

# Appearance Learning for Image-based Motion Estimation in Tomography

Alexander Preuhs, Michael Manhart, Philipp Roser, Elisabeth Hoppe, Yixing Huang, Marios Psychogios, Markus Kowarschik, and Andreas Maier, *Member, IEEE*

**Abstract**—In tomographic imaging, anatomical structures are reconstructed by applying a pseudo-inverse forward model to acquired signals. Geometric information within this process is usually depending on the system setting only, i.e., the scanner position or readout direction. Patient motion therefore corrupts the geometry alignment in the reconstruction process resulting in motion artifacts. We propose an appearance learning approach recognizing the structures of rigid motion independently from the scanned object. To this end, we train a siamese triplet network to predict the *reprojection error* (RPE) for the complete acquisition as well as an approximate distribution of the RPE along the single views from the reconstructed volume in a multi-task learning approach. The RPE measures the motion-induced geometric deviations independent of the object based on virtual marker positions, which are available during training. We train our network using 27 patients and deploy a 21-4-2 split for training, validation and testing. In average, we achieve a residual mean RPE of 0.013 mm with an inter-patient standard deviation of 0.022 mm. This is twice the accuracy compared to previously published results. In a motion estimation benchmark the proposed approach achieves superior results in comparison with two state-of-the-art measures in nine out of twelve experiments. The clinical applicability of the proposed method is demonstrated on a motion-affected clinical dataset.

**Index Terms**—rigid motion compensation, reconstruction, interventional CBCT, autofocus, appearance learning

## I. INTRODUCTION

APPEARANCE modeling [1] for interpreting images is a well examined problem in the field of computer vision. An appearance model is trained to extract invariant representations of an object of interest [2], [3], e.g., for the tracking of faces [4] or event detection [5]. Recently, Preuhs et al. [6] have applied the strategy of appearance learning for motion detection in tomographic imaging.

The key concept of tomographic imaging is the reconstruction of internal patient anatomy from a series of measured signals. This can be the relaxation properties of hydrogen atoms in *magnetic resonance imaging* (MRI) or photon attenuation in X-ray *computed tomography* (CT). When reconstructing a tomographic image from measured signals, the geometry associated with each signal only depends on the system setting,

i.e., source-detector orientation of a CT system or readout position for MRI scanners. The object itself is assumed to be static during the acquisition. As a consequence, patient motion corrupts the geometry alignment and results in motion artifacts within the reconstructed tomographic image.

Many efforts have been devoted to the problem of non-static objects, which are mainly splitted into non-rigid and rigid approaches. Rigid approaches reduce the number of unknowns to a 6 dimensional vector per measured signal, i.e., the respective rigid patient pose. However, complex movements, as apparent in heart imaging, are not reducible to such a simple model. In these cases non-rigid motion estimation must be deployed.

### A. Non-Rigid Motion Compensation

Lauritsch et al. [7] presented a gating approach, where the signal is binned to different motion states. Only similar motion states are used for reconstruction. This is extended by Taubmann et al. [8] who developed a primal-dual optimization scheme based on a spatial and temporal *total variation* (TV). Gating approaches were also presented by Larson et al. [9] and Hoppe et al. [10] for cardiac cine MRI, where the motion bin is deduced from the k-space center of each readout. Similar to gating, Fischer et al. [11] devised an MRI-based model for X-ray fluoroscopy overlays. By binning of 4-D volumes to cardiac and respiratory phases, the motion field is estimated using 3-D/3-D registration.

Recent approaches deploy image-to-image translation from motion-affected reconstructions to such without motion artifacts. Here, prior knowledge on the expected manifold of motion free reconstructions is learned [12]. Kustner et al. [13] and Latif et al. [14] propose a *conditional generative adversarial network* (cGAN) to synthesize motion free MRI reconstructions from a motion degenerated one. The same approach was presented for X-ray imaging by Xiao et al. [15].

### B. Rigid Motion Compensation

For many anatomical objects, the structure of the expected motion is already known a priori. The head, for example, is restricted by the skull to move as a rigid object. Further, many anatomies move in an approximate rigid structure during interventions, e.g., the knees or the hands. As the focus of this article is rigid motion compensation, we give a detailed overview of published methods which can be clustered into three categories.

Copyright (c) 2019 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to [pubs-permissions@ieee.org](mailto:pubs-permissions@ieee.org). A. Preuhs, P. Roser, E. Hoppe, Y. Huang and A. Maier are with the Pattern Recognition Lab, Friedrich-Alexander-Universitt Erlangen-Nrnberg, 91058 Erlangen, Germany e-mail: [alexander.preuhs@fau.de](mailto:alexander.preuhs@fau.de).

M. Psychogios is with the Neuroradiology Department, Universittsspital Basel, 4031 Basel, Switzerland.

M. Manhart and M. Kowarschik are with Siemens Healthcare GmbH, 91301 Forchheim, Germany.

2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

1) *Projection Consistency*: A computationally fast approach is projection consistency, where only the projection raw data are used, without the need for intermediate reconstructions. The main idea is that information is redundantly sampled by the forward operator with each acquired signal, e.g., the mass of the object. Powerful conditions are the *Helgason-Ludwig consistency conditions* (HLCC) [16] describing the relation between polynomials of degree  $n$  and the respective  $n^{\text{th}}$  moment of the projections for parallel-beam CT or radial sampled MR. This was devised by Yu et al. [17] to compensate motion in fan-beam geometry. A more broadly applicable approach based on the zero order HLCC and Grangeat's theorem is epipolar consistency which was applied for geometric jitter and motion compensation in *cone-beam computed tomography* (CBCT) [18], [19], [20], [21]. Similar approaches have been deployed in MRI motion compensation, where propeller trajectories measure the k-space center redundantly and compensate the motion based on this data redundancy [22].

2) *Reconstruction Consistency*: Contrary to projection consistency, reconstruction consistency solely uses tomographic images to estimate a rigid motion trajectory and is therefore often related as autofocus. The key idea is similar to the image-to-image translation approaches presented above: a motion-free reconstruction reveals some inherent properties which can be measured using an *image-quality metric* (IQM). In contrast to cGAN-based approaches, however, a motion trajectory is estimated by iterative optimization of the IQM. This ensures data integrity, which is of high importance in a clinical setting. The first application of this autofocus principle was presented by Atkinson et al. [23] for MRI reconstructions. They optimize a motion trajectory to find a reconstructed image with low entropy of the intensity histogram favoring images with high contrast structures and without motion ghosting or blur. Kingston et al. [24] presented a similar approach based on TV minimization. Subsequently various extensions were proposed [25], [26], [27], [28], including a combination of metrics as well as additional smoothness constraints.

3) *Data Consistency*: The last category is based on enforcing data fidelity, which is the consistency of the reconstruction domain with the signal domain. In CBCT this approach is used for calibrating the system geometry by minimizing a *reprojection error* (RPE) of 3-D spheres on a calibration phantom and their respective 2-D projections [29]. Using markers attached to the patient, this strategy was also investigated for motion compensation in MRI [30] and CT [31].

A second approach to enforce data fidelity is the virtual application of the forward model to an intermediate reconstruction and comparing these virtual data with the actually acquired data. Haskell et al. [32] used a SENSE forward model to maximize the data consistency with the acquired k-space data. For transmission imaging, *digitally rendered radiographs* (DRR) are commonly used to enforce consistency with the acquired projections [33], [34]. In this context, Dennerlein et al. [35] exploit directly and indirectly filtered projections to compensate for geometric misalignment.

### C. Potentials and Limitations in the State of the Art

Non-rigid approaches (see Sec. I-A) seem to be unfitting if the problem is known to be rigid. Image-to-image-based approaches do not exploit the full problem knowledge. Furthermore, their clinical applicability is limited because the consistency of the reconstructed image to the acquired data is not guaranteed and anatomic malformations can vanish [36].

Consistency conditions (see Sec. I-B1) have been used for the compensation of various other image artifacts as beam hardening, scatter correction or truncation correction [37], [38], [39], [40]. This is due to consistency being deduced from a physical model, which only holds on an approximate basis for real applications [41]. Additionally, they are insensitive to certain motion directions and their application is limited to motion patterns outside the acquisition plane [18], [21].

Image-based methods (see Sec. I-B2) currently use hand-crafted features not particularly designed for the specific task. As a consequence, they are object dependent with each object revealing a different histogram entropy or TV.

A robust approach is based on reducing the RPE using markers (see Sec. I-B3). However, this approach depends on additional marker placements, which has not found its way to clinical routine yet. Marker-free registration approaches are only working robustly if a prior reconstruction is available. Otherwise, the optimization becomes ill-posed, as the intermediate reconstruction, on which the forward model is applied, inherently reveals motion artifacts.

Deep learning has high potential to overcome some of those limitations by replacing bottlenecks of traditional methods with data-driven algorithms. For example, Bier et al. [42] tackled the problem of manual marker placement by learning anatomical landmarks directly from the projection images. The presented cGAN-based approaches potentially have the risk of vanishing anatomical malformations, however, they may solve the chicken-egg problem for marker-free registration approaches. Additionally, many applications emerged for learning-based registration [43], [44], [45]. They could potentially be extended for motion compensation scenarios.

### D. Contribution

Despite its great potential in improving rigid motion compensation algorithms, deep learning methods have caught limited attention from the research community. In Preuhs et al. [6], we have presented the concept of learning image artifacts from a single axial reconstructed slice using a simplified motion model and a vanilla network architecture. The key concept is that a certain motion state is regressed to an object-independent measure defined by the RPE. We extend this line of thinking by developing a new data-driven approach for appearance learning capable of compensating motion artifacts. Our network architecture for motion appearance learning is based on a siamese triplet network trained in a multi-task scenario. Therefore, we incorporate not only a single axial slice but make use of information from 9 slices, extracted from axial, sagittal and coronal orientations. Using a multi-task loss, we estimate both (1) an overall motion score of the reconstructed volume similar to [6] and (2) a prediction

which projections are affected by the motion. To stabilize the network prediction, we deploy a novel pre-processing scheme to compensate for training data variability. These extensions allow us to learn realistic motion appearance, composed of three translation and three rotation parameters per acquired view. We evaluate the accuracy of the motion appearance learning in dependence of the patient anatomy and also the motion type. In a rigid motion estimation benchmark, we demonstrate the performance of the appearance learning approach in comparison to state-of-the-art methods. Finally, we demonstrate its applicability to real clinical data using a motion-affected clinical scan.

