Medical Image Analysis 78 (2022) 102383

Contents lists available at ScienceDirect

Medical Image Analysis

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

Atlas-ISTN: Joint segmentation, registration and atlas construction with image-and-spatial transformer networks

Matthew Sinclair[a,b,*], Andreas Schuh[a,b], Karl Hahn[a,b], Kersten Petersen[a], Ying Bai[a], James Batten[a,b], Michiel Schaap[a,b], Ben Glocker[a,b]

[a] HeartFlow, Inc., USA
[b] Biomedical Image Analysis Group, Imperial College London, UK

ARTICLE INFO

Article history:
Received 12 March 2021
Revised 24 November 2021
Accepted 1 February 2022
Available online 10 February 2022

Keywords:
Multi-label atlas construction
Image segmentation and registration

ABSTRACT

Deep learning models for semantic segmentation are able to learn powerful representations for pixel-wise predictions, but are sensitive to noise at test time and may lead to implausible topologies. Image registration models on the other hand are able to warp known topologies to target images as a means of segmentation, but typically require large amounts of training data, and have not widely been benchmarked against pixel-wise segmentation models. We propose the Atlas Image-and-Spatial Transformer Network (Atlas-ISTN), a framework that jointly learns segmentation and registration on 2D and 3D image data, and constructs a population-derived atlas in the process. Atlas-ISTN learns to segment multiple structures of interest and to register the constructed atlas labelmap to an intermediate pixel-wise segmentation. Additionally, Atlas-ISTN allows for test time refinement of the model's parameters to optimize the alignment of the atlas labelmap to an intermediate pixel-wise segmentation. This process both mitigates for noise in the target image that can result in spurious pixel-wise predictions, as well as improves upon the one-pass prediction of the model. Benefits of the Atlas-ISTN framework are demonstrated qualitatively and quantitatively on 2D synthetic data and 3D cardiac computed tomography and brain magnetic resonance image data, out-performing both segmentation and registration baseline models. Atlas-ISTN also provides inter-subject correspondence of the structures of interest.

1. Introduction

Motivation Image segmentation and registration have long been important tools for biomedical image analysis (Maintz and Viergever, 1998; Pham et al., 2000). Deep learning models such as U-nets (Ronneberger et al., 2015) have emerged as the state-of-the-art for segmentation, with their ability to learn rich feature representations for accurate pixel-wise[1] segmentations on challenging image datasets when trained with large labelled sets of 2D and 3D images. One challenge with such segmentation models however is their sensitivity to image noise and artefacts, which can yield spurious and topologically implausible segmentations at test time. Furthermore, such undesirable predictions are made more likely with fewer training examples. Numerous recent works have sought to tackle this with post-processing (Kamnitsas et al., 2017; Larrazabal et al., 2019), anatomical constraints in training (Oktay et al., 2018), and novel regularizers or loss terms (Xu and Niethammer, 2019; Clough et al., 2019), to name a few. While such approaches have improved topological plausibility of pixel-wise model predictions, they do not ensure a consistent topology.

More recently, a growing body of research has explored the use of deep learning models for image registration for the purpose of image segmentation. Deep learning registration models typically learn to predict a dense deformation field to register a pair of images, which can be used to propagate a labelmap (of known topology) from a source to a target image for the purpose of segmentation. Most such methods rely on a single pass of a trained model at test time to predict a deformation field (de Vos et al., 2019; Dalca et al., 2018; Balakrishnan et al., 2018; Dalca et al., 2019a; 2019c; Dong et al., 2020; Mansilla et al., 2020). The accuracy of a warped segmentation to a target image has been shown to improve with test time optimization of the registration network's parameters, particularly in settings of limited training data (Balakrishnan et al., 2018; Lee et al., 2019b). Dalca et al. (2019a) proposed a framework to learn a conditional atlas image jointly with a model to

* Corresponding author at: Biomedical Image Analysis Group, Imperial College London, UK.
E-mail address: msinclair@heartflow.com (M. Sinclair).

[1] Note: 'pixel-wise' is used to refer to both 2D and 3D image settings.

https://doi.org/10.1016/j.media.2022.102383

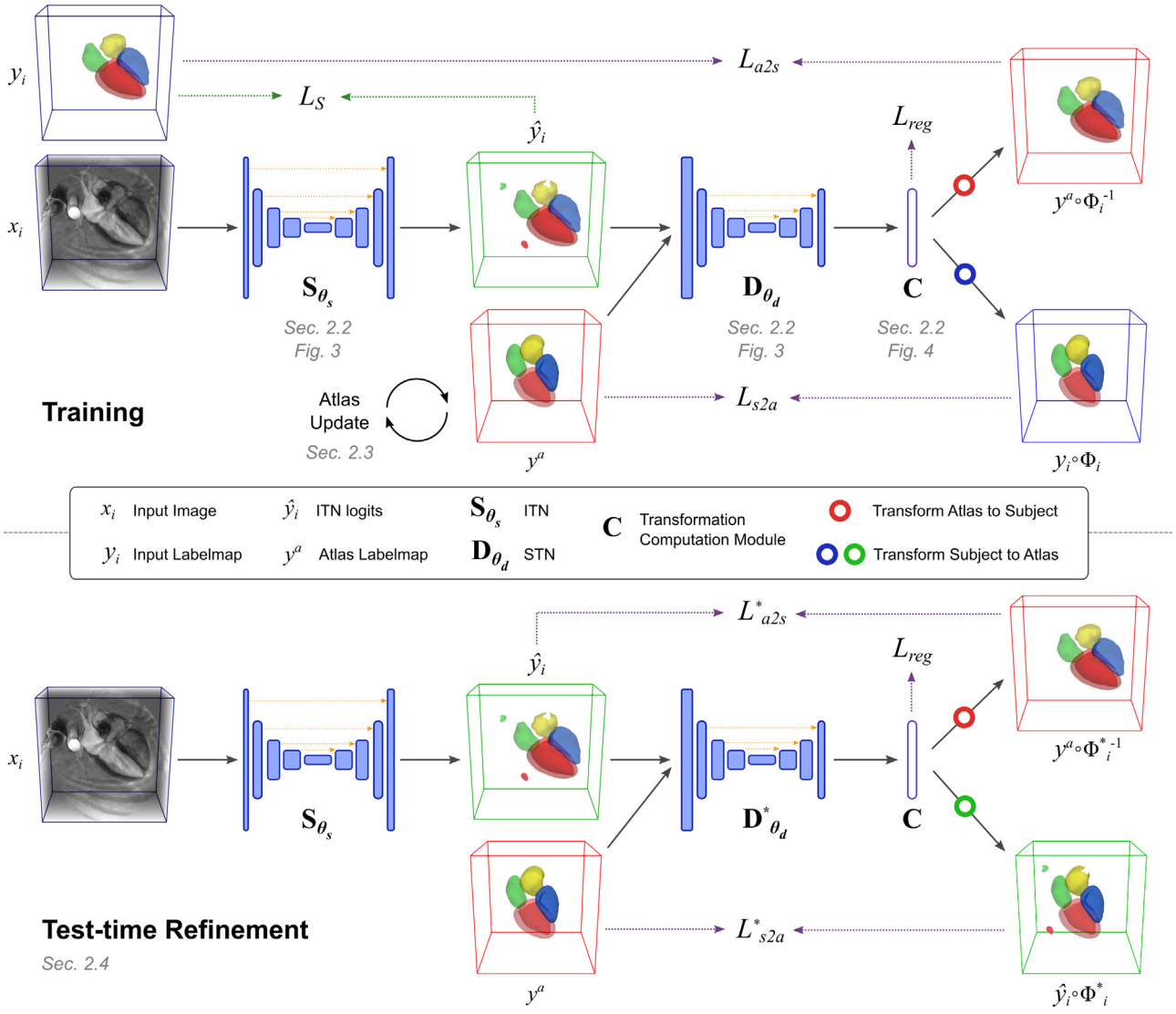

Fig. 1. An overview of the Atlas-ISTN framework in training (top) and test time refinement (bottom) illustrated with CCTA image data. TOP: A 3D image is passed to the Image Transformer Network (ITN) which predicts a voxel-wise segmentation (green box). The predicted segmentation and the atlas labelmap (red box) are passed to the Spatial Transformer Network (STN) which predicts a diffeomorphic transformation (where affine and non-rigid components are composed in C), used to warp the atlas labelmap to patient-space (red box in top right) and the patient ground-truth labelmap to atlas-space (blue box on the right). Both ITN and STN weights are optimized by leveraging ground-truth labelmaps (blue boxes) in computing a segmentation loss (L_s) and symmetric registration loss (L_{a2s} and L_{s2a}), while the atlas is updated at the end of each epoch. BOTTOM: At test time, the ITN predicts a voxel-wise segmentation which is passed along with the constructed atlas labelmap to the STN. Unlike during training, the ITN prediction is used as a target labelmap for instance-specific optimization of the STN (indicated by *), where the atlas labelmap is registered to the ITN prediction providing a refined transformation for a given subject. Spurious segmentations in the ITN prediction can be circumvented via the registration of a population-derived atlas labelmap to the ITN voxel-wise prediction. References to related sections and figures are provided in grey text. Red boxes: atlas labelmap (deformed and undeformed). Blue boxes: input image or corresponding ground-truth labelmap (deformed and undeformed). Green boxes: ITN prediction (deformed and undeformed).

perform registration of the atlas with target images. These recent image-registration driven approaches used for segmentation are however not commonly benchmarked against pixel-wise segmentation models, and have fallen short in relevant metrics when they have been (Lee et al., 2019b; Xu and Niethammer, 2019).

Inspired by features of each of these approaches, we propose Atlas-ISTN, a framework that benefits from the detailed predictions of pixel-wise segmentation while circumventing the effects of noise and artifacts via registration of a population-derived multi-class atlas labelmap constructed during training (Fig. 1, upper panel) to an initial pixel-wise segmentation. Additionally, Atlas-ISTN leverages test time refinement of model parameters to optimize for the registration of the atlas labelmap to a predicted segmentation, improving over single-pass test time performance (Fig. 1, lower panel). This framework also simultaneously preserves topology of the structures-of-interest (SoI) in the atlas and provides atlas-space correspondence between subjects for further population-level analysis, all while being contained in a single model framework with straight-forward training and test time deployment.

Related work. Medical image segmentation models can be broadly categorized into pixel-wise prediction, shape fitting and registration-based methods. A rich body of literature exists for methods which use image registration as a means to segment a target image, which build on two main approaches: multi-atlas segmentation (MAS) (Išgum et al., 2009; Kirişli et al., 2010; Iglesias and Sabuncu, 2015), and construction and registration of a statistical shape model (SSM) (Heimann and Meinzer, 2009; Young and Frangi, 2009). MAS has proven to be highly effective, and provides competitive performance with modern deep learning methods for

segmenting large structures of the heart from 3D cardiac computed tomographic angiography (CCTA) and magnetic resonance imaging (MRI), albeit in a setting with limited training data (Heinrich and Oster, 2018; Zhuang et al., 2019). MAS has also been used effectively for segmentation of heart structures in CCTA in a large-scale multicenter/multivendor evaluation (Kirişli et al., 2010). A downside of MAS is the high computational overhead at test time, which can involve registration, selection and sophisticated fusion of multiple atlas labelmaps to a target image to achieve best performance (Išgum et al., 2009).

Methods using SSMs on the other hand typically construct a template/atlas in the form of a mean image, labelmap, or mesh, which at test time is registered to a target image, with the option of leveraging a segmentation of the target image in the registration process (Medrano-Gracia et al., 2014; Bai et al., 2015). Advantages of SSMs include providing correspondence to a common atlas space for population-level analyses of shape and motion, and using a population-derived atlas tends to perform better for segmentation than warping a given training sample. For cardiac data, given the significant variability in heart orientation, size and morphology observed in CCTA and cardiac MRI, a two-stage (i) affine and (ii) non-rigid registration approach is typically used, sometimes requiring the definition of anatomical landmarks to guide the first stage of registration (Bai et al., 2015). SSMs are also often parameterized with Principal Component Analysis (PCA), and the PCA modes can be optimized to fit a SSM to a target image or segmentation (Heimann and Meinzer, 2009). Limitations of PCA representations however include over-fitting of the model to limited training data, thus not being able to accurately represent anatomies which lie outside of the training distribution and can include both significant (large-scale) and subtle (small-scale) variations in target anatomies. Both MAS and SSM-based segmentation approaches often involve workflows with multiple processing steps. For example, separate tools are typically used to optionally (1) build a SSM, (2) produce an image segmentation or landmark coordinates from an unseen target image, (3) register the atlas to a target image, segmentation, and/or landmarks (4) select and/or fuse the best atlas labelmap(s). See (Iglesias and Sabuncu, 2015) for a review of MAS and (Heimann and Meinzer, 2009; Young and Frangi, 2009) for a review of SSM works.

Prior to the emergence of powerful convolutional neural networks (CNNs) for pixel-wise segmentation (Ronneberger et al., 2015; Long et al., 2015), MAS and SSM-based methods were the dominant approaches used for biomedical image segmentation of structures known to adhere to a particular atlas geometry. An advantage of the more traditional approaches is the preservation of topology, where pixel-wise deep learning segmentation models such as the U-net (Ronneberger et al., 2015) or fully convolutional network (Long et al., 2015) can suffer from spurious and anatomically implausible segmentations at test time. Deep learning segmentation models typically improve with more training data, but are still prone to errors due to domain shift and out-of-distribution test cases (e.g. originating from different scanners, sites, acquisition protocols and caused by imaging artifacts). Among the many methods that have been proposed to tackle these challenges, we summarize a few that focus on reducing spurious segmentations and encourage anatomically plausible predictions. Post-processing steps have been commonly used, such as fully-connected conditional random fields (CRF) and connected component analysis (Kamnitsas et al., 2017), as well as shape-aware denoising auto-encoders (Larrazabal et al., 2019). Other approaches include anatomically constrained neural networks (ACNNs) where shape-regularization is enforced with a latent representation of the underlying anatomy via an autoencoder (Oktay et al., 2018; Chen et al., 2019). One approach leveraged shape priors during training by using smooth 3D segmentation masks produced via atlas-registration to motion-corrupted 2D stacks of short-axis cardiac MR images (Duan et al., 2019), which improved topological accuracy of 3D FCN predictions. Another approach involved simultaneous training of parallel network branches for registration and segmentation, providing a form of regularization on each branch (Xu and Niethammer, 2019). Loss terms which explicitly penalize topologically undesirable predictions using persistent homology (Clough et al., 2019; Byrne et al., 2020) have also been proposed, improving topological accuracy over pixel-wise segmentation models. Incorporation of point cloud prediction as an intermediate representation in a 3D segmentation network has also demonstrated improvements in topological consistency and segmentation accuracy (Ye et al., 2020). Finally, prediction of signed distance fields defining segmentation boundaries has also been proposed to improve segmentation accuracy (Li et al., 2020; Tilborghs et al., 2020). While each of these approaches have their inherent advantages, they do not guarantee a target topology at test time.

Another class of deep learning models which has received growing attention utilises a PCA shape model of surface meshes of training cases, the weights of which can be predicted directly by a CNN for a target image (Milletari et al., 2017; Bhalodia et al., 2018; Tóthová et al., 2018; 2020; Adams et al., 2020). These models have been proposed to predict 3D surface meshes both from 3D images (Bhalodia et al., 2018; Adams et al., 2020) and in settings where only sparse 2D images are available (Milletari et al., 2017; Tóthová et al., 2018; 2020). A hierarchical variational auto-encoder has also been proposed for the latter setting (Cerrolaza et al., 2018), where shape parameters are implicitly encoded in the latent variables. While such models directly encode a topologically consistent structure, as with classic PCA shape models they potentially suffer from over-constraining shape descriptors to the training set and may not be sensitive to subtle anatomical variations or out-of-distribution examples at test time. While these models show promise, with the added benefit of uncertainty quantification (Tóthová et al., 2020; Adams et al., 2020), such models have not been benchmarked against state-of-the-art 3D semantic segmentation models in the setting where dense 3D images are available.

