Computerized Medical Imaging and Graphics 106 (2023) 102204

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

Accurate volume alignment of arbitrarily oriented tibiae based on a mutual attention network for osteoarthritis analysis

Jian-Qing Zheng [a,*], Ngee Han Lim [a], Bartłomiej W. Papież [b]

[a] The Kennedy Institute of Rheumatology, University of Oxford, UK
[b] Big Data Institute, University of Oxford, UK

ARTICLE INFO

Keywords:
Image registration
Deep learning
Mutual attention
Tibiae CT

ABSTRACT

Damage to cartilage is an important indicator of osteoarthritis progression, but manual extraction of cartilage morphology is time-consuming and prone to error. To address this, we hypothesize that automatic labeling of cartilage can be achieved through the comparison of contrasted and non-contrasted Computer Tomography (CT). However, this is non-trivial as the pre-clinical volumes are at arbitrary starting poses due to the lack of standardized acquisition protocols. Thus, we propose an annotation-free deep learning method, D-net, for accurate and automatic alignment of pre- and post-contrasted cartilage CT volumes. D-Net is based on a novel mutual attention network structure to capture large-range translation and full-range rotation without the need for a prior pose template. CT volumes of mice tibiae are used for validation, with synthetic transformation for training and tested with real pre- and post-contrasted CT volumes. Analysis of Variance (ANOVA) was used to compare the different network structures. Our proposed method, D-net, achieves a Dice coefficient of 0.87, and significantly outperforms other state-of-the-art deep learning models, in the real-world alignment of 50 pairs of pre- and post-contrasted CT volumes when cascaded as a multi-stage network.

1. Introduction

Pre-clinical studies on osteoarthritis (OA) progression, such as the development of disease-modifying osteoarthritis drugs on animal models (Vincent, 2020), currently suffers from the lack of available robust quantitative biomarkers (Hosnijeh et al., 2019). The damage of cartilage is an important indicator for OA, with morphological analysis of cartilage based on the cartilage shape extracted from Magnetic Resonance Imaging (MRI) in clinical studies (Shan et al., 2014), but is currently unavailable in pre-clinical mouse models, as the thickness of cartilage is close to the resolution of MRI.

In terms of modalities available for cartilage imaging, MRI is the commonly used modality (Burton II et al., 2020; Shan et al., 2014; Grau et al., 2004; Ambellan et al., 2019), but is costly, and time-consuming (James and Gambhir, 2012), with a poor spatial resolution (maximum of 20 μm) for pre-clinical imaging of living animals where cartilage thickness is of ≤100 μm for mice (Borges et al., 2014). Computed Tomography (CT) is both cheaper (Gangwar et al., 2018; Myller et al., 2018; Maier et al., 2020), and micro-CT has a higher resolution (of 10 μm Zheng et al., 2020a), but the cartilage is not directly visible in CT scans without a contrast agent, due to its low X-ray absorption. Thus, the development of contrast agents that binds well to cartilage is an active pre-clinical and clinical research area.

In general, to provide a reliable qualitative measurement of cartilage shape and its subtle changes due to the progression of the disease over time, robust segmentation method is required. Previous attempts at cartilage segmentation from clinical imaging are based on classic segmentation methods, which require an atlas (Shan et al., 2014), or the use of markers (Grau et al., 2004). Deep learning methods have been utilized in more recent approaches to cartilage segmentation, mainly focusing on MRI clinical data (Burton II et al., 2020; Gangwar et al., 2018; Myller et al., 2018; Ambellan et al., 2019; Maier et al., 2020), with only one method designed to segment craniofacial cartilage of mice for investigation of craniofacial syndromes (Zheng et al., 2020a). The greatest challenge associated with cartilage segmentation in the development of novel CT-contrast agents (Fowkes et al., 2022) is the lack of standardized imaging protocols, resulting in scans acquired in arbitrary poses (Fig. 1). The exemplified views taken from 3D micro-CT are shown in Fig. 1.

By using the contrast agent that binds to the cartilage only, the difference between the CT scans with and without contrast could be employed to extract cartilage shape, and thus the subsequent changes

* Corresponding author.
E-mail addresses: jianqing.zheng@kennedy.ox.ac.uk, jianqing.zheng@outlook.com (J.-Q. Zheng), han.lim@kennedy.ox.ac.uk (N.H. Lim), bartlomiej.papiez@bdi.ox.ac.uk (B.W. Papież).

https://doi.org/10.1016/j.compmedimag.2023.102204
Received 23 September 2022; Received in revised form 14 February 2023; Accepted 14 February 2023
Available online 24 February 2023

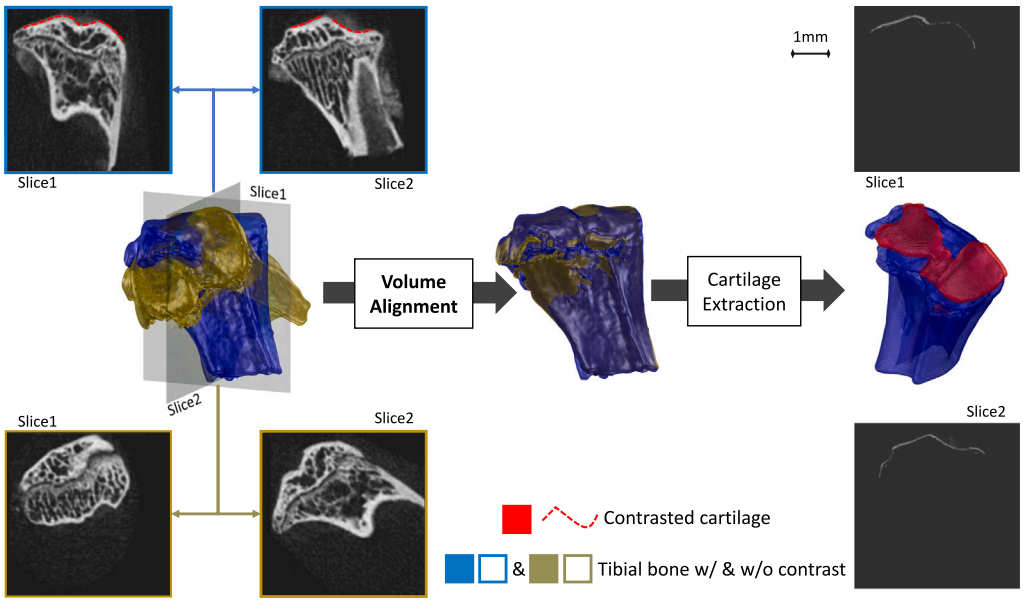

Fig. 1. Accurate 3D volume alignment is required to extract the cartilage shape due to the arbitrary positioning of tibial bone between the baseline CT scan, and the follow-up CT scan with contrast as imaging acquisition standards are not yet established at this stage of research.

to it can be tracked, as shown in Fig. 1. This, however, requires an accurate alignment of the tibial bones between the subsequent CT volume acquisitions. In the current pre-clinical setup, semi-manual alignment using *ImageJ* takes over 1 h for each pair of CT volumes, heavily limiting analysis of large-scale experiments required in pre-clinical research. Moreover, the variation in data acquisition creates difficulties in parameter tuning for semi-automatic methods, as well as the unsatisfactory alignment by classic iterative methods (Zheng et al., 2020b), which motivates the development of an accurate and automated method for alignment of 3D volumes acquired in the pre-clinical setup.

Therefore, in this paper, we propose an annotation-free framework for volume alignment using a new architecture, D-net, which is capable of estimating the arbitrary rotation and translation for tibial CT scans with varying levels of contrast.

The main contributions of this work are as follows:

1. A new architecture, D-net, based on a Siamese Encoder Decoder (SED) (Dunnhofer et al., 2020; Kwon et al., 2019), is proposed for alignment of two volumes with an arbitrary, initial orientation, outperforming the current state-of-the-art rigid registration methods, without the requirement of the prior standard template.
2. A novel Mutual Non-Local link (MNL) block containing an extended attention mechanism that covers global connections between multiple branches.
3. D-Net provides a solution for automatic alignment of tibial bones, which in turn, could allow the extraction of cartilage shape from contrasted CT volumes without costly manual annotations.

The current manuscript extends our preliminary study (Zheng et al., 2020b) with the differences as follow: (i) We provide a comprehensive description of D-Net in Section 3.2.1, and Mutual Non-Local link block in Section 3.2.2, including more folders cross-validation, two more network structure baselines, and one more experimental setting – Small-Range spatial Transformation (SRT) that is closer to the clinical application. (ii) We improved alignment accuracy by adding atrous convolution layers to improve the spatial understanding by the network as in Section 3.2.3 (so-called D-net^a), and the multi-stage approach to coarse-to-fine align the volumes as in Section 3.3.

Using pre-clinical CT data of cartilage described in Section 3.4, we show an extensive evaluation of D-net and comparison to the state-of-the-art methods in volume alignment in Section 4. The manuscript ends with a discussion and conclusion presented in Sections 5 and 6.

2. Related works

2.1. Alignment of clinical tibial scans

Volume alignment is very important in many clinical studies as a common step for data preprocessing. Therefore, many longitudinal clinical OA studies could benefit from the volume alignment of multi-scans of CT, such as trabecular bone structural analysis (Kraiger et al., 2012), temporomandibular joint osteoarthritis quantification (Paniagua et al., 2011), and knee complex shape modeling (Filip et al., 2021).

Yoo et al. (2009) worked on rigid registration between clinical scans of the knee MRI, by iteratively optimizing a mutual-information based function. Urish et al. (2013) also introduced a surface distance-based iterative registration method on volume alignment of clinical knee MRI for morphological analysis of cartilage, requiring semi-automatic segmentation of both bones and cartilage. These methods, however, have standardized imaging protocols with only a small rotation being present (e.g. only 1° rotation estimated in Yoo et al. (2009)), which is much smaller than in preclinical settings. Additionally, images of the knee from live mice need to be taken in a flexed position as a fixed straight knee joint results in torn ligaments, and the post-mortem ex-vivo tissues of our data set are scanned in solution in containers with arbitrary orientation and position. While large ranges of rotation are normally not encountered in clinical imaging, there are several instances where this occurs, such as rigid registration in surgery (Robu et al., 2018) and handheld ultrasound scanning (Namburete et al., 2018; Moser et al., 2022).

2.2. Volume alignment in pre-clinical imaging

Chow et al. (2006) designed an imaging chamber for hardware-based rigid registration of pre-clinical images in the combined micro Positron Emission Tomography (PET) and micro CT scanners, of which the application is limited by the device requirement. In terms of software-based registration for pre-clinical imaging, Baiker et al. (2011)

used classic approaches for iterative image registration on pre-clinical image alignment, where the intensity changes caused by contrast are captured by a similarity measure. However, iterative methods (Yoo et al., 2009; Baiker et al., 2011; Urish et al., 2013), usually rely on the initialization due to the non-convexity in the matching metrics function for medical images, which could be trapped in a local optimum, especially for a large-range spatial transformation as presented in Fig. 1. Due to the varying shapes of the cut sample of animals' tibiae, it is hard to find salient features consistently for feature-based registration. Therefore, a deep learning approach is explored for our task.

2.3. Deep learning for volume alignment

Several works on deep learning methods (Liao et al., 2017; Ma et al., 2017) show performance improvement over iterative image registration. However, these methods are still time-consuming in the inference phase and dependent on initialization, which further calls for rigid registration via direct regression (Haskins et al., 2020). For instance, a two-branch Siamese Encoder (SE) structure, with shared weights capturing the common features of two images, was utilized for the alignment of 2D brain images (Sloan et al., 2018). AIRNet (Chee and Wu, 2018), employing a SE structure consisting of dense structural (Zhu and Newsam, 2017) convolution layers, was applied to the affine transform estimation of 3D brain MRI. The SE structure was also used for the pre-alignment of volumes in a deformable registration framework by de Vos et al. (2019). Alternatively, affine registration can be obtained by using the Global-net (Hu et al., 2018) with the input images concatenated and fed into a one-branch encoder, termed the Mixed Encoder (ME) in our evaluation. However, the previous approaches (Sloan et al., 2018; Chee and Wu, 2018; de Vos et al., 2019; Hu et al., 2018) focused on clinical images registration, with the heavily limited capture range of rotation between ±15° (Sloan et al., 2018) and ±45.84° (0.8 rad) (Chee and Wu, 2018). They thus yield unsatisfactory results with pre-clinical data acquisition setup as shown in Section 4. A standard template, such as that used in ultrasound (Namburete et al., 2018) and MRI (Salehi et al., 2018) fetal brain imaging may aid this task. However, such a standard template is unavailable for pre-clinical cartilage imaging due to the lack of standardization.

3. Methods

The overview of the proposed image registration framework is first introduced in Section 3.1. Then the design detail of the new architecture D-net is described in Section 3.2. The coarse-to-fine rigid registration is employed in Section 3.3. Finally, the D-net is compared with other state-of-the-art methods, and those mentioned components are validated in Section 3.4.

3.1. Image registration framework

The target of 3D image (volume) registration is to estimate the spatial transformation mapping $f : \mathbb{R}^s \to \mathbb{R}^s, \boldsymbol{X}^{\mathrm{f}} \mapsto \boldsymbol{X}^{\mathrm{m}}$ between a fixed volume $\boldsymbol{X}^{\mathrm{f}} \in \mathbb{R}^s$ and a moving volume $\boldsymbol{X}^{\mathrm{m}} \in \mathbb{R}^s$, where the size of a 3D volume is denoted by $s = d \times h \times w$, and d, h, w are the thickness, height, and width. Specifically for rigid registration of two volumes, the transformation in the special Euclidean group $f := [\boldsymbol{R}\ \boldsymbol{t}] \in \mathbb{SE}(3)$ is estimated, consisting of rotation $\boldsymbol{R} \in \mathbb{SO}(3)$ and translation $\boldsymbol{t} \in \mathbb{R}^3$.

