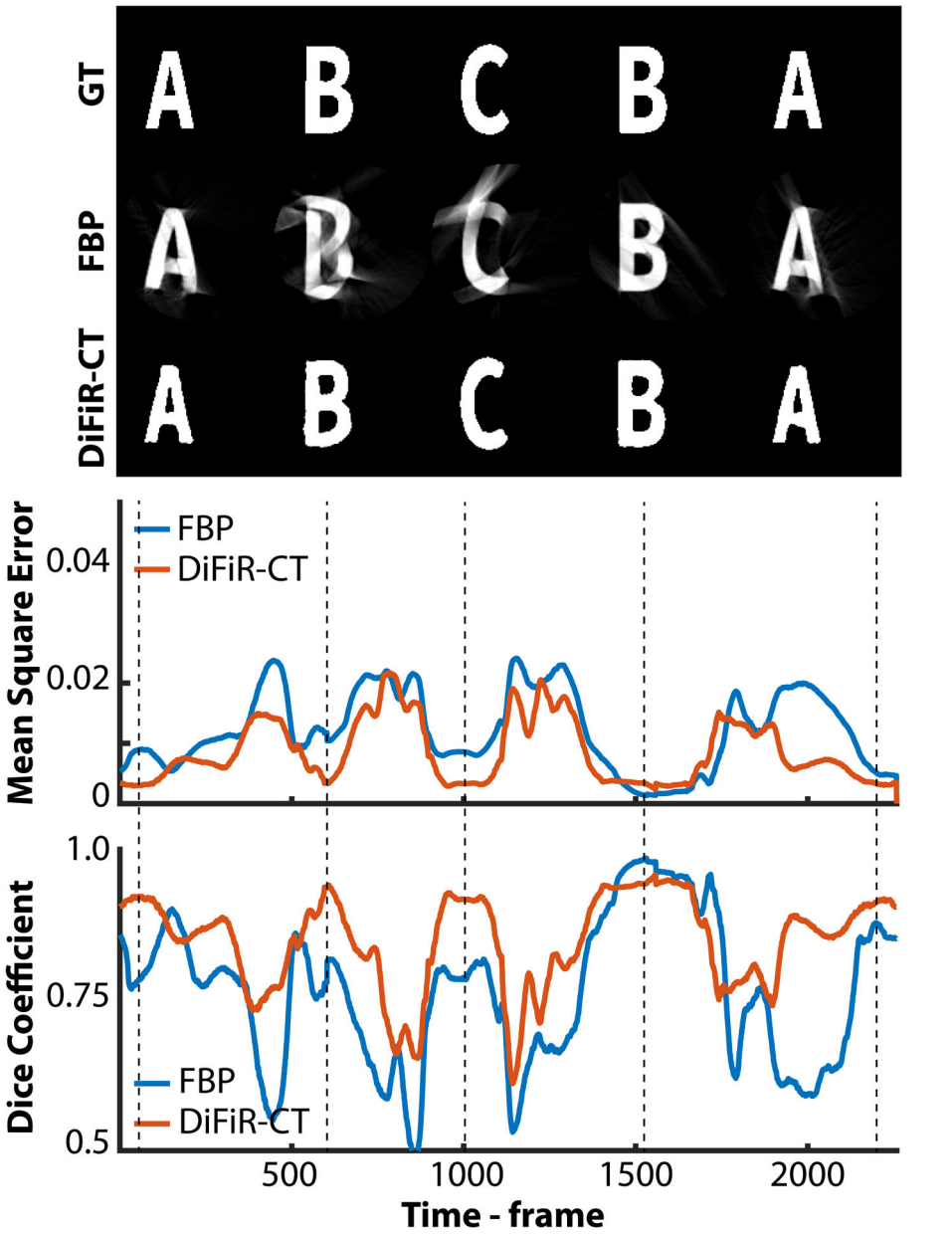

Figure 11: DiFiR-CT improves imaging of a complex change in topology.

Without modification of the framework or tuning of parameters, DiFiR-CT improved imaging of a complex scene. Further, the approach did so without estimation of motion or a prior information. The temporal location of the 5 frames is shown on the MSE and Dice plots as vertical dotted lines.