Recently, there has also been a growing interest in the field of deep learning for image registration (Haskins et al., 2020), following the seminal work on spatial transformer networks (STN) (Jaderberg et al., 2015), with a number of models proposing to use registration to propagate labelmaps to target images as a means of segmentation (Lee et al., 2019b; Balakrishnan et al., 2018; Dalca et al., 2019a; Dong et al., 2020; Mansilla et al., 2020). Prominent methods for unsupervised deep learning image registration such as DLIR (de Vos et al., 2019) and VoxelMorph (Balakrishnan et al., 2018) use encoder-decoder type CNNs to predict a dense displacement field used to register a source image to a target image. This displacement field can also be used to propagate a source labelmap to the target image. Such models can be trained in a fully unsupervised setting with loss terms including image similarity, and penalties on deformation field smoothness and magnitude. Auxiliary losses that utilize labelmaps have also been used for semi-supervised learning of image registration (Balakrishnan et al., 2018), where a subset of the training data may have labelmap annotations. VoxelMorph proved highly effective for test time image registration when trained on a large set of brain MR images, with the observation that instance-specific optimization of the predicted displacements at test time produced improved performance, particularly when fewer training examples were available (Balakrishnan et al., 2018). Mansilla et al. (2020) recently proposed AC-RegNet, an image registration model regularized with a shape-aware auto-encoder conditioned on labelmaps during training, which yielded more anatomically plausible predicted displacement fields at test time for 2D lung X-ray images compared to a pixel-wise baseline. They also demonstrated that AC-RegNet could

be used for multi-atlas segmentation. Dong et al. (2020) proposed a deep learning framework with adversarial consistency for registration of a pre-defined population-derived atlas image and labelmap to a target image. Adversarial image and labelmap pairs were used to encourage the model to predict more accurate deformations, and both affine and non-rigid transformations were predicted by separately trained CNNs. The model was trained and evaluated using a limited set of 3D echocardiography images with annotations of the left ventricular myocardium. The performance of the proposed method improved over state-of-the-art voxel-wise segmentation methods, although only 25 cases were used for training with limited data augmentation (only rotations around the z-axis), which provides a sub-optimal setting for training a voxel-wise segmentation model. Dalca et al. (2019a) proposed a framework that directly parameterizes a template (or atlas) image volume to be jointly learned with a registration model that registers the template image to target images. The learnable template could also be conditioned on parameters of interest, such as age and sex, and the method was evaluated with a dataset of brain MR images. The template image can be likened to a statistical atlas image learned from the training set. The authors demonstrated that a template *labelmap* could be constructed *after* training by registering training images to the constructed atlas *image* and propagating their labelmaps and subsequently fusing them. Image segmentation at test time was then performed by propagating the constructed template labelmap to target images, producing promising results compared to VoxelMorph (Balakrishnan et al., 2018).

Finally, the image-and-spatial-transformer network (ISTN) was proposed in (Lee et al., 2019a), where an image transformer network (ITN) is trained jointly with a spatial transformer network (STN). Given a source and target input image pair, the ITN predicts an intermediate representation for each image, such as a semantic segmentation or landmarks. These intermediate representations are passed as inputs to the STN which predicts a spatial transformation to register the image pair. Similarly to STN-only models which take images directly as inputs (Balakrishnan et al., 2018; de Vos et al., 2019), loss terms optionally include image similarity, deformation field penalties and terms which leverage ground-truth labelmaps. The use of intermediate representations in ISTNs however led to better performance in learned registration of structures of interest (SoI) compared to a STN-only model. Another advantage of ISTNs is that the STN parameters can be optimized on specific instances at test time to register the intermediate representations predicted by the STN of a source and target image pair.

Contributions. Atlas-ISTN extends the ISTN framework and draws on other proposed works to jointly construct a population-derived atlas while training a model to perform segmentation and registration. As illustrated in Fig. 1, during training, the proposed model learns to register a population-derived atlas labelmap to a predicted pixel-wise segmentation of the SoI, and vice versa, combining advantages of semantic segmentation models with atlas registration. At test time, a refinement procedure for instance-specific optimization of the STN weights is used to register the atlas labelmap to a pixel-wise segmentation. The final displacement field warps the atlas labelmap to the patient SoI, preserving the atlas topology which helps to mitigate errors in the pixel-wise, prediction while also improving upon a single pass of the Atlas-ISTN. The invertibility of the transformations also allows for patient SoI to be mapped to a common atlas space. Our contributions include:

1. An all-in-one deep learning framework for atlas construction, registration and segmentation;
2. A robust segmentation system which improves over baseline segmentation and registration models;
3. A method for on-the-fly construction of a population-derived atlas within a deep learning framework (an atlas image and labelmap constructed from CCTA data are illustrated in Fig. 2);
4. Preservation of topology for test time segmentations via registration of the constructed atlas labelmap;
5. A deep learning framework that provides inter-subject correspondences of SoI via a mapping to atlas space.

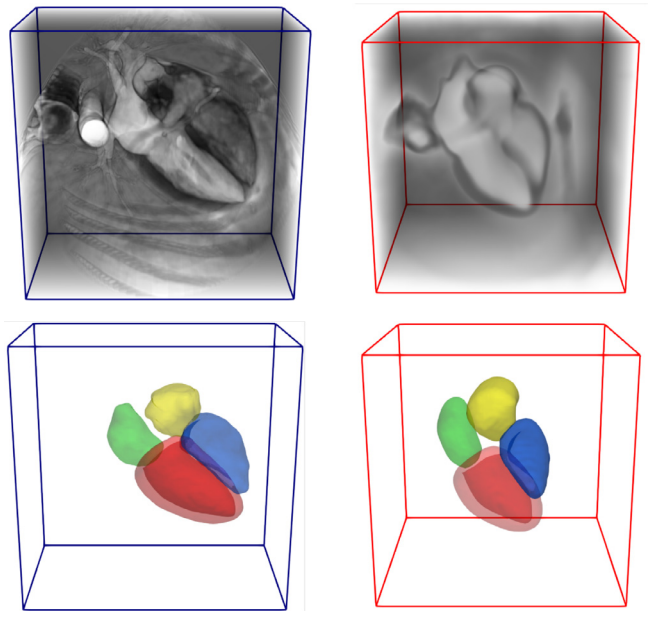

Fig. 2. Intensity projections of images (top) and labelmaps (bottom) of a CCTA training case (left) and the constructed atlas (right). Structures depicted include the left ventricle myocardium (red), right ventricle blood pool (blue), right atrial blood pool (yellow) and left atrial blood pool (green).

2. Methods

For many segmentation tasks, the topology of the target SoI is known a priori. For example, in medical image analysis, the chambers of the heart typically conform to the same spatial arrangement from subject to subject, with variation in shape, size, orientation and wall thickness. Voxel-wise segmentation models do not take advantage of this a priori knowledge, and can produce spurious segmentations or topologically inconsistent predictions, particularly with limited training data. The proposed Atlas-ISTN framework seeks to address this shortcoming by both learning a voxel-wise prediction and learning to fit a population-derived atlas labelmap to the SoI, reducing noisy predictions and ensuring topological plausibility at test time.

Similarly to ISTNs (Lee et al., 2019a), the Atlas-ISTN architecture consists of two sequential blocks: an ITN and a STN. The purpose of the ITN is to learn an intermediate representation from an input image which is useful to a downstream task, for example registration of the SoI (Lee et al., 2019a). With Atlas-ISTNs, this intermediate representation is chosen to be a semantic segmentation. Unlike in (Lee et al., 2019a) which registers pairs of input images via their intermediate representations, the STN in the Atlas-ISTN framework learns a spatial transformation to warp a constructed atlas labelmap to the ITN segmentation and vice versa.

In the following sections, we present the transformation model (Section 2.1), neural network architecture (Section 2.2), the method by which the atlas is constructed (Section 2.3), and the refinement procedure which can be used at test time to optimize the STN to improve the fit of the constructed atlas labelmap to the predicted ITN segmentation (Section 2.4).

2.1. Deformations

We model both affine and non-rigid deformations, which are commonly used in two-stage atlas registration methods particularly for cardiac image data (Lamata et al., 2014; Bai et al., 2015). We consider the following composition of an affine transformation:

$$T = M_t R_\theta D_s, \tag{1}$$

where M_t, R_θ and D_s are the translation, rotation and scaling matrices (we exclude shearing in this work). In applications with 2D image data, there are 4 degrees of freedom including translations in x and y, an in-plane rotation, and a global scaling term. In applications with 3D image data, there are 7 degrees of freedom including three translations, three rotations and a global scaling term. This affine transformation provides a coarse registration between a source and target image. Affine registration is often used as a pre-alignment step (Bai et al., 2015; Dong et al., 2020), providing a more optimal starting point for a non-rigid registration to be performed. This type of pre-alignment is also important for applications with brain MR images, so much so that it is built into standardized processing pipelines, and thus most recent work exploring learning deformations for brain image data has only utilized a non-rigid component in their registration models (Dalca et al., 2018; Balakrishnan et al., 2018; Krebs et al., 2019). While typically used as a pre-alignment, an affine transformation can also be optimized jointly with a non-rigid transformation (Stergios et al., 2018), where the affine component would be expected to account for large-scale deformations and coarse alignment of source and target images.

A diffeomorphic transformation model parameterized by a stationary velocity field (SVF) is used as the non-rigid component, which preserves the topology of the atlas after deformation, and ensures invertibility (Arsigny et al., 2006; Ashburner, 2007). The differential equation describing the evolution of a deformation generated by a SVF denoted by v, is given by

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = v(\phi^{(t)}). \tag{2}$$

The deformation at $t = 1$ ($\phi^{(1)}$) is obtained by integration of this ordinary differential equation (ODE) over unit time starting with the identity $\phi^{(0)}(x) = x$ (Ashburner, 2007). Subsequently, we denote this deformation as ϕ to simplify notation. In the theory of Lie groups, solving this ODE is equivalent to computing the exponential map of the flow field v (a member of the *Lie algebra*), i.e.,

$$\phi = \exp(v). \tag{3}$$

The inverse transformation ϕ^{-1} can thus be obtained by the exponentiation of the negative SVF. In practice, the exponential map is computed efficiently via scaling and squaring (Arsigny et al., 2006; Ashburner, 2007).

2.2. Atlas-ISTN model

Image transformer network. Given input images and ground-truth labelmaps, $M = \{x_i, y_i\}$, the ITN learns the mapping

$$\hat{y}_i = \mathbf{S}_{\theta_s}(x_i), \tag{4}$$

where $\hat{y}_i$ are the (multi-channel) logits of a labelmap prediction for sample i. Both $\hat{y}_i$ and y_i contain a background channel and as many foreground channels as structures in the training data. The ITN can be any suitable segmentation model. In this work 2D and 3D U-net models are used (Ronneberger et al., 2015; Çiçek et al., 2016), which consist of convolutional layers in an encoder-decoder format with skip connections (Fig. 3, left).

Spatial transformer network. We propose a STN which learns the mapping

$$\{v_i,\ T_i\} = \mathbf{D}_{\theta_d}(\hat{y}_i, y^a), \tag{5}$$

where the concatenation of the foreground channels of (1) the ITN logits $\hat{y}_i$ and (2) the atlas labelmap y^a make up the input tensor. The STN produces two outputs including (1) a SVF, v_i, and (2) affine transformation parameters composed into a matrix, T_i. The predicted affine parameters include translations (t) and rotations (θ) for each spatial dimension, along with a global scale (s). The matrices constructed from these respective variables are composed to form T_i (Eq. (1)), represented as a 3×4 matrix in the 3D image setting. T_i and v_i are processed by the Transformation Computation Module, $\mathbf{C}$, (see Fig. 4) to produce the final deformation fields:

$$\{\Phi_i,\ \Phi_i^{-1}\} = \mathbf{C}(v_i, T_i). \tag{6}$$

The SVF v_i is predicted at half the input image resolution, with a size of $3 \times \frac{N_x}{2} \times \frac{N_y}{2} \times \frac{N_z}{2}$ in the 3D setting, where N_k represents the size of spatial dimension k. ϕ_i and ϕ_i^{-1} are obtained via scaling and squaring followed by linear interpolation to the input image resolution, as shown schematically in Fig. 4.

Within the Transformation Computation Module, the forward and inverse transformations are the compositions:

$$\Phi_i = T_i \circ \phi_i, \tag{7}$$

$$\Phi_i^{-1} = \phi_i^{-1} \circ T_i^{-1}. \tag{8}$$

A forward pass through the network via ITN (Eq. (4)), STN (Eq. (5)) and Transformation Computation Module (Eq. (6)) can be expressed concisely as:

$$\{\Phi_i,\ \Phi_i^{-1}\} = \mathbf{C}\Big(\mathbf{D}_{\theta_d}\big(\mathbf{S}_{\theta_s}(x_i), y^a\big)\Big). \tag{9}$$

The inverse transformation resulting from a single pass (or '1-pass' for short) of the network is used to warp the atlas labelmap to patient space for sample i by:

$$\hat{y}^a_{i,j} = y^a_j \circ \Phi_i^{-1}, \tag{10}$$

where the deformations are used to warp each labelmap channel, j, independently. Similarly a labelmap for sample i is warped to atlas space by:

$$\tilde{y}_{i,j} = y_{i,j} \circ \Phi_i, \tag{11}$$

Losses

A mean squared error (MSE) loss is used for the supervised learning of the ITN weights (θ_s), i.e.,

$$L_s = \sum_i \sum_j^c \|y_{i,j} - \hat{y}_{i,j}\|^2, \tag{12}$$

where i is the sample index, j is the labelmap channel index, and c the total number of channels. Two additional MSE losses are used including the atlas-to-segmentation loss (L_{a2s}) and the segmentation-to-atlas loss (L_{s2a}):

$$L_{a2s} = \sum_i \sum_{j=2}^c \|y_{i,j} - \hat{y}^a_{i,j}\|^2, \tag{13}$$

$$L_{s2a} = \sum_i \sum_{j=2}^c \|\tilde{y}_{i,j} - y^a_j\|^2, \tag{14}$$

where $j = 1$ corresponds to the background channel. Note that y^a_j is in atlas space so does not include a sample index, i. L_{a2s} encourages accurate transformation of the atlas labelmap (y^a) to the ground-truth labelmap (y_i), despite potential noise in the ITN prediction ($\hat{y}_i$) which is an input to the STN. L_{s2a} encourages accurate

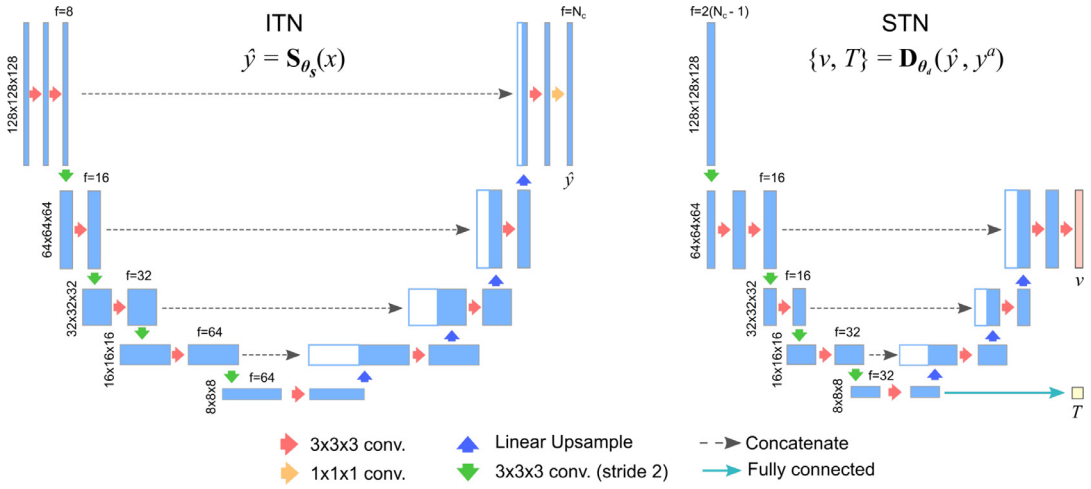

Fig. 3. An overview of the (3D) ITN and STN architectures. Note, $3 \times 3 \times 3$ conv. layers (red and green arrows) are followed by a ReLU activation, except for in the final layer of the STN. Outputs of the STN, v and T, are passed to the Transformation Computation Module block in Fig. 4.

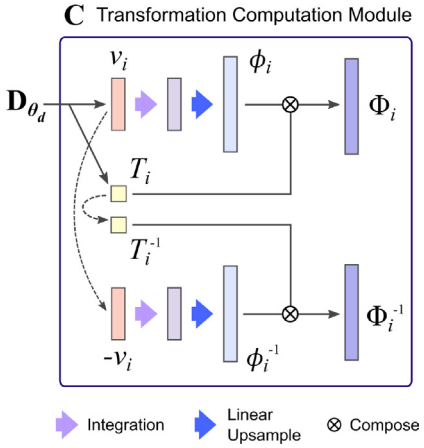

Fig. 4. The Transformation Computation Module. The predicted SVF (v) is integrated via scaling and squaring, before linear upsampling. The resulting transformation (ϕ) is composed with the predicted affine parameters (T_i). Computation of the inverse and forward transformations are shown at the top and bottom, respectively. Compositions are defined in Eqs. (7) and (8).

transformation of the ground-truth labelmap (y_i) to the atlas labelmap (y^a), which aides the atlas learning process (in Section 2.3).