To formulate the rotation representation, an n-element vector $\boldsymbol{\theta} = [\theta_{1:n}] \in \mathbb{R}^n$ is calculated from the rotation $\boldsymbol{R} = [\boldsymbol{r}_1\ \boldsymbol{r}_2\ \boldsymbol{r}_3]$ by the parameterization mapping $\omega : \mathbb{SO}(3) \to \mathbb{R}^n, \boldsymbol{R} \mapsto \boldsymbol{\theta}$, where $[\theta_{i:j}] \in \mathbb{R}^{j-i+1}$ denotes a column vector consisting of $\theta_i \cdots \theta_j$ indexed by i, j. The representation mapping $\Omega : \mathbb{R}^n \to \mathbb{SO}(3), \boldsymbol{\theta} \mapsto \boldsymbol{R}$ is used to estimate the rotation. The most intuitive way of rotation representation is a 3×3 matrix, with $n = 9$ (9d) (Sloan et al., 2018; Chee and Wu, 2018; de Vos et al., 2019; Hu et al., 2018; Namburete et al., 2018), here the parameterization mapping is calculated as:

$$\omega_{9\mathrm{d}}(\boldsymbol{R}) := [\theta_{1:9}] = [\boldsymbol{r}_1^{\top}\ \boldsymbol{r}_2^{\top}\ \boldsymbol{r}_3^{\top}]^{\top} \tag{1}$$

and the representation mapping is calculated as:

$$\Omega_{9\mathrm{d}}(\boldsymbol{\theta}) := [\boldsymbol{r}_1\ \boldsymbol{r}_2\ \boldsymbol{r}_3] = \begin{bmatrix} \mathrm{N}([\theta_{1:3}])^{\top} \\ \mathrm{N}([\theta_{4:6}] - \langle [\theta_{4:6}], \boldsymbol{r}_1 \rangle \boldsymbol{r}_1)^{\top} \\ \mathrm{N}([\theta_{7:9}] - \langle [\theta_{7:9}], \boldsymbol{r}_1 \rangle \boldsymbol{r}_1 - \langle [\theta_{7:9}], \boldsymbol{r}_2 \rangle \boldsymbol{r}_2)^{\top} \end{bmatrix}^{\top} \tag{2}$$

where $\mathrm{N}(\cdot)$ denotes the Euclidean normalization function, $\langle \cdot, \cdot \rangle$ denotes the inner production.

As $\boldsymbol{\theta} \in \mathbb{R}^9$ is redundant for rotation, the 3D orthogonalization mapping of 6D rotation representation (Zhou et al., 2019) is used with $n = 6$:

$$\omega_{6\mathrm{d}}(\boldsymbol{R}) := [\theta_{1:6}] = [\boldsymbol{r}_1^{\top}\ \boldsymbol{r}_2^{\top}]^{\top} \tag{3}$$

and

$$\Omega_{6\mathrm{d}}(\boldsymbol{\theta}) := [\boldsymbol{r}_1\ \boldsymbol{r}_2\ \boldsymbol{r}_3] = \begin{bmatrix} \mathrm{N}([\theta_{1:3}])^{\top} \\ \mathrm{N}([\theta_{4:6}] - \langle [\theta_{4:6}], \boldsymbol{r}_1 \rangle \boldsymbol{r}_1)^{\top} \\ (\boldsymbol{r}_1 \wedge \boldsymbol{r}_2)^{\top} \end{bmatrix}^{\top} \tag{4}$$

where $\wedge$ is the cross product.

The task for neural networks in rigid registration is to estimate the rotation and translation parameters $[\hat{\boldsymbol{\theta}}\ \hat{\boldsymbol{t}}]$ between the two preprocessed volumes $\boldsymbol{X}^{\mathrm{f}}$ and $\boldsymbol{X}^{\mathrm{m}}$ by networks' mapping $g : \mathbb{R}^s \times \mathbb{R}^s \to \mathbb{R}^n \times \mathbb{R}^3, (\boldsymbol{X}^{\mathrm{f}}, \boldsymbol{X}^{\mathrm{m}}) \mapsto (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{t}})$, where $\times$ between two sets denotes the Cartesian product. Thus, the one-stage rigid transformation can be estimated by:

$$\hat{f} = \left[\Omega(g^{\theta}(\boldsymbol{X}^{\mathrm{f}}, \boldsymbol{X}^{\mathrm{m}}))\ \ g^{t}(\boldsymbol{X}^{\mathrm{f}}, \boldsymbol{X}^{\mathrm{m}})\right] \tag{5}$$

where g^{θ} and g^{t} are the mappings to $\boldsymbol{R}$ and $\boldsymbol{t}$. The details of the neural network architectures are described in Section 3.2.3.

As shown in Fig. 2, in the training phase, each pair of CT volumes is preprocessed, augmented, synthesized, and then fed into the network to train it by optimizing a loss function with the synthetic transformation; in the inference phase, the preprocessed contrasted and non-contrasted CT volumes are fed to the trained one/multi-stage network to predict the spatial transformation. The implementation of multi-staging is described in Section 3.3. In this paper, the two-stage network, with the first stage for coarse alignment and the second stage for fine-refinement, is used to improve the overall accuracy of volume alignment, and compared with one-stage networks in Section 3.4.

3.2. Network design

3.2.1. D-net architecture

D-net is based on SED structure and includes the two-branch Siamese Encoder (SE) with MNL block (see the schematic structure of D-Net, Fig. 3, MNL block in red), and one-branch decoder with regression layers. Similar SED structures have been employed for tracking (Dunnhofer et al., 2020) and segmentation (Kwon et al., 2019).

In the SE part of D-net, there are two branches of six Residual-down-sampling (Res-down) blocks with shared parameters, where four latter pairs of the Res-down blocks are mutually linked by MNL. The SE is employed to extract the common features from two input volumes, and the MNL block is inserted to capture the global-range link between the features points from two branches (inter-branch), of which the implementation details are illustrated in Section 3.2.2. The decoder part consists of four Residual-up-sampling (Res-up) blocks with skip connections received from the same indexed Res-down blocks to restore the location information lost in Res-down blocks. The regression part includes several fully connected layers, with the number of neurons in the output layer equal to the parameter number for rigid spatial transformation $n + 3$. The evolution and development detail of D-net is described in Section 3.2.3.

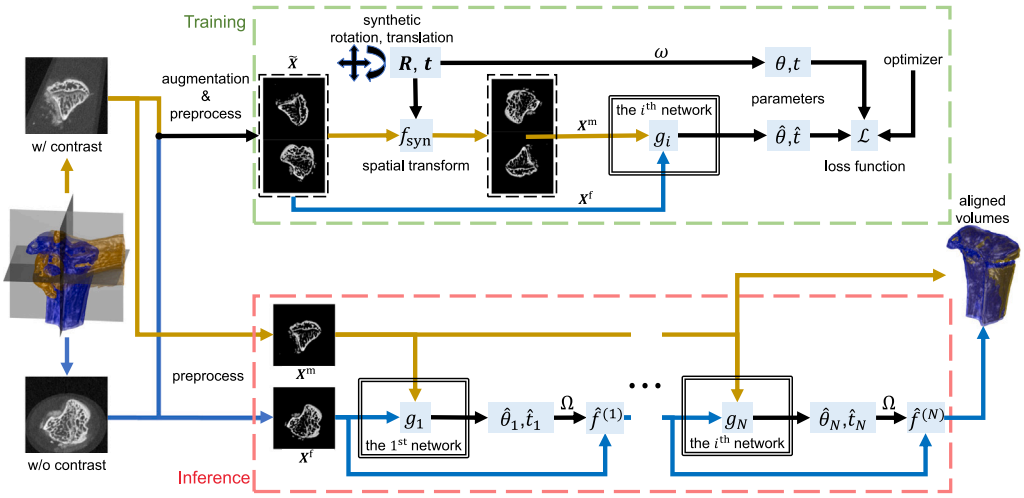

Fig. 2. The framework for volume alignment. In the training phase, both images stem from the same scan, whereas in the inference phase, the images are pre- and post-contrast of the same tibiae.

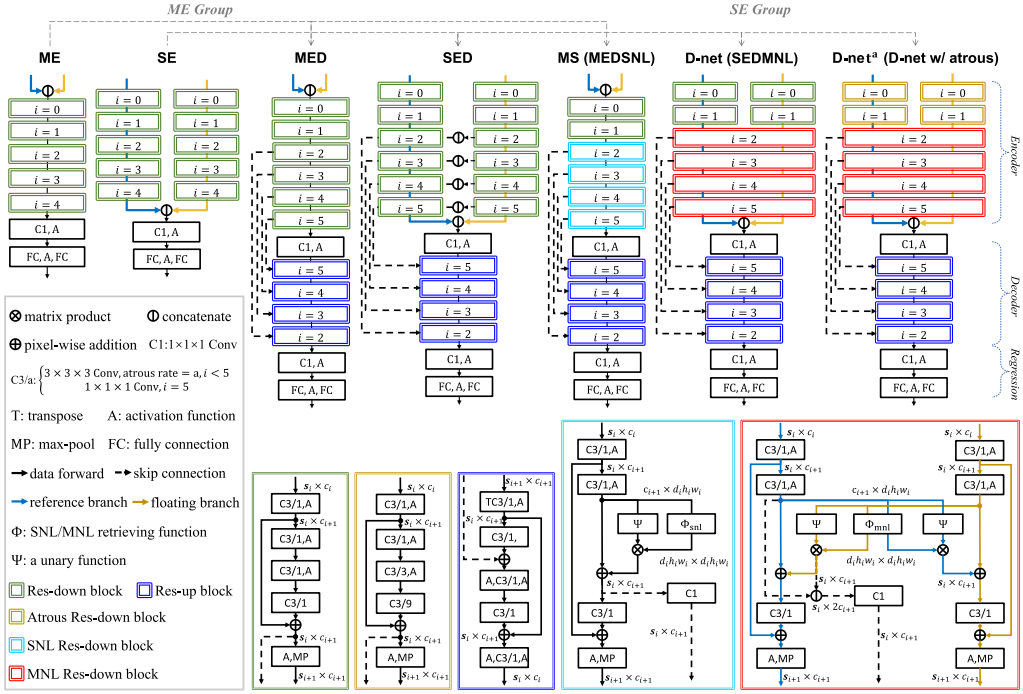

Fig. 3. The evolution of D-Net. The seven network structures have been investigated in this paper: the single branch ME (Mixed Encoder) group (ME, MED (Mixed Encoder Decoder) and MS (MED with SNL), and the two branched SE (Siamese Encoder) group (SE, SED (Siamese Encoder Decoder), D-net (SED with MNL) and D-net^a (D-net with Atrous convolution). i is the block number, $\boldsymbol{d} = (d_0 \cdots d_6) \in \mathbb{Z}_+^7$, $\boldsymbol{h} = (h_0 \cdots h_6) \in \mathbb{Z}_+^7$, $\boldsymbol{w} = (w_0 \cdots w_6) \in \mathbb{Z}_+^7$ and $\boldsymbol{c} = (c_0 \cdots c_6) \in \mathbb{Z}_+^7$ respectively denote the sequences of thickness, heights, widths, and channel number of the input volume/feature maps for each branch.

3.2.2. *Mutual non-local link*

In iterative image registration, long-range links are either estimated using a range of spatial scales, naturally captured by graph representation (Heinrich et al., 2013; Papież et al., 2016), or through a weighting function with mutual saliency which establishes unique matching between a pair of voxels (Ou et al., 2011). Here, the proposed MNL block is based on the combination of a new concept of mutual attention, which is an extended mutual link version of the conventional (self) attention mechanism (Vaswani et al., 2017), together with a Non-Local block (Wang et al., 2018).

The Non-Local mechanism is compared with direct concatenation in Fig. 4. Direct concatenation, used in ME, SE, MED and SED, fails to provide spatially long-range links between similar features. Self Non-Local (SNL), used in the concatenated branch of the encoder part of MS, succeeds in achieving the global-range link, where the features' representations between two branches are locally bounded together with both the concatenation and retrieval operations carried out on the concatenated two-branch features and their copies. The Non-Local, inter-branch links are thus coupled with the Non-Local intra-branch link in SNL, which emphasizes the features correspondence intra- rather than inter-volumes. MNL is used in D-net to search for a global-range link between two branches, and thus excludes the coupling of the intra-branch link.