We devise the proposed framework for CBCT, however, by exchanging the backward model and training data, this approach is seamlessly applicable to radial sampled MRI or *positron emission tomography* (PET). In addition, by adjusting the regression target, also for Cartesian sampled MRI.

## II. RIGID MOTION MODEL FOR CBCT

### A. Cone-Beam Reconstruction

In tomographic reconstruction we compute anatomical structures denoted by  $\mathbf{x}$  from measurements  $\mathbf{y}$  produced with a forward model  $\mathbf{A}$  by  $\mathbf{A}\mathbf{x} = \mathbf{y}$ . For X-ray transmission imaging  $\mathbf{x}$  are attenuation coefficients and  $\mathbf{y}$  are the attenuation line integrals measured at each detector pixel. The system geometry — e.g., pixel spacing, detector size and source-detector orientation — is part of the forward model  $\mathbf{A}$ . Using the pseudo-inverse

$$\mathbf{x} = \mathbf{A}^\top (\mathbf{A}\mathbf{A}^\top)^{-1} \mathbf{y} \quad (1)$$

we get an analytic solution to this inverse problem, which consists of the back-projection  $\mathbf{A}^\top$  of filtered projection data  $(\mathbf{A}\mathbf{A}^\top)^{-1} \mathbf{y}$  [46], commonly known as *filtered back-projection* (FBP). For CBCT with circular trajectories, an approximate solution is provided by the *Feldkamp-Davis-Kress* (FDK) algorithm [47]. The algorithm is regularly used for autofocus approaches [24], [27] (see Sec. I-B2) due to its low computational costs. Rit et al. [48] have further shown that even due to its approximate nature, an FDK-based motion-compensated CBCT reconstruction is capable of correcting most motion artifacts. Thus, we use the FDK reconstruction algorithm, having the benefit of only filtering the projection images once and thereafter only altering the back-projection operator for motion trajectory estimation.

It is possible to formulate the FDK algorithm using a tuple of projection matrices  $\mathbf{P} = (\mathbf{P}_0, \mathbf{P}_1, \dots, \mathbf{P}_N)$  describing the geometry of operator  $\mathbf{A}$ . The measurements  $\mathbf{y}$  are reshaped to a tuple of 2-D projection images  $\mathbf{Y} = (\mathbf{Y}_0, \mathbf{Y}_1, \dots, \mathbf{Y}_N)$ . In analogy to [47], we implement the FDK for a short scan trajectory using Parker redundancy weights  $W_i(u, v)$  [49], where  $i \in [1, 2, \dots, N]$  describes the projection index and  $(u, v)$  denotes a 2-D pixel. The first step is a weighting and filtering of the projection images

$$\mathbf{Y}'_i(u, v) = W_i(u, v) \int_{\mathbb{R}} \mathcal{F} \tilde{\mathbf{Y}}_i(\eta, v) e^{i2\pi uv} \frac{|\eta|}{2} d\eta, \quad (2)$$

with  $\mathcal{F} \tilde{\mathbf{Y}}_i$  being the 1-D Fourier transform of the  $i^{\text{th}}$  cosine weighted projection image along the tangential direction of the

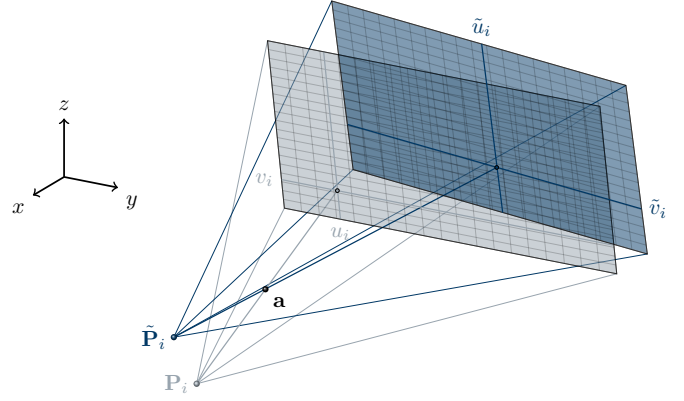


Fig. 1. Visualization of the geometry for a point  $\mathbf{a}$  and two geometries  $\mathbf{P}_i$  and  $\tilde{\mathbf{P}}_i$ . The  $L_2$  distance between the two projected points on the 2-D detector defines the RPE of the scene.

scan orbit. Thereafter, a distance-weighted voxel-based back-projection is applied mapping a homogeneous world point  $\mathbf{a} \in \mathbb{P}^3$  to a detector pixel described in the projective two-space  $\mathbb{P}^2$

$$f_{\text{FDK}}(\mathbf{a}, \mathbf{P}, \mathbf{Y}) = \sum_{i \in N} U(\mathbf{P}_i, \mathbf{a}) \mathbf{Y}'_i(\phi_u(\mathbf{P}_i \mathbf{a}), \phi_v(\mathbf{P}_i \mathbf{a})) \quad (3)$$

with  $\mathbf{P}_i$  describing the system calibration associated with  $\mathbf{Y}_i$ . (see Fig. 1). The mapping function  $\phi_\diamond : \mathbb{P}^2 \rightarrow \mathbb{R}$  is a dehomogenization

$$\phi_\diamond((x, y, w)^\top) = \begin{cases} \frac{x}{w} & \text{if } \diamond = u \\ \frac{y}{w} & \text{if } \diamond = v \end{cases}, \quad (4)$$

and  $U(\mathbf{P}_i, \mathbf{a})$  is the distance weighting according to [47].

### B. Rigid Motion Model

We assume the rigid motion to be discrete w.r.t. the acquired projections. To this end, we define the motion trajectory  $\mathbf{M}$  as a tuple of motion states  $\mathbf{M}_i \in \mathbb{SE}(3)$  describing the orientation of the patient during the acquisition of the  $i^{\text{th}}$  projection  $\mathbf{Y}_i$ . Each motion state is associated to a projection matrix  $\mathbf{P}_i$ . The motion modulated trajectory is obtained by

$$\mathbf{P} \circ \mathbf{M} = (\mathbf{P}_0 \mathbf{M}_0, \mathbf{P}_1 \mathbf{M}_1, \dots, \mathbf{P}_N \mathbf{M}_N), \quad (5)$$

where  $\circ$  is the element-wise matrix multiplication of two tuples. Typically, the motion trajectory is unknown and the task of motion compensation is to find a tuple of matrices  $\mathbf{C}_i \in \mathbb{SE}(3)$  annihilating the resulting geometry corruption produced by  $\mathbf{M}$ . The compensation is successful if an annihilating trajectory  $\mathbf{C} = (\mathbf{C}_0, \dots, \mathbf{C}_N)$  is found that suffices  $\mathbf{C} \circ \mathbf{M} = \mathbf{1}$ , with  $\mathbf{1}$  being a tuple of identities.

Each motion matrix defined in  $\mathbb{SE}(3)$  is parameterized by 3 rotations  $(r_x, r_y, r_z)$  and 3 translations  $(t_x, t_y, t_z)$ , describing Euler angles and translations along the corresponding coordinate axis, respectively. Therefore, the annihilating trajectory has  $6N$  free parameters for an acquisition with  $N$  projections. To reduce the high dimensionality, we model the trajectory using Akima splines [50]. This reduces the free parameters to  $6M$ , where  $M$  is the number of nodes typically chosen as  $M \ll N$ . Based on the expected frequency of the motion the number of spline nodes can be adapted.



### III. APPEARANCE LEARNING

Conventionally, autofocus approaches are based on hand-crafted features, selected due to their correlation with an artifact-free reconstruction. For example, entropy gives a measure on contingency. As the human anatomy consists of mostly homogeneous tissues, entropy of the gray-value histogram can be expected to be minimal if all structures are reconstructed correctly. Motion blurs the anatomy or produces ghosting artifacts distributing the gray values more randomly. A similar rationale is arguable for TV, which is also regularly used for constraining algebraic reconstruction [8]. Contrary to algebraic reconstruction, the motion estimation scenario is non-convex and optimization of a cost function based on hand-crafted image features is hardly solvable for geometric deviations exceeding a certain bound [28].

We aim to overcome this problem by designing a tailored image-based metric, which reflects the appearance of the motion structure independent of the object.

#### A. Object-Independent Motion Measure

Several metrics have been proposed to quantify image quality of motion affected reconstructions based on a given ground truth: the *structural similarity* (SSIM) [51], the  $L_2$  distance [52] or binary classification to motion-free and -affected [53]. However, they were not used for the compensation of motion, but merely for the assessment of image quality, which is of high relevance in the field of MRI to automatize prospective motion compensation techniques.

We choose the object-independent RPE for motion quantification. Its geometric interpretation is schematically illustrated in Fig. 1. The RPE measures reconstruction relevant deviations in the projection geometry and is defined by a 3-D marker position  $\mathbf{a} \in \mathbb{P}^3$  and two projection geometries  $\mathbf{P}_i, \tilde{\mathbf{P}}_i$ . We consider  $\mathbf{P}_i$  as the system calibration and  $\tilde{\mathbf{P}}_i = \mathbf{P}_i \mathbf{M}_i$  as the actual geometry due to the patient motion. Accordingly, the RPE for a patient movement at projection  $i$  is defined by

$$d_{\text{RPE}}(\mathbf{P}_i, \tilde{\mathbf{P}}_i, \mathbf{a}) = \left\| \begin{pmatrix} \phi_u(\mathbf{P}_i \mathbf{a}) \\ \phi_v(\mathbf{P}_i \mathbf{a}) \end{pmatrix} - \begin{pmatrix} \phi_u(\tilde{\mathbf{P}}_i \mathbf{a}) \\ \phi_v(\tilde{\mathbf{P}}_i \mathbf{a}) \end{pmatrix} \right\|_2^2 \quad (6)$$

where  $\phi_\diamond$  denotes the dehomogenization described in Eq. (4). Using a single marker, the RPE is insensitive to a variety of motion directions. Therefore, we use  $K = 90$  virtual marker positions  $\mathbf{a}_k$ , distributed homogeneously at three sphere surfaces with the radii 30 mm, 60 mm, and 90 mm. The high number of markers ensures that the RPE is view-independent, i.e., a displacement of a projection at the beginning of the trajectory has the same effect on the RPE as a displacement of a projection at the end of the trajectory. Accordingly, the overall RPE for a single view is

$$d_{\text{RPE}}(\mathbf{P}_i, \tilde{\mathbf{P}}_i) = \frac{1}{K} \sum_{k=1}^K d_{\text{RPE}}(\mathbf{P}_i, \tilde{\mathbf{P}}_i, \mathbf{a}_k) \quad (7)$$

As shown in Strobel et al. [29], Eq. (7) can be rewritten to a measurement matrix  $\mathbf{X}$  containing the 3-D marker positions, a vector  $\mathbf{p}$  containing the elements of  $\tilde{\mathbf{P}}_i$  and a vector  $\mathbf{d}$  containing the respective 2-D marker positions. Given at least

six markers, the components of  $\mathbf{p}$  are estimated as the solution to  $\|\mathbf{X}\mathbf{p} - \mathbf{d}\|_2^2$ . Direct application of this method for motion compensation is non-trivial, as the accurate estimation of  $\mathbf{a}_k$  is challenging. The 3-D marker positions must be estimated from projection images with corrupted geometry alignment.

