Expert Systems With Applications 213 (2023) 118866

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Fast and accurate pose estimation of additive manufactured objects from few X-ray projections

Alice Presenti [a,*], Zhihua Liang [a], Luis Filipe Alves Pereira [a,b], Jan Sijbers [a], Jan De Beenhouwer [a]

[a] imec-VisionLab, Dept. of Physics, University of Antwerp, Belgium
[b] Universidade Federal do Agreste de Pernambuco, Garanhuns, Brazil

ARTICLE INFO

Keywords:
Pose estimation
X-ray CT
Inspection
CNN
Radiography

ABSTRACT

X-ray Computed Tomography (CT) is a commonly used imaging technique for non-destructive inspection of manufactured objects. However, a full CT scan requires a long acquisition time, making this method unsuitable for inline applications. In contrast to X-ray CT, inspection can be performed directly in the projection space, using simulated X-ray projections of a reference model of the manufactured object. However, to effectively compare simulated and measured projections, an accurate 3D pose estimation of the object and consequent alignment between the measured object and the reference model are crucial. In this paper, we present a fast method to estimate the 3D pose of a measured object based on convolutional neural networks (CNNs). Through experiments on synthetic and measured data, we demonstrate that our method allows estimating the 3D pose of the object with sub-pixel accuracy. Even if very few projections are available, our approach is comparable to CT-based methods for registration, and outperforms state-of-the-art deep learning methods for radiograph-based pose estimation.

1. Introduction

X-ray CT is a non-destructive procedure widely used for quality control of additively manufactured objects. CT-based inspection involves a multi-step procedure that consists of the acquisition of projections from which a voxelized model is reconstructed, segmented and finally compared to a mesh derived from a computer-aided-design (CAD) model, after registration (Kruth et al., 2011). For effective inspection, aligning the measured object with its reference model is fundamental. It requires high-quality reconstructions, for which thousand of projections are usually needed, with a comparatively long acquisition time (Withers et al., 2021). Moreover, to prevent artifacts in the reconstructed volume, an equiangular acquisition covering at least 180° is needed. In addition, the volumetric reconstruction quality is affected by the presence of noise and artifacts (Rodríguez-Sánchez, Thompson, Körner, Brierley, & Leach, 2020), which can influence registration and, consequently, the entire quality control process. As a result, the CT procedure is not suitable for real-time inspection and applications where full rotation around the object is not feasible or cost-inefficient.

Recently, it has been shown that inspection can be performed directly in the projection space, thus avoiding the 3D reconstruction. For example, to secure the borders against smuggling, Abdolshah, Teimouri, and Rahmani (2017) classified the content of shipping containers by extracting Scale Invariant Feature Transforms (SIFT) feature vectors from its X-ray radiographs. Dong, Taylor, and Cootes (2018) investigated weld defects on aircraft components from radiographs by first extracting weld lines and then identifying defective regions. Czyzewski, Krawiec, Brzezinski, Porebski, and Minor (2021) automatically classified X-ray diffraction images with a CNN to detect seven types of anomalies in crystals.

Radiograph-based inspection, however, often lacks volumetric information. To enable 3D radiographic inspection, the measured projections can be compared to those simulated from the reference CAD model. For example, van Dael, Verboven, Zanella, Sijbers, and Nicolai (2019) built a statistical shape and non-uniform density model of apples from CT data of intact fruits. From this model, radiographs were simulated for comparison with the measured projection data. To perform 3D inspection from 2D projections, Evangelista et al. (2020) mapped X-ray projections to the CAD model of the object to directly locate the defects on the part itself.

Independent of the inspection scheme used, an accurate 3D pose estimation of the object and consequent alignment between the measured object and the reference mesh are crucial. So far, few studies have

* Corresponding author.
E-mail addresses: alice.presenti@uantwerpen.be (A. Presenti), zhihua.liang@uantwerpen.be (Z. Liang), luisfilipe.alvespereira@uantwerpen.be (L.F.A. Pereira), jan.sijbers@uantwerpen.be (J. Sijbers), jan.debeenhouwer@uantwerpen.be (J. De Beenhouwer).

https://doi.org/10.1016/j.eswa.2022.118866
Received 11 April 2022; Received in revised form 6 September 2022; Accepted 16 September 2022
Available online 22 September 2022

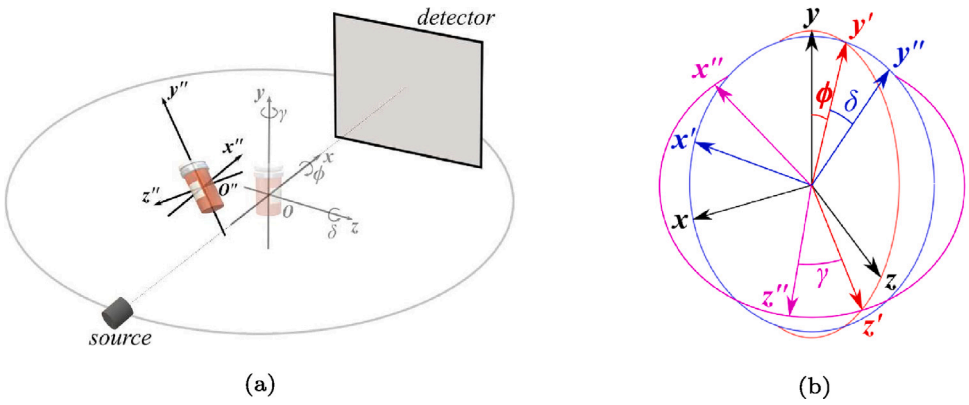


Fig. 1. The system's geometry (a) and the Euler angles γ, δ and ϕ (b).

been conducted to estimate the object pose from its radiographs. Kügler, Stefanov, and Mukhopadhyay (2018) recovered the 3D pose of screws on a skull bone by extracting pseudo landmark positions with a CNN from normalized regions of interest (ROIs). Miao, Wang, and Liao (2016) estimated the 3D pose of an object hierarchically from difference images between simulated and measured radiographs on specific ROIs. Similarly, Bui, Albarqouni, Schrapp, Navab, and Ilic (2017) optimized the pose of the object combined with a distance between the X-ray projections produced with the estimated pose and the measured X-ray images.

In our previous work (Presenti, Sijbers, & De Beenhouwer, 2021), an analytical framework for projection-based inspection was proposed. Based on simulated projections from the CAD model of the object, a few task-specific preferred orientations were selected and dynamically acquired during the inspection process. To do so, the 3D pose of the object inside the imaging system was analytically estimated from a few projections by exploiting the acquisition geometry information.

Neural networks are a powerful instrument widely adopted for 3D object pose estimation. Many networks have been designed for this purpose, and applied to the most disparate applications. For example, Peng, Liu, Huang, Zhou, and Bao (2019) extract a vector field that represents the object keypoint locations that are subsequently given to a PnP solver to predict the final pose. Kehl, Manhardt, Tombari, Ilic, and Navab (2017) use an SSD-based method for object detection and pose estimation from RGB images. Bukschat and Vetter (2020) extend the EfficientDet network, created for object detection, to 3D pose estimation.