Here we provide the general definition of SNL:

$$\boldsymbol{Y}^{\mathrm{f|m}} := \boldsymbol{\Psi}(\boldsymbol{Y}) \otimes \boldsymbol{\Phi}(\boldsymbol{Y}, \boldsymbol{Y}) \tag{6}$$

where $\boldsymbol{Y}$ denotes the reshaped input signals from one branch, and $\otimes$ denotes the matrix multiplication. Subsequently, MNL is defined as

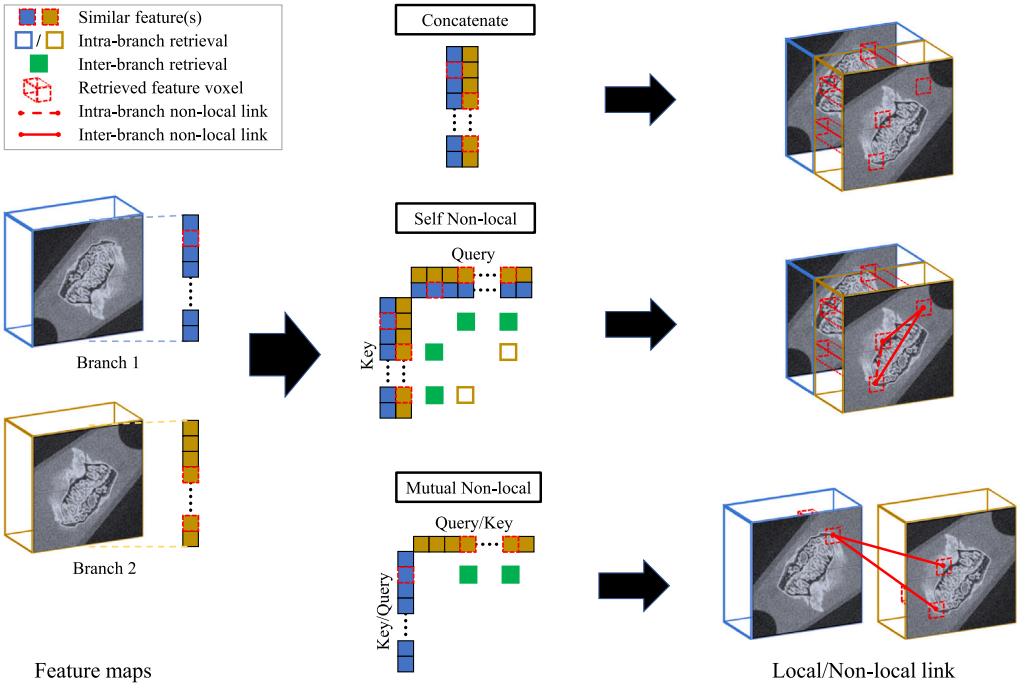

Fig. 4. Non-Local inter-branch link is required for similar features correspondence. Direct concatenation provides only local inter-branch link (top); Self Non-Local (Eq. (6)) achieves the global-range inter-branch link but coupled with intra-branch link (middle); Mutual Non-Local (Eq. (7)) obtains global-range link between two branches without the intra-branch link (bottom).

follows:

$$\begin{cases} \boldsymbol{Y}^{\mathrm{m2f}} := \Psi(\boldsymbol{Y}^{\mathrm{m}}) \otimes \Phi(\boldsymbol{Y}^{\mathrm{m}}, \boldsymbol{Y}^{\mathrm{f}}) \\ \boldsymbol{Y}^{\mathrm{f2m}} := \Psi(\boldsymbol{Y}^{\mathrm{f}}) \otimes \Phi(\boldsymbol{Y}^{\mathrm{f}}, \boldsymbol{Y}^{\mathrm{m}}) \end{cases} \tag{7}$$

where $\boldsymbol{Y}^{\mathrm{f}}, \boldsymbol{Y}^{\mathrm{m}} \in \mathbb{R}^{c \times dhw}$ denote the reshaped input signals from fixed and moving volumes, $\boldsymbol{Y}^{\mathrm{f|m}}, \boldsymbol{Y}^{\mathrm{m2f}}, \boldsymbol{Y}^{\mathrm{f2m}} \in \mathbb{R}^{c \times dhw}$ are the output signals from this block, $\Phi : \mathbb{R}^{c \times dhw} \times \mathbb{R}^{c \times dhw} \to \mathbb{R}^{dhw \times dhw}$ is a retrieval function for the similarity measurement between the two inputs, key vectors, and query vectors, to compute the compatibility weight of each query vector with the assigned key vector(s), and $\Psi : \mathbb{R}^{c \times dhw} \to \mathbb{R}^{c \times dhw}$ is a unary function for the mapping of the value vector from each retrieved vector to the assigned vector.

The instantiated SNL in MS, and MNL in D-net, are based on the embedded Gaussian similarity representation for retrieving:

$$\Phi(\boldsymbol{Y}_1, \boldsymbol{Y}_2) := \mathrm{softmax}(\boldsymbol{Y}_1{}^{\top} \otimes \boldsymbol{W}^{\top} \otimes \boldsymbol{W} \otimes \boldsymbol{Y}_2) \tag{8}$$

and the unary function for value vectors' mapping

$$\Psi(\boldsymbol{Y}) := \boldsymbol{W} \otimes \boldsymbol{Y} \tag{9}$$

where $\boldsymbol{W} \in \mathbb{R}^{c \times c}$ is a matrix of trainable weights.

3.2.3. D-net evolution and comparison with the other network structures

The structures of the seven networks used here are shown in Fig. 3. The networks can be divided into two distinct groups; the ME group based on the mixed one-branch encoder, and the SE group based on the Siamese two-branch encoder, from which the other networks evolved.

- ME (Mixed Encoder): The state-of-the-art rigid registration network "Global-net" (Hu et al., 2018) is adapted to the network structure ME with the two input volumes concatenated together and fed into one mixed branch. The mixed branch consists of several Residual-down-sampling (Res-down) blocks (Fig. 3, green blocks) as the encoder part, followed by one convolution layer, one activation layer, and two fully connected layers for regression.
- SE (Siamese Encoder): Compared to ME, the SE structure employs two weight-sharing branches of several Res-down blocks as the encoder to extract similar features from two branches, followed by regression layers. A similar structure was used by Sloan et al. (2018), Chee and Wu (2018) and de Vos et al. (2019). de Vos et al. (2019) used fewer down-sampling blocks and no residual structure, more down-sampling and fully connected layers are used in Chee and Wu (2018), and residual structure was replaced by a dense structure in Sloan et al. (2018).
- MED (Mixed Encoder Decoder): ME is modified to adopt a commonly used encoder–decoder architecture by insertion of four Residual-upsampling (Res-up) blocks (Fig. 3, blue blocks), with skip connections (Fig. 3, dashed lines) to the 4 latter Res-down blocks of the encoder, between the encoder and regression sections.
- SED (Siamese Encoder Decoder): SE with four Residual-upsampling (Res-up) blocks with skip connections to the 4 successive latter Res-down blocks of the encoder, inserted between the encoder and regression sections. Similar SED structures were used for segmentation (Kwon et al., 2019) and tracking (Dunnhofer et al., 2020), but with similar skip connections to MED for comparison.
- MS (Mixed Encoder Decoder with Self Non-Local block): MED with the additional embedded Gaussian similarity based-Non-Local block (Wang et al., 2018) paralleled into the 4 successive latter Res-down blocks (Fig. 3, light blue blocks).
- D-net (Siamese Encoder Decoder with Mutual Non-Local block): SED with the four latter pairs of the Res-down blocks linked by the mutual Non-Local block which is described in Section 3.2.2 (Fig. 3, red block).
- D-net[a] (D-net with Atrous convolution): An improved D-net with the convolution layers in the two earlier pairs of Res-down blocks replaced with the atrous convolution as described by Zhou et al. (2020).

The evolution of D-Net was necessitated by the poor performance of the initial networks in the large-range transformation task (Fig. 6), even though they were perfectly capable of handling the small-range

Table 1
The data arranged for five-fold cross validation with a subject index and disease stage for each group to evaluate the performance of the framework.

Group Id.	1	2	3	4	5
Subject Ids	1–10	11–20	21–30	31–40	41–50
Disease stage (weeks)	0,2	2,4	4,8	8,12	12,16,20

transformation (Fig. 7). The down-sampling in the encoder part of ME and SE structures is required for the long-range link to convert the common intensity features from the two volumes to the relative location information. However, the down-sampling operation also causes the loss of spatial information. Therefore the decoder part is introduced to recover the relative spatial information from the encoder part in MED and SED.

The rigid registration, with large-range initial translation and full-range initial rotation, calls for long-range links between the pairs of areas with similar features. As shown in Fig. 4, the concatenation operation in MED and SED structure limits the link range of the following convolution layers, and thus the concept of Non-Local is introduced into MS with SNL block. However, the inter-branch Non-Local link could be coupled with the intra-branch Non-Local link in SNL as explained in Section 3.2.2. The registration between two volumes requires the emphasis on the inter-branch Non-Local link, rather than the intra-branch, which limits MS. Thus, the MNL is designed to search for a global-range inter-branch link, which is used in D-net.

A further atrous convolution is employed to obtain more absolute spatial information by enlarging the receptive field according to the conclusion of Islam et al. (2019). The geometric series of atrous rate (1 3 9) within each Atrous Res-down block, as shown in Fig. 3, is the setting of Atrous-III block with the theoretical optimization of receptive field size in Zhou et al. (2020).

3.3. Multi-staging

Since the performance of the one-stage rigid registration method could be limited by the capacity of the used neural network, the multi-stage (N-stage) rigid registration is applied, which is calculated by:

$$\begin{cases} \boldsymbol{X}^{(0)} = \boldsymbol{X}^{\mathrm{f}} \\ \hat{f}^{(i)} = \left[\Omega(g_i^{\theta}(\boldsymbol{X}^{(i-1)}, \boldsymbol{X}^{\mathrm{m}}))\ g_i^{t}(\boldsymbol{X}^{(i-1)},\ \boldsymbol{X}^{\mathrm{m}})\right],\ \boldsymbol{X}^{(i)} = \hat{f}^{(i)}(\boldsymbol{X}^{(i-1)}) \\ \hat{f} = \hat{f}^{(N)} \circ \hat{f}^{(N-1)} \circ \cdots \circ \hat{f}^{(1)} \end{cases} \tag{10}$$

where $i \in [1, N] \cap \mathbb{N}$ is the stage index here, $\circ$ is the composition operation of two mappings.

In this paper, we implement both one- and two-stage frameworks. The first stage network is for the initial large translation and rotation volume alignment, and the second stage is used to refine the alignment by only dealing with the smaller range output of the first network.

3.4. Experiments

3.4.1. Data collection and processing

A total of 100 *ex vivo* micro-CT volumes were acquired using *Perkin Elmer*, Quantum FX from the tibiae of 50 subjects (mice pre- and post-contrast) with the volume size and isotropic resolution of $512 \times 512 \times 512$ pixels and $10 \times 10 \times 10\ \mu m^3$/vox. As shown in Table 1, the scanned subjects varied from 0 to 20 weeks post osteoarthritic surgery. After euthanisation, the knees of the mice were dissected out, the soft tissue was removed, and the tibias were separated from the femur and inserted into microcentrifuge tubes containing saline for the pre-contrast scan. The post-contrast scans were obtained by incubating the tibial cartilage in contrast agent for one hour, washing it in saline for a further hour, and then scanning it. Small bone fragments may exist in the volumes due to the dissection process.

Table 2
The numbers of trainable parameters (No. of Par., unit: million), thickness $\boldsymbol{d}$, height $\boldsymbol{h}$, width $\boldsymbol{w}$ and channel numbers $\boldsymbol{c}$ of each network structure in the experimental implementation with the index of i.

Networks		ME	SE	MED	SED	MS	D-net&D-net[a]
No. Par.		10.3	12.0	4.8	4.5	5.2	4.9
$\boldsymbol{d} = \boldsymbol{h} = \boldsymbol{w}$	$i = 0$	64	64	64	64	64	64
	$i = 1$	32	32	32	32	32	32
	$i = 2$	16	16	16	16	16	16
	$i = 3$	8	8	8	8	8	8
	$i = 4$	4	4	4	4	4	4
	$i = 5$	2	2	2	2	2	2
	$i = 6$	–	–	1	1	1	1
$\boldsymbol{c}$	$i = 0$	2	1	2	1	2	1
	$i = 1$	16	16	16	16	16	16
	$i = 2$	32	32	32	32	32	32
	$i = 3$	64	64	64	64	64	64
	$i = 4$	128	128	64	64	64	64
	$i = 5$	256	256	64	64	64	64
	$i = 6$	–	–	64	64	64	64

The collected data was pre-processed *in silico* for use in the framework. To obtain the scans, the tibiae (bones) were placed in solution, and the handling between the two consecutive scans of each subject would result in a different relative position and orientation to the containers. Thus, the transformations between the two scans of the solution and container were different from the subject. Both the solution and container have easily identifiable and consistent intensity values across all scans. To eliminate the impact of their different spatial transformations on the network prediction, both solution and container were removed from the volumes by thresholding. The image intensities were normalized into the 0–1 range for stable gradient propagation. Finally, the input volumes were sub-sampled with linear interpolation to the volume size of $64 \times 64 \times 64$ and resolution $80 \times 80 \times 80\ \mu m^3$/vox. The CT slices of the exemplar subjects are illustrated in Fig. 1.

For the training phase, each pre-processed CT volume $\boldsymbol{X}$ was augmented before being fed into the networks by random translation of $\sim \mathcal{U}(-0.04, 0.04)$ mm along 3 axes, and random rotation with full-range angle $\sim \mathcal{U}(-\pi, \pi)$ uniformly distributed around an arbitrary axis, and intensity value scaling and bias to extend the size of the data-set. We also set a randomly varying thresholding value within the range between the intensity of muscle and cortical bone, so that the network could focus on the alignment referring to cortical bone regardless of the level of contrast enhancement. The two inputs of networks, fixed volume $\boldsymbol{X}^{\mathrm{f}} := \tilde{\boldsymbol{X}}$ and moving volume $\boldsymbol{X}^{\mathrm{m}} := f_{\mathrm{syn}}(\tilde{\boldsymbol{X}})$ were synthesized with the augmented volume $\tilde{\boldsymbol{X}}$, and the synthetic transformation f_{syn} for training, as described in Section 3.4.3.

3.4.2. Implementation and parameter setting

The hyper-parameters setting in the encoder and decoder parts of network structures are shown in Table 2, where $i = 0$ represents the input layer. Here the setting of $\boldsymbol{d}, \boldsymbol{h}, \boldsymbol{w}, \boldsymbol{c}$ are sequentially applied after each down-sampling layer. The numbers of neurons for the two fully connected layers were set as 128 and $n + 3$ in the regression part of each network structure. The number of variables in ME and SE was increased in c i = 4 and 5 to make the networks trainable, whereas MED, SED, MS and D-nets had about the same variable number.