A regularization term is also used to encourage smoothness of the predicted non-rigid displacement fields, i.e.,

$$L_{reg} = \sum_i \|\nabla \mathbf{u}_i\|^2. \tag{15}$$

where the displacements together with an identity transform provide the mapping $\phi = Id + \mathbf{u}$. The overall loss is

$$L = L_s + \omega\,(L_{a2s} + L_{s2a} + \lambda L_{reg}), \tag{16}$$

where λ controls the effect of the regularization term, and ω controls the overall weighting of the deformation-related loss terms. The segmentation loss (L_s) provides gradients only for updating the ITN weights (θ_s) while all other loss terms also contribute to gradients which update the STN weights (θ_d). Note that since L_{reg} penalizes the gradients of the non-rigid displacements, affine transformations including rotations and scaling predicted in ϕ would be penalized. Thus L_{reg} encourages smoothness as well as allows the STN to more freely adjust for global pose and scale with the affine component of the STN prediction (T).

Other practicalities. Architecturally, the STN is similar to the ITN in terms of the encoder-decoder structure, containing the same resolution levels, as depicted in Fig. 3. Strided convolutions are used for learned downsampling in the encoders, and trilinear upsampling was used rather than transposed convolutions in the decoders to avoid checkerboard artifacts.

The foreground channels of $\hat{y}_i$ and y^a are concatenated and passed as inputs to the STN, resulting in $2(N_c - 1)$ input channels, where N_c is the number of channels (including background) of the training labelmaps. Having multiple input channels for the STN (compared to just 1 for the ITN) led to the use of more filters in the first convolutional layer of the STN to allow for a richer representation in the feature maps relating the structures across the multiple concatenated input channels. The first convolutional layer of the STN has a stride of 2, immediately leading to a spatially reduced feature map. This learned down-sampling helps to reduce memory overhead, while still learning relevant features from the full-resolution input volumes. The SVF is predicted at half the resolution of the input image, which produces smoother deformation fields and better convergence during training for the range of tasks explored, as well as reduced memory overhead.

2.3. Atlas construction

We draw inspiration from classic atlas construction work (Guimond et al., 2000; Joshi et al., 2004), which average over images registered to a common atlas space to form an atlas. The ITN and the STN are trained jointly, and the atlas labelmap (y^a) and image (x^a) are updated at the end of each epoch by warping training data to atlas space via a forward pass of Atlas-ISTN and averaging across all samples. The update procedure for both atlas labelmap and image are the same, so we denote both images (x) and labelmaps (y) by z and the respective atlas by z^a in the following. The atlas is initialized and updated by

$$z^a_{j,t} = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}(z_{i,j}), & t = 0 \\ (1-\eta)\, z^a_{j,t-1} + \eta\, \tilde{z}^a_{j,t}, & t \geq 1 \end{cases} \tag{17}$$

where i denotes the training sample index, j the channel index, t the epoch, and n the total number of training cases. The atlas is initialized (at $t = 0$) as the mean of all the (undeformed) training cases. The rate at which the atlas is updated is determined by η, and $\tilde{z}^a_t$ is the mean of the transformed training cases:

$$\tilde{z}^a_{j,t} = \frac{1}{n}\sum_{i=1}^{n}(z_{i,j} \circ \Phi_{i,t}), \qquad t \geq 1 \tag{18}$$

At the end of each training epoch, a forward pass through the network (Eq. (9)) is used to warp each training case to atlas space, from which the channel-wise mean atlas is computed (Eq. (18)). The updated atlas labelmap at the end of each epoch, ($t \geq 1$) in Eq. (17), is then used during the following epoch in the computation of the losses L_{a2s} (Eq. (13)) and L_{s2a} (Eq. (14)). The atlas labelmap and image are shown in Fig. 2 alongside the labelmap and image of a randomly selected training case from the CCTA dataset.

The most closely related work is (Dalca et al., 2019a) where a volumetric atlas of the training *images* is learned over the course of model training. In this work we jointly construct volumetric atlases of the training images and *labelmaps* during model training. This is also achieved without explicitly parameterizing atlas voxels with learnable weights as in (Dalca et al., 2019a). For the proposed Atlas-ISTN model, we focus on the use of the constructed atlas labelmap y^a to improve segmentation performance at test-time, described in the next section. The benefits of this atlas labelmap are that (1) it is derived from the whole training population as opposed to using a randomly selected atlas from the training data, and (2) the trained STN is optimized to warp this atlas labelmap to the ITN prediction, which positions the STN weights in a setting well-suited for test time refinement (as opposed to optimizing STN weights for test-time refinement from scratch).

2.4. Test time refinement

At test-time, instance-specific optimization of the STN weights is used to register the constructed atlas labelmap to the ITN prediction of a target image, referred to throughout the text as refinement. Refinement can improve over a 1-pass prediction of the model, particularly for out-of-distribution test images, while also preserving the topology of the atlas labelmap to circumvent certain errors in the ITN prediction. For a given target image, L_{a2s} (Eq. (13)) and L_{s2a} (Eq. (14)) can be repurposed to optimize the alignment between the atlas labelmap (y^a) and the ITN logits ($\hat{y}$) instead of training labels (y),

$$L^*_{a2s} = \sum_{j=2}^{c} \|\hat{y}_{i,j} - y^a_j \circ \Phi_i^{-1}\|^2, \tag{19}$$

$$L^*_{s2a} = \sum_{j=2}^{c} \|\hat{y}_{i,j} \circ \Phi_i \ - y^a_j\|^2, \tag{20}$$

where the overall refinement loss is given by:

$$L^* = L^*_{a2s} + L^*_{s2a} + \lambda^* L_{reg}, \tag{21}$$

with λ^* corresponding to weightings on the atlas-to-segmentation, segmentation-to-atlas and regularization terms, respectively. During refinement, ITN weights are fixed and only STN weights are updated. In the presence of noise in $\hat{y}_i$, larger values of λ can encourage a more rigid registration that retains more of the atlas shape, and in turn circumvent spurious segmentations or holes in $\hat{y}_i$. The deformed atlas labelmap after refinement is given by:

$$\hat{y}^{a*}_{i,j} = y^a_j \circ \Phi_i^{*-1}, \tag{22}$$

where for a given input image x_i, Φ_i^{*-1} is obtained from a forward pass of Atlas-ISTN after refinement,

$$\{\Phi_i^*, \Phi_i^{*-1}\} = \mathbf{C}\Big(\mathbf{D}^*_{\theta_d}\big(\mathbf{S}_{\theta_s}(x_i), y^a\big)\Big), \tag{23}$$

where $\mathbf{D}^*_{\theta_d}$ denotes the STN with refined weights after K refinement iterations minimizing the loss in Eq. (21). This process is illustrated in the lower panel of Fig. 1.

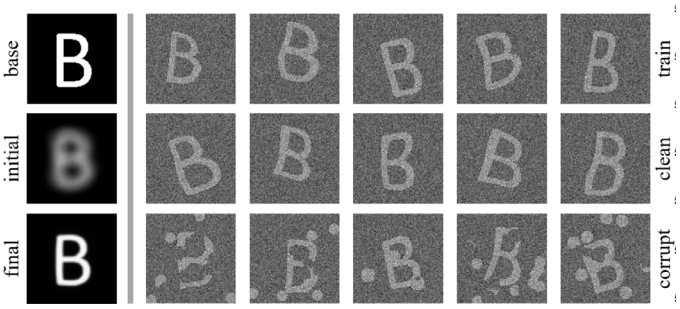

Fig. 5. The left-most column illustrates the base letter B (top), the initial atlas labelmap at the start of training defined in Eq. (18) at $t = 0$ (middle), and the recovered letter B atlas labelmap after training (bottom). The other images are examples from the training set (top row) and clean (middle row) and corrupt test set (bottom row) generated randomly from the base letter B. This toy dataset has been designed to study some key characteristics of Atlas-ISTNs with test-time refinement as described in Section 3.1.

2.5. Statistical analyses

Atlas-ISTN is evaluated on 2D and 3D datasets in the following section. Metrics used to evaluate model performance with each dataset include Dice similarity coefficient (DSC), average surface distance (ASD) and Hausdorff distance (HD). ASD and HD are computed in pixels for the synthetic dataset, and in millimetres for the 3D medical image datasets. Evaluation was performed using model predictions at the resolution of the pre-processed images. Pre-processing steps for each dataset are described in the following section. Statistical tests are used to compare performance of models trained on the 3D medical image datasets. Superiority is shown with a one-sided paired hypothesis test at a significance level of 1%, where the test statistic is the mean difference of the metric of interest. We checked the rejection of the null hypothesis using the 98% confidence interval of the test statistic which was estimated with percentage bootstrap using 10,000 repetitions.

3. Experiments and results

Overview. In this section we apply Atlas-ISTNs to three datasets to explore different aspects of the model. Firstly, a synthetic 2D dataset is used to illustrate key properties of Atlas-ISTNs, including the ability of refinement to circumvent forms of image corruption compared to baseline ITN and 1-pass predictions, and the effect of the regularization parameter λ on 1-pass and refinement predictions. Secondly, a 3D CCTA dataset is used to demonstrate the use of Atlas-ISTNs on multi-label data, as well as for model ablation studies. A large test set of 1000 cases is used to assess generalizability of Atlas-ISTNs compared to baseline U-net and STN-only models, using both limited and large training sets. Invertibility of the predicted transformations yielding inter-subject correspondence via atlas space is also demonstrated using the CCTA dataset. Finally, a complementary real world dataset of 3D brain MRI is used to explore the effects of domain shift on Atlas-ISTNs compared to baseline models.

3.1. Letter B

To illustrate some of the key properties of Atlas-ISTN for image segmentation, let us initially consider a simple toy example based on a binary image of the letter B. This example acts as proof-of-concept and allows us to demonstrate the behaviour of Atlas-ISTN in a controlled setting. We show some qualitative and quantitative results in the following.

Data description. Starting from a single, binary 2D image of the letter B, we generate warped instances of this base image using

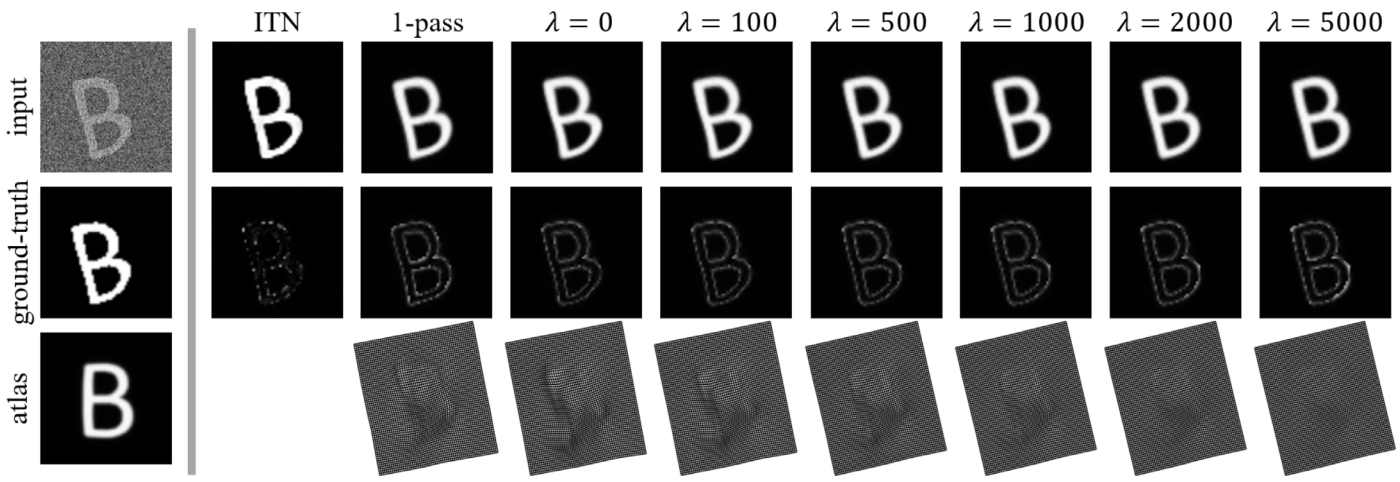

Fig. 6. Qualitative results for the 2D toy data with test data coming from the same distribution as the training data. The left-most column shows input image (top), ground-truth labelmap (middle) and atlas constructed by Atlas-ISTN (bottom). Excluding the left-most column, the top row shows the prediction of the ITN, Atlas-ISTN 1-pass, and refinement using values of λ between 0 and 5000; the middle row shows the difference between the model prediction and the ground-truth; the bottom row shows the result of deforming a regular grid by the predicted transformation. Both the ITN and 1-pass Atlas-ISTN yield accurate segmentations. Test time refinement with increasing regularization weight λ affects the smoothness of the final transformation.

random affine transformations composed with random B-spline deformations. For each warped binary image we generate a corresponding intensity image with additive Gaussian noise. We use 1,000 of these image and labelmap pairs for training an Atlas-ISTN. We further use a hold-out set of 100 pairs to test the Atlas-ISTN, once on a clean version of the test set (coming from the same distribution as the training data) and a corrupted version (with random clutter added to the intensity images). Visual examples of the training set and the clean and corrupted test sets together with base letter B, the initial and resulting constructed atlases are shown in Fig. 5. This synthetic 2D data is provided together with our source code such that the following results can be fully replicated.

Qualitative results. In the case of clean test data, we make the following observations (cf. Fig. 6): As expected, the ITN provides nearly perfect segmentations of the input images, and equally the 1-pass Atlas-ISTN predicts an accurate alignment of the constructed atlas to the ground-truth. We run test time refinement with six different regularization weights and observe the effect of increased regularization on the final transformation. This example confirms that when sufficient training data is available, and the test data comes from the same distribution, the 1-pass Atlas-ISTN is en-par with an ITN-based segmentation in terms of accuracy with the added benefit that the resulting segmentations of the Atlas-ISTN come with correspondences to the constructed atlas space. Test time refinement further allows to control the degree of deformation through the regularization weight, but may be considered optional as it may not add significant improvements in segmentation accuracy.

In reality, however, test data often does not come from the exact same distribution as the training data and this may negatively affect the predictive performance of a trained network. Test images, for example, might exhibit variations due to artifacts or pathology which were not captured in the training data. This is simulated here by testing the above Atlas-ISTN on a corrupted version of the test set. In Fig. 7, we observe that the ITN now fails to accurately segment the structures of interest yielding both many false positives and false negatives. Still, the predicted segmentation may provide useful information for subsequent refinement in our Atlas-ISTN framework. The 1-pass prediction of the Atlas-ISTN is also affected by the noisy ITN output yielding a sub-optimal alignment of the atlas, yet providing a good initial atlas alignment. Here, the benefits of the test time refinement become clear. This test-specific optimization of the STN network weights results in plausible segmentations of the corrupted test images, removing both false positives and negatives and adhering to the topology of the constructed atlas. Again, we observe the effect of the regularization weight which provides control over how closely we wish to stay to the constructed atlas up to affine transformations.

Quantitative results. Quantitative results over the 100 cases for both the clean and the corrupted test sets are summarized in Fig. 8. We observe that for the clean data (blue bars) highly accurate segmentations are obtained with all approaches and across the range of values for the regularization weight λ, indicated by high DSC and low surface distances. For the case of corrupted test data (orange bars), we can see the clear benefit of Atlas-ISTNs with test time refinement. We also observe that the results are not very sensitive to the regularization weight.

3.2. 3D cardiac CCTA

In the following, we focus on segmentation of large structures of the heart from 3D CCTA images which is an important step in both calculation of derived clinical indices such as strain (Nicol et al., 2019), as well as providing a computational domain and boundary conditions for simulations of cardiac function and coronary flow (Taylor et al., 2013; Chabiniok et al., 2016). A description of the data is first provided, followed by a brief overview of the 5 experiments which explore different aspects of the framework.