In X-ray inspection, it is common to demand 100 μm of accuracy and down to 1 μm for metrology tasks (Tan, Kiekens, Kruth, Voet, & Dewulf, 2011). This is far from the accuracy of the state-of-the-art networks for pose estimation from images, which reaches a few degrees and millimeters or centimeters (Brynte & Kahl, 2020; Giefer, Castellanos, Babr, & Freitag, 2019; Kendall, Grimes, & Cipolla, 2016). To the author's knowledge, so far very few works have been proposed for object pose estimation from X-ray radiographs. Davison et al. (2018) develop a voting-based scheme to identify landmark locations from medical images. Kügler et al. (2018) estimate the pose of screws placed close to the skull temporal bone. Their i3PosNet outputs pseudo-landmarks from which the pose is reconstructed by geometric considerations. The PoseNet architecture is utilized by Bui, Albarqouni, Schrapp, Navab, and Ilic (2017) to estimate the 3D pose of an object from its X-ray projection.

In this work, we present a framework, named *RDpose* (Radiographic Deep Pose), to estimate the 3D pose of an object from its measured projections, based on CNNs. In the RDpose framework, the ResNet-50-V2 network (He, Zhang, S., & J., 2016) pre-trained on the ImageNet dataset (Deng et al., 2009) is fine-tuned to regress the 3D pose of an object from its radiographs. To refine the estimation, the network trained with one projection input is used as a feature extractor on a multi-input, sharing weights network. In this way, the pose of the object can be iteratively improved while acquiring new projections. RDpose is expected to enable inline 3D inspection of objects from their radiographs after real-time accurate alignment between the measured object and its reference model.

2. Methods

In this section, we first describe the system geometry and the parameters that define the object pose with respect to the acquisition system. Next, the tool for creating realistic synthetic polychromatic data is introduced, as well as the source focal spot model and noise simulation. Finally, the architecture of RDpose framework is described.

2.1. The system geometry and problem formulation

Let $S = \{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ be a reference system defined with respect to the initial source and detector positions, with the $\mathbf{y}$ and $\mathbf{z}$ axis parallel to the detector plane, and the $\mathbf{x}$ axis along the source–detector direction. The center $\mathbf{O} = \{0, 0, 0\}$ of S coincides with the rotation center of the acquisition system and with the barycentre of the mesh model (see Fig. 1(a)). In an ideal setting, the object is perfectly aligned with the reference model. However, when a physical object is placed inside a real acquisition system, its position and orientation are unknown with respect to the reference coordinate system. Let $\mathbf{R_u}(\theta)$ be the rotation matrix describing the rotation by an angle θ around the axis $\mathbf{u} = (u_x, u_y, u_z)$. The orientation of the object is defined by the rotation angles ϕ, δ and γ around $\mathbf{x}, \mathbf{z}'$ and $\mathbf{y}''$, respectively, with $\mathbf{z}' = \mathbf{R}_x(\phi)\mathbf{z}$ and $\mathbf{y}'' = \mathbf{R}_{z'}(\delta)\mathbf{R}_x(\phi)\mathbf{y}$ (see Fig. 1(b)). These are the so-called intrinsic Euler rotations, defined with respect to axes updated after each rotation. The barycentre of the object is assumed to be translated by $\mathbf{t} = \{t_x, t_y, t_z\}$ with respect to the center of the reference system $\mathbf{O}$.

In industrial inline inspection, some prior knowledge about the pose of the object to be inspected is often provided, such that constraints can be imposed to its pose. In our paper, we assume the rotation around the vertical axis to be unconstrained, $\gamma \in [0, 360)°$. The other parameters are assumed to vary within narrower intervals: $\delta \in [\delta_{min}, \delta_{max}]$, $\phi \in [\phi_{min}, \phi_{max}]$, $t_x \in [t_{x_{min}}, t_{x_{max}}]$, $t_y \in [t_{y_{min}}, t_{y_{max}}]$ and $t_z \in [t_{z_{min}}, t_{z_{max}}]$.

2.2. Synthetic projection model

A frequent problem with neural networks is the demand for a large number of data points to achieve a balance between high accuracy and generalization. Training a pose regression network requires acquiring thousands of projections of the object at different orientations, which is highly impractical. Instead, our model was trained only on realistically simulated X-ray projections that account for the projection geometry, the source focal spot size and spectrum and the noise that can be expected in real projection data. Realistically simulated projections

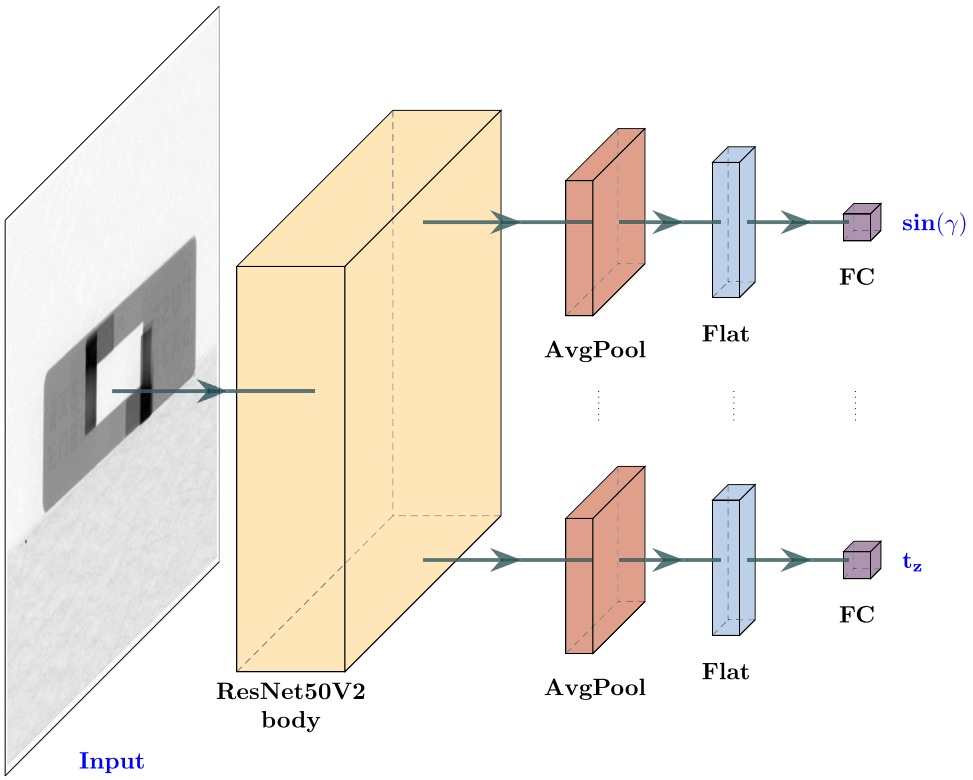


Fig. 2. The architecture of the network with one projection as input. For each output, features are extracted from the input image by the ResNet-50-V2 convolutional body and subsequently averaged and flattened. Next, a fully connected layer returns the output. Only two of the seven outputs are shown.

from a CAD model were created with a mesh projector (Marinovszki, De Beenhouwer, & Sijbers, 2018), which uses ray-tracing and models the material with energy-dependent attenuation values. To generate the synthetic projection training dataset, the system geometry parameters such as the source-object distance (SOD), the source–detector distance (SDD), the detector size and pixel dimension, the source power, the number of averages among the flat fields and the exposure time must be established. To simulate the X-ray data, similar acquisition parameters were used as for the measured data with which the network was tested. To ensure the pixel intensity of the simulated projections resembles that of the measured ones, the source spectrum must be accurately estimated. In this work, the energy spectrum of the X-ray source was calculated using Monte Carlo simulations (Nazemi et al., 2021). When simulating projections, often a point source is considered. However, the X-rays are emitted from an approximately Gaussian shaped focal spot. The finite focal spot causes the formation of penumbra at the borders of the imaged object in the detector. We relied on manufacturer supplied data for the vertical and horizontal focal spot size, as a function of the target power. Simulated projections were created with a point source positioned at the horizontal and vertical approximation of the focal spot shape, and then averaged with the weights of the corresponding Gaussian probability distributions. Finally, Poisson noise was added to the synthetic radiographs and to the synthetic flat fields. The beam intensity I_0 of the Poisson noise was estimated heuristically so as to obtain a similar SNR as the measured data.