Thus, we follow a different approach: we train a neural network to predict the RPE directly from the reconstructed images. To generate training data, we simulate rigid motion on real clinical acquisitions and compute the corresponding ground truth RPE via the virtual marker positions and their corresponding projections using Eq. (7). Thus, we aim to approximate Eq. (7) from reconstructed slice images using a neural network.

#### B. Network Architecture

Our network architecture depicted in Fig. 2 consists of two stages, a feature extraction stage followed by a regression stage. The feature extraction is driven by a siamese triplet network architecture consisting of three weight-sharing feed forward networks denoted by  $S$ . The output of the three networks is concatenated and fed to the regression network  $\mathcal{R}_t$ . The feed forward network is almost identical to the ResNet-18 architecture [54] upto the last global average pooling. We devise the network to our task by removing the last *fully connected* (FC) layer. Since the input of our network is always a tomographic reconstruction, we also remove all *batch normalization* (BN) layers. Expecting three-channel input images ranging from  $\mathbb{R}^{70 \times 216}$  to  $\mathbb{R}^{256 \times 216}$  for the different anatomical orientations, the final  $7 \times 7$  average pooling is replaced by a  $3 \times 3$  average pooling. The resulting feature maps are concatenated and represent the input to the regression network.

The regression network  $\mathcal{R}_t$  is composed of a  $1 \times 1$  convolution mapping the  $1536 \times 3 \times 3$  feature maps to  $2048 \times 3 \times 3$  feature maps followed by an  $1 \times 1$  global average pooling. The resulting feature maps are fed to four FC layers, each representing a different task  $t \in \{1, 2, 3, 4\}$ . The first FC layer maps to a single scalar output  $\mathcal{R}_1$ , the other three FC layers map to  $N$  dimensional outputs  $\mathcal{R}_2, \mathcal{R}_3$  and  $\mathcal{R}_4$ , where  $N$  represents the number of projections.

#### C. Data Generation

Motion-affected reconstructions with corresponding ground truth motion patterns are rarely available. First, contrary to spiral CT, CBCT patient data are not available from public sources and therefore difficult to obtain in general. Second, the only robust motion compensation is based on external markers, which is not used in clinical practice. The only feasible possibility is the generation of artificial motion based on motion-free acquisitions. To this end, our data-base consists of 27 clinical head CBCTs, each being ensured to have no motion artifacts by a medical expert. The data are acquired with a clinical CBCT system (Artis Q, Siemens Healthcare GmbH, Forchheim, Germany). After filtering, the high resolution projection images are down-sampled to low resolution projection images  $\mathbf{Y}_i \in \mathbb{R}^{320 \times 413}$  using an average filter. This improves the computational performance of the method and matches the voxel size of the volume reconstructed from these

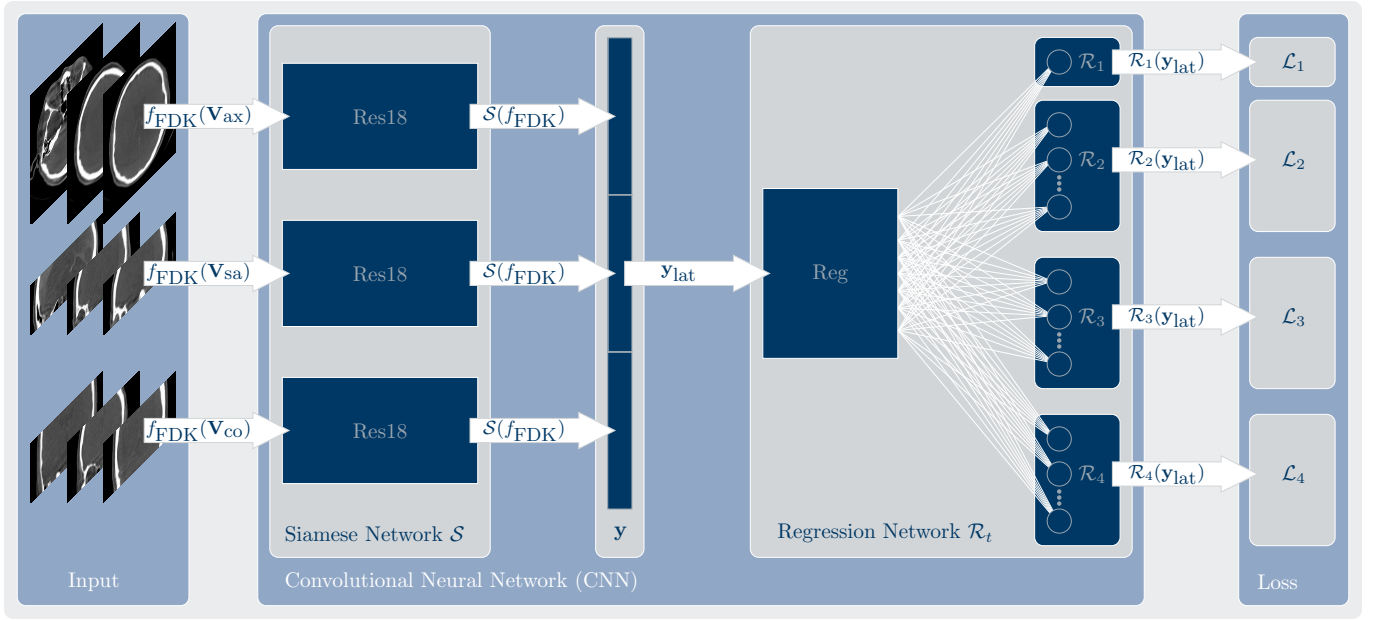


Fig. 2. Flowchart of network architecture. The input to the siamese triplet network are three slices of different anatomical orientations. The concatenated output is fed to the multi-task regression network. Based on the four outputs the respective loss is computed.

images. Down-sampling does barely affect the accuracy of the autofocus method [23].

The first step of the data generation is an alignment process of the 27 clinical CBCT scans to a mean shape. We perform this semi-automatically based on a symmetry plane alignment [18]. The result of the alignment process is a single rigid transformation which is incorporated into the system trajectory  $\mathbf{P}$ .

The second step is the generation of rigid motion which can be realized by two approaches: (1) the reconstructed volume is reprojected on a motion modulated trajectory using DRRs and then reconstructed again using  $\mathbf{P}$  or (2) the calibrated system trajectory  $\mathbf{P}$  is virtually altered by a motion trajectory and reconstructed. As DRRs are simulated projections, they cannot model the complexity of a real system and alter resolution and noise characteristics, where the latter is known to be critical for *convolutional neural networks* (CNN) [36], [46]. Therefore, we decide to choose strategy (2) where projection characteristics of a real clinical setting are preserved. Further, note that  $\mathbb{P}^3$  is diffeomorphic to  $\mathbb{SE}(3)$ , as a consequence a rigid motion can be analogously expressed by a transformation of the system geometry ( $\mathbf{P}_i \mathbf{T}_i \mathbf{a}$ ) or a transformation of the object  $\mathbf{P}_i (\mathbf{T}_i \mathbf{a})$ .

The motion generation is applied as follows: First, the system calibration  $\mathbf{P}$  is altered by a motion trajectory  $\mathbf{M}$ , giving the effective trajectory  $\mathbf{E} = \mathbf{P} \circ \mathbf{M}$ . The motion trajectory contains a random misalignment in one of the 6 motion splines. The length of misaligned motions is chosen to be distributed over a third of the trajectory and restricted to views unaffected by the Parker redundancy weighting (see Sec. II-A). The redundancy weighting alters the appearance of motion artifacts — e. g., any translations of the last few views would barely affect the reconstruction quality as those projections are mostly outfaded — making a consistent mapping of

artifact pattern to RPE infeasible. Secondly, two values are computed, the corresponding RPE per view using Eq. (7) and the motion affected reconstruction using Eq. (3). Note that the RPE is only computed based on virtual marker positions which is possible because we know the system calibration  $\mathbf{P}$  and motion trajectory  $\mathbf{M}$  during training. The volume is reconstructed on nine slices distributed in triplets of axial ( $\mathbb{R}^{216 \times 256}$ ), coronal ( $\mathbb{R}^{70 \times 216}$ ) and sagittal ( $\mathbb{R}^{70 \times 256}$ ) slices with an isotropic voxel size of 0.84 mm. The respective slices are distributed in a volume ( $\mathbb{R}^{256 \times 216 \times 70}$ ). The field-of-view is selected such that no truncation artifacts in longitudinal direction are present if reconstructed from a typical clinical scan.

