MELBA Journal Machine Learning for Biomedical Imaging

Robust deformable image registration using synthetic data and transfer learning

Iris D. **Kolenbrander** [1,2], Matteo Maspero [3,4], Josien P.W. Pluim [1]

1 IMAG/e Group, Department of Biomedical Engineering, Eindhoven University of Technology, The Netherlands
2 Eindhoven Artificial Intelligence Systems Institute, Eindhoven University of Technology, The Netherlands
3 Computational Imaging Group for MR Diagnostics & Therapy, University Medical Center Utrecht, The Netherlands
4 Department of Radiotherapy,University Medical Center Utrecht, The Netherlands

Abstract

Deep learning models have recently achieved accuracy comparable to traditional iterative methods for deformable image registration. However, their performance often degrades when test data deviate from the training distribution, for example due to changes in imaging protocols or subject demographics. While data augmentation can improve robustness, it may not account for all types of domain shifts, particularly those that are unpredictable. To address this limitation, recent approaches have explored training on synthetic images with diverse structures and intensities to promote generalization across modalities and anatomical regions. Nevertheless, the robustness of such models to other types of domain shifts, and their adaptability to specific registration tasks, remains unclear. In this study, we investigate whether transfer learning can enhance registration model robustness under unforeseen domain shifts. We first train a generic, weakly supervised model on synthetic data, then fine-tune it on target domains. Specifically, we examine: (1) how design factors in synthetic data influence cross-domain performance (e.g., brain MRI vs. lung CT); (2) the efficiency of fine-tuning compared to training from scratch; and (3) the robustness of fine-tuned models on previously unseen datasets. Our results show that the fine-tuned model achieved the most consistent registration performance across datasets. It also required fewer training epochs and less target domain data than from-scratch training. Furthermore, while the use of random geometric shapes and dynamic contrast in the synthetic data was critical for cross-domain performance, we found that different target domains may benefit from different synthetic data. These findings support transfer learning as a promising strategy to improve robustness to diverse domain shifts in deformable image registration.

Keywords

Deep learning, deformable image registration, synthetic data, transfer learning, robustness, domain shifts.

Article informations

https://doi.org/10.59275/j.melba.2025-5dab

Received: 2024-09-17, Published 2025-08-06

Corresponding author: i.d.kolenbrander@tue.nl

1. Introduction

Deformable image registration is crucial in medical image analysis. It aligns two medical images, one fixed and one moving, by estimating a nonlinear spatial transformation, which is often parametrized as the displacement vector field (DVF). Traditional registration methods update the DVF iteratively by optimizing an objective function that balances image similarity and regularization. This process can take several minutes per image pair.

In contrast, deep learning models can register images in seconds, making them well-suited for time-critical applications such as image-guided intervention, surgery, and adaptive radiotherapy (Terpstra et al., 2021; Chehab et al., 2023; Kolenbrander et al., 2024). These models are trained on large datasets of image pairs and once trained, they can predict the DVF directly in a single forward pass. This enables fast and accurate registration on new image pairs (Balakrishnan et al., 2019; Hering et al., 2021; Hoffmann et al., 2022; Mok et al., 2024).

Despite this advantage, deep learning-based registration faces a significant challenge in generalizing to unseen data variations. Differences in imaging protocols, scanner types, anatomical sites, or populations can introduce discrepancies that affect registration (Ketcha et al., 2019; Mok et al., 2024). Addressing these domain shifts is essential for ensuring its reliable usage in clinical settings.

Data augmentation is often used to address this challenge (Chlap et al., 2021). Common methods include geometric transformations, e.g., affine transformations, and intensity modifications, e.g., gamma transformations, contrast adjustments, and noise injection (Terpstra et al., 2021; de Vos et al., 2024). However, the effectiveness of data augmentation may diminish when domain shifts are unpredictable because the augmented data may not accurately represent new, unseen variations. This can lead to suboptimal model performance.

Recently, Hoffmann et al. (2022) proposed generating synthetic training data to improve the robustness of registration models. Their approach involved training a model on synthetic images with diverse structures and intensities to promote generalization across different MRI contrasts and anatomical regions, including the heart and brain. While this method enables models to handle substantial domain shifts, such as changes in imaging modality, many clinical scenarios require models that are fine-tuned for specific registration tasks. In addition, it is unclear whether this approach improves robustness to other types of domain shifts, such as variations in scanner types, imaging protocols, image quality, or patient demographics.

This is where transfer learning can be valuable. It aims to transfer knowledge from a generic model to a specific domain, producing a model that is both robust and domain-specific. Typically, this involves pre-training a model on a source dataset and then fine-tuning it on a target domain dataset. One of the main advantages is its reduced dependency on large, diverse, and labeled datasets from the target domain, a benefit demonstrated across various medical image analysis tasks (Oliveira and dos Santos, 2018; Kang and Gwak, 2019; Li et al., 2024). For instance, models pre-trained on the ImageNet dataset, such as AlexNet and ResNet, have been successfully applied to disease classification tasks (Wang et al., 2021; Talo et al., 2019). In 3D semantic segmentation, where annotated medical datasets are often scarce, pre-training becomes even more critical. To address this challenge, Li et al. (2024) constructed a large, annotated CT dataset for supervised pre-training. Their findings showed that pre-training on a related dataset improved performance on the target task and enabled the model to segment previously unseen structures and classify diseases more accurately.

Research on transfer learning in deformable image registration is still limited. Some studies focus on adapting a model to a single unseen image pair during testing (instance optimization), which improves the registration of out-of-distribution samples (Ferrante et al., 2018; Guan et al., 2021; Zhu et al., 2021; Wang et al., 2022; Mok et al., 2024). However, this optimization adds around a 30-second delay to the registration process (Mok et al., 2024; Wang et al., 2022). Transferring models to entire datasets has received little attention. One recent effort in this direction is UniGradICON (Tian et al., 2024), a universal registration model that was trained using a composite dataset containing over 5,000 images from multiple modalities (CT and MRI) and anatomical regions (lung, knee, brain, and abdomen). As one of the first universal registration models, UniGradICON paves the way toward transfer learning. In this work, we focus on a transfer learning approach that does not require a large composite dataset of acquired images for pre-training.