3.2.1. Data description

3109 3D CCTA images from multiple sites around the world including scanners from a range of manufacturers were used, consisting of commercial cases which were submitted to HeartFlow, Inc. for the FFR_{CT} Analysis[2] (Taylor et al., 2013). All cases were processed through the FFR_{CT} Analysis product pipeline, one output of which is the segmentation of the left ventricle myocardium (LVM). Segmentation of the LVM involves the manual inspection and correction of an automated segmentation produced by a Random Forest + shape fitting model. 109 high quality 3D images were selected for further 3D annotation of the large structures of the heart using in-house tools. In addition to the LVM, these included the

[2] https://www.heartflow.com/heartflow-ffrct-analysis/

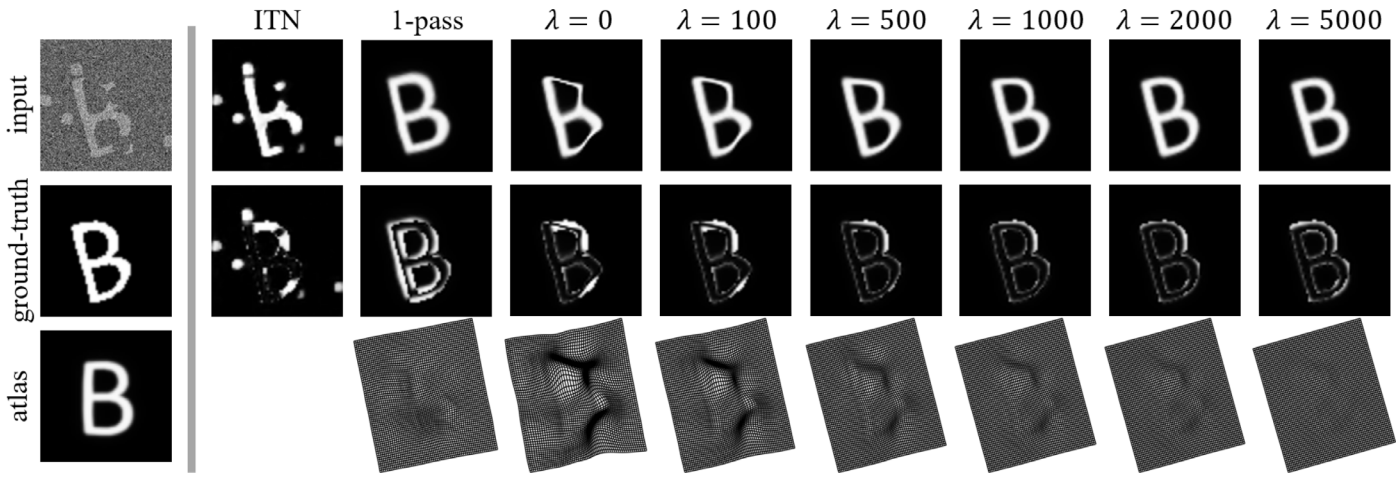

Fig. 7. Qualitative results for the 2D toy data with corrupted, out-of-distribution test data. The left-most column shows corrupted input image (top), ground-truth labelmap (middle) and atlas constructed by Atlas-ISTN (bottom). Excluding the left-most column, the top row shows the prediction of the ITN, Atlas-ISTN 1-pass, and refinement using values of λ between 0 and 5000; the middle row shows the difference between the model prediction and the ground-truth; the bottom row shows the result of deforming a regular grid by the predicted transformation. The ITN yields many false positives and false negatives and is topologically implausible. The 1-pass Atlas-ISTN yields a reasonable atlas alignment despite the corrupted input data. Test time refinement with increasing regularization weight λ can yield accurate and topologically plausible segmentations.

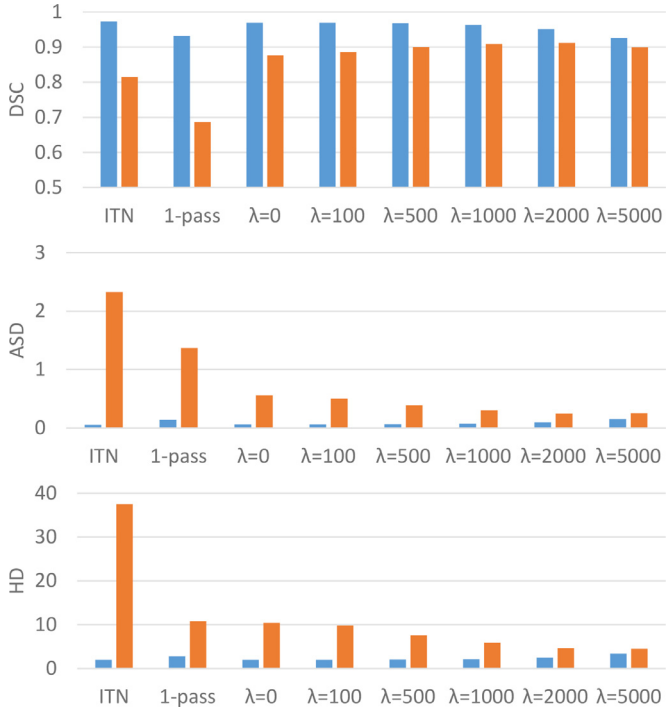

Fig. 8. Quantitative results for the 2D toy data with test data from the same distribution as the training data (blue bars), and corrupted, out-of-distribution test data (orange bars). The plots show the results using the segmentation metrics DSC (top), ASD (middle), and HD (bottom). On the within-distribution test data all metrics indicate good segmentation accuracy for all approaches and across different regularization weights. For the out-of-distribution data, Atlas-ISTNs with test time refinement out-perform the ITN and the 1-pass prediction by a significant margin with good robustness to the selection of the regularization weight λ.

left ventricle blood pool (LV), the right ventricle blood pool (RV), the right atrial blood pool (RA) and the left atrial blood pool (LA) (illustrated in Figs. 2 and 9).

Images were acquired in accordance with SCCT guidelines (Abbara et al., 2016). In-plane image resolution ranged from 0.235 mm to 0.499 mm with a mean ± SD of 0.400 ± 0.056 mm. Through-plane resolution ranged from 0.250 mm to 0.800 mm with a mean ± SD of 0.448 ± 0.158 mm. Image intensities were clipped to [−1000, 1000] Hounsfield units, and linearly rescaled to the range [−0.5, 0.5]. All images were isotropically resampled to ensure through-plane resolution (in the z direction) matched the already isotropic in-plane resolution, and images were subsequently downsampled by a factor of 4. This greatly reduced computational overhead while producing models with similar performance compared to models trained on higher resolution images. The thinnest cardiac structure considered is the LVM, with a mean thickness of 12 mm, and the mean isotropic resolution of the processed images was 1.600 ± 0.224 mm. While in-plane dimensions are fixed, images were either padded (for the vast majority of cases) or cropped in the z-dimension[3] to obtain volumes of size $128 \times 128 \times 128$.

The 109 cases with large structure annotations (LSA) were randomly split into 80 training, 10 validation, and 19 testing cases. A further 1000 cases with just the LVM label were used for additional testing to provide a more diverse test set on which to assess the performance of Atlas-ISTN. In all experiments, Atlas-ISTN is trained using the same 80 cases with all labelled structures, except in *Exp. 4* where the remaining 2000 cases with the LVM label only are also used for training.

3.2.2. Overview of experiments

Exp. 1: Comparison with Baseline Segmentation Model explores the value of refinement using the Atlas-ISTN framework compared to a U-net, as well as the ITN and 1-pass predictions of Atlas-ISTN on a limited test set with multiple labels (illustrated in Fig. 9), and on a large test set with only the LVM label.

Exp. 2 (Ablation study): Comparison with Baseline Deformation Models explores the value of using an intermediate representation for learned registration of SoI, and the value of explicitly learning a semantic segmentation as the intermediate representation.

Exp. 3 (Ablation study): Comparison of Framework Variants explores the impact of certain design choices, including (1) optimizing a trained versus untrained STN during refinement, (2) using a population-derived versus fixed atlas at test-time, and (3) including versus excluding the affine component in the STN.

[3] This was performed in a way that ensured SoI were in the cropped region.

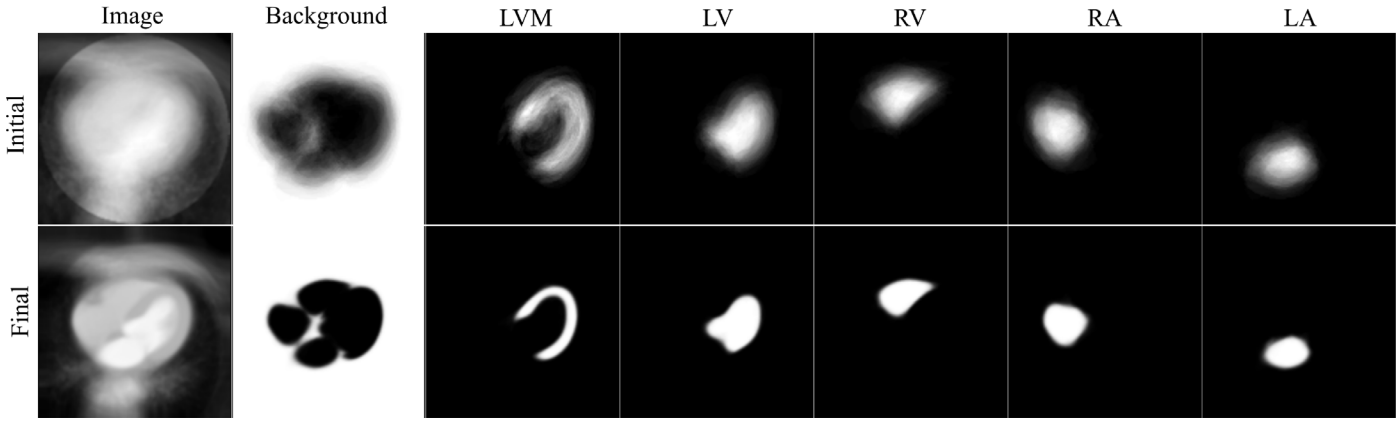

Fig. 9. An axial slice through the initial (top row) and final (bottom row) atlas image (first column) and the 6 channels of the atlas labelmap produced when training Atlas-ISTN on multi-label CCTA data. The atlas image and labelmap are constructed from the training data as described in Section 2.3.

Exp. 4: Upper bound LVM model explores the impact of training with a large dataset on performance of Atlas-ISTN compared to baseline segmentation and registration models.

Exp. 5: Inter-Subject Correspondence assesses the invertibility of the transformations produced by Atlas-ISTN, and demonstrates how inter-patient correspondence of SoI is achieved via a mapping to atlas space.

Model settings. PyTorch (Paszke et al., 2019) was used for all model development. Models were trained using NVIDIA Tesla V100 SXM2 GPUs (with 32 GB memory) for 800 epochs when training without data augmentation and 1200 epochs when using data augmentation (to reach convergence on the validation set). The full Atlas-ISTN model was trained with mini-batch size of 8, requiring ≈22 GB of GPU memory. The Adam optimizer was used with a learning rate of 1×10^{-3}, and an exponential learning rate decay with a half-life of 400 epochs was used. Weighting variables for the training loss in Eq. (16) were set to $\omega = 1$ and $\lambda = 800$. The atlas update rate was $\eta = 0.01$ in Eq. (17). At the start of training, the STN was susceptible to predicting large affine transformations in T, which could cause the atlas and target labelmaps to have no overlap, thus failing to back-propagate useful gradients. Initializing training with only the SVF prediction from the STN removed this instability, and the affine component was introduced after 200 epochs. Note, alternative schedules for phasing in the affine component in training would also be feasible. On-the-fly spatial augmentations were also used during training in all experiments except for the initial models in *Exp. 1*, which included translations (range: −8 to +8 voxels), rotations (range: −15 to 15 degrees in x, y, z) and scaling (range: 0.9 to 1.1 image resolution).

A regularization weight of $\lambda^* = 800$ was used for the refinement loss in Eq. (21) and the number of refinement iterations was $K = 100$. Refinement often would reach convergence within 30–50 iterations, but 100 iterations were used to ensure convergence was reached for all test samples. Refinement, performed with a single image at a time, required ≈3 GB of GPU memory and ≈10 s runtime for 100 iterations. A parameter sweep was not performed to optimize hyper-parameters, but rather 2 or 3 settings were compared for some parameters including λ, learning rate and epoch at which the affine component was introduced. We kept these hyper-parameters fixed across all experiments, since all models are variants of the Atlas-ISTN model (e.g. the ITN or STN components) and displayed similar convergence behaviour.

3.2.3. Experimental results

Exp. 1: comparison with baseline segmentation model

A baseline U-net segmentation model was trained by optimizing only the weights of the ITN, $\mathbf{S}_{\theta_s}$, with the segmentation loss L_s. A simple and commonly used post-processing step of retaining only the largest connected component of the U-net prediction ('U-net$_{1cc}$') was used as an additional comparison. An Atlas-ISTN model was trained on 80 multi-label samples (described in Section 3.2.1) and comparisons were made with various outputs of this model. These included (1) 'ITN', the prediction of the ITN (Eq. (4)), which has an identical architecture to the U-net but is trained as part of the Atlas-ISTN model, (2) '1-pass', the warped atlas labelmap predicted from the first pass of the Atlas-ISTN model (Eq. (10)), and (3) 'Refine', the warped atlas labelmap after test time refinement of the STN weights (Eq. (23)).

Fig. 9 shows the initial and final atlas image and multi-channel labelmap resulting from training Atlas-ISTN. The SoI in the final atlas image and labelmap are noticeably sharper, while the background structures in the atlas image remain fairly homogeneous. This is to be expected given that we are optimizing for the segmentation and alignment of the structures depicted in the labelmap. The final atlas image and labelmap are also shown in 3D in Fig. 2.

Models were initially trained with no data augmentation, and a significant drop in performance was observed from the ITN to 1-pass DSC (Table 1, left). This likely stems from the types of image features required to effectively train an ITN compared to a STN for CCTA data. The encoder-decoder architecture learns both global and local image features, and the ITN can more heavily rely on local image features to make accurate voxel-wise predictions. The STN performance however depends more heavily on learning global image features, given that the predicted deformation must operate across the entire image to transform SoI to significantly different orientations, scales and shapes for the CCTA data. The bottom two rows of Fig. 10 show the outputs of Atlas-ISTN on particularly challenging cases, where significant global and local deformations are required to register the undeformed atlas to the target SoI.

Table 1 shows the results of the U-net, U-net$_{1cc}$ and Atlas-ISTN outputs using the high quality 19 case LSA test set, for which labels of all chambers were available. For models trained with no data augmentation, we observe slight improvements of the ITN over the U-net for LVM, LV and RV DSC (1.1%, 0.7% and 0.6%, respectively), but not for other structures. The 1-pass performance of Atlas-ISTN falls short of the ITN across all metrics for the aforementioned reasons. Refinement produces the best results across almost all metrics, and although DSC improves by a moderate 0.1 – 0.8% over the ITN, performance on ASD and HD metrics improves significantly. U-net$_{1cc}$ also improves over the U-net in terms of HD and to a lesser extent ASD, but is almost always out-performed by refinement.

Table 1
Comparison with U-net and U-net with single largest connected component (U-net$_{1cc}$) baselines with and without spatial augmentation on the high quality 19 case LSA test set. Arrows indicate direction of metric improvement. Bold numbers are the best and second best, with the best also underlined, for a given metric and augmentation setting. Note that U-net and U-net$_{1cc}$ models with augmentation are identical for LVM and LA labels. Statistically significant ($p < 0.01$) improvement of the best or second best model over a given model is indicated by superscripts $*$ and †, respectively.

		No augmentation					With augmentation				
				Atlas-ISTN					Atlas-ISTN		
Label	Metric	U-net	U-net$_{1cc}$	ITN	1-pass	Refine	U-net	U-net$_{1cc}$	ITN	1-pass	Refine
LVM	↑ DSC	0.883*	0.883*	**0.893**	0.803*†	**0.894**	**0.911**	**0.911**	0.909	0.896*†	0.911
	↓ ASD	0.202*†	0.190*	**0.169**	0.401*†	**0.165**	0.137	**0.136**	0.138	0.169*†	**0.136**
	↓ HD	16.401*†	**6.260***	6.842*	7.775*†	**5.255**	6.150	4.862	**4.785**	5.313*†	**4.544**
LV	↑ DSC	0.936*	0.936*	**0.941**	0.896*†	**0.943**	**0.950**	**0.950**	0.948*	0.942*†	**0.950**
	↓ ASD	0.123	0.123	**0.113**	0.235*†	**0.109**	0.091	0.091	**0.091**	0.108*†	**0.089**
	↓ HD	7.157	7.132	**7.124**	8.619*†	**6.938**	6.283	6.283	**5.981**	6.527*	**5.973**
RV	↑ DSC	0.894	0.895	**0.900**	0.846*†	**0.898**	0.903	0.903	**0.906**	0.897*†	**0.906**
	↓ ASD	0.344*†	**0.287**	0.309	0.483*†	**0.284**	0.267	0.267	**0.263**	0.270	**0.258**
	↓ HD	29.082*†	**10.773**	38.093*†	12.224*†	**10.683**	13.034	10.879	11.793*†	**10.269**	**10.647**
RA	↑ DSC	**0.862**	**0.862**	0.857	0.825*†	0.860	0.883	0.883	**0.884**	0.873*†	**0.883**
	↓ ASD	**0.363**	**0.362**	0.511*†	0.521*†	0.383	0.292	0.291	**0.288**	0.313*†	**0.288**
	↓ HD	17.263	**13.459**	35.670*†	13.736	**13.082**	14.593*†	**12.243**	12.862*†	12.468	**12.187**
LA	↑ DSC	0.886*†	0.886*†	**0.899**	0.846*†	**0.900**	0.911	0.911	**0.917**	0.892*†	**0.913**
	↓ ASD	0.344*	0.338*	**0.304**	0.494*†	**0.286**	**0.236**	**0.236**	**0.230**	0.297*†	0.238
	↓ HD	15.149*†	**12.647***	23.031*†	13.396*	**11.725**	**11.182**	**11.182**	12.032	11.740*	**11.037**

Table 2
Comparison with U-net baseline with and without spatial augmentation on the 1000 case LVM test set. Arrows indicate direction of metric improvement. Bold numbers are the best and second best, with the best also underlined, for a given metric and given augmentation setting. Statistically significant ($p < 0.01$) improvement of the best or second best model over a given model is indicated by superscripts $*$ and †, respectively.