3.4.3. Training and validation strategy

The Euclidean distance based-loss function with respect to translation $\boldsymbol{t}$, $\hat{\boldsymbol{t}}$ and rotation $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$ is calculated as:

$$\mathcal{L} = \alpha \frac{\|\boldsymbol{t} - \hat{\boldsymbol{t}}\|_2^2}{\|\boldsymbol{t}\|_2^2 + \epsilon} + \beta \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2 \tag{11}$$

where $\|\cdot\|_2 := \sqrt{(\sum \cdot)^2}$ is L2-norm, α and β are the weighting parameters of relative translation and rotation respectively, and $\epsilon = 0.1$ was used to avoid the singularity. The network parameters were initialized

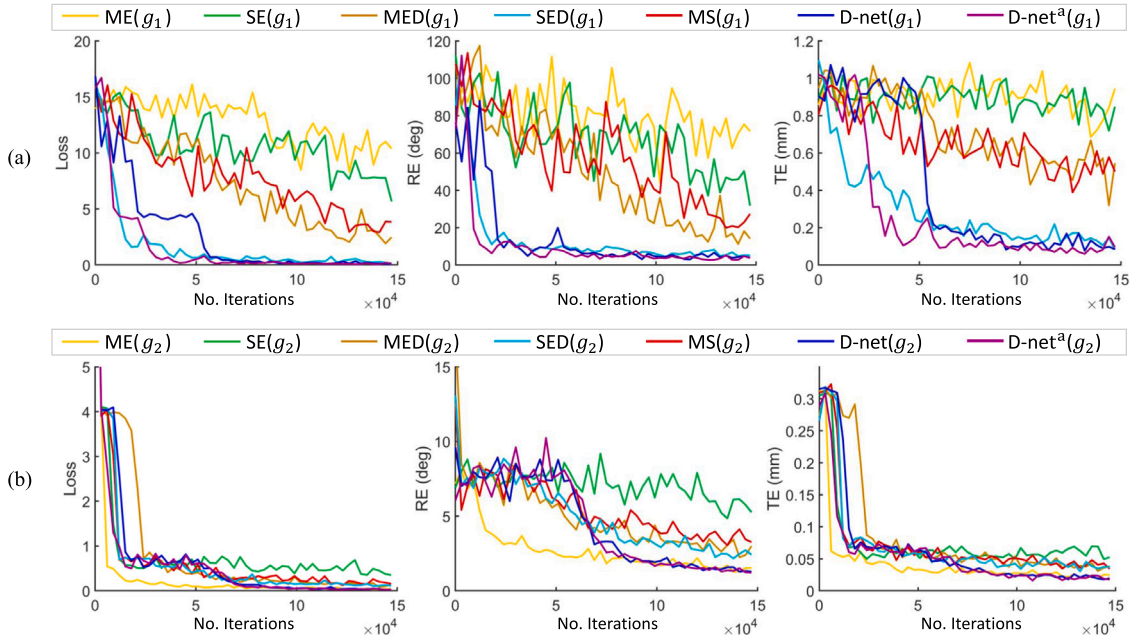

Fig. 5. The training curves of the networks for large (a) and small (b) range transformation, evaluated using the loss function (left) RE (mid) and TE (right) shows SED, D-net and D-net^a achieve the best performance in the training phase in both SRT and LRT.

randomly with normal distribution $\mathcal{N}(0, 0.01)$. Momentum Stochastic Gradient Descend was used for optimization with the learning rate of 0.0001 and the learning momentum of 0.9.

All the networks described in Section 3.2.3 were trained with 150k iteration. Two different training settings were designed with varying synthetic translations and rotations:

(1) Large-Range Transform (LRT) training: the synthetic translation $t_1, t_2, t_3 \sim \mathcal{U}(-0.96, 0.96)$ mm, which is almost 1/5 of the image size 5.12 mm, and the rotation of angle $\sim \mathcal{U}(-\pi, \pi)$ around an arbitrary axis uniformly distributed in the sphere surface. Rotations were represented in 6D. The weights were set as $\alpha = \beta = 0.5$.
(2) Small-Range Transformation (SRT) training: the synthetic translation $t_1, t_2, t_3 \sim \mathcal{U}(-0.32, 0.32)$ mm and rotation of angle $\sim \mathcal{U}(-\pi/12, \pi/12)$ (similar to Sloan et al. (2018)) around an arbitrary axis uniformly distributed in the sphere surface. Rotations were represented in 6D. The parameters were set as $\alpha = 0.5$ and $\beta = 1$. The β term is different in SRT compared to LRT to compensate for the unequal contributions of translation vs rotation.

Two different testing settings were used to evaluate the networks: synthetic testing and real testing:

(1) Synthetic test: the two input volumes were synthesized from the same pre-processed volume and varying known spatial transformations. Over the five-fold cross-validation, a total of 12 100 cases were synthesized for each registration method, with 121 different spatial transformations applied on 100 CT volumes, which were obtained by combining 11 initial translations and 11 initial rotations between the two input volumes. The synthetic transformations were generated similarly to the training procedure.
(2) Real test: each pair of contrasted and non-contrasted CT volumes was pre-processed and fed into each network crossing the five-fold validation.

The ground truth of translation and rotation is known in the synthetic testing, but unknown in the real test.

Different strategies were used for the different targets:

(1) To optimize the networks in terms of LRT and SRT: All networks described in 3.2.3 were compared and tested separately using the one-stage framework in synthetic testing for performance comparison and validation of each network design. g_1 was additionally tested using the real tests.
(2) To prove the superiority of multi-staging: D-net trained with LRT was used as the first-stage alignment network g_1 to feed all the g_2 networks trained with SRT. The cascaded two-stage networks $g_2 \oplus g_1$ were tested with synthetic and real test settings.

Our CT dataset was divided as in Table 1 to allow five-fold cross-validation in all cases with different disease stages, in order to prove the applicability to to progression of the disease.

All experiments were performed in Python using TensorFlow on an Nvidia 2080 Ti GPU 11 GB VRAM.

3.4.4. Evaluation

The Euclidean distance of Translation Error (TE) is calculated by the L2-norm of the difference between the predicted and expected translation:

$$TE = \|\boldsymbol{t} - \hat{\boldsymbol{t}}\|_2 \tag{12}$$

The Rotation Error (RE) between the predicted and expected rotation is calculated by:

$$RE = \arccos\left(\frac{\mathrm{tr}(\boldsymbol{R}^\top \Omega(\hat{\theta})) - 1}{2}\right) \tag{13}$$

The Dice Similarity Coefficient (DSC) between the corresponding tibial bones is calculated by:

$$\mathrm{DSC}(\boldsymbol{V}_1, \boldsymbol{V}_2) = \frac{2|\boldsymbol{V}_1 \cap \boldsymbol{V}_2|}{|\boldsymbol{V}_1| + |\boldsymbol{V}_2|} \tag{14}$$

Here the DSC is calculated for evaluation of registration via $\mathrm{DSC}(\boldsymbol{V}^{\mathrm{p}}_{\mathrm{bone}}, \boldsymbol{V}^{\mathrm{m}}_{\mathrm{bone}})$, where $\boldsymbol{V}^{\mathrm{p}}_{\mathrm{bone}}$ and $\boldsymbol{V}^{\mathrm{m}}_{\mathrm{bone}}$ are the volumes of the cortical bones extracted from the predicted outputs $\hat{f}(\boldsymbol{X}^{\mathrm{f}})$ and the expected ones $\boldsymbol{X}^{\mathrm{m}}$ via thresholding and manual correction.

To evaluate registration performance, TE, RE, and DSC were calculated in synthetic tests. In the Real test, only DSC was used because the ground truth transformations are unknown.

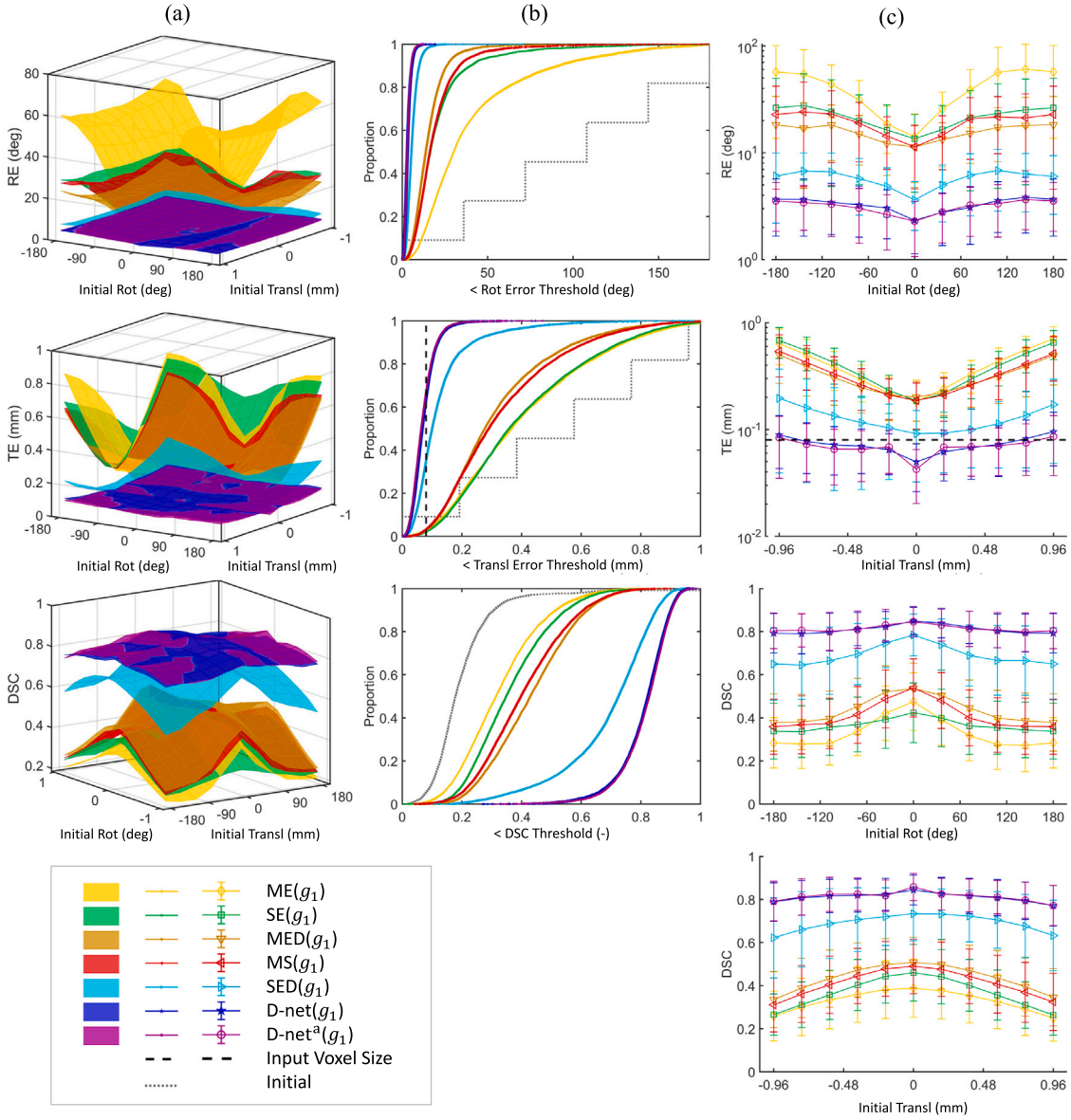

Fig. 6. D-nets outperform other networks in large range transformation (LRT) in synthetic tests. (a) The average errors, (b) empirical cumulative distribution functions, and (c) avg±std of in RE, TE, and DSC for 5-fold cross-validation using one-stage networks trained with large-range transformation.

To quantify the robustness of framework with the respect to the initialization in the real tests, the regression coefficient (RC) which is the one-order coefficient in linear regression is used to evaluate the linear trend between the DSCs before and after registration. It is calculated by:

$$RC = \frac{\mathrm{cov}(DSC, DSC^{\mathrm{init}})}{\mathrm{var}(DSC^{\mathrm{init}})} \tag{15}$$

where the initial DSC^{init} is calculated by $DSC^{\mathrm{init}} := \mathrm{DSC}(\boldsymbol{V}^{\mathrm{f}}, \boldsymbol{V}^{\mathrm{m}})$, $\boldsymbol{V}^{\mathrm{f}}$ denotes the tibial bone volume extracted from the input $\boldsymbol{X}^{\mathrm{f}}$, $\mathrm{cov}(\cdot,\cdot)$ is the covariance between two input samples, and $\mathrm{var}(\cdot)$ is the variance function.

4. Results

All the networks were trained with LRT (Fig. 5(a)) and SRT (Fig. 5(b)). D-nets (dark blue and purple) converged fastest in LRT and converges after ME in SRT settings. D-net^a further improves the training efficiency compared to vanilla D-net, converging on average 10k iterations faster.

Table 3
Average values of Translation Error (TE/μm), Rotation Error (RE/°) and Dice Similarity Coefficient (DSC/%) in Synthetic test (Syn) for one-stage networks trained on large-range transform (g_1) over five-fold cross-validation.

g_1	Init	ME	SE	MED	SED	MS	D-net	D-net^a
TE	523.6	408.3	405.7	322.7	128.5	334.5	72.8	**69.7**
RE	98.2	41.6	21.9	15.8	5.8	19.6	3.3	**3.2**
DSC	20.2	32.9	36.5	43.4	69.1	41.0	81.1	**81.4**

The TE, RE, and DSC results of all networks were confirmed to have a normal distribution via the Kolmogorov–Smirnov test with $p < 1\%$.