#### D. Motion Learning

The overall goal of the motion learning process is to train a network that is capable of approximating Eq. (7) only based on a tomographic reconstruction. Therefore, nine slices of the tomographic reconstruction are used as input to the network. To keep the computational effort on a minimum and still capture all types of motion artifacts, the input of the network are triplets of three slices oriented in *axial* (ax), *coronal* (co) and *sagittal* (sa) direction. We denote the respective coordinates of the nine slices by  $\mathbf{V} = (\mathbf{V}_{\text{ax}}, \mathbf{V}_{\text{co}}, \mathbf{V}_{\text{sa}})$ , where  $\mathbf{V}_{\diamond}$  is a set of coordinates defined in  $\mathbb{P}^3$ . Thus,  $f_{\text{FDK}}(\mathbf{V}_{\text{ax}}, \mathbf{E}, \mathbf{Y})$  will denote the reconstruction of three slices in axial direction with effective trajectory  $\mathbf{E}$ . Let us define a triplet of reconstructed slice images for a set of projections  $\mathbf{Y}$  reconstructed with the effective trajectory  $\mathbf{E}$  as  $\mu_{\text{FDK}}^{\diamond}(\mathbf{E}) = f_{\text{FDK}}(\mathbf{V}_{\diamond}, \mathbf{E}, \mathbf{Y})$  and further, let the tuple of all reconstructed slices be  $\mu_{\text{FDK}}(\mathbf{E}) = (\mu_{\text{FDK}}^{\text{ax}}(\mathbf{E}), \mu_{\text{FDK}}^{\text{co}}(\mathbf{E}), \mu_{\text{FDK}}^{\text{sa}}(\mathbf{E}))$ . Then, the input to the regression network is computed as

$$\mathbf{y}_{\text{lat}}(\mu_{\text{FDK}}(\mathbf{E})) = \cup_{\diamond \in \{\text{ax}, \text{co}, \text{sa}\}} \mathcal{S}(\mu_{\text{FDK}}^{\diamond}(\mathbf{E})) , \quad (8)$$

where  $\cup$  denotes concatenation. Thus, each feed forward network processes three slices of the same anatomical orientation and the result is concatenated representing the latent space  $\mathbf{y}_{\text{lat}}(\boldsymbol{\mu}_{\text{FDK}}(\mathbf{E}))$ . The loss function  $l$  is based on a multi-task loss

$$l(\boldsymbol{\mu}_{\text{FDK}}(\mathbf{E}), \mathbf{E}) = \sum_{t=1}^4 \|\mathcal{R}_t(\mathbf{y}_{\text{lat}}(\boldsymbol{\mu}_{\text{FDK}}(\mathbf{E}))) - \mathcal{L}_t(\mathbf{E})\|_2^2, \quad (9)$$

with

$$\mathcal{L}_t(\mathbf{E}) = \begin{cases} \frac{1}{N} \sum_{i=1}^N d_{\text{RPE}}(\mathbf{E}_i) & \text{if } t = 1 \\ (d_{\text{RPE}}(\mathbf{E}_1), \dots, d_{\text{RPE}}(\mathbf{E}_N)) & \text{if } t = 2 \\ (d_{\text{RPE}}(\mathbf{E}_1^{\text{ip}}), \dots, d_{\text{RPE}}(\mathbf{E}_N^{\text{ip}})) & \text{if } t = 3 \\ (d_{\text{RPE}}(\mathbf{E}_1^{\text{op}}), \dots, d_{\text{RPE}}(\mathbf{E}_N^{\text{op}})) & \text{if } t = 4 \end{cases}. \quad (10)$$

Here, we assume that  $\mathbf{E}$  is implemented such that it can be decomposed into  $\mathbf{P}$  and  $\mathbf{M}$  allowing to compute the RPE using Eq.(7).  $\mathbf{E}^{\text{op}}$  and  $\mathbf{E}^{\text{ip}}$  refer to in-plane and out-plane motion. Assuming the system is rotating around the  $z$  axis, in-plane motion is within the acquisition plane and represented by parameters  $(r_z, t_x, t_y)$  and out-plane motion is stepping out the acquisition plane and represented by parameters  $(r_y, r_z, t_z)$ . We use this distinction, because in-plane motion is better detectable in axial slices, whereas out-plane motion is better detectable in coronal and sagittal slices.

For optimization we use the ADAM optimizer with a learning rate of  $10^{-4}$  and a batch size of 32. To avoid over-fitting, we use the validation set for early stopping. The residual network  $\mathcal{S}$  is initialized using pre-trained weights learned for the ImageNet classification task. The regression network  $\mathcal{R}_k$  is randomly initialized.

#### IV. EXPERIMENTS AND RESULTS

In this section we evaluate the network performance w.r.t. three aspects: (1) the behavior of the network in its core task, i.e., the regression of the RPE, (2) the performance of the network in a motion compensation benchmark in comparison to state-of-the-art methods, and (3) the applicability of the proposed method to motion-affected clinical data.

##### A. Network Accuracy

Using the data generation proposed in Sec. III-C, we generate 9001 different motion affected reconstructions. The amplitude of the applied motion is in the range of  $0^\circ$  to  $15^\circ$  mm, i.e., *mean RPEs* (mRPE) are in a range of 0 mm to 0.74 mm. Using a 21-4-2 split, this provides us with a total of 189021 samples for training, 36004 samples for validation and 18002 samples for testing. The number of spline nodes is set to  $M = 20$ . Following the training described in Sec. III-D, we achieve the optimal validation loss after  $12 \times 10^3$  iterations (see Tab. I).

1) *Ablation Study*: To inspect the network performance as well as the influence of the pre-processing, Tab. I displays the respective best validation loss values for alterations in the network structure or input data. The most important performance boost is obtained by the pre-processing step of aligning the respective reconstructions and slight improvements are

TABLE I  
BEST PERFORMING VALIDATION LOSS FOR DIFFERENT NETWORK SETTINGS.

	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_4$
Proposed	<b>0.0098</b>	0.2493	<b>0.1615</b>	<b>0.1981</b>
Proposed with BN	0.0138	0.2835	0.2145	0.2412
Proposed no alignment	0.0436	0.5644	0.5815	0.8192
Proposed no pre-training	0.0171	0.4350	0.2941	0.3279
Proposed dual task	0.0146	<b>0.2481</b>	x	x
Proposed with DenseNet	0.0205	0.3331	0.1929	0.2818

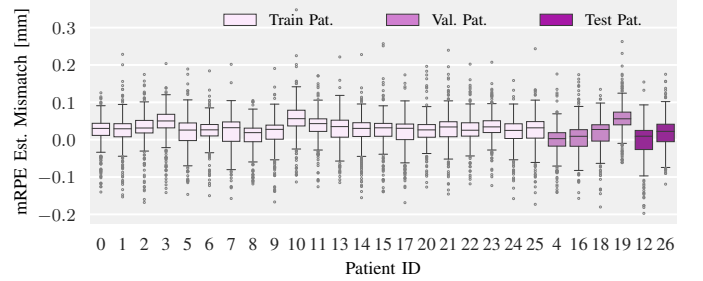


Fig. 3. Boxplots showing the deviation of the mRPE ( $\mathcal{L}_1$ ) from the ground truth for each patient. The boxplots are grouped to training (Train Pat.), validation (Val. Pat.) and test (Test Pat.) patients. All outliers are displayed as circles.

obtained by removing the BN. Further, a pre-training of  $\mathcal{S}$  on ImageNet increases the accuracy. Without the distinction of in-plane and out-plane motion (dual-task learning), the accuracy of  $\mathcal{L}_2$  decreases slightly, however, the mRPE ( $\mathcal{L}_1$ ) accuracy increases by  $\approx 50\%$ . A replacement of the residual network architecture  $\mathcal{S}$  with a pre-trained DenseNet [55] worsens the accuracy.

2) *Patient and Motion Variability*: An important aspect of motion appearance learning is the independence to the patient anatomy, similar motions applied to different patients should be predicted alike. Therefore, following the data generation presented in Sec. III-C we generate 300 additional motion shapes per patient ranging from mRPEs of 0 mm to 0.7 mm. Note, that the simulated motion is random and therefore not part of the training set. Consequently, the applied motion was never seen by the network. The results depicted in Fig. 3 show the patient-wise accuracy in predicting the mRPE ( $\mathcal{L}_1$ ). Most of the outliers are within a range of 0.2 mm, and no outlier is exceeding an error of 0.35 mm. While the accuracy is high with an mRPE of 0.013 mm, there is a slight tendency of overestimating the mRPE. The inter-patient variability of the estimation is small with a standard deviation of 0.022 mm. From the patients never seen during training, we can observe a good generalization of the learned features. The tendency to overestimate the mRPE is even slightly less observable. Besides the mRPE the network further predicts three *view-wise RPEs* (vRPE) split to in-plane motion ( $\mathcal{L}_2$ ), out-plane motion ( $\mathcal{L}_3$ ) and both ( $\mathcal{L}_4$ ). The accuracy for this task is depicted in Fig. 4. Comparing the accuracy of the vRPE estimations to the mRPE we observe higher deviations and a higher number of outliers in the vRPE estimations. The accuracy of the in-plane vRPE is higher than for the out-plane vRPE. In-plane motion is mostly distributed in axial slices, which can be

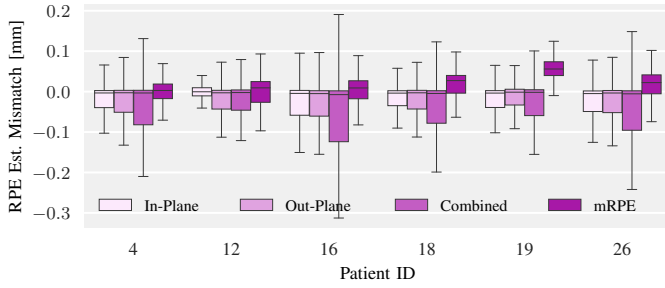


Fig. 4. Boxplots visualizing the deviations to the ground truth for the in-plane ( $\mathcal{L}_3$ ), out-plane ( $\mathcal{L}_4$ ) and combined ( $\mathcal{L}_2$ ) vRPE and the mRPE ( $\mathcal{L}_1$ ). The evaluation is based on the four validation and two test patients.

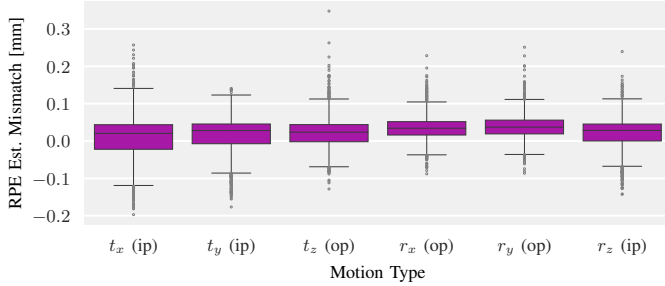


Fig. 5. Boxplots showing the mRPE deviations from the ground truth for motions clustered to the 6 motion directions. The evaluation is based on the four validation and two test patients. All outliers are displayed as circles.

reconstructed without significant cone-beam artifacts and are best suitable for motion prediction. The patient-wise deviations are more pronounced compared to the mRPE, however, still on a reasonable low level.

Figure 5 shows the mRPE clustered w. r. t. the motion directions for all patients. All motion directions can be predicted with similar accuracy, however, a slight tendency is observable that out-plane motion is predictable with less deviations.

In conclusion, the experiments have shown the patient independence of the proposed appearance learning approach, as well as the independence of the motion direction. This provides us a method that has no inherent limitations to certain motion patterns as apparent, e. g., in epipolar consistency [21], which is sensitive to out-plane motion, but barely applicable to in-plane motion.