		No augmentation					With augmentation				
				Atlas-ISTN					Atlas-ISTN		
Label	Metric	U-net	U-net$_{1cc}$	ITN	1-pass	Refine	U-net	U-net$_{1cc}$	ITN	1-pass	Refine
LVM	↑ DSC	0.840*†	0.850*†	**0.863***	0.683*†	**0.869**	0.884*†	**0.885***	0.883*†	0.850*†	**0.888**
	↓ ASD	0.973*†	0.417*	**0.367***	1.207*†	**0.256**	0.301*†	**0.224***	0.342*†	0.311*†	**0.212**
	↓ HD	38.046*†	**10.566***	22.948*†	11.763*†	**6.120**	9.854*†	**6.440***	13.046*†	6.579*	**5.644**

When models are trained with on-the-fly data augmentation, there is a significant improvement across all metrics for all models compared to without augmentation. Most marked is the improvement in 1-pass performance, where for example the 1-pass DSC improves across all channels in absolute terms by between 4.3–8.7%, while the refinement DSC improves by 0.8–1.7%, leaving a smaller gap between 1-pass and refine metrics. Given that the images in this test set were selected for their high quality, all models produce accurate results and it is perhaps unsurprising that we see only modest improvement with Atlas-ISTN.

To assess performance on a more diverse dataset originating from a wide range of scanners and sites, the same models are run on 1000 test cases each containing a 3D annotation of just the LVM. The results are summarized in Table 2. All metrics are noticeably worse compared to the LVM metrics on the 19 high quality test cases in Table 1 as a reflection of the more diverse and challenging images in the 1000 case test set. For the models trained without data augmentation, the ITN shows an improvement over the U-net across all metrics, and refinement further improves on all metrics. Data augmentation during training improves all metrics significantly, most noticeably for the result of 1-pass DSC with an increase in absolute terms of 14.8% compared to an increase of 2.8% for the refinement DSC.

We observe that U-net$_{1cc}$ improves over the U-net performance both with and without augmentation, and also out-performs the ITN model trained with spatial augmentations. Refinement still produces the best results across both augmentation settings, despite a lower ITN performance compared to the U-net and U-net$_{1cc}$ with augmentation. Additionally, the result of refinement preserves topology and encourages smoothness of the target structures while the single largest connected component of an ITN prediction may not (see examples in Fig. 11). Statistically significant improvement across all metrics is achieved with refinement compared to all other models.

The improvement achieved with refinement over the ITN prediction is also illustrated in Fig. 12. Generally we see that there is seldom any degradation in the metrics, while outliers are generally corrected for, particularly for the metrics HD and ASD.

Exp. 2 (Ablation study): comparison with baseline deformation models Most related approaches in the literature propose the use of a single pass of a STN at test time to make a registration prediction, which can be used to propagate a labelmap (Balakrishnan et al., 2018; Dalca et al., 2019a; Dong et al., 2020). These methods also use images as inputs to the STN, while Atlas-ISTN enforces an explicit intermediate representation (semantic segmentation) predicted by an ITN as input to the STN. We make comparisons to the following models:

STN: This model is effectively Atlas-ISTN without the ITN component, where the inputs to the STN are the concatenation of (i) a target image, x_i, and (ii) the atlas image, x^a (see top left and right panels in Fig. 2, respectively, for an example). This model is representative of STN-only models in the literature, e.g. de Vos et al. (2019); Balakrishnan et al. (2018); Dalca et al. (2019a); Dong et al. (2020). The closest work conceptually is (Dalca et al., 2019a), where a learned atlas image and a target image are used as inputs to a STN. Differently to (Dalca et al., 2019a) however, our atlas is constructed via registration of training images (Section 2.3), and our loss terms depend on labelmaps rather than intensity im-

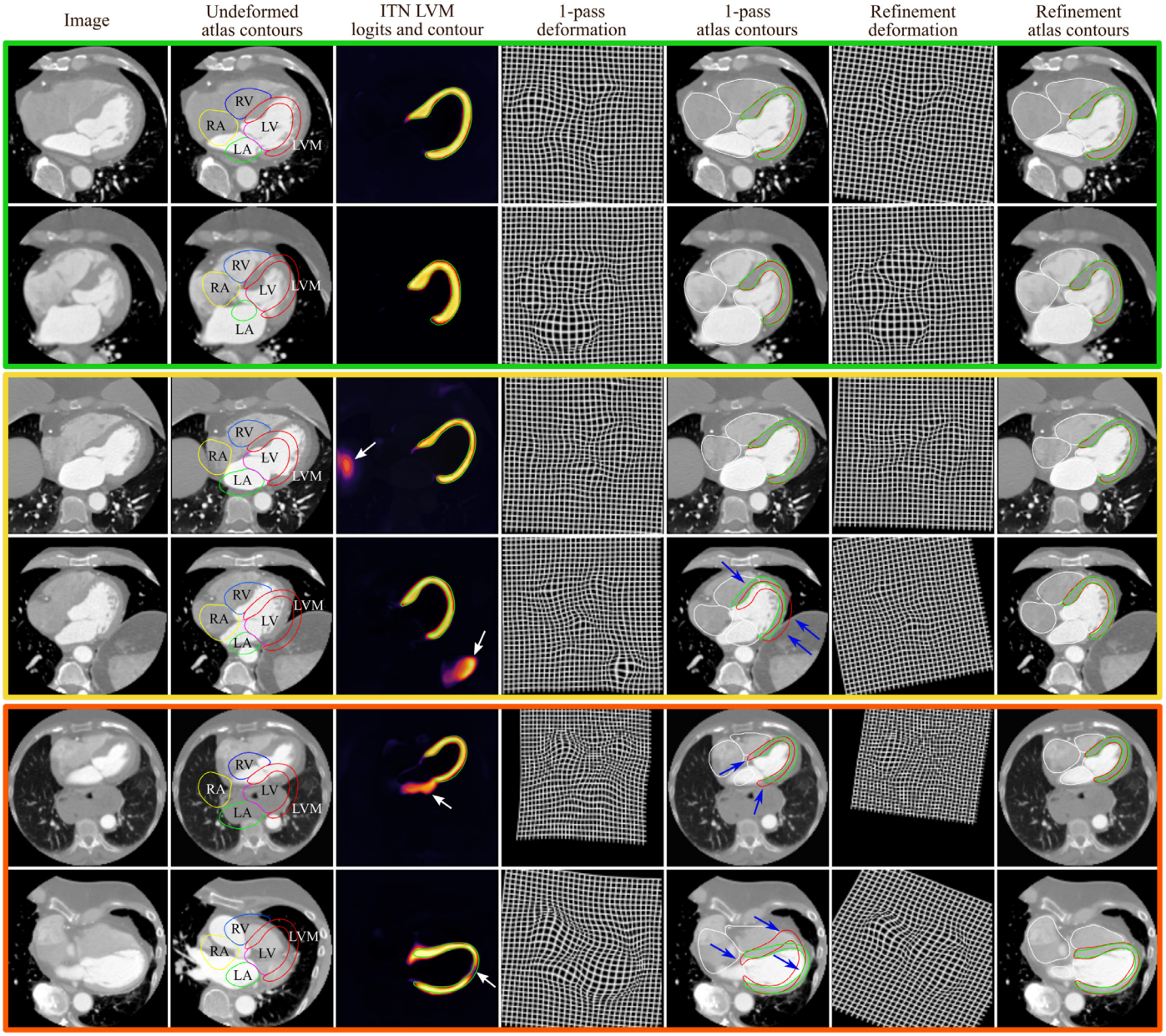

Fig. 10. Test cases in order of increasing difficulty (top to bottom), displaying an axial slice from a CCTA image. In columns 3, 5 and 7, manual and predicted contours of the LVM are shown in green and red, respectively, while white contours are used for other predicted structures for clarity of the LVM comparison. Column descriptions are provided above. White arrows highlight false positive and false negative LVM predictions from the ITN. Blue arrows highlight errors between the 1-pass deformed LVM and the manual LVM contours. Rows 1–2 contain examples for which ITN, 1-pass and refinement predictions all perform well, representative of a significant proportion of test cases. Rows 3–4 show cases where the ITN produces spurious segmentations, but refinement is able to circumvent this. In row 4 however, the 1-pass LVM contour is visibly offset, while the refinement produces a better fit. Rows 5–6 show challenging cases, where row 5 contains an example where the heart is particularly small in the field-of-view. On top of this, the ITN predicts a large spurious segmentation extending from the base of the LVM. The deformed contours after 1-pass do not fit the target structures accurately, but after refinement the LVM and other chambers are well positioned. Row 6 shows a case which required a significant rotation and translation of the atlas, so much so that in column 2 a different axial slice is shown to include the undeformed atlas. Additionally the LVM wall is particularly thin, where the ITN predicts a hole near the apex, and an over-segmention in the basal septal region. The significant transformation proves challenging for a single pass of the model, resulting in poor alignment of the atlas to the target SoI. The deformed contours after refinement however align well with the target SoI and bridges the ITN's predicted hole in the LVM (albeit not perfectly fitting to the manual LVM contour). Notice for cases which require larger global transformations (rows 3–6), the deformed grid after refinement contains a more noticeable affine transformation, and the non-rigid component appears to deform the grid less compared to the 1-pass. Best viewed in online version.

ages. The purpose of our model is to provide an atlas-based segmentation, while the focus of (Dalca et al., 2019a) was to produce a conditional atlas image. Note, since we are making a comparison between the concepts underlying Atlas-ISTN versus a STN-only model, we use our own implementation of the STN to control for differences that would arise from architecture and hyperparameter choices of other publicly available implementations.

Atlas-ISTN $_{-L_s}$: Since the STN model above reduces overall capacity of the network compared to Atlas-ISTN by using only the STN component of the model, the Atlas-ISTN$_{-L_s}$ model investigates the effect of keeping the entire Atlas-ISTN architecture (hence model capacity) the same and simply training without the segmentation loss L_s. This removes the constraint on the ITN to produce a semantic segmentation as input to the STN. The inputs to the STN

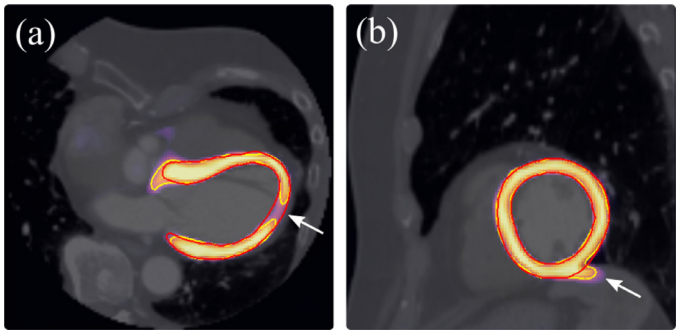

Fig. 11. Axial (a) and sagittal (b) planes showing examples where refinement corrects for (a) holes (false negatives) and (b) spurious segmentations (false positives) in the LVM channel predicted by the ITN. Red contour: LVM label of deformed atlas labelmap after refinement. Yellow contour: ITN prediction of the LVM channel. Heatmap: ITN logits of LVM channel. In both (a) and (b), the ITN prediction contour consists of only a single connected component of the LVM.

Table 3
Comparison of 1-pass performance of models trained with data augmentation on the 1000 case LVM test set. Bold numbers are the best and second best, with the best also underlined, for a given metric and given augmentation setting. Note further improvement for Atlas-ISTN is achieved with refinement (Table 2). Statistically significant ($p < 0.01$) improvement of the best or second best model over a given model is indicated by superscripts $*$ and †, respectively.

Label	Metric	STN 1-pass	Atlas-ISTN$_{-L_s}$ 1-pass	Atlas-ISTN 1-pass
LVM	↑ DSC	0.822*†	**0.839***	**0.850**
	↓ ASD	0.413*†	**0.368***	**0.311**
	↓ HD	7.471*†	**7.302***	**6.579**

in this model are still the concatenation of (i) the ITN prediction, $\hat{y}_i$, and (ii) the atlas labelmap, y^a.

In both models, the atlas labelmap and image are updated as usual (Section 2.3), and the loss terms L_{s2a}, L_{a2s} and L_{reg} are used as described in Section 2.2. Both models were trained with data augmentation and with the same hyper-parameters as in *Exp. 1*. Atlas-ISTN$_{-L_s}$ however used batch-normalization after each convolutional layer in the ITN to prevent vanishing gradients during training (given gradients were no longer back-propagated from the ITN predictions via L_s but instead only from losses using the STN predictions). The final constructed atlas image and labelmap for both of these models were similar in appearance to those produced by Atlas-ISTN.

Table 3 presents the results of the 1-pass performance of the models, with metrics computed for the LVM label using the 1000 case test set. Note, the Atlas-ISTN 1-pass results reported in Table 3 are repeated from Table 2 for convenience. Atlas-ISTN outperforms both the Atlas-ISTN$_{-L_s}$ and STN models across all metrics with gaps of about 1.1% and 2.7% in DSC, respectively. Furthermore, Atlas-ISTN$_{-L_s}$ and STN could not benefit from test time refinement of the STN weights by warping the atlas labelmap to the ITN logits as done by Atlas-ISTN. We found that test time refinement using an MSE loss on image intensities worsened performance, which is explained in the Discussion.

Exp. 3 (Ablation study): comparison of framework variants

Here we investigate the effects of ablating certain components from the Atlas-ISTN model. The following models were compared:

Independent: In this setting, a U-net was first trained (the same model as 'U-net' in Table 2). For test time refinement, the U-net predictions were concatenated with an atlas labelmap as inputs to a STN with randomly initialized weights (i.e. untrained). In previously explored settings where a STN had been trained, a 1-pass of the network can already align the atlas labelmap quite well with the target SoI. A 1-pass of a randomly initialized STN by contrast produces a near identity transformation. This setting investigates the importance of training a STN before performing refinement. Note, since a STN was not trained, an atlas labelmap was not constructed, and so a randomly selected case from the training data was used to provide the atlas labelmap (this particular case is shown in Fig. 2).

Fixed: This model involved training Atlas-ISTN without constructing an atlas on-the-fly, but instead using a randomly selected case from the training data to provide a fixed atlas labelmap and image throughout training and in refinement. This was the same randomly selected case used for refinement in the 'Independent' setting. This model allows us to investigate the effect of constructing an atlas during training versus using a pre-selected atlas.

SVF: This model involved training Atlas-ISTN without the affine component predicted by the STN, predicting only the non-rigid SVF component. This model lets us investigate the impact of having both affine and non-rigid components predicted by the STN.

All models above were trained with the same hyper-parameters and with data augmentation as described in *Exp. 1*. It was observed that the U-net and ITN predictions of the Atlas-ISTN model variants each produced similar performance. Additionally, the ITN of a given model could be paired with the STN and atlas labelmap of a different model at test time for refinement without significant differences in 1-pass or refinement performance. In light of this, we substitute the U-net trained independently for the ITN in all model variants to make a head-to-head comparison of the 1-pass and refinement results, once again using $K = 100$ iterations for refinement.

Table 4 shows the results of the baseline U-net, together with the 1-pass and refinement results of the 'Independent', 'Fixed', and 'SVF' models as well as the full Atlas-ISTN ("SVF + affine"), where metrics are computed for the LVM label using the 1000 case test set. A 1-pass of the untrained STN for the 'Independent' model re-

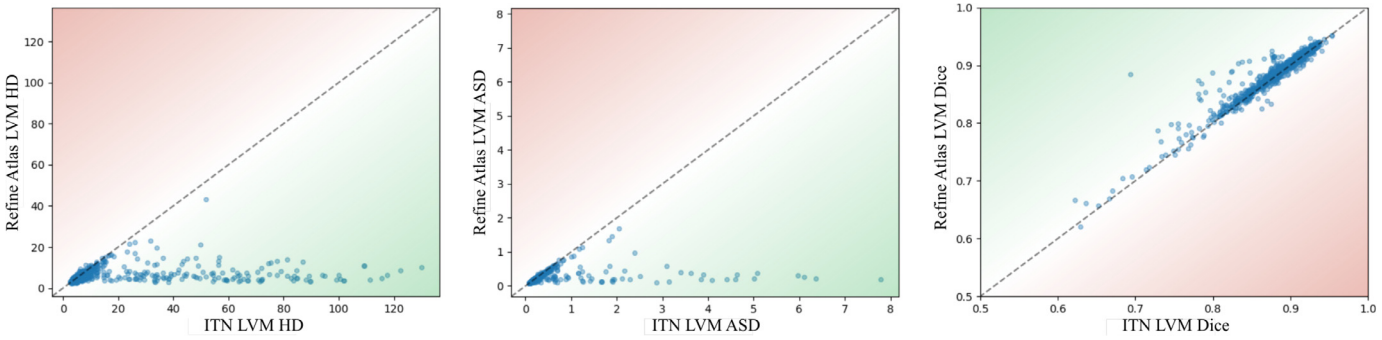

Fig. 12. Scatter plots showing HD (left), ASD (middle) and DSC (right) results of Atlas-ISTN on the 1000 case test set, comparing the LVM label from the ITN (x-axis) versus refinement (y-axis). The green/red gradients indicate increase/decrease in performance with refinement. HD and ASD of the ITN predictions almost always improve with refinement. Degradation observed for some cases in terms of DSC is always small, whereas improvements in DSC can be significant.