4.1. One-stage network

4.1.1. Results in synthetic test

In this section, we report the results of the quantitative analysis performed using the synthetic data set from the same pre-processed volume and varying known spatial transformations. For each pair of volumes, TE, RE, and DSC were calculated with results shown in Figs. 6,

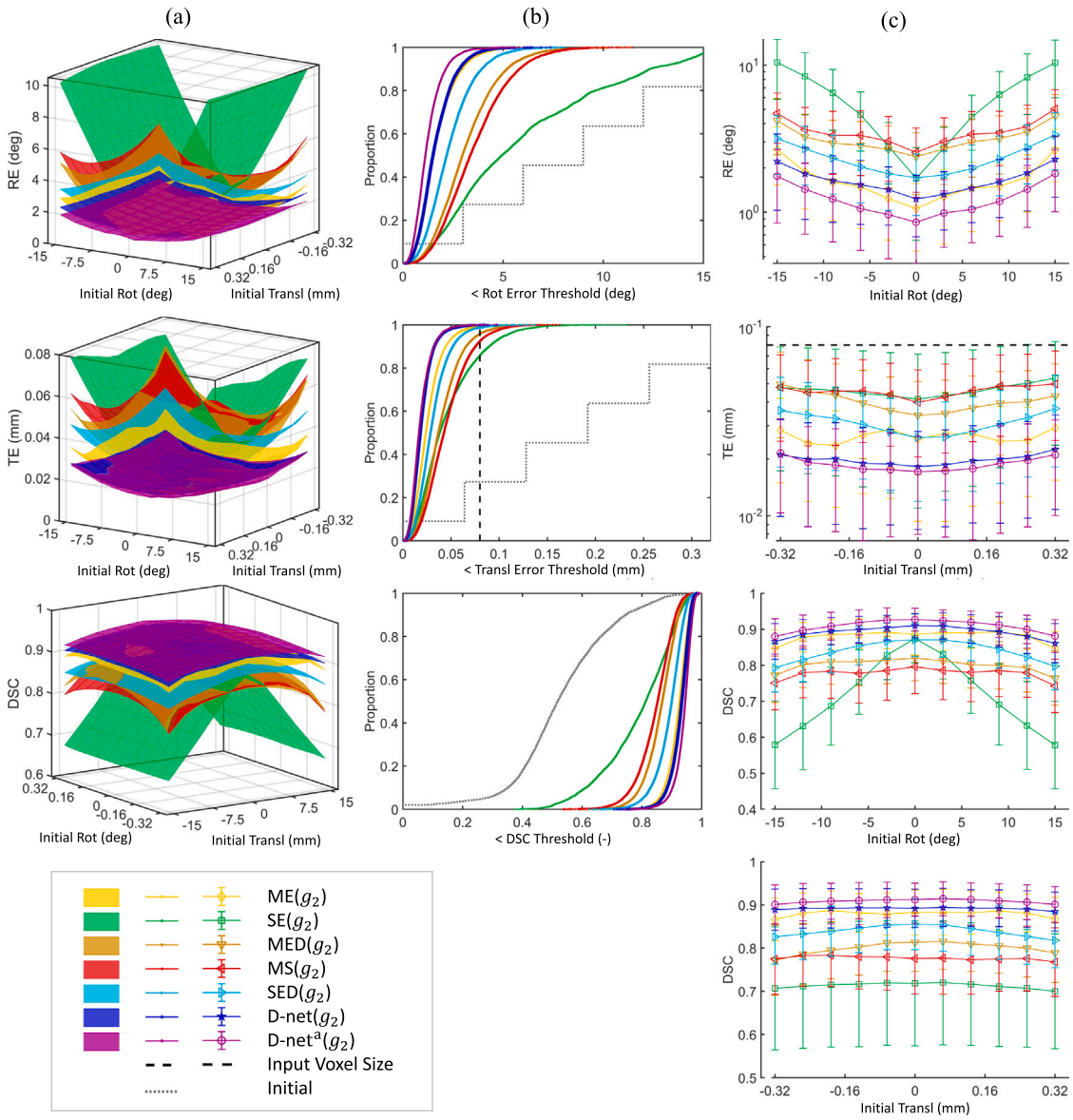

Fig. 7. D-nets are also top performers in small-range transformation (SRT) in synthetic tests. (a) The average errors, (b) empirical cumulative distribution functions, and (c) avg±std of in RE, TE, and DSC for 5-fold cross-validation using one-stage networks trained with small-range transformation.

Table 4
Average values of Translation Error (TE/μm), Rotation Error (RE/°) and Dice Similarity Coefficient (DSC/%) in Synthetic test (Syn) for one-stage networks trained on small-range transform (g_2) over five-fold cross-validation.

g_2	Init	ME	SE	MED	SED	MS	D-net	D-net^a
TE	174.5	26.4	46.3	40.3	31.0	45.7	19.8	**18.7**
RE	8.2	1.7	6.0	3.2	2.4	3.6	1.7	**1.3**
DSC	54.3	92.0	78.9	86.3	89.1	84.5	92.8	**93.9**

7, Tables 3 and 4. We found that both D-Nets are the only networks to estimate the full range of rotation with the voxel-level translation error.

In particular, D-Nets with LRT setting achieved the lowest average TE, RE and the highest average DSC within the varying range of initial rotation and initial translation. The flat shape of the D-nets' surfaces in Fig. 6(a) indicates the robust performances at extreme rotations. SED is the second-best network in the LRT setting, the following third-tier includes MED and MS, and ME and SE are the bottom ones. Fig. 6(b) shows D-nets lower than sub-voxel TE in around 68% cases. Fig. 6(c) shows that only D-nets achieved the voxel-level average TE in the synthetic test in LRT (Fig. 6(c)). The average TE, RE, and DSC are summarized in Table 3. The D-Nets outperform the other networks, and there is no significant difference between D-net and D-net^a (ANOVA with null hypothesis rejected by $p < 1\%$). The performance of SED is better than all others except D-nets.

Similarly as suggested in the previous section, synthetic results for each network with the SRT setting are shown in Fig. 7. The top-tier networks include ME and D-nets, and the second-tier networks are MED, SED, and MS, which are all better than SE. Fig. 7(a) shows D-net^a obtained the lowest mean RE and the highest mean DSC within the varying range of initial rotation and initial translation, where the average TE of D-net^a is also one of the lowest and outperforms single-stage D-net, where the valley shape of the SE's surfaces indicates the least robust performance over the small range. Fig. 7(b) shows over 99% of cases of ME, D-net and D-net^a achieved sub-voxel TE. All the SRT-trained networks achieved sub-voxel average TE with varying initial translation (Fig. 7(c) 2nd graph), whereas SE showed sensitivity to starting rotation, in contrast to the other networks ((Fig. 6(c) top graph). Those quantitative results of average TE, RE, and DSC are

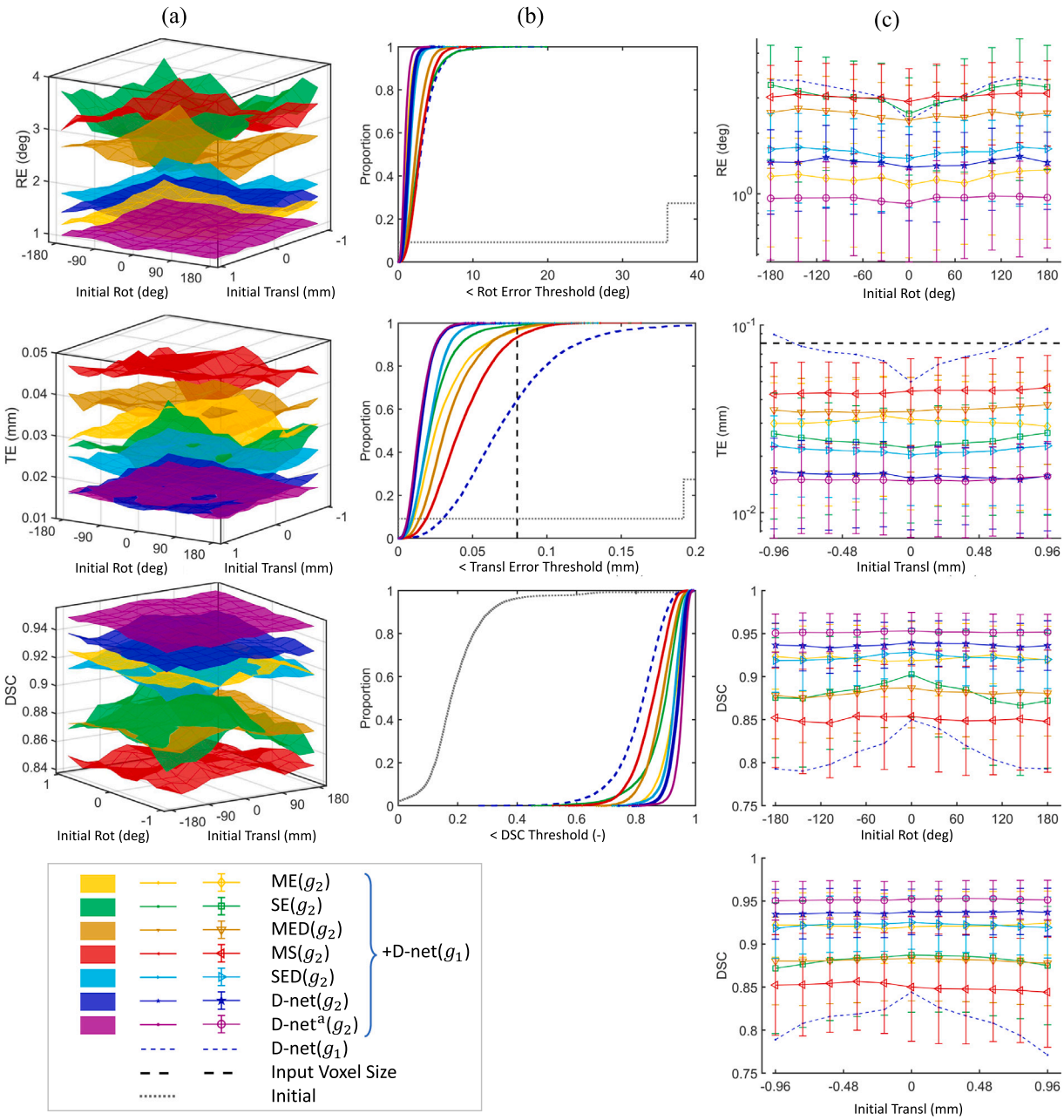

Fig. 8. Cascaded LRT with SRT D-Nets improve alignment further. The average values' surfaces (a), empirical cumulative distribution functions (b), and avg±std (c) of in RE, TE, and DSC for two-stage frameworks of D-net(g_1) cascaded with SRT(g_2) networks show the improvement of multi-staging.

shown in Table 4. The significance test using ANOVA reveals that D-net^a outperforms all others. D-net and ME are the next best networks. This illustrates that D-net is one of the best in the SRT setting and that the Atrous convolution improves the performance compared to vanilla D-net in the SRT synthetic setting.

4.1.2. Results in real test

In this section, we report the quantitative results of each one-stage network with the LRT setting performed on the real data set. For each pair of volumes with and without contrast, DSCs were calculated with results shown in Fig. 9(a) and Table 6.

The results of the Real test as shown in Fig. 9(a) yield consistent accuracy in the synthetic results, including the DSC plotted for each case in the five folders (of cross-validation), the empirical cumulative distribution function with the approximating shape to that in synthetic test (Fig. 6(b)), and the linearly regressed trend line for each network. Those quantitative values of average DSC are shown in Table 6 and the RC with respect to the initial DSC is illustrated in row 2 of Table 7. The ANOVA in the real test shows consistent results to the synthetic with LRT setting. The results above illustrate the improvement of the decoder and MNL, the superiority of D-nets in the LRT setting, and the rationality to use D-net as the first stage network g_1 in a multi-stage registration framework.

4.2. Two-stage network

4.2.1. Results in synthetic test

Previously, Fig. 6(b) shows that over 99.9% of cases' TE and RE of D-nets are within the range of SRT, $[0\text{ mm}, 0.32\text{ mm}]$ and $[0°, 15°]$, and thus prove the rationality of two-stage networks using the LRT-trained D-net cascaded with SRT-trained networks described in Section 3.4.