This study investigates whether transfer learning, based on synthetic data training, can improve the robustness of deformable image registration models under unforeseen domain shifts. Our contributions are threefold: (1) We study how design factors in the synthetic data influence cross-domain performance, including different imaging modalities (CT and MRI) and anatomical regions (lung and brain); (2) We investigate the efficiency of fine-tuning versus training from scratch in lung CT registration; (3) We assess the robustness of fine-tuned models on previously unseen datasets in lung CT and brain MRI registration. The transfer learning approach is compared against training from scratch and data augmentation.

2. Methods

We first describe the deep learning models used in our study, followed by the transfer learning strategy, including the pre-training and fine-tuning stages. We then present the synthetic data used for pre-training and the data used for fine-tuning and performance evaluations. Finally, we describe the evaluation metrics and methods used for comparison. Our code is available at `https://github.com/iriskolenbrander/robust-DIR`.

2.1 Deep learning models

The U-Net architecture is used throughout this study (Ronneberger et al., 2015), using a single U-Net to understand the impact of synthetic data design (Section 3.1), and a cascaded U-Net in transfer learning experiments (Section 3.2) (Figure 1). The U-Nets predict a displacement vector field (DVF: ϕ) that aligns the moving image to the fixed image. The DVF is parameterized as a stationary velocity vector field (Balakrishnan et al., 2019), which represents a constant-in-time (stationary) flow and produces a smooth and invertible DVF when integrated. This integration uses the five-step scaling and squaring method (Arsigny et al., 2006).

2.1.1 Single U-Net:

The U-Net has four convolution blocks in the encoder (256 filters) and three in the decoder (256), skip connections, and a registration head consisting of four convolution blocks (256) (Figure 1). Each convolution block contains a 3D convolution with a kernel size of three, followed by LeakyReLU activation (slope=-0.2). The registration head produces a stationary velocity field at half the resolution of the input images. This vector field is integrated and resampled to obtain the full-resolution DVF.

2.1.2 Cascaded U-Net:

The cascaded U-Net contains two U-Nets with configurations similar to the single U-Net. The first U-Net (64 filters) operates at half the image resolution, and a second U-Net (128 filters) operates at full image resolution (Figure 1). Its output is upsampled to a full-resolution DVF, which deforms the input to the second U-Net. The two predicted DVFs are combined into a single DVF. Several studies have shown that cascaded models improve registration compared to single networks (Mok and Chung, 2020; Jiang et al., 2020; Hering et al., 2021).

2.2 Transfer learning

The transfer learning approach involves pre-training deep learning registration models on synthetic data, followed by fine-tuning on a specific target dataset. Both stages used weak supervision.

2.2.1 Model pre-training

The model is first trained on synthetic data, comprising gray value moving and fixed images (M, F) and corresponding segmentation labels (L_M, L_F) (section 2.3.1 details the synthetic data). The segmentation labels are used solely during training and are not required as input to the model at inference. The objective function ($\mathcal{L}$) includes a Dice loss term ($\mathcal{L}_{Dice}$) that maximizes the overlap between the fixed and deformed moving segmentation labels (Equation 1). It also includes a regularization term ($\mathcal{L}_{reg}$), which is the L2-norm of the DVF gradients scaled by the regularization weight (λ).

$$\mathcal{L} = \mathcal{L}_{Dice}(L_F, L_M, \phi) + \lambda \mathcal{L}_{reg}(\phi), \tag{1}$$

Let L_F^j and L_M^j be the one-hot representations of the label j in the fixed and moving label maps L_F and L_M. The Dice loss is the average Dice loss of all labels $j \in [1, J]$:

$$\mathcal{L}_{Dice}(L_F, L_M, \phi) = 1 - \frac{2}{J}\sum_{j=1}^{J} \frac{|L_F^j \odot \phi(L_M^j)| + 1\mathrm{e}{-5}}{|L_F^j| + |\phi(L_M^j)| + 1\mathrm{e}{-5}} \tag{2}$$

$$\mathcal{L}_{reg}(\phi) = \frac{1}{P}\sum_{p=1}^{P} ||\nabla\phi(p)||_2, \tag{3}$$

Here, $\odot$ denotes voxel-wise multiplication, $|L^j|$ represents the sum of all non-zero voxels in the label map, $\nabla\phi(p) = (\frac{\partial\phi((p)}{\partial x}, \frac{\partial\phi((p)}{\partial y}, \frac{\partial\phi((p)}{\partial z})$, and P is the number of voxels in the DVF.

2.2.2 Model fine-tuning

The model is fine-tuned by further training it on the lung CT training set (detailed in section 2.3.2) or the brain MRI training set (detailed in section 2.3.3) using the same training procedure described in Section 2.2.1. The objective function is similar to equation 1 with the Dice loss maximizing the overlap of the target domain segmentations, i.e., segmentations of the pulmonary veins and arteries for lung CT fine-tuning and brain structure segmentations for brain MRI fine-tuning.

2.2.3 Training details

We train the models in 1000 epochs (unless otherwise specified) using the Adam optimizer and a batch size of one on a PC with an Nvidia A100-SXM4 40GB vRAM GPU. Model convergence is verified based on the average validation Dice: when the change in Dice running average (averaged over three epochs) was below 0.0001 for four consecutive epochs. We tune the learning rate and the regularization weight in 17 validation runs with a random search in the ranges [0.00005-0.001] and [0-10], finding the optimal learning rate of 0.0001 and regularization weight of 0.75.