3) *RPE Trajectory Prediction*: Using the same data as in Sec. IV-A2, we investigate in this experiment the performance of the estimated vRPE for motion classification. The predicted

vRPEs are interpreted as soft-classifiers, where we define a view to be in the motion-free class (negative) if the average predicted value for view  $i$  satisfies

$$\frac{1}{2}\mathcal{R}_2 + \frac{1}{4}\mathcal{R}_3 + \frac{1}{4}\mathcal{R}_4 \leq 0.1. \quad (11)$$

In Fig. 6 the accuracy is displayed encoded to *false negative* (FN), *false positive* (FP), and the combination of FP, *true positive* (TP) and *true negative* (TN). If the predicted value is used as an indicator function (see Sec. IV-B) in a motion estimation scenario a low FN rate is important. Regions classified as motion free will receive little attention within the optimization. On the opposite, a FP classification does not worsen the result and is therefore non-critical. These properties are satisfied as observable from Fig. 6. The FN rate is  $\approx 0\%$  and the FP rate is  $\approx 25\%$ . Note that the peaks in the FP curve are due to the spline nodes, where transitions from motion-affected to motion-free views arise with increased frequency.

## B. Motion Estimation Benchmark

1) *Autofocus*: The motion estimation benchmark is based on the four validation patients and two test patients. We apply a known motion trajectory  $\mathbf{M}$  to the projection matrices  $\mathbf{P}$  and evaluate the performance of six metrics (see IV-B3) to find the annihilating trajectory  $\mathbf{C}$  (see II-B). We describe the trajectory as a function of six motion spline nodes  $\mathbf{m} = (\mathbf{m}_{t_x}, \mathbf{m}_{t_y}, \mathbf{m}_{t_z}, \mathbf{m}_{r_x}, \mathbf{m}_{r_y}, \mathbf{m}_{r_z})$ . Each element of  $\mathbf{m} \in \mathbb{R}^{6 \times M}$  describes the respective spline node within the trajectory. Thus,  $\mathbf{m}_{r_y, 420}$  describes the rotation around the  $y$ -axis at acquisition view 420. Then, the motion curve vector  $\mathbf{t}(\mathbf{m}) = (t_x, t_y, t_z, r_x, r_y, r_z)$  is computed by evaluating the spline for each acquired view. For example  $t_x = (\eta_{\mathbf{m}_{t_x}}(0), \eta_{\mathbf{m}_{t_x}}(1), \dots, \eta_{\mathbf{m}_{t_x}}(N))^T$ , with  $\eta_{\mathbf{m}_{t_x}}(i)$  denoting the spline evaluation at position  $i$  based on the spline nodes  $\mathbf{m}_{t_x}$  as proposed in [50]. From the six motion curves described by  $\mathbf{t}$  we can directly compute the annihilating trajectories denoted by  $\mathbf{C}(\mathbf{t}(\mathbf{m}))$ . Note, that the motion trajectory itself is generated in an equal way.

The components of  $\mathbf{m}$  are found by optimizing the IQM  $f_{\text{IQM}}$

$$\hat{\mathbf{m}} = \underset{\mathbf{m}}{\operatorname{argmin}} f_{\text{IQM}}(f_{\text{FDK}}(\mathbf{V}, \mathbf{P} \circ \mathbf{C}(\mathbf{m}), \mathbf{Y})) \quad (12)$$

2) *Optimization*: Equation (12) is optimized using the gradient free downhill simplex algorithm [56]. We optimize only one node at a time iterating over all nodes in sequential order. We use a coarse to fine strategy by defining 5 stages. In the first three stages we define a starting stepsize of  $1^\circ/\text{mm}$  for the simplex and set the number of iterations to 2. This allows a rough estimate of the trajectory. In the last two stages, we set the number of iterations to 100 with initial stepsize of  $0.5^\circ/\text{mm}$ . The optimization is finished if either the maximum number of iterations is exceeded or the improvement in  $\mathbf{m}$  is below  $0.001^\circ/\text{mm}$ .

3) *Image Quality Metrics*: We define three IQMs denoted by Ent, Tv and Cnn. Ent and Tv refer to the histogram entropy and TV of the slice images, respectively. We implement Ent following the methodology of Herbst et al. [28] and Tv as

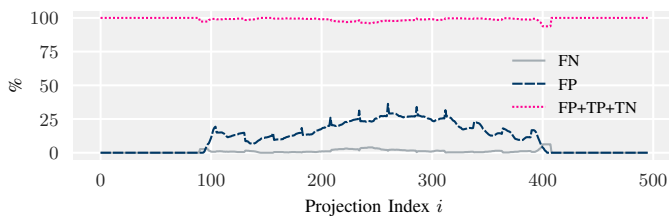


Fig. 6. Binary soft-classification results plotted over the respective views. FN relates to a motion-affected region that is classified by the network as motion-free. FP relates to a misclassification to the motion-affected class. TP and TN are correctly predicted views.



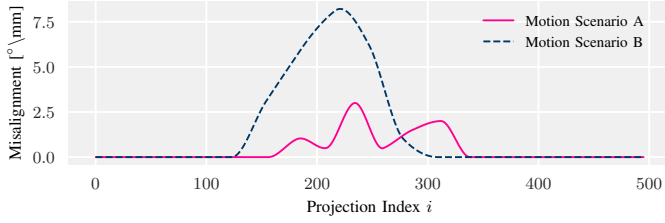


Fig. 7. The motion trajectories for both motion scenarios. The motion is applied respectively to the motion axis under investigation. The curves are generated using 20 spline nodes and 17 spline nodes for scenario A and B, respectively.

proposed in Kingston et al. [24]. We selected these two metrics due to their popularity in the literature. Ent was found to be superior for geometry alignment in a study by Wicklein et al. [25]. Cnn is our proposed method. In addition we define Ent+, Tv+ and Cnn+, each denoting an initial optimization with either Ent, Tv or Cnn followed by a fine-tuning of the annihilating trajectory with an additional optimization stage using Ent (for Tv+ and Cnn+) or Tv (for Ent+). For Ent and Tv the implementation of  $f_{IQM}$  is straightforward: the histogram entropy or total variation of the nine reconstructed slices are calculated. Following Wicklein et al. [25] we use a bone window for the histogram calculation. Their studies showed that restricting the histogram calculation to values within a bone window improves the method's performance, because only relevant image features are captured. The implementation of  $f_{IQM}$  for Cnn and Cnn+ uses an additional indicator function  $\mathbf{1}_{\mathbb{M}}$ , where  $\mathbb{M}$  describes the set of views satisfying Eq. (11). Thus, the IQM for Cnn is defined by

$$f_{IQM} = \mathcal{R}_1(\mathbf{y}_{lat}(\mathbf{V}, \mathbf{P} \circ \mathbf{C}, \mathbf{Y})) \text{ s.t. } \mathbf{1}_{\mathbb{M}}|\mathbf{t}(\mathbf{m})| = \mathbf{0}, \quad (13)$$

with  $|\cdot|$  denoting element wise absolute value. Note that  $\mathbf{1}_{\mathbb{M}}$  is not updated during the iterations.

4) *Motion Scenarios*: We design two motion scenarios (scenario A and scenario B) differing in their motion shapes and the number of spline nodes used for both, the motion trajectory and annihilating trajectory. Scenario A uses 20 spline nodes for the motion curves and the same number of spline nodes for the annihilating curves. Scenario B uses 17 spline nodes for the motion curves and 40 spline nodes for the annihilating curves. The two motion amplitudes are depicted in Fig. 7. For both scenarios the respective motion is applied to one of the 6 motion axes, respectively. In each scenario, we optimize only for the axis which is affected by the motion.

5) *Motion Estimation Results*: To quantify the performance, we measure the mismatch of the respective motion curves and estimated annihilating curves. For a complete compensation, we need  $\mathbf{M} \circ \mathbf{C} = \mathbf{1}$ , which is the case if the motion curve and the annihilating curve add up to zero. The second metric measures the reconstruction quality by computing the SSIM of the respective motion-compensated reconstruction and the ground truth.

Figure 8 and Fig. 9 show the misalignment of the motion curve for motion scenario A and B, respectively, averaged over the 4 validation and 2 test patients for Ent, Tv, Cnn and Cnn+. Numeric results showing the misalignment for all six

metrics are displayed in Tab. II and Tab. IV and numeric results showing the SSIM values for all six metrics are displayed in Tab. III and Tab. V, respectively. Selected reconstructions for both motion scenarios are presented in Fig. 10.

The proposed method performs well in both scenarios. In motion scenario A, the state-of-the-art methods perform similar to the network-based solution. In the majority of cases, the network-based results are superior. However, in almost 50% of the experiments, either Tv or Ent achieves the best results. A fine-tuning of the traditional metrics (Ent+, Tv+) barely improves the results and can lead to a degeneration of the performance. The margin by which the network-based method outperforms both state-of-the-art metrics is much higher than vice-versa. Figure 10 shows that the network-based approach is further capable of dealing with metal artifacts. In this case, the network without post-optimization using the entropy (Cnn) achieves the best results. Note, that our training set also includes patients with metal artifacts.

For scenario B, our method is constantly outperforming the state-of-the-art metrics, both in terms of SSIM and measured misalignment. By an additional post-optimization with the entropy-based compensation, the best results are achieved with Cnn+. As can be seen from Fig. 10, the network is capable of approximating the true motion curve but ignores small high-frequency motions. These motions are then eliminated by the entropy. However, deploying entropy alone produces mediocre results because the optimization gets stuck in local minima. Tv performs worst w.r.t. the misalignment of the annihilating and motion curves, however, the SSIM is comparable to the entropy-based procedure.

### C. Motion-Affected Clinical Data

1) *Data and Preprocessing*: To demonstrate the effectiveness of the proposed method in clinical practice, we apply it on a motion-affected clinical dataset. Similar to the acquired data used for the network training and evaluation, the patient was scanned with a clinical CBCT system (Artis Q, Siemens Healthcare GmbH, Forchheim). The projections were down-sampled and aligned following the same procedure (i.e., step 1 of data generation) as presented in Sec. III-C.

2) *Motion Compensation Scheme*: We model the annihilating trajectory with an Akima spline consisting of 20 spline nodes equally distributed over the trajectory. We adapt the optimization scheme from Sec. IV-B2. We sequentially optimize for all six motion parameters in the following sequence  $(t_z, t_x, t_y, r_x, r_y, r_z)$ . To optimize for motion we use Cnn and Cnn+.