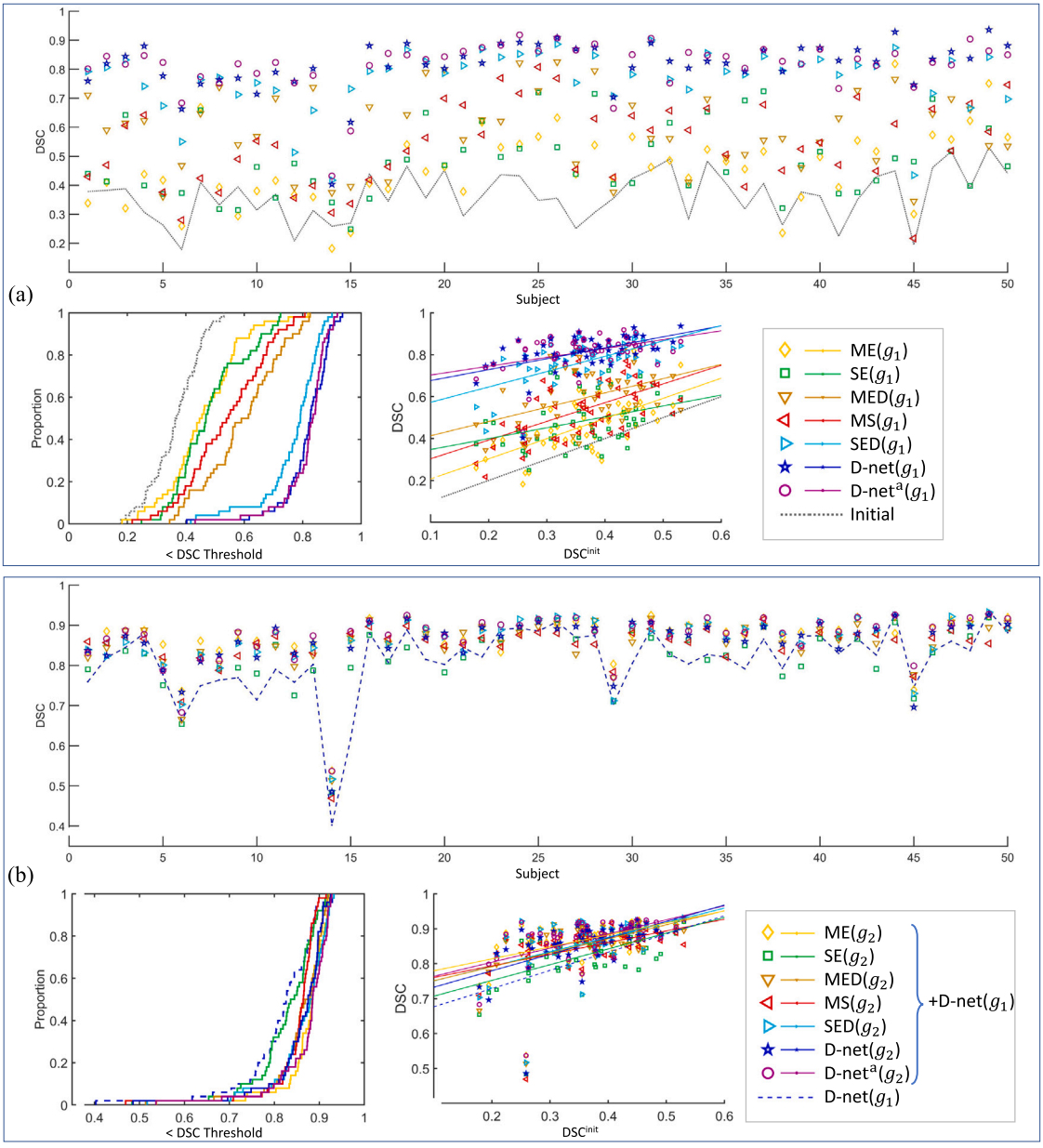

Fig. 9. Real test results of single and multi-stage D-net. The DSC results of the 50-case real test, including the empirical cumulative distribution functions, and the DSC with trend line with respect to the varying initial DSC for one-stage networks trained with large-range transformation (a) and for two-stage networks (b) confirm the consistent conclusion of synthetic test that D-nets achieve the best performance in the LRT setting and also one of the best performance as the second stage network in a two-stage network.

Table 5
Average values of Translation Error (TE/μm), Rotation Error (RE/°) and Dice Similarity Coefficient (DSC/%) in Synthetic test (Syn) for two-stage networks $g_2 \oplus g_1$ (g_1 cascaded with g_2) over five-fold cross-validation.

g_2 ($\oplus g_1$:	D-net)	ME	SE	MED	SED	MS	D-net	D-net^a
TE	72.8	30.5	24.3	35.3	21.5	44.2	15.8	**15.0**
RE	3.3	1.2	3.1	2.5	1.6	3.1	1.4	**0.9**
DSC	81.1	92.1	88.2	88.1	92.2	85.0	93.6	**95.2**

The quantitative results of the Synthetic dataset are presented in this section for the two-stage network of each network with the SRT-setting listed above cascaded to the LRT-trained D-net. For each pair of volumes with and without contrast, TE, RE and DCS were calculated with results shown in Fig. 8 and Table 5.

Fig. 8(a) illustrates that D-net^a+D-net (the LRT-trained D-net cascaded with SRT-trained D-net^a) obtained the lowest average RE, the highest average DSC, and the comparable TE with D-net+D-net in the varying range of initial spatial transformation. Fig. 8(b) and (c) present all the two-stage networks outperform one-stage D-net(g_1) on RE, TE and DSC, with, as shown in Fig. 8(b), $RE < 5°$ and sub-voxel TE in over 97% cases and in all cases for the best one, D-net^a+D-net. In addition, it is presented in Fig. 8(c) that all the two-stage networks performed more robustly than the one-stage D-net with varying initial spatial transformation and all achieved the sub-voxel average TE in the LRT setting. Those quantitative results of average TE, RE, and DSC are shown in Table 6. Using ANOVA for the statistic test, it is found that all two-stage networks outperform the one-stage D-net(g_1), and D-net^a+D-net is the best one.

4.2.2. Results in real test

Furthermore, the results of the real test for each two-stage network as shown in Fig. 9(b) and Fig. 10 yield a consistent conclusion in the synthetic results, including DSC plotted for each case, the empirical

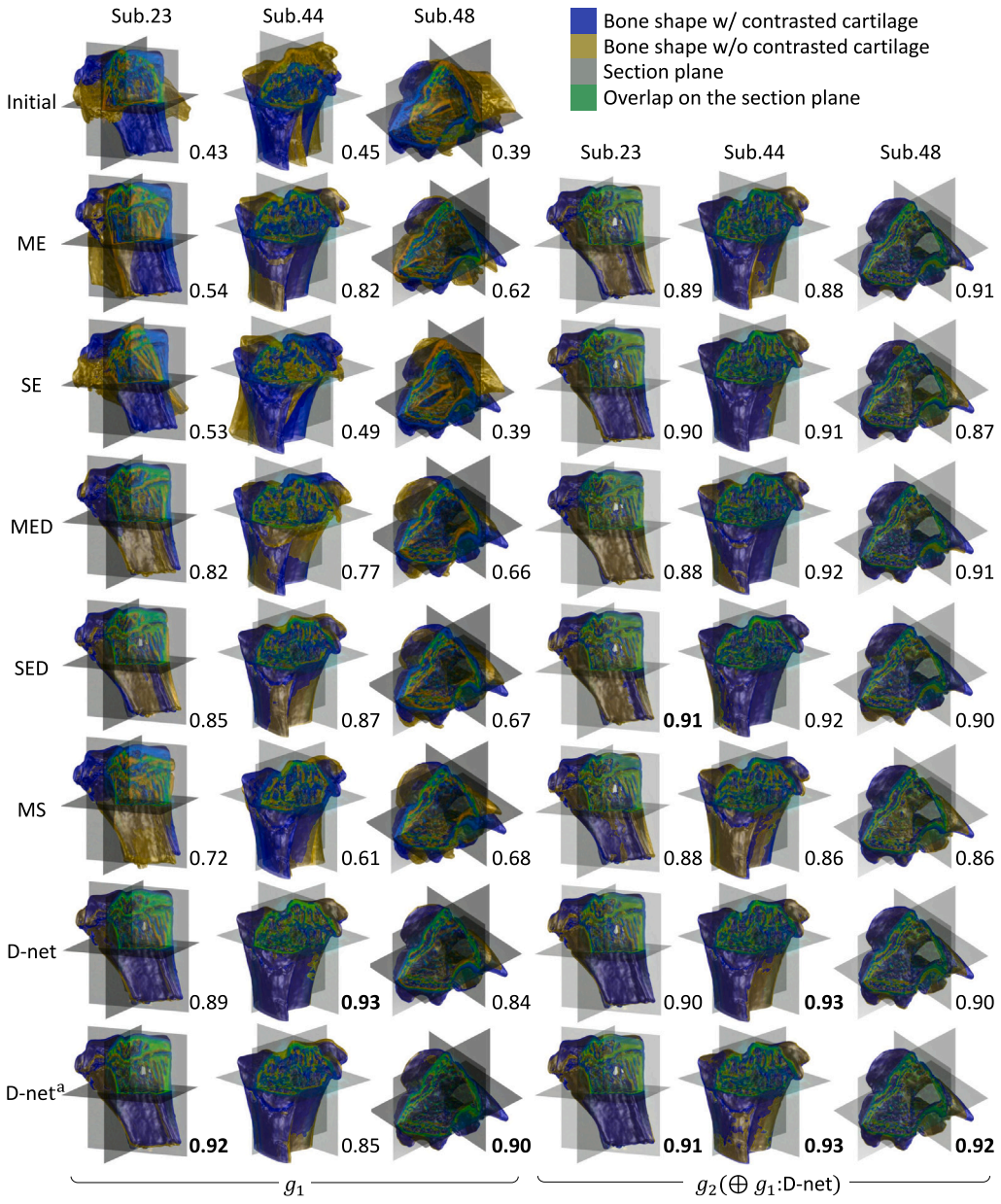

Fig. 10. D-net superiority in single and multi-stage exemplar alignments of volumes used in the real experiment. Pre- and post-contrast volumes of Subject Id 23, 44, and 48 are shown aligned by the (g_1) LRT-trained networks and by the ($g_2 \oplus g_1$) cascaded networks. DSC values are indicated in the bottom right corner of each alignment.

Table 6
Average and standard deviation values (avg±std) of Dice Similarity Coefficient (DSC/%) for one-stage networks g_1 and two-stage networks $g_2 \oplus g_1$ in the Real test over five-fold cross-validation.

g_1	Init	ME	SE	MED	SED	MS	D-net	D-net[a]
avg	36.4	46.1	48.5	59.3	76.6	53.9	**81.4**	**81.4**
std	±8.4	±12.9	±12.2	±13.3	±10.5	±13.9	±8.8	±8.4
g_2 ($\oplus g_1$:	D-net)	ME	SE	MED	SED	MS	D-net	D-net[a]
avg	81.4	**87.0**	82.6	85.7	86.1	84.9	85.7	**87.1**
std	±8.8	±6.2	±7.6	±6.8	±7.2	±6.6	±7.2	±6.7

Table 7
Regression Coefficient (RC) of DSC with respect to initial DSC and the output of g_1 in the real test respectively for one-stage networks g_1 and g_2 as well as the RC of DSC with respect to initial DSC for two-stage networks $g_2 \oplus g_1$(g_1 cascaded with g_2) over five-fold cross-validation.

	ME	SE	MED	SED	MS	D-net	D-net[a]
g_1	0.91	0.49	0.63	0.65	0.85	0.41	**0.32**
g_2	0.41	0.68	0.49	0.57	**0.32**	0.53	0.52
$g_2(\oplus g_1$: D-net)	0.25	0.36	0.30	0.32	**0.22**	0.37	0.31

cumulative distribution function with the approximating shape to that in the synthetic test (Fig. 8(b)), and the linearly regressed trend line for each network. D-net[a]+D-net obtained DSC higher than 80% on over 90% of cases. Those quantitative values of average DSC are shown in Table 5, and the RC with respect to initial DSC is shown in row 3 of Table 7. The ANOVA results with hypothesis rejection by $p < 1\%$ show that D-net[a]+D-net significantly outperforms others except ME+D-net, and all two-stage networks except SE+D-net outperform one-stage D-net(g_1). They present the improvement of multi-staging in volume

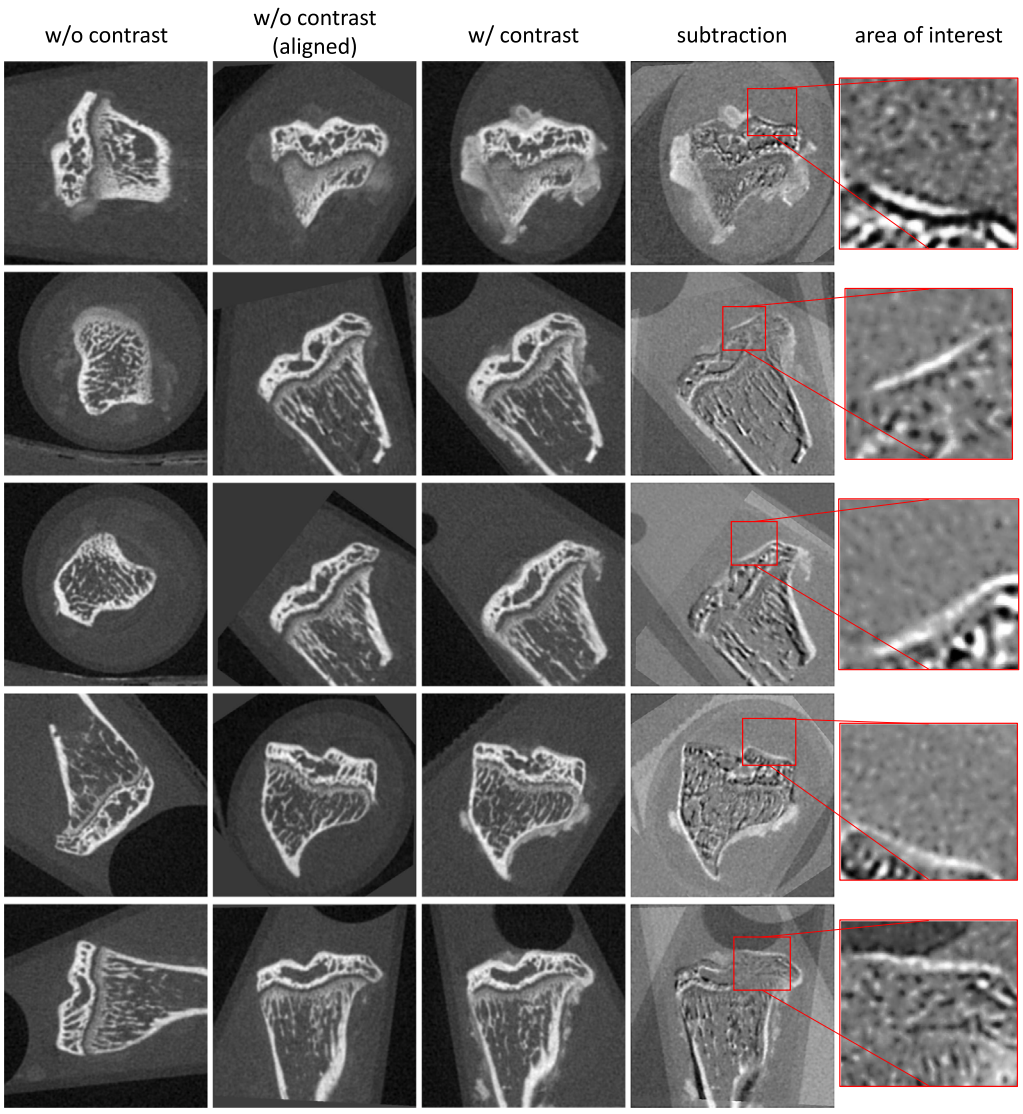

Fig. 11. Five examples of the CT images, with (w/) and without (w/o) contrast as well as the subtracted images (subtraction) after alignment via two-stage D-net^{a}, shows the cartilage area highlighted after the accurate alignment.

alignment and the superiority of using D-net^{a} as the fine-refinement network.