2.3. The network architecture

The aim of the RDpose framework is to estimate the 3D pose of an object starting from a single projection image. Depending on the required accuracy for a specific application, the pose estimation can be iteratively improved while acquiring additional projections. The core of the framework is formed by a ResNet-50-V2 network, which was used here as a feature extractor. The RDpose framework consists of multiple networks, each making use of the feature extractor, for which the input and output layers were adapted to learn from an increasing number of input images.

2.3.1. Single-input network

The single-input network is expected to estimate the 3D pose of the measured object from a single projection image. A sketch of the network architecture that accepts as input the first acquired projection and outputs the 3D pose parameters of the object is shown in Fig. 2. ResNet-50-V2, implemented in Keras (high-level API built on Tensorflow) (Abadi et al., 2015; He et al., 2016), and excluding its classification head, was used as a feature extractor. For every pose parameter, a fully connected network was attached to the feature extractor composed of an average pooling layer with pooling size 3, a flattening layer and, finally, a fully connected layer to the output (see Fig. 2). The feature extractor parameter weights were initialized to those obtained from the ImageNet dataset. During training, the network is optimized with our X-ray dataset (see Section 3.2 for details).

As explained in Section 2.1, the angle γ is within the interval $[0, 360)°$. Mapping this parameter into $[0, 1]$ or $[-1, 1]$, codomains of common activation functions, would heavily penalize a predicted value close to 360° when the true value is around 0°. To circumvent this problem, we opted for estimating the sine and cosine of γ, parameters that together uniquely define the rotation angle and are by definition in $[-1, 1]$. The network thus returns seven separate outputs: the sine and cosine of γ, the angles δ and ϕ and the translations t_x, t_y and t_z. As activation functions of the final layers, the hyperbolic tangent function was used for the sine and cosine of γ and the sigmoid function for the other parameters.

2.3.2. Multi-input network

To improve the estimate obtained from a single projection, multiple projections acquired at different angular views can be fed to a multi-input network. A fixed angular distance between the consecutive input

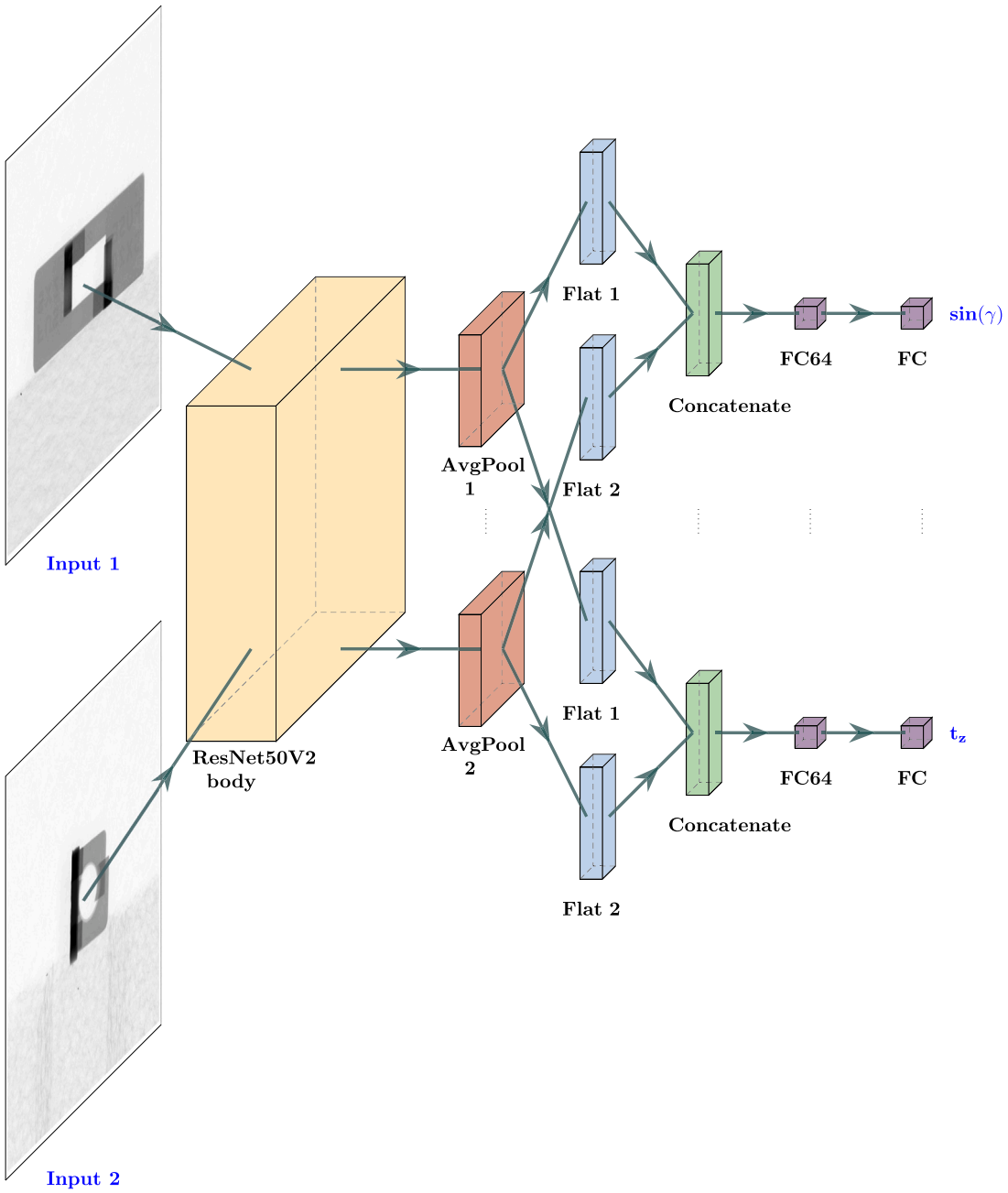


Fig. 3. The architecture of the network with two projections as input. For each output, features are extracted from the input images by the ResNet-50-V2 convolutional body and subsequently averaged, flattened and concatenated. Next, two fully connected layers return the output. Only two of the seven outputs are shown.

images was chosen, so that the network could learn the angular relation between the projection images. In the RDpose framework, the single-input network was extended to multiple projections by concatenating the features extracted from each input image. As an example, the summary scheme of the dual-input network is shown in Fig. 3. The feature extractor structure is the same as the one used for the single-input network. Its parameter weights were initialized to those obtained after the single-input network training, and further optimized (see Section 3.2 for details). The feature extractor was therefore applied to the different inputs by sharing weights. Then, for every output, the features were averaged with a pooling size 3, flattened and concatenated. Concatenation is followed by a fully connected layer with 64 nodes and, finally, a fully connected layer with the sigmoid or tangent hyperbolic activation function.