Table 4
Comparison of 1-pass and refinement results of Atlas-ISTN variants with a U-net baseline on the 1000 case LVM test set, trained with data augmentation. All models use the independently trained U-net as the ITN at test time. The column 'Id' refer to the initial alignment (with identity transformation) of the constructed atlas to put the resulting DSC, ASD and HD into context. Arrows indicate direction of metric improvement. Bold numbers are the best and second best, with the best also underlined, for a given metric. Statistically significant ($p < 0.01$) improvement of the best or second best model over a given model is indicated by superscripts $*$ and †, respectively.

				Independent		Fixed		SVF		SVF+affine	
Label	Metric	U-net	Id	Refine	Refine $K = 200$	1-pass	Refine	1-pass	Refine	1-pass	Refine
LVM	↑ DSC	**0.884***	0.204*†	0.770*†	0.820*†	0.848*†	0.879*†	0.854*†	0.883*	0.842*†	**0.886**
	↓ ASD	0.301*†	9.775*†	1.483*†	1.035*†	0.313*†	0.232*†	0.298*†	**0.218***	0.328*†	**0.213**
	↓ HD	9.854*†	32.552*†	11.312*†	9.405*†	6.980*†	6.196*†	6.549*†	**5.539**	6.567*†	**5.506**

sults in a near identity transform. The column 'Id' shows the result of an identity transform applied to the fixed atlas labelmap (i.e. the atlas labelmap used for 'Independent' and 'Fixed' models). Refinement of the STN weights from scratch for 'Independent' performs significantly worse than all other models. Standard refinement (with $K = 100$) and even refinement with double the number of iterations ($K = 200$) resulted in considerably worse results compared to the other models likely due to falling into bad local minima (e.g. registering to large spurious segmentations near the undeformed fixed atlas), or not reaching convergence. The best 1-pass results are obtained with the 'SVF' model, followed by 'Fixed' and Atlas-ISTN. Despite this, refinement with the proposed Atlas-ISTN out-performs all other models, followed by refinement with 'SVF', and both out-perform refinement with 'Fixed' across all metrics. This highlights the benefit of using both affine and non-rigid components in refinement, and using a population-derived atlas labelmap as opposed to using a fixed labelmap. Compared to voxel-wise segmentation of the U-net, refinement with Atlas-ISTN, 'Fixed', and 'SVF' models all improve in terms of ASD and HD, while Atlas-ISTN also marginally improves DSC.

We observe that the Atlas-ISTN 1-pass degrades slightly (by 0.8% DSC) in this experiment compared to using a jointly trained ITN (Table 2), although the refinement result overall is quite similar. In practice, the outputs of refinement would be used and not the 1-pass as the final segmentation.

Exp. 4: upper bound LVM model To estimate an upper bound on the performance of Atlas-ISTN for the LVM label, an Atlas-ISTN model is trained using 2000 additional training cases which contain only the LVM label. Only the LVM label is used during training, and as a result the constructed atlas does not contain any other foreground labels.The original 80 LSA cases are also included in the training process with the LVM labels only. The atlas is still constructed from these original 80 cases at the end of each epoch for faster convergence of the atlas construction during training. At each epoch, the 80 LSA cases are passed to the network with on-the-fly spatial augmentation, while another 80 cases are sampled randomly without replacement from the 2000 case dataset without augmentation. The number of epochs was halved from 1200 to 600 so that the models were trained with the same number of iterations as previous models. Epoch-dependent parameters were adjusted accordingly, including learning rate decay half-life reduced from 400 to 200 epochs, and the affine component introduced after 100 instead of 200 epochs. Hyper-parameters were the same as for previously trained models otherwise. U-net and STN-only models were also trained in this way.

Results on the 1000 case test set for these LVM-only models are presented in Table 5. Interestingly, the ITN out-performs the U-net across all metrics, with a 0.7% DSC increase. The ITN also performs better than the 1-pass and refinement for DSC and ASD. Refinement improves over the ITN, 1-pass and U-net in terms of HD, and performs similarly to the U-net for DSC and ASD. The better HD of the refinement model compared to the ITN reflects the observation that the ITN is still susceptible to producing topological errors,

Table 5
Results on the 1000 case LVM test set of U-net, STN, and Atlas-ISTN models trained with an additional 2000 cases with LVM label only. Bold numbers are the best and second best, with the best also underlined, for a given metric. Statistically significant ($p < 0.01$) improvement of the best or second best model over a given model is indicated by superscripts $*$ and †, respectively.

LVM			Atlas-ISTN		
Metric	U-net	STN	ITN	1-pass	Refine
↑ DSC	0.911*	0.890*†	**0.918**	0.902*†	**0.911***
↓ ASD	0.160*	0.217*†	**0.150**	0.186*†	**0.158***
↓ HD	5.464*	6.143*†	**5.380**	5.836*†	**5.093**

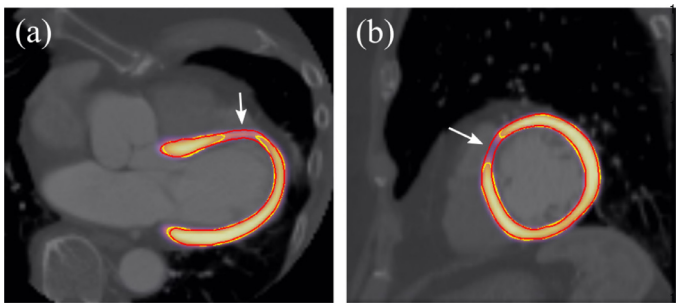

Fig. 13. Axial (a) and sagittal (b) planes showing examples where refinement corrects for holes (false negatives) in the LVM channel predicted by the ITN, for the Atlas-ISTN trained with an additional 2000 cases. Red contour: LVM label of deformed atlas labelmap after refinement. Yellow contour: ITN prediction of the LVM channel. Heatmap: ITN logits of LVM channel. In both (a) and (b), the ITN prediction contour consists of only a single connected component of the LVM.

which can be rectified by refinement (see Fig. 13 for an example). It should be noted that although the ITN out-performs refinement on DSC and ASD, the performance of both models is very close to human-level variability.

The improvements observed across all models compared to previous experiments reflect the use of a significantly larger training dataset, which is more representative of the population from which the test data was drawn. Compared to the Atlas-ISTN trained with spatial augmentation using the 80 cases with all structures (Table 2), an improvement of about 3% in LVM DSC is observed. The improvement in 1-pass performance is about 5% between the two models. The gap in performance between the 1-pass result and refinement is just under 1% DSC for the Atlas-ISTN with the extended training set, compared to a gap of 3.8% for the model trained on 80 cases, demonstrating the effect that a significantly larger training set can have on closing this gap. Finally, while the STN model also significantly improves compared to using limited training samples (Table 3), its performance still falls short of the Atlas-ISTN 1-pass as before, which itself still falls short of the refinement results.

Exp. 5: inter-subject correspondence. In addition to segmentation, Atlas-ISTN provides correspondence of the SoI across sub-

Table 6
Invertibility results on the 19 case test set. IC-DSC: inverse consistency Dice similarity coefficient, averaged over all labels, MICE: masked inverse consistency error (in terms of voxels).

	Atlas-ISTN	
Metric	1-pass	Refine
↑ IC-DSC	0.997	0.996
↓ MICE	0.0440	0.0574

jects, which can be used for example to propagate the location of anatomical landmarks not originally in the training data, or to assess inter-subject variability. Registration from atlas space to patient space and vice versa are jointly optimized via a symmetric loss, exploiting the invertibility of the chosen SVF and affine model. The use of this transformation model provides an inherently inverse consistent registration between any two subjects via atlas space, as described in (Joshi et al., 2004). Specifically, if we consider the registered atlas labelmap (after refinement) as the most accurate estimate of the SoI for each subject at test-time, as demonstrated in the above results, the composition of the transformations to and from atlas space for two subjects provides an inherently inverse consistent mapping of the SoI between these two subjects. Extending Eq. (22) and ignoring the channel index c, the mapping of subject k to subject i via atlas space is given by:

$$y_i^a = y^a \circ \Phi_i^{-1} = (\hat{y}_k^a \circ \Phi_k) \circ \Phi_i^{-1}, \tag{24}$$

and the mapping of subject i to subject k by:

$$y_k^a = y^a \circ \Phi_k^{-1} = (\hat{y}_i^a \circ \Phi_i) \circ \Phi_k^{-1}, \tag{25}$$

where y^a is the atlas labelmap, $\hat{y}_i^a$ is the deformed atlas labelmap (representing the SoI) to patient space for sample i, and Φ_i is the transformation from patient space to atlas space.

Theoretically, the properties of the diffeomorphic transformation model ensure that these mappings are inverse consistent, though error can arise from numerical precision, the integration of velocity fields, and discrete grid interpolation. Since the SoI under consideration for a given subject is the deformed atlas labelmap, it follows that if the composed transformation from atlas space to patient space and back introduces minimal error, then the composition of transformations from one image to another via atlas space (Eqs. (24) and (25)) will similarly have minimal error.

Two metrics are proposed to estimate the error associated with composing transformations to and from atlas space, including masked inverse consistency error (MICE) and inverse consistency DSC (IC-DSC), which are described below. An atlas labelmap deformed by both inverse and forward transformations for a given subject i is first defined, $y_{IC}^a = (y^a \circ \Phi_i^{-1}) \circ \Phi_i$. A grid, G, of size $N_x \times N_y \times N_z \times 3$, with voxel values corresponding to (x, y, z) voxel indices is also defined along with a twice deformed grid, $G_{IC} = (G \circ \Phi_i^{-1}) \circ \Phi_i$. MICE is the mean absolute displacement error in terms of voxels computed over the voxels masked by the SoI of the undeformed atlas labelmap (y^a), i.e. the mean of $\|G - G_{IC}\|$ within the voxels of the atlas SoI. The voxels of the SoI are computed by taking the argmax of the atlas labelmap and defining a foreground label mask, i.e. with voxels ≥ 1. IC-DSC is computed between the atlas labelmap (y^a) and the twice deformed atlas labelmap (y_{IC}^a). Table 6 shows that MICE is approximately 0.05 voxels, and IC-DSC is extremely close to 1. This indicates that the predicted SoI resulting from both 1-pass or refinement of Atlas-ISTN for one subject can be mapped to atlas space and subsequently to another subject with minimal error.

3.3. 3D brain MRI

While cardiac CT is our main focus in this work, we also conducted experiments for the task of brain structure segmentation in T1-weighted 3D MRI scans. Brain MRI has been the predominant type of data in related work on learning-based image registration (Balakrishnan et al., 2018; Dalca et al., 2019a; Zhao et al., 2019; Hoffmann et al., 2020), including our own work on structure-guided image registration (Lee et al., 2019a). Experiments on brain MRI study the specific aspect of generalization across images from different sites and scanners. We train the Atlas-ISTN on data from one site, and compare the segmentation results with and without test time refinement when testing on data from several other sites. Additionally, we have made some specific choices to complement the cardiac CT application. We opted for a binary segmentation problem for brain MRI (compared to multi-class for cardiac CT) by merging sub-cortical structures into one labelmap, and we use a SVF-only transformation model assuming rigidly pre-aligned scans, as it is the most common setup in the learning-based image registration literature on brain imaging.

Data description. We utilize brain MRI data from three publicly available imaging studies. We use data from the UK Biobank imaging study (UKBB)[4] (Sudlow et al., 2015; Miller et al., 2016; Alfaro-Almagro et al., 2018), the Cambridge Centre for Ageing and Neuroscience study (Cam-CAN) (Shafto et al., 2014; Taylor et al., 2017), and the IXI dataset.[5] Both UKBB and Cam-CAN use a similar imaging protocol with Siemens 3T scanners. IXI contains subsets from three different clinical sites, namely Guy's Hospital (IXI-Guys) using a Philips 1.5T system, Hammersmith Hospital (IXI-HH) using a Philips 3T scanner, and Institute of Psychiatry (IXI-IoP) using a GE 1.5T system. While the UKBB data is provided with pre-processed images and segmentations, we apply the following pipeline to the Cam-CAN and IXI data in order to match these as closely as possible to UKBB: (1) Skull stripping with ROBEX v1.2[6] (Iglesias et al., 2011); (2) Bias field correction with N4ITK[7] (Tustison et al., 2010); (3) Sub-cortical brain structure segmentation using FSL FIRST[8] (Patenaude et al., 2011). A very similar pipeline had been employed for UKBB with the same automatic segmentation for extracting brain structures. For computational reasons, we resample all brain scans to an isotropic 2 mm voxel size, and image size of $64 \times 64 \times 64$. Intensities are normalized within the brain masks to zero mean unit variance, where voxels outside the mask are set zero.

We merge the 15 individual brain structures (brain stem, left/right thalamus, caudate, putamen, pallidum, hippocampus, amygdala, accumbens) obtained from the FSL FIRST algorithm into a single binary label map, similar to Lee et al. (2019a). In line with related work on learning-based registration for brain images, we pre-align all scans rigidly to MNI[9] standard space using drop2,[10] and hence the task of the Atlas-ISTN here is to recover the non-rigid deformation between the images and the to-be-learned brain atlas. We use 100 scans from UKBB for training, 20 for validation, and 200 scans each from UKBB and Cam-CAN and all 581 scans from IXI (with Guys $n = 322$, HH $n = 185$, IoP $n = 74$) for testing the segmentation performance of Atlas-ISTN with and without refinement and the same baselines as for cardiac CT including U-net, STN-only, and ITN.

[4] UK Biobank Resource under Application Number 12579
[5] https://brain-development.org/ixi-dataset/
[6] https://www.nitrc.org/projects/robex
[7] https://itk.org
[8] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FIRST
[9] MNI stands for Montreal Neurological Institute
[10] https://github.com/biomedia-mira/drop2

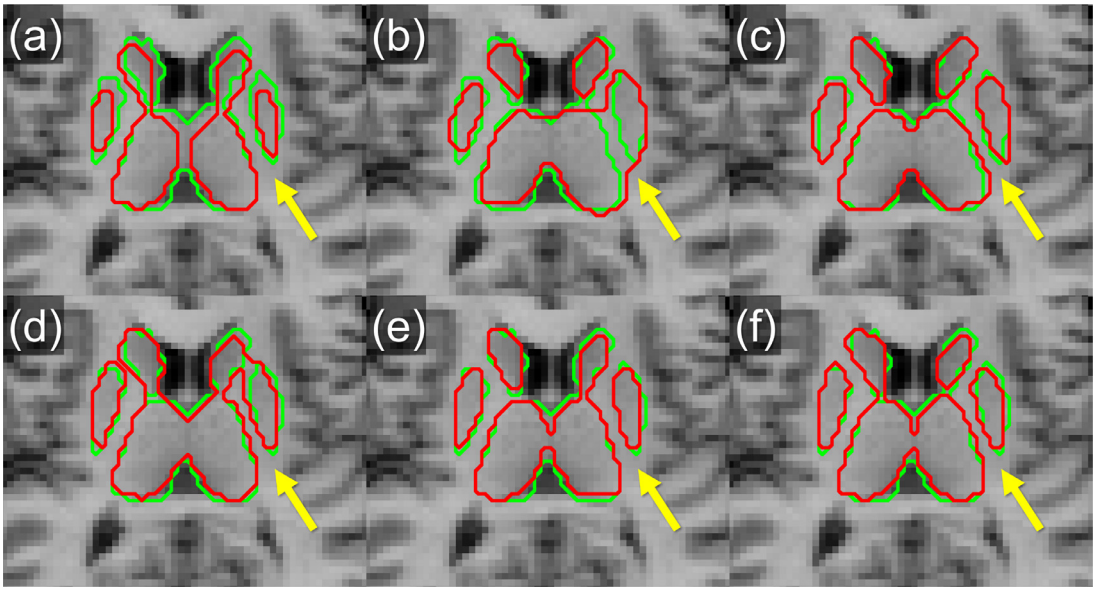

Fig. 14. Qualitative result for the brain structure segmentation. The figure shows one exemplary case from the IXI-IoP test set to highlight a key issue stemming from domain shift. IXI-IoP has the largest differences in imaging characteristics compared to the UKBB training data, yielding generally worse performance for all methods as shown in Table 7. The figure shows in (a) the initial atlas alignment before registration, (b) U-net, (c) STN-only, (d) ITN, (e) 1-pass, (f) refine. The U-net indeed struggles to delineate the boundaries between neighbouring structures (yellow arrows). This is due to poor image contrast and is unlikely to improve with additional data augmentation during training. This is a fundamental issue where an atlas-based approach with topological constraints can be beneficial.