5. Discussion

In this paper, an annotation-free deep learning framework is proposed for the accurate and automatic volume alignment of pre- and post-contrasted cartilage CT scans of tibial bones through the estimation of the spatial transformation without standard position and orientation. It includes a multi-stage neural network, which contains the novel D-net. Deep learning was required as classical image registration methods explored previously were insufficient (Zheng et al., 2020b). The proposed two-stage framework achieved an average TE of 15 μm and RE of 1° in synthetic tests. In the real test on matching individual pre- with post-contrasted scans, an average DSC of 87% was achieved, with an inference time of ≤0.2 s per pair of volumes. The example pairs of CT images with and without contrast enhancement before and after alignment are shown in Fig. 11. The subtracted figures show the cartilage area (bright region) is highlighted and could be distinguished after the accurate alignment.

In terms of the improvement of accuracy and robustness in network structures with large-range transform, the superiorities of using a decoder part, and adding MNL are respectively validated by the over 0.3 DSC increment of the SED compared with SE, as well as the over 0.1 incremental DSC by D-nets compared with SED. D-nets are shown outperforming all other networks with comparable or fewer trainable variables in the LRT setting, and are among the top performing networks in the more clinically applicable SRT setting. D-net could also be applicable to other parametric registration (i.e. affine registration) and even non-parametric registration (i.e. deformable registration) without the usual requirement of initial alignment.

In addition, despite the MNL used between the two-branch encoder part of D-net in this paper for global-range links of similar inter-branch features, it can be also employed in other multi-branch network structures with different purposes (Kwon et al., 2019; Dunnhofer et al., 2020), such as sharing a common latent space with varying mapping for multi-task learning.

Furthermore, the improvement of Atrous convolution in terms of training efficiency is presented by the training speed boost of D-net^{a}(g_1) from D-net(g_1) with the LRT setting. This training speed boost was not observed in the SRT setting, showing that in D-Net, the SRT setting might not benefit from as large a receptive field as the LRT setting. The longer distance between every two corresponding voxels in the LRT setting plays into the larger receptive field provided by the Atrous

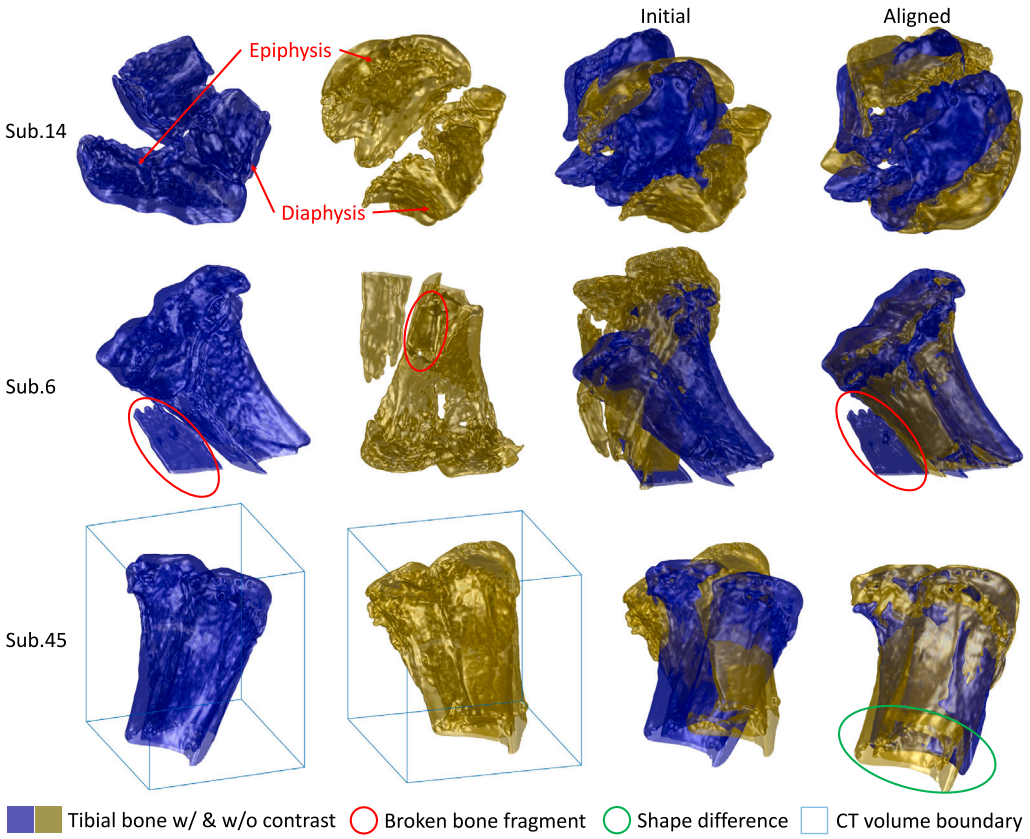

Fig. 12. Segmented surfaces for the exemplar tibial bones causing problems in the real test experiment, with subject 14 due to the epiphysis and the diaphysis separated with the epiphysial plate broken, subject 6 due to fragments broken from tibial bone, and subject 2 due to the shape difference caused by the CT boundary.

convolution. This validates the effect of the receptive field on this task and potentially other spatial information-based tasks. That is also one reason to recommend D-net^a(g_2) for more time efficiency.

The two-stage network D-net^a(g_2)+D-net(g_1), which is the combination of the best one-stage networks with LRT g_1 and SRT g_2 achieves the best overall accuracy with a DSC of 0.87 as shown in Tables 3, 4. Multi-staging of D-net(g_1) with any of the SRT g_2 networks was also superior compared to just D-net(g_1) as shown in Table 7 in terms of accuracy and robustness.

In terms of the application perspective, there are also several special situations concerned when we used and evaluated our framework in real cases, which are illustrated in Fig. 12. It is notable in Fig. 9 that all the one- and two-stage networks failed on the subject (sub.) 14, resulting from the separation of the epiphysis and diaphysis parts with the epiphysial plate broken as shown in Fig. 12. This shape has never been observed in the others, and thus never been learned by networks. The bone could be broken during the cutting process. Another similar situation of broken bone is also observed in sub. 6 and sub. 29 but the rough shape of the main body is well preserved. Therefore the results show that the framework is still able to deal with it although not as well as in other cases due to the multi-body rigid transformation caused by fragments. These two problems explain the cause of the outliers results in the real test and also one potential reason that the results of the real test are not as good as the synthetic test, which could be avoided during the data sampling. In addition, the shape difference is also found in sub. 45, because the portion of the sample is out of the field of view in the scan. It has decreased the superior boundary of the evaluation based on volume overlap via DSC but should not really affect the piratical use since the cartilage part is always kept in the field of view.

In the future work, the morphological analysis will be explored based on this framework for further osteoarthritis diagnosis and assessment based on the cartilage shapes. In terms of the technique proposed in this paper, D-net could potentially be adopted for clinical or pre-clinical registration tasks in other modalities (e.g. ultrasound) and other organs (e.g. fetal brain). Although this work mainly focuses on volume alignment with the LRT setting problem, our proposed D-net^a also achieves the best alignment in the SRT setting, which is close to the clinical setting with standard protocols. Thus, it could be further improved and extended for more biomedical applications.

6. Conclusion

Overall, we proposed a multi-stage deep learning framework based on a new Siamese-based network, D-net, for the accurate volume alignment of pre- and post-contrasted CT scans of tibial bones toward cartilage shape extraction. As far as the authors' knowledge, it is also the first implementation of volume alignment with a full capture range of rotation and without annotation or poses template. The optimization of the components in this framework, network structures, and multi-staging, are validated, showing that D-net outperforms the current state-of-the-art deep learning alignment methods with 87% DSC, and that the framework performs a plausible alignment of the contrasted and non-contrasted CT scans of tibial bones. In the future, we will improve the accuracy by focusing on the failure cases resulting from the bone fragments and the deviation from the limited field of view.

CRediT authorship contribution statement

Jian-Qing Zheng: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Visualization. **Ngee Han Lim:** Conceptualization, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Bartłomiej W. Papież:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jian-Qing Zheng, Ngee Han Lim, Bartlomiej Papiez has patent Neural network for cartilage thickness quantification pending to Intellectual Property Office UK. N.H.L. is a named inventor on a patent for radiopaque compounds containing diiodotyrosine (WO2018020262A1, EP3490614A1), the analysis of which would benefit from this work.

Data availability

The data that has been used is confidential.

Acknowledgments

This work was supported by a Kennedy Trust for Rheumatology Research Studentship, UK, the Centre for OA Pathogenesis Versus Arthritis (Versus Arthritis grant 21621) and the Rutherford Fund at Health Data Research UK. The authors acknowledge Patricia das Neves Borges as the researcher who collected the pre-clinical CT dataset, as part of the National Centre for Replacement, Refinement, and Reduction of Animals in Research, UK (NC3R grant NC/M000141/1).

Appendix A. Model detail

See Table A.8.

Appendix B. Model efficiency

See Table B.9.

Table A.8
Network structure detail for D-net and D-net[a], where ker denotes kernel size, dila denotes dilation rate, conv denotes convolution, deconv denotes deconvolution or transpose convolution, norm denotes the batch normalization, and act denotes the leaky ReLU activation function with 0.01 negative slope.

Block	Layer (s)	ker	dila		Channels	In	Out
			Dnet	Dnet[a]			
Res-down or Atrous Res-down	conv,norm,act	3	1	1	1/16	$\boldsymbol{I}^{f}$ $\boldsymbol{I}^{m}$	$r1^{f}$ $r1^{m}$
	conv,norm,act	3	1	3	16/16	$r1^{f}$ $r1^{m}$	$f1^{f}$ $f1^{m}$
	conv,norm,act	3	1	9	16/16	$f1^{f}$ $f1^{m}$	$f1^{f}$ $f1^{m}$
	conv,norm	3	1	1	16/16	$f1^{f}$ $f1^{m}$	$f1^{f}$ $f1^{m}$
	act	–	–	–	–	$f1^{f}$ + $r1^{f}$ $f1^{m}$ + $r1^{m}$	$s1^{f}$ $s1^{m}$
	downsample	–	–	–	–	$s1^{f}$ $s1^{m}$	$s1^{f}$ $s1^{m}$
Res-down or Atrous Res-down	conv,norm,act	3	1	1	16/32	$s1^{f}$ $s1^{m}$	$r2^{f}$ $r2^{m}$
	conv,norm,act	3	1	3	32/32	$r2^{f}$ $r2^{m}$	$f2^{f}$ $f2^{m}$
	conv,norm,act	3	1	9	32/32	$f2^{f}$ $f2^{m}$	$f2^{f}$ $f2^{m}$
	conv,norm	3	1	1	32/32	$f2^{f}$ $f2^{m}$	$f2^{f}$ $f2^{m}$
	act	–	–	–	–	$f2^{f}$ + $r2^{f}$ $f2^{m}$ + $r2^{m}$	$s2^{f}$ $s2^{m}$
	downsample	–	–	–	–	$s2^{f}$ $s2^{m}$	$s2^{f}$ $s2^{m}$
MNL Res-down	conv,norm,act	3	1	1	16/32	$s2^{f}$ $s2^{m}$	$r3^{f}$ $r3^{m}$
	conv,norm,act	3	1	1	32/32	$r3^{f}$ $r3^{m}$	$f3^{f}$ $f3^{m}$
	conv	1	1	1	32/32	$f3^{f}$ $f3^{m}$	$a3^{f}$ $a3^{m}$
	mut-attn,conv	1	1	1	32/32	$a3^{m} \otimes \Phi(a3^{m},a3^{f})$ $a3^{f} \otimes \Phi(a3^{f},a3^{m})$	$a3^{f}$ $a3^{m}$
	conv,norm	3	1	1	32/32	$f3^{f}$ + $a3^{f}$ $f3^{m}$ + $a3^{m}$	$f3^{f}$ $f3^{m}$
	act	–	–	–	–	$f3^{f}$ + $r3^{f}$ $f3^{m}$ + $r3^{m}$	$s3^{f}$ $s3^{m}$
	downsample	–	–	–	–	$s3^{f}$ $s3^{m}$	$s3^{f}$ $s3^{m}$

(*continued on next page*)

Table A.8 (*continued*).