3. Experiments

Our RDpose framework was tested on both simulated and realistic data to evaluate the accuracy and precision of pose estimation. Moreover, our method was compared to conventional CT-based registration and to the state-of-the-art radiograph-based registration (Kendall et al., 2016). In our experiments, we tested the RDpose framework for one, two and three input images. Details about the generation of training and test data are given in Section 3.1. Subsequently, the hyperparameters chosen for the training of the networks are reported in Section 3.2.

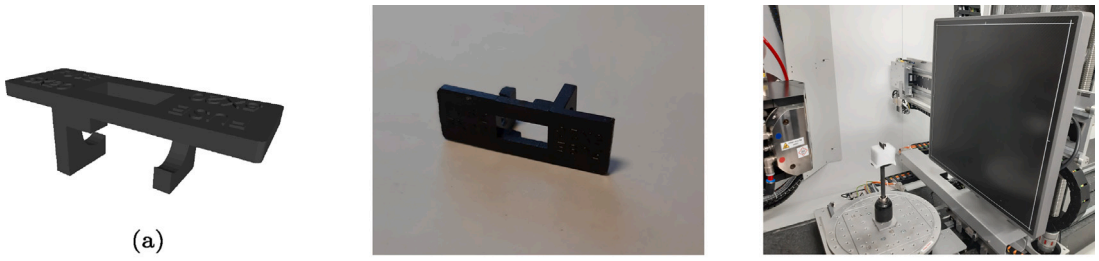


Fig. 4. A render of the fuse cover used in our experiments (a), and pictures of the real object (b) and the object inside the X-ray system (c).

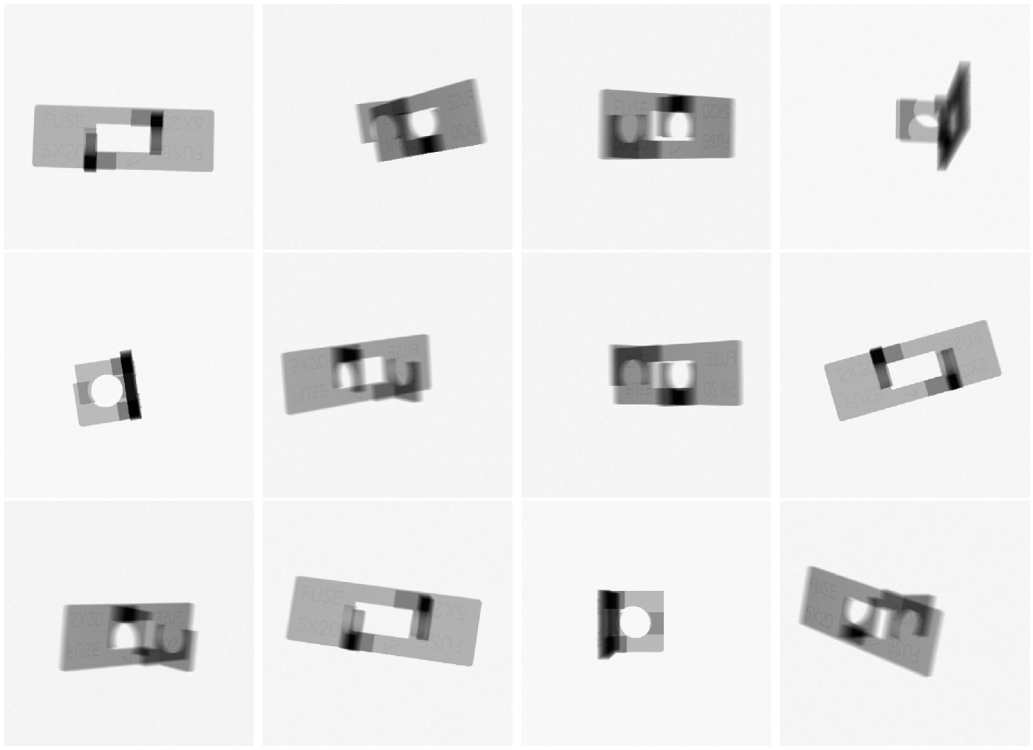


Fig. 5. Couples of simulated projections in our datasets. In the first row, the images corresponding to an acquisition angle of 0°, and in the second and third rows, the images after a 90° and 225° rotation, respectively.

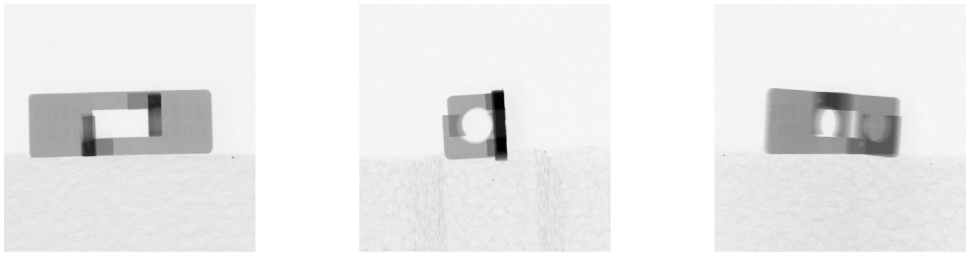
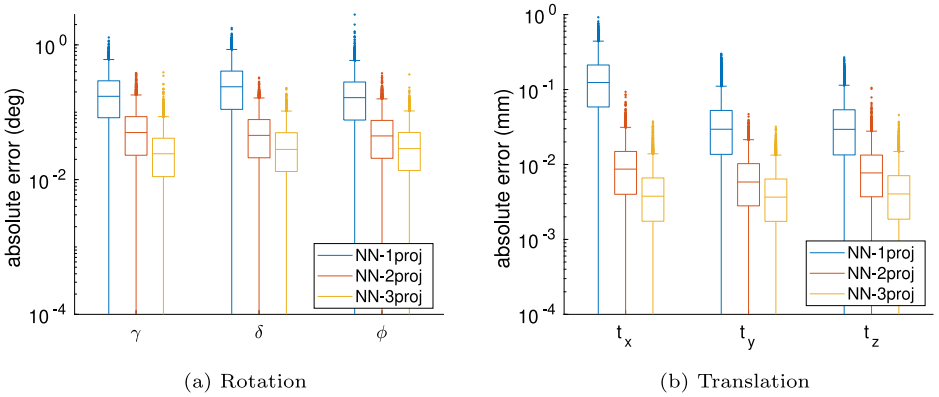


Fig. 6. Measured projections of the fuse cover acquired at (a) 0°, (b) 90° and (c) 225°.

Finally, in Section 3.3, the experiments conducted to validate RDpose are discussed.

3.1. Training and test data

The RDpose framework was evaluated with both simulated and real data of a fuse cover made of glass fiber reinforced nylon (see Fig. 4). To train the single-input network, 288 000 projections were simulated from the mesh model of the fuse cover with the system parameters as shown in the first row of Table 1. In order to reduce the memory footprint during training, the pixel size of the simulated images was chosen four times larger in each dimension than for the measured (real) data (see the second row of Table 1). The object pose was perturbed by randomly varying $\gamma \in [0, 360)°$, $\delta, \phi \in [-16, 16]°$, $t_x, t_y, t_z \in [-4, 4]$ mm. These parameter ranges were selected from the analysis of the measured data. For the multi-input networks, a second



(a) Rotation (b) Translation

Fig. 7. Absolute difference between the ground truth and the estimated rotation (a) and translation (b) for the networks with one, two and three input images.