2.3 Data

2.3.1 Synthetic data

Based on the method of Hoffmann et al. (2022), we generate images with random structures and intensities (Figure 2). We begin with generating random label maps. Then, each label map is deformed using a pair of deformations, ϕ_F and ϕ_M, to form a pair of fixed and moving label maps. Finally, these are turned into gray value images.

Step 1. Label map generation: We generate 100 3D label maps with a size of $192 \times 192 \times 192$. Each label map contains 26 (J) labels that can occur multiple times within a single map. The maps are saved to disk for reuse in each epoch. The label maps contain shapes of random geometry, rectangles, or spheres (Figure 2). To study the impact of shape size, we also created three separate datasets of random geometry with small, medium and large random shapes. The methods to create the different label maps are detailed in Appendix A.

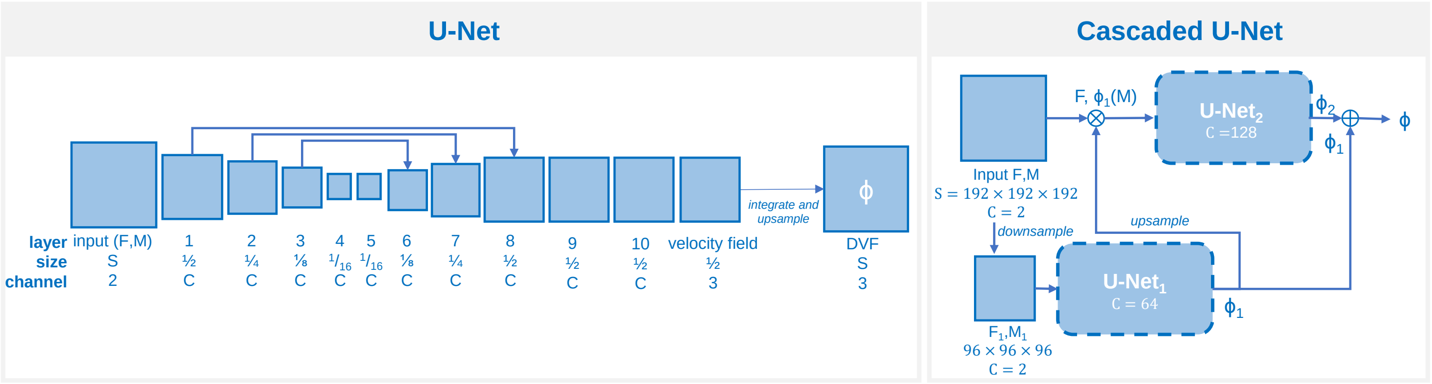

Figure 1: Neural network architectures: A single U-Net was used for experiments involving synthetic data design, while a cascaded U-Net (two U-Nets) was used in experiments focused on transfer learning. Each U-Net has C channels (filters) in each convolutional block. The $\otimes$ operation indicates applying the DVF (ϕ) to an image.

Step 2. Deformation generation: Two DVFs, ϕ_F and ϕ_M, are generated using the method of Hoffmann et al. (2022) to deform each label map, thereby obtaining the fixed and moving label maps, respectively. Each DVF starts as three grids with sizes $s \times s \times s \times 3$ for $s \in [24, 12, 6]$. The grid values are sampled from $\mathcal{N}(0, \sigma_\phi^2)$ with σ_ϕ sampled between 0 and 3. The three grids are upsampled to full-sized stationary velocity fields through linear interpolation, which are summed and integrated through scaling and squaring (step=5) to produce the DVF. Each epoch, a new set of ϕ_F and ϕ_M deformations is generated on the fly to increase the variety of synthetic data. Although both ϕ_F and ϕ_M are individually diffeomorphic, the composed transformation between the fixed and moving labels is not necessarily diffeomorphic.

Step 3. Image generation: A 3D gray value image is created from each fixed and moving label map. We assign to each label j in the label map a Gaussian distribution $\mathcal{N}(\mu_j, \sigma_j^2)$ from which the image intensities are sampled independently. For each label $j \in [1, J]$, μ_j and σ_j are sampled from uniform distributions $\mathcal{U}(a_\mu, b_\mu)$ and $\mathcal{U}(a_\sigma, b_\sigma)$. The images are generated on the fly for each training epoch by sampling new values for μ_j and σ_j for each image pair. We study three configurations for setting μ_j and σ_j (Figure 2):

- Static: The same μ_j and σ_j values generate the fixed and moving images within an image pair, resulting in similar fixed and moving image appearances. Here, a_μ and b_μ are set to 0.1 and 1.0, and a_σ and b_σ to 0.02 and 0.1. We expect this configuration to be optimal for mono-modal registration.

- Dynamic: Different μ_j and σ_j values generate the fixed and moving images within an image pair, resulting in different fixed and moving image appearances. Similar to the static configuration, a_μ and b_μ are set to 0.1 and 1.0, and a_σ and b_σ to 0.02 and 0.1. We expect this configuration to be critical for multi-modal registration.

- Dynamic*: Similar to the dynamic configuration, but with larger σ_j values sampled from $\mathcal{U}(a_\sigma, b_\sigma)$ with $a_\sigma = 0.05$ and $b_\sigma = 0.2$. This configuration results in more uniformly distributed intensities with fewer and less intense peaks and allows us to study the effect of the distribution on the registration. This is important because we observed that acquired images of various types (including CT and MR) naturally have flatter intensity distributions than those generated with dynamic contrast.

Unlike Hoffmann et al. (2022), we do not include image corruptions such as synthetic bias fields and contrast augmentation to generalize the process beyond MRI, as our experiments also include other image types. We also remove the simulated partial volume effects, further simplifying the generation of images. We verified that removing these corruptions did not affect the registration (Appendix D.1).