3) *Motion Compensation Results*: Figure 11 displays reconstructed slice images from the motion-affected clinical dataset (None) as well as motion-compensated reconstructions (Cnn, Cnn+). As ground-truth reconstructions are not available, only a qualitative inspection is possible. In slice 1 we observe motion artifacts especially at the borders of the temporal bones as well as near the nasal cavities and ethmoid bone. The anatomy contours can be well recovered using Cnn or Cnn+. As can be observed from the difference images (dCnn, dCnn+), streaks at the bone contours are



TABLE II

MEAN MISALIGNMENT [ $^{\circ}$ \mm] BETWEEN ANNIHILATING TRAJECTORY AND GROUND-TRUTH TRAJECTORY FOR MOTION SCENARIO A.

	$t_x$ (ip)	$t_y$ (ip)	$t_z$ (op)	$r_x$ (op)	$r_y$ (op)	$r_z$ (ip)
None	0.45	0.45	0.45	0.45	0.45	0.45
Ent	0.69	0.20	0.13	<b>0.10</b>	0.33	0.33
Ent+	1.07	0.16	0.16	0.12	0.35	0.39
Tv	0.97	<b>0.12</b>	0.45	0.32	0.46	0.69
Tv+	0.95	0.14	0.46	0.27	0.46	0.69
Cnn	<b>0.27</b>	0.24	0.15	0.24	0.21	0.18
Cnn+	0.28	0.20	<b>0.13</b>	0.19	<b>0.19</b>	<b>0.14</b>

TABLE III

SSIM VALUES NORMALIZED TO THE RANGE [0,100] FOR MOTION SCENARIO A. THE SSIM IS COMPUTED BETWEEN THE GROUND-TRUTH RECONSTRUCTION AND THE RESPECTIVE COMPENSATED RECONSTRUCTION. IN BRACKETS, THE SSIM IS COMPUTED IN A VOLUME-OF-INTEREST. THE VOLUME OF INTEREST COVERS THE NASAL BONES ONLY.

	$t_x$	$t_y$	$t_z$	$r_x$	$r_y$	$r_z$
None	58 (64)	81 (89)	75 (72)	81 (68)	79 (86)	68 (66)
Ent	53 (67)	94 (97)	<b>95 (95)</b>	97 (97)	89 (97)	76 (81)
Ent+	50 (65)	94 (97)	95 ( <b>95</b> )	<b>97 (98)</b>	89 (97)	74 (82)
Tv	51 (66)	<b>97 (98)</b>	78 (77)	88 (85)	82 (93)	66 (70)
Tv+	49 (67)	96 (98)	77 (78)	90 (89)	83 (94)	63 (71)
Cnn	<b>69 (81)</b>	90 (95)	92 (92)	90 (87)	90 (97)	83 (87)
Cnn+	69 ( <b>81</b> )	92 (97)	94 (95)	92 (92)	<b>92 (98)</b>	<b>86 (91)</b>

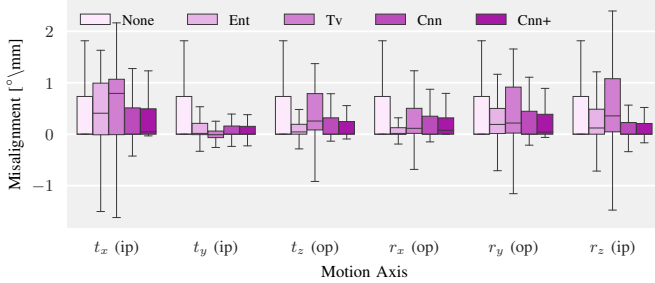


Fig. 8. Boxplot for motion scenario A, showing the misalignment of the annihilating curve to the motion curve plotted for each of the four different IQMs used in the motion estimation benchmark.

eliminated. In slice 2 the motion artifacts are severe in the orbital bone structures. The streak artifacts are reduced in both motion-compensated reconstructions, restoring the homogeneous regions. A small residual motion is still observable with the motion-compensated reconstructions, however, the image quality could be improved substantially. From the difference images between the Cnn-based and Cnn+-based compensated reconstructions (dCnnCnn+) we see that the entropy-based fine-tuning barely affects the reconstruction quality.

## V. DISCUSSION

We propose an appearance learning approach that can be deployed for image-based motion compensation. For that purpose, we devise a framework that learns the mRPE as well

TABLE IV

MEAN MISALIGNMENT [ $^{\circ}$ \mm] BETWEEN ANNIHILATING TRAJECTORY AND GROUND-TRUTH TRAJECTORY FOR MOTION SCENARIO B.

	$t_x$ (ip)	$t_y$ (ip)	$t_z$ (op)	$r_x$ (op)	$r_y$ (op)	$r_z$ (ip)
None	1.52	1.52	1.52	1.52	1.52	1.52
Ent	1.70	1.18	1.24	1.09	1.32	1.43
Ent+	1.93	1.06	1.26	1.09	1.33	1.57
Tv	2.01	1.34	1.54	1.38	1.55	1.83
Tv+	1.95	1.17	1.43	1.18	1.48	1.79
Cnn	1.20	0.54	1.15	0.84	0.76	0.83
Cnn+	<b>1.14</b>	<b>0.45</b>	<b>0.90</b>	<b>0.67</b>	<b>0.69</b>	<b>0.62</b>

TABLE V

SSIM VALUES NORMALIZED TO THE RANGE [0,100] FOR MOTION SCENARIO B. THE SSIM IS COMPUTED BETWEEN THE GROUND-TRUTH RECONSTRUCTION AND THE RESPECTIVE COMPENSATED RECONSTRUCTION. IN BRACKETS, THE SSIM IS COMPUTED IN A VOLUME-OF-INTEREST. THE VOLUME OF INTEREST COVERS THE NASAL BONES ONLY.

	$t_x$	$t_y$	$t_z$	$r_x$	$r_y$	$r_z$
None	49 (49)	65 (71)	66 (61)	69 (45)	66 (68)	54 (56)
Ent	46 (47)	66 (76)	66 (61)	72 (51)	67 (71)	55 (55)
Ent+	46 (47)	68 (78)	66 (59)	72 (52)	67 (71)	55 (55)
Tv	48 (48)	67 (74)	65 (60)	70 (47)	66 (68)	53 (55)
Tv+	45 (46)	67 (76)	65 (59)	71 (48)	65 (69)	52 (54)
Cnn	46 (51)	75 (85)	67 (66)	73 (58)	72 (80)	54 (57)
Cnn+	<b>50 (54)</b>	<b>81 (90)</b>	<b>70 (67)</b>	<b>78 (67)</b>	<b>75 (84)</b>	<b>66 (65)</b>

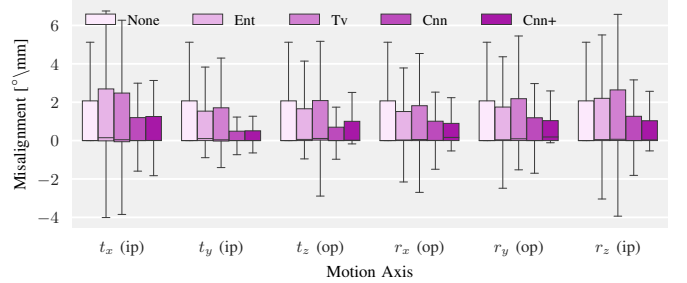


Fig. 9. Boxplot for motion scenario B, showing the misalignment of the annihilating curve to the motion curve plotted for each of the four different IQMs used in the motion estimation benchmark.

as vRPEs from reconstructed slice images. Exact computation of the RPE allows for geometric calibration for high-quality CBCT [29]. Hence, given a 100% network accuracy, minimizing the network-predicted mRPE would yield highly accurate motion parameters. The axis with lowest accuracy in predicting the mRPE is also the axis with lowest performance in the motion compensation benchmark.

We further show that we can learn general features applicable to all three types of translations and rotations. The learned features are even less dependent on the motion axis than traditional methods. For example, Tv shows superior results in compensating translation along the y-axis as observable from motion experiment A.

Autofocus methods are characterized by optimizing an IQM in the reconstruction domain. Inherently, information from

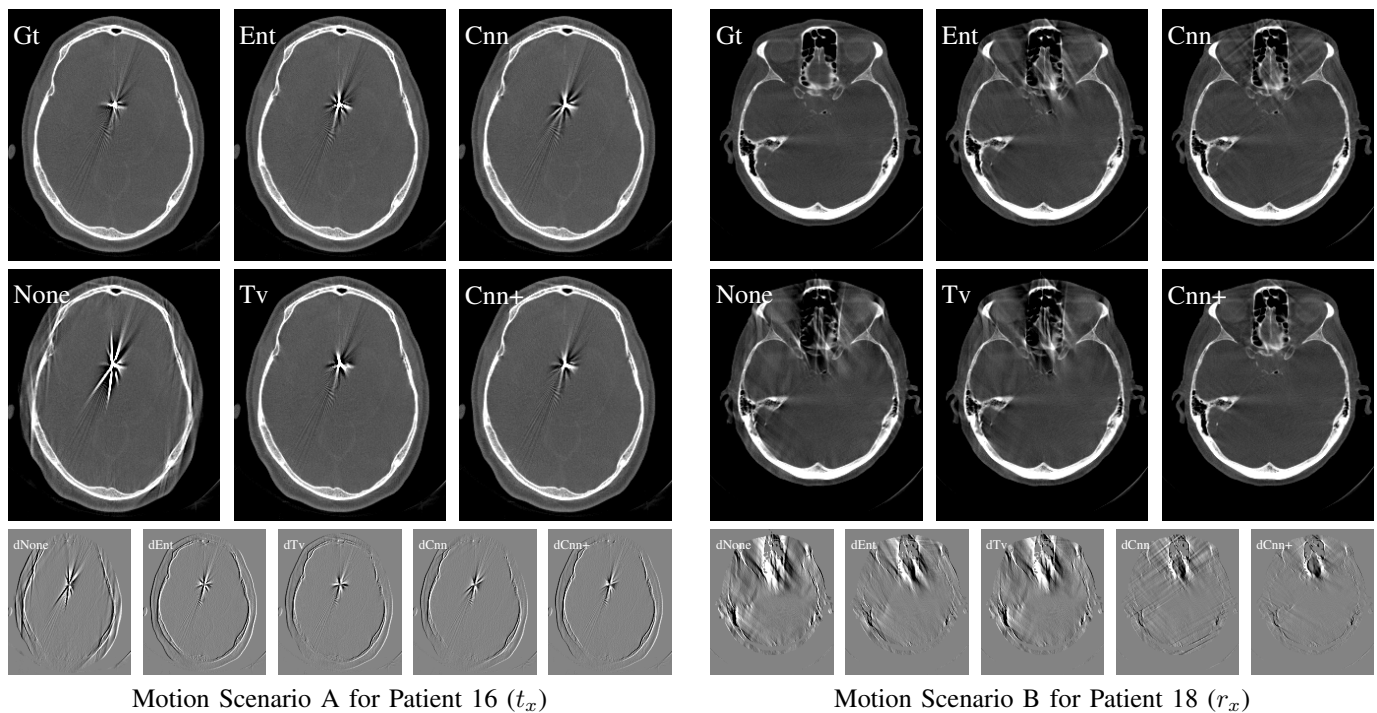


Fig. 10. Selected reconstructions (HU [50-2000]) from the motion benchmark. Left block: Motion scenario A using 20 spline nodes to model the annihilating trajectory, and right block: Motion scenario B using 40 spline nodes to model the annihilation trajectory. The respective bottom row displays the difference images to the Ground truth (Gt). The deviation of the annihilating curve to the negative motion curve is None = 0.44 mm, Ent = 0.74 mm, Tv = 0.69 mm, Cnn = 0.37 mm and Cnn+ = 0.45 mm for motion scenario A, and None = 1.52 mm, Ent = 0.62 mm, Tv = 1.23 mm, Cnn = 0.36 mm and Cnn+ = 0.31 mm for motion scenario B.