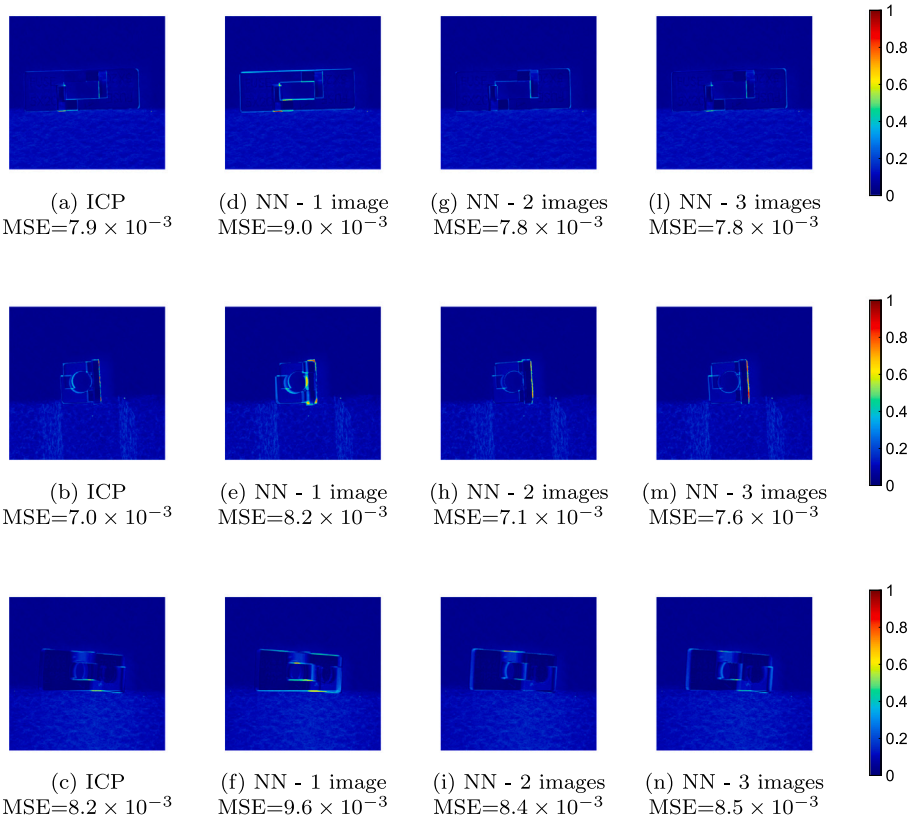


Fig. 8. Projection difference between the measured ground truth radiograph of the object acquired at position 1 and the simulated radiographs with pose parameters as estimated with (a)–(c) ICP, (d)–(f) the NN with one image as input, (g)–(i) the NN with two images as input and (l)–(n) the NN with three images as input. In the first row, it is shown the acquisition at 0°, while in the second and third rows the ones at 90° and 225°, respectively.

and a third dataset, each composed of 288 000 simulated projections, were simulated by rotating the reference mesh by first 90° and then 225° around the center of the reference system, while maintaining the object and acquisition system parameters identical to those of the first acquisition. Poisson distributed projections were simulated using a beam intensity of $I_0 \in [32\,000, 34\,000]$. The beam intensity range was estimated heuristically so as to obtain a similar SNR to the measured data. A few simulated projection images of all three datasets are shown in Fig. 5.

To evaluate the accuracy of the method, 7 200 simulated projections were generated, with the system parameters and beam intensity range of I_0 identical to those of the training data. The object pose was also randomly perturbed within the same intervals as for the training data. For each of the projections in the dataset, an acquisition at angular distance of 90° and one at 225° were simulated as the second and third input for the multi-input networks, respectively.

3.2. The network hyper-parameters

In our experiments, the weighted mean absolute error (MAE) between labels and predictions was used as a loss function with weights $w = \{0.35, 0.35, 0.05, 0.05, 0.1, 0.05, 0.05\}$ for the $\sin\gamma$, $\cos\gamma$, δ, ϕ, t_x, t_y and t_z outputs, respectively. The sine and cosine of γ and the translation t_x received higher weights being out of plane parameters (see Fig. 1(a)), which are more difficult to regress from a single projection. Both the single-input and the multi-input networks were fine-tuned on our datasets by training only the newly introduced layers for the first 20 epochs, while freezing the feature extractor. Subsequently, all the

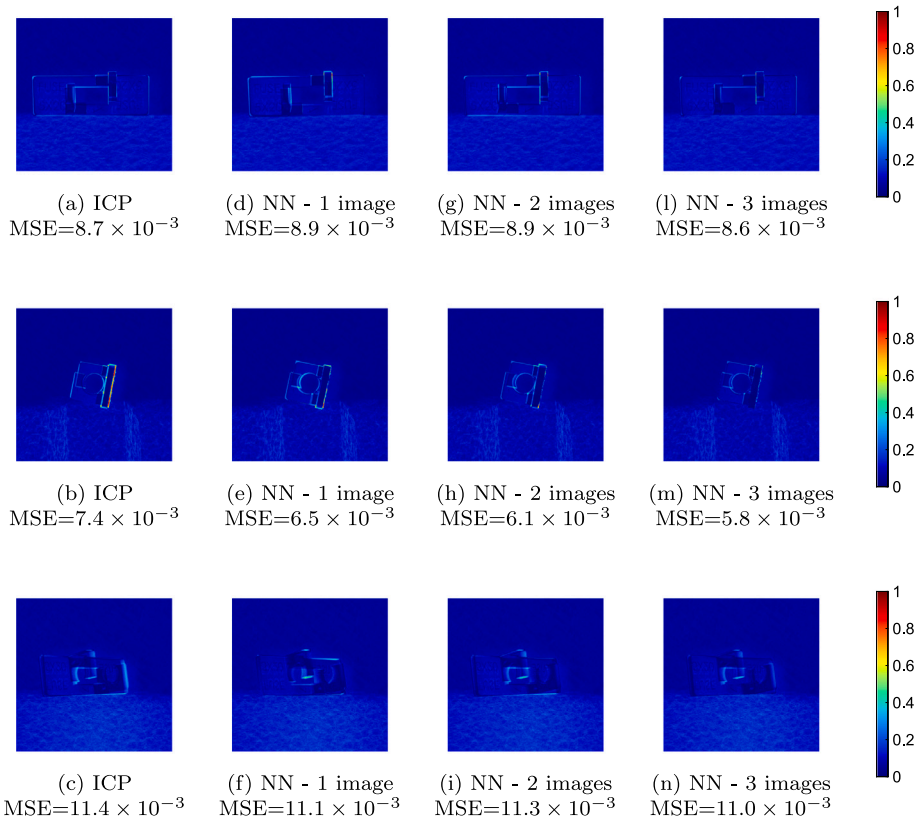


Fig. 9. Projection difference between the measured ground truth radiograph of the object acquired at position 2 and the simulated radiographs with pose parameters as estimated with (a)–(c) ICP, (d)–(f) the NN with one image as input, (g)–(i) the NN with two images as input and (l)–(n) the NN with three images as input. In the first row, it is shown the acquisition at 0°, while in the second and third rows the ones at 90° and 225°, respectively.

Table 1
The system geometry used in our experiments.