Model settings. Here, a SVF-only transformation model is employed as all scans are rigidly pre-aligned to MNI. The Atlas-ISTN model for brain MRI is based on the same ITN and STN architectures as in the cardiac experiments with the coarsest resolution level removed due to the smaller size of the input images. The model is trained for 800 epochs and an exponential learning rate decay with a half-life of 400 epochs. As there is less variation between scans, we found fewer epochs are necessary compared to the cardiac data. Weighting variables λ and ω for the training loss in Eq. (16) were set empirically, with $\lambda = 500$. While in cardiac experiments a fixed value of $\omega = 1$ was used, for the brain experiments a 'fade-in' function (see Eq. (26)) was used to initially favour the segmentation loss L_s, with weighting on the deformation-related loss terms coming into full effect after about 200 epochs. This slightly improved the performance of the ITN and in turn refinement results for the brain experiments (observed on the UKBB validation data).

$$\omega = \frac{1}{1 + e^{-(t-200)/25}}, \tag{26}$$

where t is the epoch index. This approach brought the performance of the ITN closer to that of the U-net, as it possibly reduced the effects of competing gradients from the deformation and segmentation loss terms early in training (competing gradients in multi-task models are studied in (Yu et al., 2020)). The atlas update rate was also the same as for cardiac experiments, $\eta = 0.01$. $\lambda^* = 500$ was used for the refinement loss in Eq. (21), and the number of refinement iterations was $K = 50$.

Baselines. We compare segmentation performance of Atlas-ISTN with two baselines, a U-net and a STN-only model (simply denoted as STN) described earlier, i.e. an Atlas-ISTN without the ITN, where an intensity image and the atlas image are passed directly to the STN, by-passing an intermediate representation altogether, and at test time predicting the final transformation in a single forward pass.

Experimental results. The quantitative results are presented in Table 7 with qualitative results shown in Fig. 14, and with a sensitivity analysis regarding the regularization weight λ in Fig. 15. Overall, the U-net baseline performed well, out-performing the STN baseline on all datasets and performing similarly to the ITN. Atlas-ISTN performed the best on all test datasets across all metrics, with the exception of HD on the Cam-CAN and IXI-HH dataset. Test time refinement almost always improved over the ITN and Atlas-ISTN 1-pass performance across all metrics as well. In Fig. 14, the U-net struggles to delineate the boundaries between neighbouring structures due to poor image contrast. These errors are generally corrected with the STN, 1-pass and refinement methods.

The STN model under-performed compared to the U-net and Atlas-ISTN models generally on all datasets across all metrics. The STN model did not generalize as well to other datasets, with the performance gap between STN and the other models increasing for datasets less similar to the training dataset. For the least similar dataset compared to the training data, IXI-IoP, the STN model DSC was 0.813 compared to 0.846, 0.839 and 0.862 for the Atlas-ISTN's ITN, 1-pass and refinement results, respectively. The Atlas-ISTN 1-pass results out-performed the STN model across almost all metrics for all datasets, which could be attributed to the intermediate representation from the ITN provided to the STN. Test time refinement of Atlas-ISTN also produced greater improvements over the ITN and 1-pass results for less similar datasets. On the UKBB dataset, the increase in DSC compared to the ITN and 1-pass were just 0.4% and 0.3%, respectively, while for IXI-IoP it was 1.6% and 2.3%, respectively.

The sensitivity analysis in Fig. 15 shows robustness to the choice of the regularization weight. For DSC and ASD, Atlas-ISTN with test time refinement achieves the best performance for the entire range of λ values, while for HD it performs similarly to the STN and Atlas-ISTN 1-pass. This also highlights that the improvement for Atlas-ISTN with test time refinement is obtained consistently and independent of the specific strength of regularization.

Table 7
Quantitative results for the brain MRI experiments where segmentation methods are trained on 100 cases from UKBB and tested on different datasets from three imaging studies, UKBB, Cam-CAN and IXI. The columns 'Id' refer to the initial alignment (with identity transformation) of the constructed atlas to put the resulting DSC, ASD and HD into context. Bold numbers are the best and second best per row, with the best also being underlined. Statistically significant ($p < 0.01$) improvement of the best or second best model over a given model is indicated by superscripts $*$ and †, respectively.

	UKBB ($n = 200$)						Cam-CAN ($n = 200$)					
				Atlas-ISTN						Atlas-ISTN		
Metric	U-net	STN	Id	ITN	1-pass	Refine	U-net	STN	Id	ITN	1-pass	Refine
↑ DSC	**0.900***	0.876*†	0.773*†	0.898*†	0.889*†	**0.902**	0.869*†	0.863*†	0.765*†	0.866*†	**0.873**	**0.877**
↓ ASD	**0.208***	0.263*†	0.553*†	0.213*†	0.232*†	**0.202**	0.369*†	0.315*	0.597*†	0.388*†	**0.291**	**0.301**
↓ HD	5.868*	5.851*	7.399*†	5.976*†	**5.743**	**5.651**	9.131*†	**5.862**	7.266*†	8.715*†	**6.311***	6.926*†
	IXI all ($n = 581$)						IXI-Guys ($n = 322$)					
				Atlas-ISTN						Atlas-ISTN		
Metric	U-net	STN	Id	ITN	1-pass	Refine	U-net	STN	Id	ITN	1-pass	Refine
↑ DSC	**0.880***	0.851*†	0.741*†	0.880*	0.874*†	**0.890**	0.897*	0.874*†	0.769*†	**0.898***	0.890*†	**0.906**
↓ ASD	**0.261***	0.343*†	0.683*†	0.264*	0.279*†	**0.235**	0.217*	0.278*†	0.574*†	**0.215***	0.233*†	**0.197**
↓ HD	7.006*†	6.104	7.610*†	6.780*†	**6.075**	**6.022**	5.640*	5.588*	6.976*†	5.779*†	**5.458***	**5.367**
	IXI-HH ($n = 185$)						IXI-IoP ($n = 74$)					
				Atlas-ISTN						Atlas-ISTN		
Metric	U-net	STN	Id	ITN	1-pass	Refine	U-net	STN	Id	ITN	1-pass	Refine
↑ DSC	**0.866***	0.827*†	0.707*†	0.861*†	0.859*†	**0.875**	0.844*	0.813*†	0.708*†	**0.846***	0.839*	**0.862**
↓ ASD	**0.296***	0.403*†	0.809*†	0.309*†	0.310*†	**0.269**	0.368*	0.480*†	0.845*†	**0.367***	0.401*	**0.316**
↓ HD	7.945*†	**6.418**	8.206*†	7.896*†	**6.530**	6.656	10.599*†	**7.568**	8.882*†	8.343*†	7.618	**7.290**

Fig. 15. Sensitivity analysis for the effect of the regularization weight λ for the application of 3D brain segmentation. The Atlas-ISTN with test time refinement achieves overall best performance for the metrics DSC (top) and ASD (middle) across the range of regularization weights, and performs similar on HD (bottom) compared to the STN baseline and 1-pass prediction. The Atlas-ISTN outperforms the U-net baseline on all three metrics. The results shown here are the averages over the UKBB, Cam-CAN and IXI test datasets. U-net results are independent of λ and thus constant.

4. Discussion

Results from experiments with synthetic 2D data, real 3D CCTA and T1-weighted brain MRI scans illustrate the benefits of the Atlas-ISTN framework.

Experiments with synthetic 2D data demonstrated that improvements in terms of DSC, ASD and HD were achieved with test-time refinement on corrupted, out-of-distribution test images compared to both the ITN and Atlas-ISTN 1-pass, and that this improvement was insensitive to the choice of the weighting (λ) of the regularization term (Fig. 8). The parameter λ could also be adjusted for test-time refinement to adjust the rigidity of the deformation used to obtain a final registration of the atlas. Since only the non-rigid component is penalized by the regularization, higher values of λ result in deformations that rely increasingly on the affine parameters. This flexibility allows for scenarios where one might want to restrict the non-rigid deformation from adhering too closely to potentially noisy predictions of the ITN, and retain more of the underlying shape of the atlas. On clean test data, the Atlas-ISTN performs en-par with a baseline U-net but with the added benefit of yielding atlas correspondences.

For the application to real 3D data, values of λ were selected empirically. Generally, values of λ that were too high resulted in worse performance with test time refinement, and values that were too low could produce undesirably sharp gradients in the predicted non-rigid deformations and less robustness to spurious segmentations from the ITN. Adjustment of λ between training and test time was not extensively explored on the real datasets, although the experiments with synthetic data demonstrate the potential benefits.

The experiments with both CCTA and T1-weighted brain MRI data demonstrate the improved performance of Atlas-ISTN over segmentation-only and registration-only baseline models. The Atlas-ISTN 1-pass out-performed the STN-only model in both applications, where a larger gap in performance was observed for

data further from the training distribution (Table 7). The Atlas-ISTN 1-pass also out-performed the 1-pass of an Atlas-ISTN model without an imposed intermediate representation (Table 3). This indicates that the use of semantic segmentations as intermediate representations in Atlas-ISTN are advantageous for the 1-pass registration, providing robustness to variability and noise in the input images, reinforcing the findings of (Lee et al., 2019b). One limitation of the analysis is that downsampled images were used both in our models and evaluation. While this had limited impact for the applications considered, in applications which involve fine structures relative to the original image resolution, such as coronary vessels in CCTA images, a model which processes higher resolution images would be important.

Obtaining and annotating large sets of 3D medical image data is a common challenge. Most of our experiments with real 3D data involved training with datasets of fewer than 100 samples. In experiments where models were trained on a significantly larger dataset of CCTA cases with LVM labels only, the performance gap between 1-pass and test time refinement of Atlas-ISTN narrowed significantly (Table 5), with ITN and test time refinement of Atlas-ISTN still out-performing the 1-pass result and the STN-only model. This suggests that learned (1-pass) registration models generally require more training data to reach the performance of the ITN and subsequent test time refinement of Atlas-ISTN, particularly for datasets which may have significant inherent spatial variability like CCTA data. This demonstrates the advantage of using test time refinement with Atlas-ISTN for models trained with a limited size annotated dataset. Furthermore, when auxiliary information in the form of labelmaps is available, it can be used not only in the loss (Balakrishnan et al., 2018) but also to generate intermediate representations using an ITN.

While hyper-parameters were tuned, an exhaustive search of architectures and parameters was not undertaken. Between the 3D cardiac and brain experiments, minimal changes were made to hyper-parameters, with modifications to λ and ω providing some slight performance gains (Eq. (16)). We assessed Atlas-ISTNs using a largely consistent implementation across three different tasks, including 2D and 3D data, multi-class and binary labels, different modalities and multiple baselines. That our implementation works consistently well is evidence that the concept of Atlas-ISTNs can be beneficial in a range of settings.

A U-net was chosen as a baseline and as the ITN component, but any suitable segmentation model (or image-to-image architecture) can be used as the ITN. Improvements in the ITN performance are also likely to result in improvements in performance of test time refinement, as shown throughout the experiments. Conversely, refinement accuracy is also limited by ITN prediction accuracy, particularly in scenarios where the ITN systematically over-segments or under-segments the SoI. Refinement would not be able to correct such errors.

Different loss functions were also not extensively explored. However, separate experiments were performed using a cross-entropy loss to train the ITN and U-net instead of a MSE loss, producing comparable results (not presented here). This is consistent with recent work showing that squared losses for vision tasks perform similarly to a cross-entropy loss (Hui and Belkin, 2021). Additionally, using a MSE loss for L_s avoided the need to carefully weight loss terms in training since MSE losses were also used for L_{a2s} and L_{s2a}. Use of losses which encourage topological consistency of voxel-wise predictions (Tilborghs et al., 2020; Byrne et al., 2020) could also be explored in the Atlas-ISTN framework in future work.

In this work, improving segmentation accuracy over baseline methods was facilitated by the construction of the atlas labelmap. The transformation of the constructed atlas labelmap to the patient SoI in test time refinement also improved performance compared to the use of a pre-selected, fixed atlas labelmap (Table 4). An important advantage of atlas-based segmentation is the preservation of the atlas topology, which should contain smooth, contiguous labels without spurious components. Empirically, atlas labelmaps constructed for all explored datasets met these criteria, including with single label structures using 2D synthetic and 3D MRI data, and multiple cardiac structures using 3D CCTA data. However, a caveat of averaging the labelmaps of multiple structures from a set of co-registered images, as used in the past for cardiac (Bai et al., 2015) and brain (Joshi et al., 2004; Cabezas et al., 2011) atlas formation, is that there are no *explicit* constraints that guarantee the topological consistency of the final atlas labelmap. It is possible that for certain settings, such as using data with more complex structures and large variability in ground-truth labels, the proposed atlas construction process may not produce a topologically consistent atlas labelmap. In such a setting however, corrections could be made to the atlas labelmap to rectify any topological inconsistencies, for example by using more sophisticated fusion strategies (Iglesias and Sabuncu, 2015), or even by manually editing the atlas labelmap at the end of training. This would only need to be done once, as subsequent test time registration with diffeomorphisms would preserve the topology of the corrected atlas. Another limitation of atlas based segmentation methods in general is where anomalous or pathological morphology renders the atlas topology unsuitable. This may occur for example in subjects with situs inversus, or where a significant alteration to the anatomy of an organ is caused by disease.

While we propose a method to generate a population-derived atlas by incorporating ideas from classic atlas construction methods (Joshi et al., 2004) into a deep learning framework, we do not explore the use of constraints that can be imposed to ensure a mean shape (Joshi et al., 2004; Dalca et al., 2019a; Bône et al., 2020), or preserve high resolution detail in the atlas image (Guimond et al., 2000). This may be explored further in future work. We also do not extensively explore the use of image loss terms which are typically part of atlas construction frameworks, particularly for atlas images. Challenges of using image similarity loss terms for learning based registration with CCTA data in particular are discussed below.

The presented use of loss terms driven only by labelmaps and not by image intensities aims to accurately register the SoI. Unlabelled structures in the image are not of interest for the application of atlas-based segmentation. However, further exploration of image similarity loss terms could allow for registration not just of labelled SoI but also of unlabelled structures in the intensity images. Image similarity loss terms have been used in unsupervised and semi-supervised (i.e. using labelmaps and intensity images in loss terms) learning settings with brain MRI (Balakrishnan et al., 2018; de Vos et al., 2019; Dalca et al., 2019a; 2019b; Xu and Niethammer, 2019), chest X-ray (Mansilla et al., 2020) and knee MRI (Xu and Niethammer, 2019) for example. Semi-supervised learning for cardiac CCTA presents several unique challenges, including large variability in the extent of visible anatomy, significant variability in the shape of the field-of-view, tissue-level intensity variations due to differences in contrast timing, artifacts due to implants and differences between scanners and acquisition protocols. Accounting for these factors would be important when introducing image similarity losses for CCTA image registration. As demonstrated with Atlas-ISTN, an intermediate representation of SoI segmentations helps to mitigate such variations, with the proposed method out-performing the baseline U-net and registration models when assessed on 1000 CCTA test cases from a wide range of sites around the world.

Our findings on brain MRI may further be of interest in the context of domain shift in learning-based image registration. The accuracy of the 1-pass predictions became significantly worse for data from different sites highlighting the benefit of incorporating

test-time refinement, in particular, when only limited amounts of training data are available, a point we made in our earlier work (Lee et al., 2019a).

The Atlas-ISTN framework could also be adapted for other potential applications. New structures such as landmarks, or representations such as meshes, can be added directly into the atlas after training. Additional labels can also be learned by the ITN without contributing to the atlas construction (e.g. structures which may not have strong one-to-one mappings between cases, such as coronary trees). Inter-subject correspondence via atlas space also provides the opportunity for population shape and motion analysis.

5. Conclusions

Atlas-ISTN provides a framework to jointly learn image segmentation and registration, while simultaneously generating a population-derived atlas used in the model training process. Registration of the atlas labelmap via test time refinement provides a topologically consistent and accurate segmentation of the target structures. We have demonstrated quantitatively and qualitatively the improvement in segmentation performance of the proposed Atlas-ISTN model over baseline segmentation-only and registration-only models using synthetic 2D and real 3D datasets. Through ablation studies, we have also demonstrated the importance of design choices, including the use of both affine and non-rigid components in the transformation model for refinement, the value of using intermediate representations of SoI for registration, and the advantage of using a population-derived atlas compared to a pre-selected atlas from the training data. Furthermore, Atlas-ISTN shows greater improvement over segmentation and registration baselines on test data further from the training distribution, particularly when trained with limited data. Atlas-ISTN may benefit segmentation applications where a known topology is expected, and where inter-subject correspondences may be of interest.