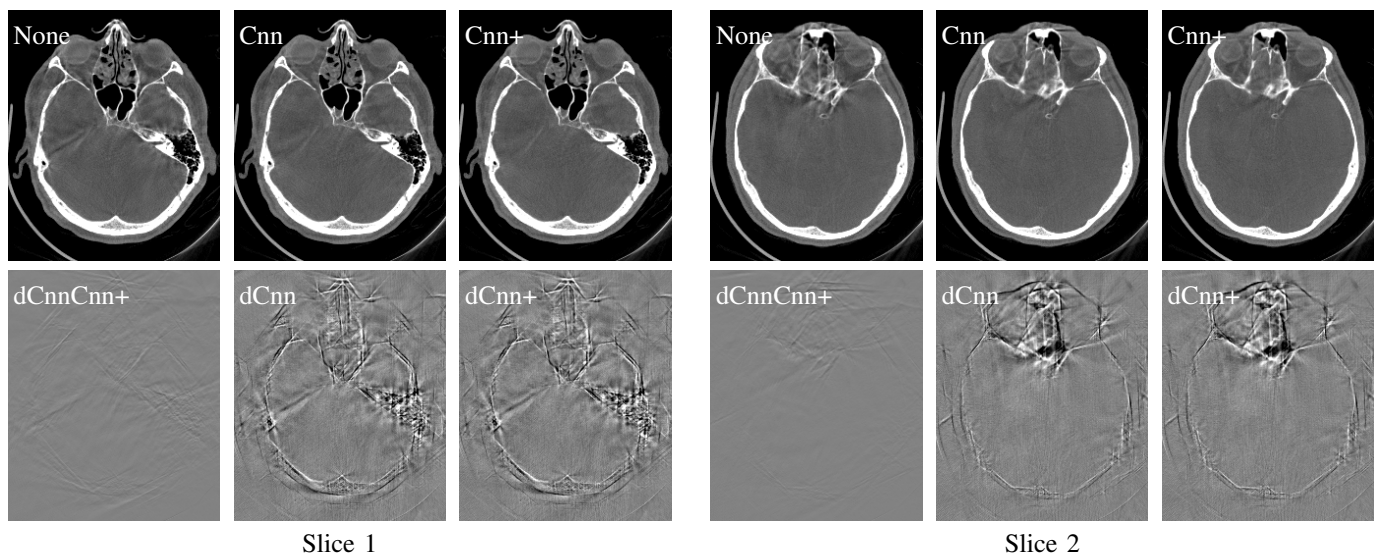


Fig. 11. Top row: two reconstructed slices (HU [50-2000]) of a motion-affected clinical dataset (None) and motion-compensated data using Cnn or Cnn+, respectively. Bottom row: difference images of motion compensated and motion-affected reconstruction (dCnn, dCnn+) as well as difference image of motion compensated reconstructions using Cnn and Cnn+, respectively (dCnnCnn+). Display windows within each row are equal.

which part of the trajectory a misalignment is expected can only be deduced from the gradients within the optimization. We aim to overcome this by learning an initial estimate about the distribution of the expected motion. The view-wise prediction can be used as a soft-classifier to steer the optimization. The FN range is close to 0%, ensuring that the optimization cannot be worsened by applying the soft-classifier for optimization steering.

We use randomly generated motions that are limited in their frequency due to the deployed splines. However, the network is capable to generalize to unseen motion frequencies. In motion scenario B we compensate the motion with a spline controlled by 40 spline nodes, whereas in training, only motion patterns generated with 20 spline nodes were shown to the network. Hence, the network is capable to generalize to higher frequency motion patterns. However, we can observe, that very small but high-frequent motion artifacts are barely accounted for by the network (see Fig. 10, scenario B). These types of motion patterns were not part of the network training. Besides those small motion artifacts, the overall motion trajectory can be estimated well by the network. This property is synergistic with traditional IQMs. After fine-tuning with the entropy-based IQM the fine motion artifacts are eliminated.

Traditional IQMs measure the artifact strength by a limited set of image features. If the reconstruction is not corrupted by motion, the image shows homogeneous soft-tissue areas and clear bone-boundaries. This results in a low TV value and low histogram entropy. Motion corrupts the homogeneous regions and blur bone-edges increasing the histogram entropy and TV value. However, the image features recognized by TV and entropy are not directly linked to the patient motion strength and therefore are susceptible to local minima. Thus, both metrics are successful if the motion is small but fail if the motion is large. This is shown by the experiments, where both metrics, TV and Ent, perform well in scenario A. For larger motions as apparent in motion scenario B, both metrics fail. In those scenarios, the learned metric Cnn outperforms the traditional methods in all six experiments.

Besides being only trained on synthetically generated motion, the network generalizes well to real clinical motion. We demonstrate this using a clinical motion-affected scan. Due to the rigid structure of the motion, a transformation of the object can be equivalently described by a static object and a transformation of the system geometry. This allows realistic generation of motion artifacts from artifact-free CBCT acquisitions.

## VI. CONCLUSION AND OUTLOOK

Our proposed method can be used in a variational manner for image-based autofocus techniques. The result is always based on the acquired raw-data and ensures data integrity. This is a strong advantage to all other learning-based approaches found in our literature review. Current learning-based approaches perform an image-to-image translation, without any guarantee for the consistency with the acquired raw data. In contrast, using the proposed method the images are always reconstructed from the raw data minimizing the risk for generating clinical images leading to improper diagnosis.

The experiments show that motion artifacts can be learned by a neural network and that our learning-based approach can outperform state-of-the-art IQMs in a motion estimation benchmark. We devised the approach based on the FDK algorithm and artificial motion. Using a motion-affected clinical dataset, we further demonstrate that the method translates to real clinical motion. The FDK is suitable for autofocus approaches [27], [24] due to its computational efficiency. A possible extension, however, would be a reconstruction algorithm, capable of reconstructing arbitrary trajectories [57]. The FDK assumes two fundamental properties: (1) homogeneous object in the direction perpendicular to the acquisition plane and (2) equally sampled trajectories along an arc. If any of those assumptions are not met, the reconstruction reveals cone-beam artifacts or intensity inhomogeneities. Therefore, it can only compute approximate solutions for motion compensation.

Although our experiments are tailored for head CBCT, the concept is neither limited to rigid head motion nor to transmission imaging. By replacing the filtered back-projection with the inverse model for MRI — e.g., non-uniform Fourier transform [58] — the approach can be directly trained for propeller trajectories in MRI. By additionally replacing the RPE-based regression metric with an appropriate metric (e.g., energy of a spline deformation field), also Cartesian sampled MRI can be tackled. Similar strategies are thinkable for PET.

**Disclaimer:** The concepts and information presented in this article are based on research and are not commercially available.

## REFERENCES

- [1] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *Comput Vis ECCV*, pp. 484–498, 1998.
- [2] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans Pattern Anal Mach Intell*, no. 5, pp. 564–575, 2003.
- [3] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," *Proc BMVC*, p. 6, 2006.
- [4] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans Pattern Anal Mach Intell*, vol. 33, no. 8, pp. 1619–1632, 2010.
- [5] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," *Proc BMVC*, 2015.
- [6] A. Preuhs, M. Manhart, P. Roser, B. Stimpel, C. Syben, M. Psychogios, M. Kowarschik, and A. Maier, "Image quality assessment for rigid motion compensation," *MedNeurIPS*, 2019.
- [7] G. Lauritsch, J. Boese, L. Wigstrom, H. Kemeth, and R. Fahrig, "Towards cardiac C-arm computed tomography," *IEEE Trans Med Imaging*, vol. 25, no. 7, pp. 922–934, 2006.
- [8] O. Taubmann, G. Lauritsch, G. Krings, and A. Maier, "Convex temporal regularizers in cardiac C-arm CT," *Procs CT Meeting*, pp. 545–548, 2016.
- [9] A. C. Larson, R. D. White, G. Laub, E. R. McVeigh, D. Li, and O. P. Simonetti, "Self-gated cardiac cine MRI," *Magn Reson Med*, vol. 51, no. 1, pp. 93–102, 2004.
- [10] E. Hoppe, J. Wetzel, C. Forman, G. Kördörfer, M. Schneider, P. Speier, M. Schmidt, and A. Maier, "Free-breathing, self-navigated and dynamic 3D multi-contrast cardiac cine imaging using cartesian sampling and compressed sensing," *Proc ISMRM*, 2019.
- [11] P. Fischer, A. Faranesh, T. Pohl, A. Maier, T. Rogers, K. Ratnayaka, R. Lederman, and J. Hornegger, "An MR-based model for cardio-respiratory motion compensation of overlays in X-ray fluoroscopy," *IEEE Trans Med Imaging*, vol. 37, no. 1, pp. 47–60, 2017.
- [12] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen, "Image reconstruction by domain-transform manifold learning," *Nature*, vol. 555, no. 7697, p. 487, 2018.