Object	SOD	SDD	Proj size	Pixel size	Resolution
Fuse cover (simul.)	108.340 mm	650 mm	354 × 360	0.600 mm	0.100 mm
Fuse cover (real)	108.340 mm	650 mm	1416 × 1440	0.150 mm	0.025 mm

Table 2
The 95% confidence interval of the pose parameters estimated by the three-inputs network on simulated and measured images.

Parameter	Simulated	Measured
γ	$(-1.688 \pm 0.017)°$	$(-1.876 \pm 0.012)°$
δ	$(-3.236 \pm 0.005)°$	$(-3.417 \pm 0.012)°$
ϕ	$(-1.287 \pm 0.005)°$	$(-1.407 \pm 0.009)°$
t_x	(1248 ± 2) µm	(1224 ± 5) µm
t_y	(461 ± 0) µm	(423 ± 1) µm
t_z	(-1206 ± 1) µm	(-1218 ± 2) µm

networks layers were trained for an additional 80 epochs. The learning rate was initialized to 10^{-4} and reduced to 10^{-5} and 0.5×10^{-6} after 20 and 40 epochs, respectively.

3.3. Comparison to state-of-the-art

To evaluate the precision of RDpose, a scan of the real object composed of 2159 projections equiangularly distributed over 360° was acquired. To limit the size of the network, the high resolution projections, acquired with the geometry settings as in the second row of Table 1, were downsampled four times in each dimension with a bicubic interpolation. After this operation, the image size corresponded to the one of the simulated data, which had a virtual detector pixel size that was four times larger than the pixel size of the real detector. In addition, projections of the mesh model at a perturbed pose were simulated at the same angles as for the real data acquisition. Among the acquired projections, couples of projections at angular distance of 90° and 255° were used as input for the three-inputs network. For each prediction, the estimated local pose was transformed to the global coordinate system (where the x axis coincides with the source–detector direction at the first acquisition). The accuracy of the angular distance between two consecutive acquisitions is system dependent. The manufacturer of the acquisition system used in our experiments specified that the axial and radial run-out at 15 cm is lower than 1 µm for the rotation axis. Therefore, we can assume that the error on the angular distance between two consecutive acquisition is < 1 µm.

To compare RDpose to a conventional registration approach, three other full CT scans of 2159 projections equiangularly spanning 360° were acquired by placing the real object in different poses. An example of measured projections of the object at 0°, 90° and 225° is shown in Fig. 6. Then, 2D slices were reconstructed with the Feldkamp, Davis and Kress (FDK) algorithm (Feldkamp, Davis, & Kress, 1984). The resulting 3D volume was then manually thresholded and an STL model was extracted by using the VTK library (Schroeder, Martin, & Lorensen, 2006). Finally, the reference mesh was aligned to the reconstructed one with an iterative closest point (ICP) algorithm. Among the acquired radiographs, only those at 0°, 90° and 225° rotation, and downsampled four times, were used as input of our networks.

We compared our results to PoseNet (Kendall et al., 2016), a network based on GoogLeNet (Szegedy et al., 2014). In this architecture,

Table 3
The MSE between the vertices of the reconstructed mesh and the nominal mesh oriented as estimated by ICP and RDpose.

Experiment	ICP	NN - 1 image	NN - 2 images	NN - 3 images
position 1	**0.13**	0.30	0.19	0.24
position 2	0.30	**0.14**	0.17	0.16
position 3	0.22	0.52	0.32	**0.17**
position 4	**0.14**	0.21	0.19	0.20

Table 4
The mean MSE between the measured projections and the simulated ones ± the 95% CI. All the values have to be multiplied by a factor 10^{-3}.

Experiment	ICP	NN - 1 image	NN - 2 images	NN - 3 images
position 1	$\mathbf{6.0 \pm 0.041}$	7.6 ± 0.045	6.1 ± 0.038	6.1 ± 0.037
position 2	6.1 ± 0.033	5.9 ± 0.037	6.0 ± 0.038	$\mathbf{5.8 \pm 0.037}$
position 3	7.0 ± 0.049	9.5 ± 0.053	7.5 ± 0.046	$\mathbf{6.9 \pm 0.050}$
position 4	$\mathbf{5.2 \pm 0.032}$	5.6 ± 0.028	5.5 ± 0.035	$\mathbf{5.2 \pm 0.033}$

Table 5
The performance of RDpose compared to PoseNet.

	PoseNet	RDpose 1 input	RDpose 2 inputs	RDpose 3 inputs
$\mathbf{t}_{MAE}$	98 μm	77 μm	9 μm	**5 μm**
$\mathbf{q}_{MSE}$	6×10^{-4}	7×10^{-4}	6×10^{-4}	$\mathbf{1 \times 10^{-7}}$

inception modules were repeated, and auxiliary outputs connected to intermediate layers were introduced to improve the network performance. During training, the auxiliary outputs loss was included in the network loss function. Auxiliary outputs were not used in our implementation of PoseNet, and the translation parameters and the quaternion defining the rotation were directly regressed. PoseNet was trained for 100 epochs, with a learning rate of 10^{-3}, and the MSE as a loss function. As Bui et al. (2017), the performance of the networks was validated in terms of the MSE between the estimated quaternions and the ground truth ones ($\mathbf{q}_{MSE}$) and the MAE between the estimated and the ground truth translation vectors ($\mathbf{t}_{MAE}$). Finally, we compared our network to EPro-PNP by Chen et al. (2022), a state-of-the-art probabilistic PnP method for end-to-end pose estimation. The network was trained on 256×256 downscaled input images and returned a 3×3 rotation matrix and a 3×1 translation vector.

4. Results and discussion

The absolute difference between the ground truth rotation and translation of the object and the pose estimated from simulated test data is shown in Figs. 7(a) and 7(b), respectively, for the network with one, two and three projections as input. The mean errors obtained by the single-input network are $\overline{\gamma} = 0.21°$, $\overline{\delta} = 0.26°$, $\overline{\phi} = 0.20°$, $\overline{t_x} = 0.15$ mm, $\overline{t_y} = 0.04$ mm and $\overline{t_z} = 0.04$ mm. For the two-inputs network, the accuracy increased, with mean errors of $\overline{\gamma} = 0.06°$, $\overline{\delta} = 0.05°$, $\overline{\phi} = 0.05°$, $\overline{t_x} = 11$ μm, $\overline{t_y} = 7$ μm and $\overline{t_z} = 10$ μm. Finally, the three-inputs network showed a further increase in accuracy, with mean errors of $\overline{\gamma} = 0.03°$, $\overline{\delta} = 0.03°$, $\overline{\phi} = 0.04°$, $\overline{t_x} = 5$ μm, $\overline{t_y} = 5$ μm and $\overline{t_z} = 5$ μm. The precision of the three-inputs network is shown in Table 2. Here, the 95% confidence intervals (CI) of the parameters are reported for both simulated and measured data. From this experiment, we can conclude that RDpose showed a high precision, i.e. the estimates over multiple measurements returned values in a narrow range.

The parameters sine and cosine of γ are not independent, but are related by the orthogonality rule. The orthogonality of our predicted parameters was evaluated by calculating $(\sin\gamma)^2 + (\cos\gamma)^2$, obtaining the value of $(1.0015 \pm 0.0053)°$ for the one input network. This shows that the designed network learns the relation between the two parameters from the observation of the labels. Therefore, explicit modeling of the orthogonality loss is not required.