2.3.2 Lung CT

3D thoracic CTs were obtained from three open-access datasets: NLST (Team, 2013; Aberle et al., 2011), DIR-Lab (Castillo et al., 2009a,b), and DIR-Lab-COPDgene (Castillo et al., 2013). The NLST dataset is used for training (fine-tuning), validation, and testing. The DIR-Lab and DIR-Lab-COPDgene datasets are used as hold-out test sets to evaluate robustness across datasets. These datasets originated from different sources, and the imaging parameters varied (Table B.1, Appendix B).

1. **NLST:** The dataset includes longitudinal pairs of low-dose CTs of 210 adults at high risk for lung cancer. We use a dataset version with readily preprocessed images, which had voxel sizes of $1.5 \times 1.5 \times 1.5$ mm and image sizes of $224 \times 192 \times 224$ voxels (Learn2Reg, 2023). Further preprocessing of the images included cropping to

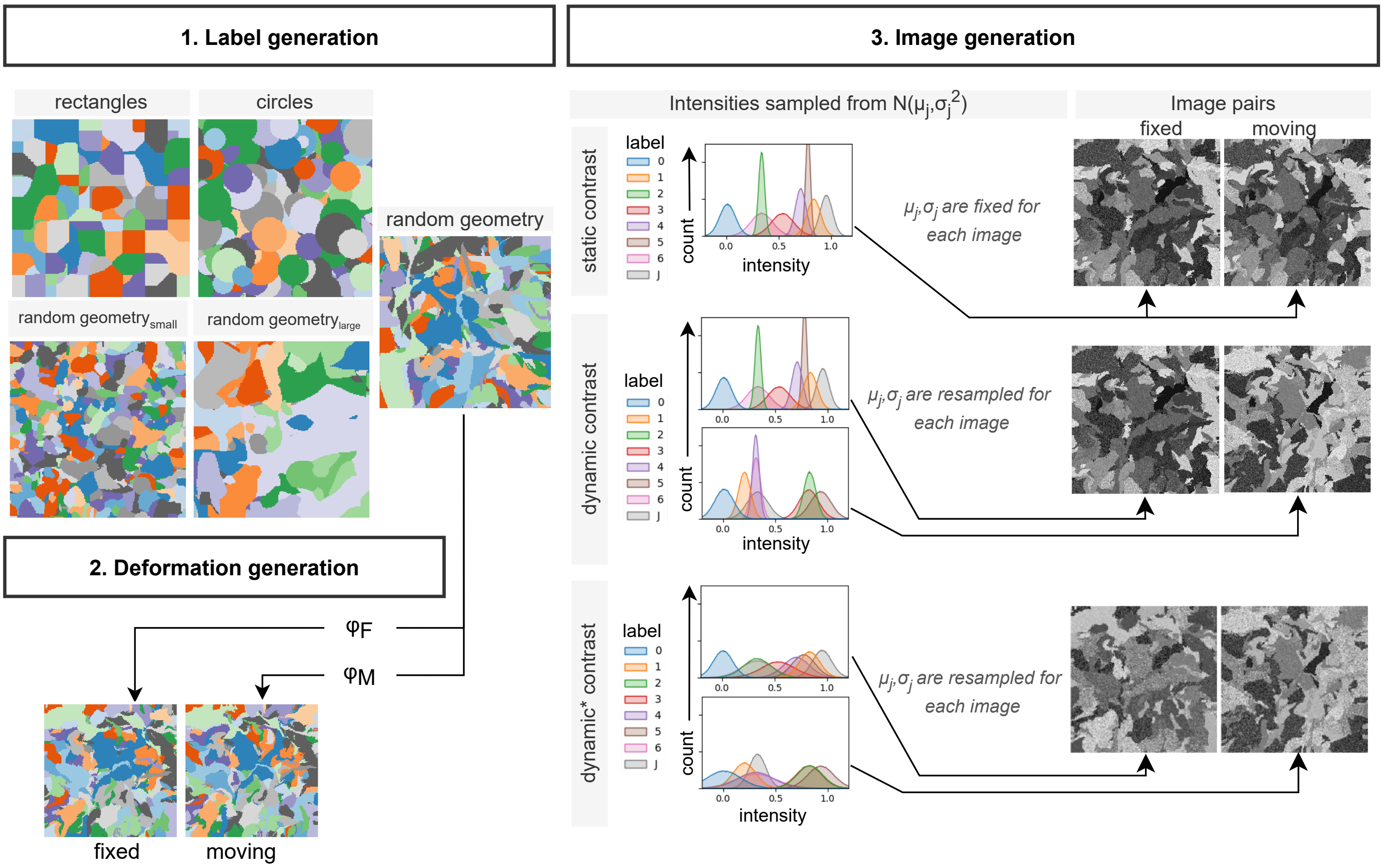
1. Label generation
rectangles
circles
random geometry
random geometry$_{small}$
random geometry$_{large}$
2. Deformation generation
φF
φM
fixed
moving
3. Image generation
Intensities sampled from N(μj,σj²)
Image pairs
fixed
moving
static contrast
dynamic contrast
dynamic* contrast
label
0
1
2
3
4
5
6
J
count
intensity
0.0
0.5
1.0
μj,σj are fixed for each image
μj,σj are resampled for each image
μj,σj are resampled for each image

Figure 2: Synthetic data: **(1)** Label maps are generated containing structures of different shapes (random geometry, rectangles, and circles) and sizes (medium, small, and large); **(2)** A label map is deformed with synthetic DVFs (ϕ_F and ϕ_M) to form a pair of a fixed and a moving label map; **(3)** The label maps are turned into gray value images by sampling their intensities from label-specific Gaussian distributions $\mathcal{N}(\mu_j, \sigma_j^2)$ for $j \in [1, J]$, where μ_j and σ_j are sampled based on static, dynamic, and dynamic* configurations.