- [13] T. Küstner, K. Armanious, J. Yang, B. Yang, F. Schick, and S. Gatidis, "Retrospective correction of motion-affected MR images using deep learning frameworks," *Magn Reson Med*, vol. 82, no. 4, pp. 1527–1540, 2019.
- [14] S. Latif, M. Asim, M. Usman, J. Qadir, and R. Rana, "Automating motion correction in multishot MRI using generative adversarial networks," *MedNeurIPS*, 2018.
- [15] K. Xiao, Y. Han, Y. Xu, L. Li, X. Xi, H. Bu, and B. Yan, "X-ray cone-beam computed tomography geometric artefact reduction based on a data-driven strategy," *Appl Opt*, vol. 58, no. 17, pp. 4771–4780, 2019.
- [16] S. Helgason, "The radon transform," *Birkhauser, Boston, Massachusetts*, 1980.
- [17] H. Yu and G. Wang, "Data consistency based rigid motion artifact reduction in fan-beam CT," *IEEE Trans Med Imaging*, vol. 26, no. 2, pp. 249–260, 2007.
- [18] A. Preuhs, A. Maier, M. Manhart, M. Kowarschik, E. Hoppe, J. Fotouhi, N. Navab, and M. Unberath, "Symmetry prior for epipolar consistency," *Int J Comput Assist Radiol Surg*, vol. 14, no. 9, pp. 1541–1551, 2019.
- [19] A. Preuhs, A. Maier, M. Manhart, J. Fotouhi, N. Navab, and M. Unberath, "Double your views – exploiting symmetry in transmission imaging," *Med Image Comput Comput Assist Interv*, 2018.
- [20] N. Maass, F. Dennerlein, A. Aichert, and A. Maier, "Geometrical jitter correction in computed tomography," *Proc CT-Meeting*, pp. 338–342, 2014.
- [21] R. Frysch and G. Rose, "Rigid motion compensation in C-arm CT using consistency measure on projection data," *Med Image Comput Comput Assist Interv*, pp. 298–306, 2015.
- [22] J. G. Pipe, "Motion correction with propeller MRI: application to head motion and free-breathing cardiac imaging," *Magn Reson Med*, vol. 42, no. 5, pp. 963–969, 1999.
- [23] D. Atkinson, D. L. Hill, P. N. Stoye, P. E. Summers, and S. F. Keevil, "An autofocus algorithm for the automatic correction of motion artifacts in MR images," *Proc IPMI*, pp. 341–354, 1997.
- [24] A. Kingston, A. Sakellariou, T. Varslot, G. Myers, and A. Sheppard, "Reliable automatic alignment of tomographic projection data by passive auto-focus," *J Med Phys*, vol. 38, no. 9, pp. 4934–45, 2011.
- [25] J. Wicklein, H. Kunze, W. A. Kalender, and Y. Kyriakou, "Image features for misalignment correction in medical flat-detector CT," *J Med Phys*, vol. 39, no. 8, pp. 4918–4931, 2012.
- [26] C. Rohkohl, H. Bruder, K. Stierstorfer, and T. Flohr, "Improving best-phase image quality in cardiac CT by motion correction with MAM optimization," *J Med Phys*, vol. 40, no. 3, p. 031901, 2013.
- [27] A. Sisniega, J. W. Stayman, J. Yorkston, J. Siewerdsen, and W. Zbijewski, "Motion compensation in extremity cone-beam CT using a penalized image sharpness criterion," *Phys Med Biol*, vol. 62, no. 9, p. 3712, 2017.
- [28] M. Herbst, C. Luckner, J. Wicklein, J.-P. Grunz, T. Gassenmaier, L. Ritschl, and S. Kappler, "Misalignment compensation for ultra-high-resolution and fast CBCT acquisitions," *Proc SPIE*, vol. 10948, pp. 406–412, 2019.
- [29] N. K. Strobel, B. Heigl, T. M. Brunner, O. Schuetz, M. M. Mitschke, K. Wiesent, and T. Mertelmeier, "Improving 3D image quality of X-ray C-arm imaging systems by using properly designed pose determination systems for calibrating the projection geometry," *J Med Imaging*, pp. 943–954, 2003.
- [30] M. B. Ooi, S. Krueger, W. J. Thomas, S. V. Swaminathan, and T. R. Brown, "Prospective real-time correction for arbitrary head motion using active markers," *Magn Reson Med*, vol. 62, no. 4, pp. 943–954, 2009.
- [31] K. Müller, M. Berger, J. Choi, S. Datta, S. Gehrisch, T. Moore, M. P. Marks, A. K. Maier, and R. Fahrig, "Fully automatic head motion correction for interventional C-arm systems using fiducial markers," *Proc Fully3D*, pp. 1–4, 2015.
- [32] M. W. Haskell, S. F. Cauley, and L. L. Wald, "Targeted motion estimation and reduction (TAMER): data consistency based motion mitigation for MRI using a reduced model joint optimization," *IEEE Trans Med Imaging*, vol. 37, no. 5, pp. 1253–1265, 2018.
- [33] M. Berger, K. Müller, A. Aichert, M. Unberath, J. Thies, J.-H. Choi, R. Fahrig, and A. Maier, "Marker-free motion correction in weight-bearing of the knee joint," *J Med Phys*, vol. 43, no. 3, pp. 1235–1248, 2016.
- [34] S. Ouadah, W. Stayman, J. Gang, T. Ehtiati, and J. Siewerdsen, "Self-calibration of cone-beam CT geometry using 3D2D image registration," *Phys Med Biol*, vol. 61, no. 7, p. 2613, 2016.
- [35] F. Dennerlein and A. Jerebko, "Geometric jitter compensation in cone-beam CT through registration of directly and indirectly filtered projections," *Proc NSS/MIC*, pp. 2892–2895, 2012.
- [36] Y. Huang, A. Preuhs, G. Lauritsch, M. Manhart, X. Huang, and A. Maier, "Data consistent artifact reduction for limited angle tomography with deep learning prior," *Workshop on MLMIR*, 2019.
- [37] S. Abdurahman, R. Frysch, R. Bismark, S. Melnik, O. Beuing, and G. Rose, "Beam hardening correction using cone beam consistency conditions," *IEEE Trans Med Imaging*, vol. 37, no. 10, pp. 2266–2277, 2018.
- [38] M. Hoffmann, T. Würfl, N. Maaß, F. Dennerlein, A. Aichert, and A. K. Maier, "Empirical scatter correction using the epipolar consistency condition," *Proc CT-Meeting*, 2018.
- [39] T. Würfl, N. Maaß, F. Dennerlein, X. Huang, and A. K. Maier, "Epipolar consistency guided beam hardening reduction-ecc<sup>2</sup>," *Proc Fully 3D*, 2017.
- [40] D. Punzet, R. Frysch, and G. Rose, "Extrapolation of truncated C-arm CT data using grangeat-based consistency measures," *Proc CT-Meeting*, pp. 218–221, 2018.
- [41] A. Preuhs, N. Ravikumar, M. Manhart, B. Stimpel, E. Hoppe, C. Syben, M. Kowarschik, and A. Maier, "Maximum likelihood estimation of head motion using epipolar consistency," *Proc BVM*, pp. 134–139, 2019.
- [42] B. Bier, K. Aschoff, C. Syben, M. Unberath, M. Levenston, G. Gold, R. Fahrig, and A. Maier, "Detecting anatomical landmarks for motion estimation in weight-bearing imaging of knees," *Workshop on MLMIR*, pp. 83–90, 2018.
- [43] H. Liao, W.-A. Lin, J. Zhang, J. Zhang, J. Luo, and S. K. Zhou, "Multiview 2D/3D rigid registration via a point-of-interest network for tracking and triangulation," *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, pp. 12 638–12 647, 2019.
- [44] D. Toth, S. Miao, T. Kurzendorfer, C. A. Rinaldi, R. Liao, T. Mansi, K. Rhode, and P. Mountney, "3D/2D model-to-image registration by imitation learning for cardiac procedures," *Int J Comput Assist Radiol Surg*, vol. 13, no. 8, pp. 1141–1149, 2018.
- [45] C.-R. Chou, B. Frederick, G. Mageras, S. Chang, and S. Pizer, "2D/3D image registration using regression learning," *Comput Vis Image Und*, vol. 117, no. 9, pp. 1095–1106, 2013.
- [46] A. K. Maier, C. Syben, B. Stimpel, T. Würfl, M. Hoffmann, F. Schebesch, W. Fu, L. Mill, L. Kling, and S. Christiansen, "Learning with known operators reduces maximum error bounds," *Nat Mach Intell*, vol. 1, no. 8, pp. 373–380, 2019.
- [47] L. Feldkamp, L. Davis, and J. Kress, "Practical cone-beam algorithm," *J Opt Soc Am A*, vol. 1, no. 6, pp. 612–619, 1984.
- [48] S. Rit, D. Sarrut, and L. Desbat, "Comparison of analytic and algebraic methods for motion-compensated cone-beam CT reconstruction of the thorax," *IEEE Trans Med Imaging*, vol. 28, no. 10, pp. 1513–1525, 2009.
- [49] D. L. Parker, "Optimal short scan convolution reconstruction for fan beam CT," *J Med Phys*, vol. 9, no. 2, pp. 254–257, 1982.
- [50] H. Akima, "A new method of interpolation and smooth curve fitting based on local procedures," *JACM*, vol. 17, no. 4, pp. 589–602, 1970.
- [51] S. Arroyo-Camejo, B. Odry, X. Chen, K. Nael, L. Liu, D. Grodzki, and M. Nadar, "Towards contrast-independent automated motion detection using 2D adversarial denesets," *Proc ISMRM*, 2019.
- [52] S. Braun, X. Chen, B. Odry, B. Mailhe, and M. Nadar, "Motion detection and quality assessment of MR images with deep convolutional denesets," *Proc ISMRM*, 2018.
- [53] K. Meding, A. Loktyushin, and M. Hirsch, "Automatic detection of motion artifacts in MR images using CNNs," *Proc ICASSP*, pp. 811–815, 2017.
- [54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, pp. 770–778, 2016.
- [55] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, pp. 4700–4708, 2017.
- [56] D. M. Olsson and L. S. Nelson, "The nelder-mead simplex procedure for function minimization," *Technometrics*, vol. 17, no. 1, pp. 45–51, 1975.
- [57] M. Defrise and R. Clack, "A cone-beam reconstruction algorithm using shift-variant filtering and cone-beam backprojection," *IEEE Trans Med Imaging*, vol. 13, no. 1, pp. 186–195, 1994.
- [58] J. A. Fessler, "On NUFFT-based gridding for non-cartesian MRI," *J Magn Reson*, vol. 188, no. 2, pp. 191–195, 2007.