Measured projections were compared to simulated ones with the reference mesh oriented and positioned as estimated by our networks and by the ICP alignment. Figs. 8 and 9 show the difference images for the two positions underlined in Table 3. The 2159 projections of the 4 experiments were compared to simulated ones after alignment. For memory reasons, this comparison was performed on four times down-sampled projections. The mean MSE over the projections and the 95% CI are reported on Table 4. In Table 3, the MSE between the vertices of the reconstructed mesh from the measured data and those of the reference mesh oriented as estimated with ICP and with the neural networks is reported for the four experiments. Color map representations of the L^1-norm between the vertices of the meshes after registration are shown in Fig. 10, for the two positions underlined in Table 3.

From these results, we might conclude that RDpose framework and ICP are comparable. By observing Fig. 10, one may notice a discrepancy between the colormap and the MSE. The first thing to bear in mind when looking at the colormap images is that the reference mesh of the analyzed object did not correspond exactly to the measured object. Secondly, the two registration methods seemed to concentrate on two distinguished parts of the object. Moreover, and maybe more significant, the vertices in the CAD model were not equidistributed. As the colormap was created with a color gradient between two vertices of a triangle, the area of the triangle impacts our perception of the registration quality. In other words, if a vertex with a high error belongs to a triangle with a large surface, this will visually result in a larger red area than for a smaller triangle. Similarly, also the MSE measure might not be an appropriate indicator for registration accuracy. Indeed, if many vertices are concentrated on an area where the distance to the extracted mesh for the object is high, this will have a high negative impact on the MSE measure. However, despite the difficulty of interpreting and comparing these results, two main arguments point in our favor: first, the ICP registration was achieved from a full CT acquisition, while our results were based on a maximum of three 2D projections; second, the ICP algorithm was applied to a mesh extracted from high-resolution projections, in contrast to our method where, for computational reasons, the registration was performed from four times down-sampled projections.

Results of the comparison between RDpose and the PoseNet network (Kendall et al., 2016), are reported in Table 5. Our approach outperformed PoseNet both in terms of the t_{MAE} and the q_{MSE}. Moreover, the advantage of using multiple projections is clearly shown, allowing for a more accurate estimation of the object pose. As far as we know, the use of more than one projection as input is an innovative approach for deep learning radiograph-based 3D registration.

Table 6 shows the comparison between RDPose and EPro-PnP networks. From this result, the advantage of a multi-input network is evident compared to a single input network which is the state-of-the-art on pose estimation. When comparing the two architectures on a

Table 6
The mean absolute error of RDpose and EPro-PnP.

	γ	δ	φ	t_x	t_y	t_z
EPro-PnP	0.21°	0.06°	0.06°	30 µm	**4 µm**	**4 µm**
RDPose, 1 input	0.21°	0.26°	0.20°	150 µm	40 µm	40 µm
RDPose, 2 inputs	0.06°	0.05°	0.05°	11 µm	7 µm	10 µm
RDPose, 3 inputs	**0.03°**	**0.03°**	**0.04°**	**5 µm**	5 µm	5 µm

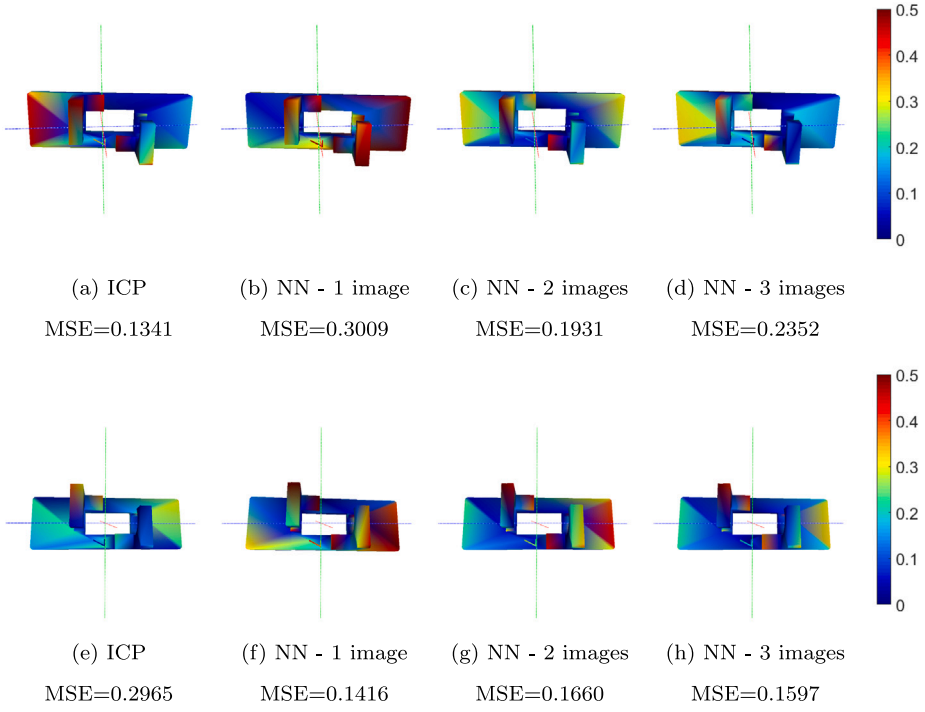


Fig. 10. L^1-norm between the reconstructed mesh and the nominal one oriented as estimated with ICP ((a) and (e)), the NN with one image as input ((b) and (f)), the NN with two images as input ((c) and (g)) and the NN with three images as input ((d) and (h)). The images in the first row refers to position 1, while those in the second row refers to position 2.

single input, instead, our method has a worse performance. However, as previously mentioned, the innovative idea of the RDpose approach is the iterative refinement of the object pose, rather than the use of a specific network architecture.

5. Conclusion

In this paper, we presented RDpose, a CNN-based method to estimate the 3D pose of an object from its X-ray projections. By using a pre-trained ResNet50-V2 network, the pose was regressed from a single projection image and iteratively refined when more projections were acquired. With the RDpose framework, real-time, few-view, and limited angular acquisition X-ray inspection becomes feasible. By skipping the 3D reconstruction, indeed, very few X-ray projections are sufficient for an accurate registration. Our methodology was validated with a test case glass fiber reinforced object for both simulated and measured data. In our experiments, the networks were trained solely with realistic data simulated from the object reference model. Compared to conventional 3D volumetric registration, RDpose achieved a similar accuracy on pose estimation without the need for a CT reconstruction. Moreover, our multi-input networks outperformed the state-of-the-art CNN-based methods for 3D pose estimation from radiographs, taking advantage of the 3D information obtained from different orientations.

CRediT authorship contribution statement

Alice Presenti: Methodology, Software, Validation, Data curation, Writing – original draft, Writing – review & editing. **Zhihua Liang:** Methodology, Software, Writing – review & editing. **Luis Filipe Alves Pereira:** Methodology, Software, Writing – review & editing. **Jan Sijbers:** Methodology, Writing – review & editing, Funding acquisition, Project administration. **Jan De Beenhouwer:** Methodology, Writing – review & editing, Funding acquisition, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This research is funded by the FWO, Belgium SBO project MetroFlex (S004217N), the FWO SBO project (S003421N), the European Commission through the INTERREG Vlaanderen Nederland program project Smart*Light (0386), and the Flemish Government under the “Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen” programme.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. URL: https://www.tensorflow.org/, Software available from tensorflow.org.