$192 \times 192 \times 192$ voxels, clipping the gray value intensities to the range [-1000; 200] Hounsfield Units (HU), and scaling to the range [-1; 1]. The data is divided into 170, 20, and 20 image pairs for training, validation, and testing. The dataset provided automatic segmentations of the lungs, pulmonary veins and pulmonary arteries, along with matched points within the lungs (approximately 1000-3500 points per image pair) (Isensee et al., 2021; Heinrich et al., 2015). Since these points are matched automatically, we will refer to them as keypoints instead of landmarks. We use the segmentations for weakly supervised model training and the keypoints for evaluation.

2. **DIR-Lab:** The dataset contains 4D CTs of 10 adults treated for thoracic malignancies. Each 4D CT consists of 10 breathing phases, from which we include the end-expiration and end-inspiration phases (3D CTs), used as fixed and moving images, respectively. The images are resampled to voxel sizes of $1.5 \times 1.5 \times 1.5$ mm, center-cropped to a standard size of $192 \times 192 \times 192$ voxels, clipped to the range [-1000; 200] HU, and scaled to the range [-1; 1]. For evaluation, the dataset provides 300 manually annotated landmark pairs within the lungs. Rough lung masks were obtained by thresholding below -250 HU and extracting the largest connected component. These masks were used exclusively for iterative registration (Section 2.4.1).

3. **DIR-Lab-COPDgene:** The dataset contains paired expiratory (fixed) and inspiratory (moving) breath-hold CTs of 10 adults with chronic obstructive pulmonary disease. We applied the same preprocessing steps as for DIR-Lab, including lung mask extraction used for iterative registration. As with DIR-Lab, the dataset provides 300 paired manual landmarks for evaluation.

2.3.3 Brain MRI

The 3D brain MR images from three open-access datasets were used: The Open Access Series of Imaging Studies (OASIS) (Marcus et al., 2007; Hoopes et al., 2021), the Information eXtraction from Images (IXI) dataset (IXI, 2024), and the paired3T dataset (Chen et al., 2023). OASIS is used for training (fine-tuning), validation, and testing. The other two datasets are used as hold-out test sets to evaluate robustness across datasets or to evaluate the synthetic data design.

1. **OASIS (T1-weighted MRI):** We used a readily preprocessed dataset (Marcus et al., 2007; Hoopes et al., 2021) with T1-weighted MRIs acquired at 1.5 Tesla of 414 subjects, including healthy adults and adults diagnosed with mild to moderate Alzheimer's disease. The images have a size of $180 \times 192 \times 180$ after skull-stripping, affine alignment to a standard atlas (FreeSurfer (Fischl, 2012)), resampling of voxel sizes to $1 \times 1 \times 1$ mm, cropping, and gray value scaling to the range [-1; 1]. We zero-pad these images to $192 \times 192 \times 192$ voxels. The data is divided into 312 images for training, and 51 images each for validation and testing. In the training set, each subject is used as a moving image and registered to a randomly selected fixed image, creating 312 image pairs. For validation and testing, one randomly selected subject serves as the fixed image, and all other subjects are registered to it, creating 50 image pairs in each set.

2. **IXI (T1-weighted MRI):** We used 78 T1-weighted MRIs of healthy adults, which were acquired at 1.5 or 3 Tesla and collected from three different hospitals (IXI, 2024). The images range in size from $130 - 150 \times 256 \times 256$ with voxel sizes of $1.2 \times 0.94 \times 0.94$. We preprocessed the images, including skull stripping, affine alignment to a standard atlas (FreeSurfer (Fischl, 2012)), resampling of voxel sizes to $1 \times 1 \times 1$ mm, cropping to $192 \times 192 \times 192$ voxels, and gray value scaling to the range [-1; 1]. We included 26 scans from each hospital, from which we randomly selected one subject as the fixed image, to which all other scans from the same hospital were registered, resulting in 75 image pairs.

3. **Paired3T (T1 and T2-weighted MRI):** The dataset contains paired T1 and T2-weighted MRIs of 10 healthy adults (without reported history of neurological or psychiatric diseases), which were acquired at 3 Tesla (Chen et al., 2023). The purpose of this dataset was two-fold: 1) To evaluate multi-modal registration (T1 to T2) in Section 3.1; 2) To perform mono-modal registration (T1 to T1) to evaluate cross-dataset robustness in Section 3.2.2. We perform inter-subject registration, using all unique subject combinations, resulting in 45 image pairs for evaluation. In Section 3.1, we obtain the from-scratch model by performing a three-fold cross-validation with training sets of 6 subjects each, where each subject serves as the fixed image once and all other subjects (moving images) are registered to it, resulting in 30 training image pairs.

   The provided images have a size of $256 \times 304 \times 308$ and voxel sizes of $0.65 \times 0.65 \times 0.65$ after face information removal and rigid alignment (T1 to T2). We performed skull-stripping, affine alignment of the moving images to the fixed image (Elastix (Klein et al., 2009)), resampling of voxel sizes to $1 \times 1 \times 1$ mm, cropping to $192 \times 192 \times 192$ voxels, and gray value scaling to the range [-1; 1].

All datasets provide semi-automatic labels for 35 cortical and subcortical brain structures, which were obtained using FreeSurfer (Fischl, 2012). We use a selection of structures for validation, which include the cerebral cortex, brainstem, lateral ventricle, and hippocampus labels, representing varying structure sizes and shapes and different brain areas.

2.3.4 Synthetic deformation model

We augment the training data on the fly during fine-tuning to increase the data diversity and ensure a good baseline registration. Specifically, we generate synthetic deformations to create virtual new subjects with augmented brain MRIs or lung CTs. There is a 0.5 probability of generating a new subject; otherwise, we use the original real subject.