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

CRediT authorship contribution statement

Matthew Sinclair: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Andreas Schuh:** Conceptualization, Methodology, Software, Writing – review & editing. **Karl Hahn:** Conceptualization, Methodology, Writing – review & editing. **Kersten Petersen:** Conceptualization, Formal analysis, Writing – review & editing. **Ying Bai:** Conceptualization, Writing – review & editing. **James Batten:** Conceptualization, Writing – review & editing. **Michiel Schaap:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision. **Ben Glocker:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision.

Acknowledgements

This research was funded by HeartFlow, Inc.; James Batten was supported by the UKRI CDT in AI for Healthcare (Grant No. P/S023283/1).

References

Abbara, S., Blanke, P., Maroules, C.D., Cheezum, M., Choi, A.D., Han, B.K., Marwan, M., Naoum, C., Norgaard, B.L., Rubinshtein, R., Schoenhagen, P., Villines, T., Leipsic, J., 2016. SCCT guidelines for the performance and acquisition of coronary computed tomographic angiography: a report of the society of cardiovascular computed tomography guidelines committee: endorsed by the North American society for cardiovascular imaging (NASCI). J. Cardiovasc. Comput. Tomogr. 10 (6), 435–449. doi:10.1016/j.jcct.2016.10.002.

Adams, J., Bhalodia, R., Elhabian, S., 2020. Uncertain-deepssm: from images to probabilistic shape models. In: Reuter, M., Wachinger, C., Lombaert, H., Paniagua, B., Goksel, O., Rekik, I. (Eds.), International Workshop on Shape in Medical Imaging. Springer International Publishing, Cham, pp. 57–72.

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D.C., Zhang, H., Dragonu, I., Matthews, P.M., Miller, K.L., Smith, S.M., 2018. Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. NeuroImage 166, 400–424. doi:10.1016/J.NEUROIMAGE.2017.10.034.

Arsigny, V., Commowick, O., Pennec, X., Ayache, N., 2006. A log-euclidean framework for statistics on diffeomorphisms. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 4190. LNCS, pp. 924–931. doi:10.1007/11866565_113.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38 (1), 95–113. doi:10.1016/j.neuroimage.2007.07.007.

Bai, W., Shi, W., de Marvao, A., Dawes, T.J., O'Regan, D.P., Cook, S.A., Rueckert, D., 2015. A bi-ventricular cardiac atlas built from 1000+ high resolution MR images of healthy subjects and an analysis of shape and motion. Med. Image Anal. 26 (1), 133–145. doi:10.1016/j.media.2015.08.009.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2018. VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging 38 (8), 1788–1800. doi:10.1109/tmi.2019.2897538.

Bhalodia, R., Elhabian, S.Y., Kavan, L., Whitaker, R.T., 2018. Deepssm: A deep learning framework for statistical shape modeling from raw images. In: Reuter, M., Wachinger, C., Lombaert, H., Paniagua, B., Lüthi, M., Egger, B. (Eds.), Shape in Medical Imaging. Springer International Publishing, Cham, pp. 244–257.

Bône, A., Vernhet, P., Colliot, O., Durrleman, S., 2020. Learning joint shape and appearance representations with metamorphic auto-encoders. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12261. LNCS, pp. 202–211. doi:10.1007/978-3-030-59710-8_20.

Byrne, N., Clough, J.R., Montana, G., King, A.P., 2020. A persistent homology-based topological loss function for multi-class CNN segmentation of cardiac MRI. In: MICCAI STACOM, pp. 1–11.

Cabezas, M., Oliver, A., Lladó, X., Freixenet, J., Bach Cuadra, M., 2011. A review of atlas-based segmentation for magnetic resonance brain images. Comput. Methods Prog. Biomed. 104 (3), 158–177. doi:10.1016/j.cmpb.2011.07.015.

Cerrolaza, J.J., Li, Y., Biffi, C., Gomez, A., Sinclair, M., Matthew, J., Knight, C., Kainz, B., Rueckert, D., 2018. 3D fetal skull reconstruction from 2DUS via deep conditional generative networks. In: Frangi Alejandro, F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing, Cham, pp. 383–391.

Chabiniok, R., Wang, V.Y., Hadjicharalambous, M., Asner, L., Lee, J., Sermesant, M., Kuhl, E., Young, A.A., Moireau, P., Nash, M.P., Chapelle, D., Nordsletten, D.A., 2016. Multiphysics and multiscale modelling, data-model fusion and integration of organ physiology in the clinic: ventricular cardiac mechanics. Interface Focus 6 (2). doi:10.1098/rsfs.2015.0083.

Chen, C., Biffi, C., Tarroni, G., Petersen, S., Bai, W., Rueckert, D., 2019. Learning shape priors for robust cardiac mr segmentation from multi-view images. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11765. LNCS, pp. 523–531. doi:10.1007/978-3-030-32245-8_58.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Springer International Publishing, Cham, pp. 424–432.

Clough, J.R., Oksuz, I., Byrne, N., Schnabel, J.A., King, A.P., 2019. Explicit topological priors for deep-learning based image segmentation using persistent homology. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11492. LNCS, pp. 16–28. doi:10.1007/978-3-030-20351-1_2.

Dalca, A., Rakic, M., Guttag, J., Sabuncu, M., 2019. Learning conditional deformable templates with convolutional networks. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc., pp. 806–818.

Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R., 2019. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Med. Image Anal. 57, 226–236. doi:10.1016/j.media.2019.07.006.

Dalca, A.V., Guttag, J.V., Sabuncu, M.R., 2018. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018. IEEE Computer Society, pp. 9290–9299. doi:10.1109/CVPR.2018.00968.

Dalca, A.V., Yu, E., Golland, P., Fischl, B., Sabuncu, M.R., Eugenio Iglesias, J., 2019. Unsupervised deep learning for bayesian brain MRIsegmentation. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention–MICCAI 2019. Springer International Publishing, Cham, pp. 356–365.

Dong, S., Luo, G., Tam, C., Wang, W., Wang, K., Cao, S., Chen, B., Zhang, H., Li, S., 2020. Deep atlas network for efficient 3D left ventricle segmentation on echocardiography. Med. Image Anal. 61, 101638. doi:10.1016/j.media.2020.101638.

Duan, J., Bello, G., Schlemper, J., Bai, W., Dawes, T.J.W., Biffi, C., de Marvao, A., Doumoud, G., O'Regan, D.P., Rueckert, D., 2019. Automatic 3D bi-ventricular segmentation of cardiac images by a shape-refined multi- task deep learning approach. IEEE Trans. Med. Imaging 38 (9), 2151–2164. doi:10.1109/tmi.2019.2894322.

Guimond, A., Meunier, J., Thirion, J.P., 2000. Average brain models: a convergence study. Comput. Vis. Image Underst. 77 (2), 192–210. doi:10.1006/cviu.1999.0815.

Haskins, G., Kruger, U., Yan, P., 2020. Deep learning in medical image registration: a survey. Mach. Vis. Appl. 31 (1), 1–18. doi:10.1007/s00138-020-01060-x.

Heimann, T., Meinzer, H.P., 2009. Statistical shape models for 3D medical image segmentation: a review. Med. Image Anal. 13 (4), 543–563. doi:10.1016/j.media.2009.05.004.

Heinrich, M.P., Oster, J., 2018. MRI whole heart segmentation using discrete non-linear registration and fast non-local fusion. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10663. LNCS, pp. 233–241. doi:10.1007/978-3-319-75541-0_25.

Hoffmann, M., Billot, B., Iglesias, J. E., Fischl, B., Dalca, A. V., 2020. Learning image registration without images. arXiv:2004.10282

Hui, L., Belkin, M., 2021. Evaluation of neural architectures trained with square loss vs. cross-entropy in classification tasks. arXiv:2006.07322

Iglesias, J.E., Liu, C.-Y., Thompson, P.M., Tu, Z., 2011. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imaging 30 (9), 1617–1634.

Iglesias, J.E., Sabuncu, M.R., 2015. Multi-atlas segmentation of biomedical images: asurvey. Med. Image Anal. 24 (1), 205–219. doi:10.1016/j.media.2015.06.012.

Išgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M.A., Van Ginneken, B., 2009. Multi-atlas-based segmentation with local decision fusion-application to cardiac and aortic segmentation in CT scans. IEEE Trans. Med. Imaging 28 (7), 1000–1010. doi:10.1109/TMI.2008.2011480.

Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K., 2015. Spatial Transformer Networks. In: Neural Information Processing Symposium, vol. 2, pp. 2017–2025. doi:10.1145/2948076.2948084.

Joshi, S., Davis, B., Jomier, M., Gerig, G., 2004. Unbiased diffeomorphic atlas construction for computational anatomy. NeuroImage 23, S151–S160.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78. doi:10.1016/j.media.2016.10.004.

Kirişli, H.A., Schaap, M., Klein, S., Papadopoulou, S.L., Bonardi, M., Chen, C.H., Weustink, A.C., Mollet, N.R., Vonken, E.J., Van Der Geest, R.J., Van Walsum, T., Niessen, W.J., 2010. Evaluation of a multi-atlas based method for segmentation of cardiac CTA data: a large-scale, multicenter, and multivendor study. Med. Phys. 37 (12), 6279–6291. doi:10.1118/1.3512795.

Krebs, J., Delingette, H., Mailhe, B., Ayache, N., Mansi, T., 2019. Learning a probabilistic model for diffeomorphic registration. IEEE Trans. Med. Imaging 38 (9), 2165–2176. doi:10.1109/TMI.2019.2897112.

Lamata, P., Sinclair, M., Kerfoot, E., Lee, A., Crozier, A., Blazevic, B., Land, S., Lewandowski, A.J., Barber, D., Niederer, S., Smith, N., 2014. An automatic service for the personalization of ventricular cardiac meshes. J. R. Soc. Interface 11 (91). doi:10.1098/rsif.2013.1023.

Larrazabal, A. J., Martinez, C., Ferrante, E., 2019. Anatomical Priors for Image Segmentation via Post-processing with Denoising Autoencoders (i), 585–593. arXiv:1906.02343

Lee, M., Oktay, O., Schuh, A., Schaap, M., Glocker, B., 2019. Image-and-spatial transformer networks for structure-guided image registration. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 337–345. doi:10.1007/978-3-030-32245-8_38.

Lee, M.C.H., Petersen, K., Pawlowski, N., Glocker, B., Schaap, M., 2019. TETRIS: template transformer networks for image segmentation with shape priors. IEEE Trans. Med. Imaging 1. doi:10.1109/tmi.2019.2905990.

Li, S., Zhang, C., He, X., et al., 2020. Shape-aware semi-supervised 3D semantic segmentation for medical images. In: Martel, A.L., et al. (Eds.), MICCAI 2020, vol. 12261. Springer International Publishing, pp. 552–561. doi:10.1007/978-3-030-59710-8_54.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Maintz, J.B., Viergever, M.A., 1998. A survey of medical image registration. Med. Image Anal. 2 (1), 1–36. doi:10.1016/S1361-8415(01)80026-8.

Mansilla, L., Milone, D. H., Ferrante, E., 2020. Learning Deformable Registration of Medical Images with Anatomical Constraints (i). arXiv:2001.07183

Medrano-Gracia, P., Cowan, B.R., Ambale-Venkatesh, B., Bluemke, D.A., Eng, J., Finn, J.P., Fonseca, C.G., Lima, J.A., Suinesiaputra, A., Young, A.A., 2014. Left ventricular shape variation in asymptomatic populations: the multi-ethnic study of atherosclerosis. J. Cardiovasc. Magn. Reson. 16 (1), 1–10. doi:10.1186/s12968-014-0056-2.

Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L.R., Griffanti, L., Douaud, G., Okell, T.W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P.M., Smith, S.M., 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat. Neurosci. 19 (11), 1523–1536. doi:10.1038/nn.4393.

Milletari, F., Rothberg, A., Jia, J., Sofka, M., 2017. Integrating statistical prior knowledge into convolutional neural networks. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10433. LNCS, pp. 161–168. doi:10.1007/978-3-319-66182-7_19.

Nicol, E.D., Norgaard, B.L., Blanke, P., Ahmadi, A., Weir-McCall, J., Horvat, P.M., Han, K., Bax, J.J., Leipsic, J., 2019. The future of cardiovascular computed tomography: advanced analytics and clinical insights. JACC 12 (6), 1058–1072. doi:10.1016/j.jcmg.2018.11.037.

Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., De Marvao, A., Dawes, T., O'Regan, D.P., Kainz, B., Glocker, B., Rueckert, D., 2018. Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE Trans. Med. Imaging 37 (2), 384–395. doi:10.1109/TMI.2017.2743464.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: an imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc., pp. 8024–8035.

Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M., 2011. A Bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage 56 (3), 907–922.

Pham, D.L., Xu, C., Prince, J.L., 2000. Current methods in medical image segmentation. Annu. Rev. Biomed. Eng. 2 (1), 315–337. doi:10.1146/annurev.bioeng.2.1.315.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9351, pp. 12–20. doi:10.1007/978-3-319-24574-4.

Shafto, M.A., Tyler, L.K., Dixon, M., Taylor, J.R., Rowe, J.B., Cusack, R., Calder, A.J., Marslen-Wilson, W.D., Duncan, J., Dalgleish, T., et al., 2014. The cambridge centre for ageing and neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. BMC Neurol. 14 (1), 204.

Stergios, C., Mihir, S., Maria, V., Guillaume, C., Marie-Pierre, R., Stavroula, M., Nikos, P., et al., 2018. Linear and deformable image registration with 3D convolutional neural networks. In: Stoyanov, D., et al. (Eds.), RAMBO 2018/BIA 2018/TIA 2018. Springer International Publishing, pp. 13–22. doi:10.1007/978-3-030-00946-5_2.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R., 2015. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12 (3), e1001779. doi:10.1371/journal.pmed.1001779.

Taylor, C.A., Fonte, T.A., Min, J.K., 2013. Computational fluid dynamics applied to cardiac computed tomography for noninvasive quantification of fractional flow reserve: scientific basis. J. Am. Coll. Cardiol. 61 (22), 2233–2241. doi:10.1016/j.jacc.2012.11.083.

Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., Tyler, L.K., Henson, R.N., et al., 2017. The cambridge centre for ageing and neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. NeuroImage 144, 262–269. doi:10.1016/j.neuroimage.2015.09.018.

Tilborghs, S., Dresselaers, T., Claus, P., Bogaert, J., Maes, F., 2020. Shape constrained CNN for cardiac MR segmentation with simultaneous prediction of shape and pose parameters. arXiv:2010.08952

Tóthová, K., Parisot, S., Lee, M., Puyol-Antón, E., King, A., Pollefeys, M., Konukoglu, E., 2020. Probabilistic 3D surface reconstruction from sparse MRI information. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention–MICCAI 2020. Springer International Publishing, Cham, pp. 813–823.

Tóthová, K., Parisot, S., Lee, M.C.H., Puyol-Antón, E., Koch, L.M., King, A.P., Konukoglu, E., Pollefeys, M., 2018. Uncertainty quantification in CNN-based surface prediction using shape priors. In: Reuter, M., Wachinger, C., Lombaert, H., Paniagua, B., Lüthi, M., Egger, B. (Eds.), International Workshop on Shape in Medical Imaging. Springer International Publishing, Cham, pp. 300–310.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320.

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. doi:10.1016/j.media.2018.11.010.

Xu, Z., Niethammer, M., 2019. DeepAtlas: Joint Semi-supervised Learning of Image Registration and Segmentation (Dl), 420–429. arXiv:1904.08465

Ye, M., Huang, Q., Yang, D., Wu, P., Yi, J., Axel, L., Metaxas, D., 2020. PC-U Net: learning to jointly reconstruct and segment the cardiac walls in 3D from CT data. arXiv:2008.08194

Young, A.A., Frangi, A.F., 2009. Computational cardiac atlases: from patient to population and back. Exp. Physiol. 94 (5), 578–596. doi:10.1113/expphysiol.2008.044081.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C., 2020. Gradient surgery for multi-task learning.

Zhao, A., Balakrishnan, G., Durand, F., Guttag, J. V., Dalca, A. V., 2019. Data augmentation using learned transformations for one-shot medical image segmentation.

Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., Yang, X., Heng, P.A., Mortazi, A., Bagci, U., Yang, G., Sun, C., Galisot, G., Ramel, J.Y., Brouard, T., Tong, Q., Si, W., Liao, X., Zeng, G., Shi, Z., Zheng, G., Wang, C., MacGillivray, T., Newby, D., Rhode, K., Ourselin, S., Mohiaddin, R., Keegan, J., Firmin, D., Yang, G., 2019. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Med. Image Anal. 58, 101537. doi:10.1016/j.media.2019.101537.