Abdolshah, M., Teimouri, M., & Rahmani, R. (2017). Classification of X-Ray images of shipping containers. *Expert Systems with Applications*, *77*, 57–65. http://dx.doi.org/10.1016/j.eswa.2017.01.030, URL: https://www.sciencedirect.com/science/article/pii/S0957417417300362.

Brynte, L., & Kahl, F. (2020). Pose proposal critic: Robust pose refinement by learning reprojection errors. CoRR abs/2005.06262, URL: https://arxiv.org/abs/2005.06262.

Bui, M., Albarqouni, S., Schrapp, M., Navab, N., & Ilic, S. (2017). X-Ray PoseNet: 6 dof pose estimation for mobile X-Ray devices. In *2017 IEEE winter conference on applications of computer vision* (pp. 1036–1044). http://dx.doi.org/10.1109/WACV.2017.120.

Bui, M., Albarqouni, S., Schrapp, M., Navab, N., & Ilic, S. (2017). X-Ray PoseNet: 6 DoF pose estimation for mobile X-Ray devices. In *2017 IEEE winter conference on applications of computer vision* (pp. 1036–1044). http://dx.doi.org/10.1109/WACV.2017.120.

Bukschat, Y., & Vetter, M. (2020). EfficientPose: An efficient, accurate and scalable end-to-end 6D multi object pose estimation approach. arXiv preprint arXiv:2011.04307.

Chen, H., Wang, P., Wang, F., Tian, W., Xiong, L., & Li, H. (2022). EPro-PnP: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. http://dx.doi.org/10.48550/ARXIV.2203.13254, arXiv, URL: https://arxiv.org/abs/2203.13254.

Czyzewski, A., Krawiec, F., Brzezinski, D., Porebski, P. J., & Minor, W. (2021). Detecting anomalies in X-ray diffraction images using convolutional neural networks. *Expert Systems with Applications*, *174*, Article 114740. http://dx.doi.org/10.1016/j.eswa.2021.114740, URL: https://www.sciencedirect.com/science/article/pii/S0957417421001810.

Davison, A. K., Lindner, C., Perry, D. C., Luo, W., Cootes, T. F., et al. (2018). Landmark localisation in radiographs using weighted heatmap displacement voting. In *International workshop on computational methods and clinical applications in musculoskeletal imaging* (pp. 73–85). Springer.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *CVPR09*.

Dong, X., Taylor, C., & Cootes, T. (2018). Automatic inspection of aerospace welds using X-Ray images. In *2018 24th international conference on pattern recognition*. http://dx.doi.org/10.1109/ICPR.2018.8545738.

Evangelista, D., Terreran, M., Pretto, A., Moro, M., Ferrari, C., & Menegatti, E. (2020). 3D mapping of X-Ray images in inspections of aerospace parts. In *2020 25th IEEE international conference on emerging technologies and factory automation, vol. 1* (pp. 1223–1226). http://dx.doi.org/10.1109/ETFA46521.2020.9212135.

Feldkamp, L. A., Davis, L. C., & Kress, J. W. (1984). Practical cone-beam algorithm. *Journal of the Optical Society of America A*, *1*(6), 612–619. http://dx.doi.org/10.1364/JOSAA.1.000612, URL: http://www.osapublishing.org/josaa/abstract.cfm?URI=josaa-1-6-612.

Giefer, L. A., Castellanos, J. D. A., Babr, M. M., & Freitag, M. (2019). Deep learning-based pose estimation of apples for inspection in logistic centers using single-perspective imaging. *Processes*, *7*, *424*, http://dx.doi.org/10.3390/pr7070424.

He, K., Zhang, X., S., R., & J., S. (2016). Identity mappings in deep residual networks. CoRR abs/1603.05027, URL: http://arxiv.org/abs/1603.05027.

Kehl, W., Manhardt, F., Tombari, F., Ilic, S., & Navab, N. (2017). Ssd-6D: Making rgb-based 3D detection and 6D pose estimation great again. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1521–1529).

Kendall, A., Grimes, M., & Cipolla, R. (2016). PoseNet: A convolutional network for real-time 6-DOF camera relocalization. arXiv:1505.07427.

Kruth, J. P., Beartscher, M., Carmignato, S., Schmitt, R., Chiffre, L. D., & Weckmann, A. (2011). Computed tomography for dimensional metrology. *CIRP Annals-Manufacturing Technology*, *60*, 821–842. http://dx.doi.org/10.1016/j.cirp.2011.05.006.

Kügler, D., Stefanov, A., & Mukhopadhyay, A. (2018). i3PosNet: Instrument pose estimation from X-ray. CoRR abs/1802.09575, URL: http://arxiv.org/abs/1802.09575.

Marinovszki, A., De Beenhouwer, J., & Sijbers, J. (2018). An efficient CAD projector for X-ray projection based 3D inspection with the ASTRA toolbox. In *8th conference on industrial computed tomography*.

Miao, S., Wang, Z. J., & Liao, R. (2016). A CNN regression approach for real-time 2D/3D registration. *IEEE Transactions on Medical Imaging*, *35*(5), 1352–1363. http://dx.doi.org/10.1109/TMI.2016.2521800.

Nazemi, E., Six, N., Iuso, D., De Samber, B., Sijbers, J., & De Beenhouwer, J. (2021). Monte-Carlo-based estimation of the X-ray energy spectrum for CT artifact reduction. *Applied Sciences*, *11*(7), http://dx.doi.org/10.3390/app11073145, URL: https://www.mdpi.com/2076-3417/11/7/3145.

Peng, S., Liu, Y., Huang, Q., Zhou, X., & Bao, H. (2019). Pvnet: Pixel-wise voting network for 6DoF pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4561–4570).

Presenti, A., Sijbers, J., & De Beenhouwer, J. (2021). Dynamic few-view X-ray imaging for inspection of CAD-based objects. *Expert Systems with Applications*, *180*, Article 115012. http://dx.doi.org/10.1016/j.eswa.2021.115012, URL: https://www.sciencedirect.com/science/article/pii/S095741742100453X.

Rodríguez-Sánchez, Á., Thompson, A., Körner, L., Brierley, N., & Leach, R. (2020). Review of the influence of noise in X-ray computed tomography measurement uncertainty. *Precision Engineering*, *66*, 382–391. http://dx.doi.org/10.1016/j.precisioneng.2020.08.004, URL: https://www.sciencedirect.com/science/article/pii/S0141635920306012.

Schroeder, W., Martin, K., & Lorensen, B. (2006). *The visualization toolkit* (4th ed.).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., et al. (2014). Going deeper with convolutions. CoRR abs/1409.4842, URL: http://arxiv.org/abs/1409.4842.

Tan, Y., Kiekens, K., Kruth, J.-P., Voet, A., & Dewulf, W. (2011). Material dependent thresholding for dimensional X-ray computed tomography. In *DGZFP-Proceedings BB 128-CD*.

van Dael, M., Verboven, P., Zanella, A., Sijbers, J., & Nicolai, B. (2019). Combination of shape and X-ray inspection for apple internal quality control: In silico analysis of the methodology based on X-ray computed tomography. *Postharvest Biology and Technology*, *148*, 218–227. http://dx.doi.org/10.1016/j.postharvbio.2018.05.020.

Withers, P. J., Bouman, C., Carmignato, S., Cnudde, V., Grimaldi, D., Hagen, C. K., et al. (2021). X-ray computed tomography. *Nature Reviews Methods Primers*, *1*(1), 1–21.