Each on-the-fly generated subject consists of a synthetic fixed–moving image pair. The synthetic moving image is created by deforming the original moving image with ϕ_{aug}, a full-resolution DVF of size $192 \times 192 \times 192 \times 3$ voxels, obtained by upsampling a coarse $6 \times 6 \times 6 \times 3$ grid with displacement values between -4 and 4 voxels via third-order B-spline interpolation. The synthetic fixed image is created through ϕ_{M2F}, which models typical anatomical motion (e.g., expiratory breathing motion in lung CT or inter-patient variation in brain MRI) based on a statistical deformation model from Corral Acero et al. (2019) (detailed in Appendix C). The combined DVF of ϕ_{aug} and ϕ_{M2F} deforms the original moving image to obtain the synthetic fixed image.

2.4 Evaluation

2.4.1 Comparisons

We compared transfer learning with training from scratch, both with and without data augmentation, as well as with baseline iterative registration.

From-scratch training: From-scratch training is achieved using random (Kaiming) initialization of the model's trainable parameters and the same training procedure as for fine-tuning.

Data augmentation: Data augmentation included contrast and noise augmentations. Contrast augmentation involved scaling the image intensities (v) with a probability of 0.5 during training ($v = v * (1 + factor)$), where the factor is randomly sampled between -0.5 and +0.5. Noise augmentation involved adding Gaussian noise ($\mathcal{N}(0, 0.2)$) to the image intensities with a probability of 0.5.

Iterative registration: Lung CT registration was performed in Elastix (Klein et al., 2009) using the approach from Staring et al. (2010), which optimizes the normalized cross-correlation (NCC) and bending energy term (weight factor 0.05) in two stages via stochastic gradient descent in a multiresolution scheme of five resolutions. In the first stage, the optimization runs 1000 iterations in each of five resolutions, where the images are downsampled with factors of 16, 8, 4, 2, and 1, and the B-spline grid spacing is set to 80, 80, 40, 20, and 10 mm. In the second stage, the NCC is computed inside the lung mask of the fixed image, and the optimization runs 2000 iterations in each resolution with downsample factors of 4, 3, 2, 1, and 1, and B-spline grid spacings of 80, 40, 20, 10, and 5 mm.

Brain MRI registration was performed in ANTs with symmetric normalization, using NCC as the objective function and a multi-resolution scheme (Avants et al., 2008, 2009). The optimization runs 100 iterations in each of four resolutions, where the images are downsampled with factors of 8, 4, 2, and 1.

2.4.2 Metrics

The accuracy of lung CT registration is evaluated using the keypoint/landmark error, computed as the Euclidean distance for every pair of moving and fixed points. Brain MRI registration is evaluated using the 95th percentile Hausdorff distance (HD_{95}) and the average symmetric surface distance (ASSD) of the fixed and deformed moving contours. We also evaluate visual image alignment after registration, using the structural similarity index measure (SSIM) in all mono-modal registration tasks. SSIM was computed with a window size of 11 for brain registration and 21 for lung registration, to better accommodate larger misalignments. Furthermore, we evaluate the standard deviation of the logarithm of the DVF's Jacobian determinant (SDLogJ) and the percentage of non-positive values of the Jacobian determinant (indicative of undesired tissue folding).

2.4.3 Statistical analysis

We assessed whether evaluation metrics differed significantly between methods. For keypoint and landmark errors, values were first averaged across all landmarks per subject. Overall differences were tested using repeated measures ANOVA on rank-transformed data. When significant differences were found with ANOVA, pairwise comparisons were conducted using the Wilcoxon signed-rank test. Bonferroni correction was applied to account for multiple comparisons, based on the number of metrics multiplied by the number of method pairs.

3. Experiments

3.1 Training on synthetic data

This section identifies critical factors in the design of synthetic data that make a model applicable to different target domains. Specifically, we study the effect of different label map structures and intensity sampling configurations. A single U-Net is trained (using different synthetic datasets) in 200 epochs and evaluated on lung CT and brain MRI registration.

3.1.1 Structures

We compare three different structure shapes, random geometric, rectangular, and circular, and three different structure sizes of random geometric shapes (Section 2.3.1). We use dynamic contrast as the intensity distribution configuration and evaluate the resulting models on the NLST (lung CT) and OASIS (brain MRI) validation sets.

3.1.2 Intensity distribution

We study the impact of different intensity configurations on image registration. Specifically, our objective is to understand how the original method of Hoffmann et al. (2022), i.e., with different appearances for fixed and moving images (mimicking a multi-modal registration problem), affects mono-modal and multi-modal registration tasks. To address this, we train models on synthetic data of dynamic, dynamic*, and static configurations (Section 2.3.1) using random geometric shapes (medium size). We evaluate the resulting models on the NLST (lung CT), OASIS (brain $\mathrm{MRI}_{\text{T1-T1}}$), and Paired3T (brain $\mathrm{MRI}_{\text{T1-T2}}$) validation sets.

3.2 Transfer learning

This section investigates the efficiency and robustness of transfer learning. The cascaded U-Net is pre-trained in 1000 epochs (using the optimal synthetic data configuration) and fine-tuned in 300 epochs on either the lung CT training set ($\mathrm{NLST}_{\text{train}}$) or the brain MRI training set ($\mathrm{OASIS}_{\text{train}}$). For comparison, from-scratch cascaded U-Nets are trained in 300 epochs on the same datasets.

3.2.1 Efficiency

We investigate the efficiency of transfer learning in lung CT registration. The validation Dice (of the pulmonary veins and arteries' segmentations) was monitored during training to evaluate the convergence speed of both fine-tuned and from-scratch models. Furthermore, we examine the impact of different training subsets with dataset sizes (N) of 10, 20, 30, 50, 75, 100, and 170 (the complete $\mathrm{NLST}_{\text{train}}$ dataset) image pairs on the validation keypoint error. We also include a scenario with "infinitely" many synthetic training subjects ($N = \infty$), created on the fly using the statistical deformation model (Section 2.3.4). Each experiment is carried out three times (except for $N = 170$ and $N = \infty$) with different random seeds to minimize the potential bias of stochastic processes.