Block	Layer (s)	ker	dila		Channels	In	Out
			Dnet	Dnet[a]			
MNL Res-down	conv,norm,act	3	1	1	32/64	$s3^f$ $s3^m$	$r4^f$ $r4^m$
	conv,norm,act	3	1	1	64/64	$r4^f$ $r4^m$	$f4^f$ $f4^m$
	conv	1	1	1	64/64	$f4^f$ $f4^m$	$a4^f$ $a4^m$
	mut-attn,conv	1	1	1	64/64	$a4^m \otimes \Phi(a4^m,a4^f)$ $a4^f \otimes \Phi(a4^f,a4^m)$	$a4^f$ $a4^m$
	conv,norm	3	1	1	64/64	$f4^f + a4^f$ $f4^m + a4^m$	$f4^f$ $f4^m$
	act	–	–	–	–	$f4^f + r4^f$ $f4^m + r4^m$	$s4^f$ $s4^m$
	downsample	–	–	–	–	$s4^f$ $s4^m$	$s4^f$ $s4^m$
MNL Res-down	conv,norm,act	3	1	1	32/64	$s4^f$ $s4^m$	$r5^f$ $r5^m$
	conv,norm,act	3	1	1	64/64	$r5^f$ $r5^m$	$f5^f$ $f5^m$
	conv	1	1	1	64/64	$f5^f$ $f5^m$	$a5^f$ $a5^m$
	mut-attn,conv	1	1	1	64/64	$a5^m \otimes \Phi(a5^m,a5^f)$ $a5^f \otimes \Phi(a5^f,a5^m)$	$a5^f$ $a5^m$
	conv,norm	3	1	1	64/64	$f5^f + a5^f$ $f5^m + a5^m$	$f5^f$ $f5^m$
	act	–	–	–	–	$f5^f + r5^f$ $f5^m + r5^m$	$s5^f$ $s5^m$
	downsample	–	–	–	–	$s5^f$ $s5^m$	$s5^f$ $s5^m$
MNL Res-down	conv,norm,act	3	1	1	32/64	$s5^f$ $s5^m$	$r6^f$ $r6^m$
	conv,norm,act	3	1	1	64/64	$r6^f$ $r6^m$	$f6^f$ $f6^m$
	conv	1	1	1	64/64	$f6^f$ $f6^m$	$a6^f$ $a6^m$
	mut-attn,conv	1	1	1	64/64	$a6^m \otimes \Phi(a6^m,a6^f)$ $a6^f \otimes \Phi(a6^f,a6^m)$	$a6^f$ $a6^m$
	conv,norm	3	1	1	64/64	$f6^f + a6^f$ $f6^m + a6^m$	$f6^f$ $f6^m$
	act	–	–	–	–	$f6^f + r6^f$ $f6^m + r6^m$	$s6^f$ $s6^m$
	downsample	–	–	–	–	$s6^f$ $s6^m$	$s6^f$ $s6^m$
	conv	1	1	1	128/64	$s6^f \Vert s6^m$	s6
Res-up	deconv,act	1	–	–	64/64	s6	r7
	conv,norm	3	1	1	64/64	r7	f7
	act	–	–	–	–	$f7 + f6^f$	f7
	conv,norm,act	3	1	1	64/64	f7	f7
	conv,norm	3	1	1	64/64	f7	f7
	act	–	–	–	–	f7 + r7	s7
Res-up	deconv,act	1	–	–	64/64	s7	r8
	conv,norm	3	1	1	64/64	r8	f8
	act	–	–	–	–	$f8 + f5^f$	f8
	conv,norm,act	3	1	1	64/64	f8	f8
	conv,norm	3	1	1	64/64	f8	f8
	act	–	–	–	–	f8 + r8	s8
Res-up	deconv,act	1	–	–	64/64	s8	r9
	conv,norm	3	1	1	64/64	r9	f9
	act	–	–	–	–	$f9 + f4^f$	f9
	conv,norm,act	3	1	1	64/64	f9	f9
	conv,norm	3	1	1	64/64	f9	f9
	act	–	–	–	–	f9 + r9	s9

(*continued on next page*)

Table A.8 (*continued*).

Block	Layer (s)	ker	dila		Channels	In	Out
			Dnet	Dnet[a]			
Res-up	deconv,act	1	–	–	64/64	s9	r10
	conv,norm	3	1	1	64/64	r10	f10
	act	–	–	–	–	f10 + f3[f]	f9
	conv,norm,act	3	1	1	64/64	f10	f10
	conv,norm	3	1	1	64/64	f10	f10
	act	–	–	–	–	f10 + r10	s10
	conv	1	1	1	64/3	s10	s10
	dense,act	–	–	–	$\frac{3DHW}{64}$/128	s10	s10
	dense	–	–	–	128/9	s10	$(\hat{\theta}, \hat{t})$

Table B.9
The time and computation efficiency of each model is evaluated with the number of Parameters (#Par), Float Operations (FLOPs) and the time cost per image pair (Time/Img).

	ME	SE	MED	SED	MS	D-net	D-net[a]
#Par (10^6)	10.28	12.05	4.85	4.54	5.24	4.95	4.95
FLOPs (10^9)	23.26	46.06	23.05	44.55	26.36	48.24	48.24
Time/Img	0.08 s	0.07 s	0.12 s	0.11 s	0.13 s	0.14 s	0.15 s

References

Ambellan, F., Tack, A., Ehlke, M., Zachow, S., 2019. Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative. Med. Image Anal. 52, 109–118.

Baiker, M., Staring, M., Löwik, C.W., Reiber, J.H., Lelieveldt, B.P., 2011. Automated registration of whole-body follow-up MicroCT data of mice. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 516–523.

Borges, P.D.N., Forte, A., Vincent, T., Dini, D., Marenzana, M., 2014. Rapid, automated imaging of mouse articular cartilage by microCT for early detection of osteoarthritis and finite element modelling of joint mechanics. Osteoarthr. Cartil. 22 (10), 1419–1428.

Burton II, W., Myers, C., Rullkoetter, P., 2020. Semi-supervised learning for automatic segmentation of the knee from MRI with convolutional neural networks. Comput. Methods Programs Biomed. 189, 105328.

Chee, E., Wu, Z., 2018. Airnet: Self-supervised affine registration for 3d medical images using neural networks. arXiv preprint arXiv:1810.02583.

Chow, P.L., Stout, D.B., Komisopoulou, E., Chatziioannou, A.F., 2006. A method of image registration for small animal, multi-modality imaging. Phys. Med. Biol. 51 (2), 379.

Dunnhofer, M., Antico, M., Sasazawa, F., Takeda, Y., Camps, S., Martinel, N., Micheloni, C., Carneiro, G., Fontanarosa, D., 2020. Siam-U-Net: encoder-decoder siamese network for knee cartilage tracking in ultrasound images. Med. Image Anal. 60, 101631.

Filip, K., Zacharaki, E.I., Moustakas, K., 2021. Regularized multi-structural shape modeling of the knee complex based on deep functional maps. Comput. Med. Imaging Graph. 89, 101890.

Fowkes, M.M., Das Neves Borges, P., Cacho-Nerin, F., Brennan, P.E., Vincent, T.L., Lim, N.H., 2022. Imaging articular cartilage in osteoarthritis using targeted peptide radiocontrast agents. Plos One 17 (5), e0268223.

Gangwar, T., Calder, J., Takahashi, T., Bechtold, J.E., Schillinger, D., 2018. Robust variational segmentation of 3D bone CT data with thin cartilage interfaces. Med. Image Anal. 47, 95–110.

Grau, V., Mewes, A., Alcaniz, M., Kikinis, R., Warfield, S.K., 2004. Improved watershed transform for medical image segmentation using prior information. IEEE Trans. Med. Imaging 23 (4), 447–458.

Haskins, G., Kruger, U., Yan, P., 2020. Deep learning in medical image registration: a survey. Mach. Vis. Appl. 31 (1), 8.

Heinrich, M.P., Jenkinson, M., Papież, B.W., Glesson, F.V., Brady, M., Schnabel, J.A., 2013. Edge-and detail-preserving sparse image representations for deformable registration of chest MRI and CT volumes. In: International Conference on Information Processing in Medical Imaging. pp. 463–474.

Hosnijeh, F.S., Bierma-Zeinstra, S., Bay-Jensen, A., 2019. Osteoarthritis year in review 2018: biomarkers (biochemical markers). Osteoarthr. Cartil. 27 (3), 412–423.

Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C.M., Emberton, M., et al., 2018. Weakly-supervised convolutional neural networks for multimodal image registration. Med. Image Anal. 49, 1–13.

Islam, M.A., Jia, S., Bruce, N.D., 2019. How much position information do convolutional neural networks encode? In: International Conference on Learning Representations.

James, M.L., Gambhir, S.S., 2012. A molecular imaging primer: modalities, imaging agents, and applications. Physiol. Rev. 92 (2), 897–965.

Kraiger, M., Martirosian, P., Opriessnig, P., Eibofner, F., Rempp, H., Hofer, M., Schick, F., Stollberger, R., 2012. A fully automated trabecular bone structural analysis tool based on T2*-weighted magnetic resonance imaging. Comput. Med. Imaging Graph. 36 (2), 85–94.

Kwon, D., Ahn, J., Kim, J., Choi, I., Jeong, S., Lee, Y.-S., Park, J., Lee, M., 2019. Siamese U-Net with healthy template for accurate segmentation of intracranial hemorrhage. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 848–855.

Liao, R., Miao, S., de Tournemire, P., Grbic, S., Kamen, A., Mansi, T., Comaniciu, D., 2017. An artificial agent for robust image registration. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31.

Ma, K., Wang, J., Singh, V., Tamersoy, B., Chang, Y.-J., Wimmer, A., Chen, T., 2017. Multimodal image registration with deep context reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 240–248.

Maier, J., Monroy, L.C.R., Syben, C., Jeon, Y., Choi, J.-H., Hall, M.E., Levenston, M., Gold, G., Fahrig, R., Maier, A., 2020. Multi-channel volumetric neural network for knee cartilage segmentation in cone-beam CT. In: Bildverarbeitung für die Medizin 2020. Springer, pp. 67–72.

Moser, F., Huang, R., Papież, B.W., Namburete, A.I., st Consortium, I., et al., 2022. BEAN: Brain extraction and alignment network for 3D fetal neurosonography. NeuroImage 119341.

Myller, K.A., Honkanen, J.T., Jurvelin, J.S., Saarakkala, S., Töyräs, J., Väänänen, S.P., 2018. Method for segmentation of knee articular cartilages based on contrast-enhanced CT images. Ann. Biomed. Eng. 46 (11), 1756–1767.

Namburete, A.I., Xie, W., Yaqub, M., Zisserman, A., Noble, J.A., 2018. Fully-automated alignment of 3D fetal brain ultrasound to a canonical reference space using multi-task learning. Med. Image Anal. 46, 1–14.

Ou, Y., Sotiras, A., Paragios, N., Davatzikos, C., 2011. DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting. Med. Image Anal. 15 (4), 622–639.

Paniagua, B., Cevidanes, L., Walker, D., Zhu, H., Guo, R., Styner, M., 2011. Clinical application of SPHARM-PDM to quantify temporomandibular joint osteoarthritis. Comput. Med. Imaging Graph. 35 (5), 345–352.

Papież, B.W., Szmul, A., Grau, V., Brady, J.M., Schnabel, J.A., 2016. Non-local graph-based regularization for deformable image registration. In: Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging. Springer, pp. 199–207.

Robu, M.R., Ramalhinho, J., Thompson, S., Gurusamy, K., Davidson, B., Hawkes, D., Stoyanov, D., Clarkson, M.J., 2018. Global rigid registration of CT to video in laparoscopic liver surgery. Int. J. Comput. Assist. Radiol. Surg. 13 (6), 947–956.

Salehi, S.S.M., Khan, S., Erdogmus, D., Gholipour, A., 2018. Real-time deep pose estimation with geodesic loss for image-to-template rigid registration. IEEE Trans. Med. Imaging 38 (2), 470–481.

Shan, L., Zach, C., Charles, C., Niethammer, M., 2014. Automatic atlas-based three-label cartilage segmentation from MR knee images. Med. Image Anal. 18 (7), 1233–1246.

Sloan, J.M., Goatman, K.A., Siebert, J.P., 2018. Learning rigid image registration-utilizing convolutional neural networks for medical image registration. In: 11th International Joint Conference on Biomedical Engineering Systems and Technologies,. SCITEPRESS-Science and Technology Publications, pp. 89–99.

Urish, K.L., Williams, A.A., Durkin, J.R., Chu, C.R., Group, O.I., 2013. Registration of magnetic resonance image series for knee articular cartilage analysis: data from the osteoarthritis initiative. Cartilage 4 (1), 20–27.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30, 5998–6008.

Vincent, T.L., 2020. Of mice and men: converging on a common molecular understanding of osteoarthritis. Lancet Rheumatol. 2 (10), e633–e645.

de Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I., 2019. A deep learning framework for unsupervised affine and deformable image registration. Med. Image Anal. 52, 128–143.

Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803.

Yoo, J.H., Kim, S.K., Hong, H., Shim, H., Kwoh, C.K., Bae, K.T., 2009. Automatic bone registration in MR knee images for cartilage morphological analysis. In: Medical Imaging 2009: Image Processing, Vol. 7259. pp. 815–822.

Zheng, J.-Q., Lim, N.H., Papież, B.W., 2020b. D-net: Siamese based network for arbitrarily oriented volume alignment. In: International Workshop on Shape in Medical Imaging. pp. 73–84.

Zheng, H., Perrine, S.M.M., Pitirri, M.K., Kawasaki, K., Wang, C., Richtsmeier, J.T., Chen, D.Z., 2020a. Cartilage segmentation in high-resolution 3D Micro-CT images via uncertainty-guided self-training with very sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 802–812.

Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H., 2019. On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5745–5753.

Zhou, X.-Y., Zheng, J.-Q., Li, P., Yang, G.-Z., 2020. ACNN: a full resolution DCNN for medical image segmentation. In: 2020 IEEE International Conference on Robotics and Automation. ICRA, pp. 8455–8461.

Zhu, Y., Newsam, S., 2017. Densenet for dense flow. In: 2017 IEEE International Conference on Image Processing. ICIP, pp. 790–794.