3.2.2 Robustness

We assess model robustness across datasets of the same modality and anatomical region but acquired in different studies. More specifically, the datasets originate from different institutions and were acquired with different scanners and imaging protocols (Table B.1, Appendix B). NLST and OASIS, while used for fine-tuning, serve as in-distribution datasets, while the others are considered out-of-distribution.

4. Results

4.1 Training on synthetic data

4.1.1 Structures

Training with random geometric shapes resulted in the best overall registration (Table 1). In lung CT registration, random shapes achieved a median (interquartile range, IQR) keypoint error of 3.5 mm (2.0–7.4), which was not significantly different from the 4.3 mm (2.5–7.9) observed with both circles and squares. In brain MRI registration, random shapes resulted in lower HD_{95} and ASSD values than circles and squares. For instance, the cerebral cortex had a median HD_{95} of 2.2 mm (2.2–2.3), compared to 3.0 mm (2.6–3.2) and 2.6 mm (2.3–2.9) when using circles and squares, respectively ($p<0.05$). Similarly, the brainstem's median ASSD was 0.6 mm (0.5–0.6), outperforming circles (1.0 mm, 0.8–1.1) and squares (0.8 mm, 0.8–0.9) ($p<0.05$)).

While shape size had a smaller impact, medium-sized shapes provided the best alignment for most brain structures (Table 1). In the lungs, they achieved a keypoint error of 3.5 mm (2.0-7.4), which was smaller than the 3.8 mm (1.9-9.1) of large shapes ($p<0.05$) and not significantly different from the 3.2 mm (1.4-9.9) of small shapes ($p=0.06$) (Table 1). Based on these findings, medium-sized random geometric shapes were used for the remainder of the study. Folding was 0.00 (0.00-0.00) for all shape and size configurations.

4.1.2 Intensity distribution

Static contrast achieves the best mono-modal lung and brain registrations (NLST and OASIS) (Table 2). For example, it achieved a median (IQR) keypoint error of 2.9 mm (1.4-6.6) in lung CT registration compared to 3.5 mm (2.0-7.5) with dynamic contrast ($p<0.05$) and 4.8 mm (3.0-8.9) with dynamic* contrast ($p<0.05$). In brain $MRI_{T1\text{-}T1}$ registration of the cerebral cortex, it achieved HD_{95} and ASSD values of 1.4 mm (1.4-1.6) and 0.6 mm (0.6-0.7), compared to 2.2 mm (2.2-2.3) and 0.9 mm (0.9-0.9) for dynamic contrast ($p<0.05$). However, static contrast fails in multi-modal brain $MRI_{T1\text{-}T2}$ registration (Paired3T); for example, the brainstem HD_{95} increased from 4.5 mm (4.0-5.9) before registration to 11.7 mm (10.2-13.3) after registration ($p<0.05$). In addition, we observed unstable training with static contrast beyond 200 epochs.

Dynamic contrast, on the other hand, showed more consistent performance across both mono-modal (NLST, OASIS) and multi-modal (Paired3T) registration tasks (Table 2). For instance, it achieved a median HD_{95} of 3.6 mm (3.0–4.2) of the lateral ventricle, compared to the 14.7 mm (13.8–15.2) of static contrast ($p<0.05$). The consistency of dynamic contrast likely stems from the model's ability to learn structural correspondences independent of intensity. Dynamic* contrast, despite having intensity distributions similar to CT and MRI, does not further improve registration compared to dynamic contrast and even degrades performance in lung CT ($p<0.05$). This may be due to increased image noise, introducing discrepancies between the training and validation noise levels. Folding was 0.00 (0.00-0.00) for all contrast configurations. We used dynamic contrast for all subsequent experiments due to its consistent performance and training stability.

4.2 Transfer learning

4.2.1 Efficiency

The lung CT registration model is fine-tuned in 22 epochs when pre-trained on synthetic data, compared to 47 epochs required for training from scratch (Figure 3a). The fine-tuned model achieves reasonable accuracy (median error < 2 mm) with just 20 image pairs, whereas training from scratch requires 75 image pairs, representing a 73% reduction in training data (Figure 3b). Notably, pre-training alone leads to keypoint errors similar to those achieved with training from scratch with around 30 image pairs, even without fine-tuning (solid line at 0 image pairs). Finally, the model trained with the complete training dataset and the on-the-fly-generated synthetic subjects (∞) achieves the best registration accuracy, making it the model of choice for further evaluation.

4.2.2 Robustness

The fine-tuned models achieve the most consistent registration across unseen lung CT and brain MRI datasets (Figure 4 and Table D.2, Appendix D.2).

The fine-tuned model outperformed from-scratch models in DIR-Lab-COPDgene, improving the median (IQR) landmark error from 3.5 mm (1.8-8.5) and 6.0 mm (2.5-

Table 1: The effect of the structures' shape and size in the synthetic dataset on brain and lung registration. The best median (IQR) values, excluding the from-scratch model, are highlighted in **bold**. *p $<$ 0.05

Lung CT (NLST)			
	Keypoint error (mm) ↓	SSIM ↑	SDLogJ ↓
Before reg.	6.9 (4.7–11.1)	0.31 (0.23-0.37)	-
From-scratch	2.2 (1.3–4.6)	0.55 (0.50-0.60)	0.33 (0.30-0.35)
Circles	**4.3 (2.5–7.9)***	0.57 (0.50-0.63)	0.56 (0.53-0.62)
Squares	**4.3 (2.5–7.9)***	0.56 (0.51-0.61)	0.49 (0.47-0.54)
Random	**3.5 (2.0–7.4)***	0.56 (0.53-0.61)	0.35 (0.32-0.36)
Random$_{\text{small}}$	**3.2 (1.4–9.9)***	**0.59 (0.56-0.67)***	**0.30 (0.28-0.33)***
Random$_{\text{large}}$	3.8 (1.9–9.1)	0.55 (0.51-0.62)	**0.30 (0.29-0.31)***

Brain MRI (OASIS)							
		Cerebral cortex	Brainstem	Lat. Ventricle	Hippocampus	SSIM ↑	SDLogJ ↓
Before reg.	HD$_{95}$ ↓	3.0 (3.0–3.3)	2.8 (2.2–3.2)	4.2 (3.3–6.2)	3.0 (2.7–3.7)	0.69 (0.68-0.70)	-
	ASSD ↓	1.1 (1.1–1.1)	1.2 (1.0–1.3)	1.4 (1.2–2.5)	1.2 (1.1–1.5)		
From-scratch	HD$_{95}$ ↓	1.4 (1.4–1.4)	1.0 (1.0–1.0)	1.0 (1.0–1.0)	1.2 (1.2–1.2)	0.83 (0.82-0.84)	0.47 (0.44-0.52)
	ASSD ↓	0.6 (0.6–0.7)	0.3 (0.3–0.4)	0.3 (0.3–0.3)	0.4 (0.4–0.4)		
Circles	HD$_{95}$ ↓	3.0 (2.6–3.2)	2.8 (2.2–3.0)	**2.4 (2.0–3.0)***	2.9 (2.6–3.3)	0.77 (0.75-0.77)	0.39 (0.39-0.43)
	ASSD ↓	1.0 (0.9–1.0)	1.0 (0.8–1.1)	**0.8 (0.7–0.9)***	1.0 (0.9–1.1)		
Squares	HD$_{95}$ ↓	2.6 (2.3–2.9)	2.2 (2.2–2.8)	**2.5 (1.9–4.0)***	2.5 (2.1–3.2)	**0.77 (0.75-0.78)***	0.40 (0.39-0.42)
	ASSD ↓	0.9 (0.9–1.0)	0.8 (0.8–0.9)	**0.8 (0.7–1.1)***	0.9 (0.8–1.1)		
Random	HD$_{95}$ ↓	**2.2 (2.2–2.3)***	**1.4 (1.4–1.7)***	**2.2 (1.8–3.9)***	**2.2 (1.9–2.6)***	**0.77 (0.75-0.78)***	0.23 (0.23-0.24)
	ASSD ↓	**0.9 (0.9–0.9)***	**0.6 (0.5–0.6)***	**0.8 (0.7–1.3)***	**0.8 (0.7–1.0)***		
Random$_{\text{small}}$	HD$_{95}$ ↓	2.4 (2.3–2.6)	**1.4 (1.4–1.7)***	2.3 (1.8–5.4)	**2.2 (1.9–2.7)***	0.76 (0.74-0.77)	0.26 (0.25-0.28)
	ASSD ↓	0.9 (0.9–1.0)	**0.6 (0.5–0.6)***	**0.8 (0.6–1.3)***	**0.8 (0.7–1.0)***		
Random$_{\text{large}}$	HD$_{95}$ ↓	2.2 (2.2–2.5)	**1.4 (1.4–2.0)***	2.8 (2.2–5.4)	**2.2 (2.0–2.7)***	0.76 (0.74-0.77)	**0.22 (0.22-0.23)***
	ASSD ↓	0.9 (0.9–1.0)	**0.6 (0.5–0.7)***	0.9 (0.7–1.6)	**0.8 (0.7–1.0)***		

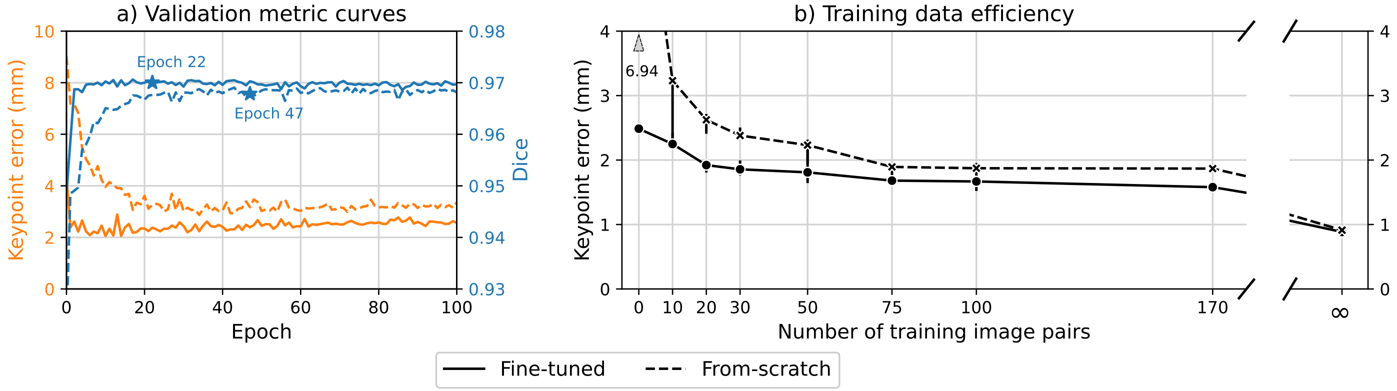

Figure 3: The training and data efficiency of transfer learning. **a)** The mean validation metrics (Dice and error) in the first 100 epochs of model training on lung CT data. The stars ($\star$) mark the epochs corresponding to convergence (when the change in the Dice running average was below 0.0001 for four consecutive epochs). **b)** The median validation error achieved with different subsets of the training data (10, 20, 30, 50, 75, 100, 170 (all $NLST_{train}$) image pairs). ∞ denotes the scenario including on-the-fly generated synthetic subjects. The markers and error bars indicate the median values and the minimum and maximum of three repeated experiments.

13.3) (with and without data augmentation) to 2.3 mm (1.3-4.7) (p<0.05) (Figure 4a and Table D.2a, Appendix D.2). These substantial improvements (62% and 34% error reductions) are likely related to the substantial train-test discrepancy in respiratory motion. In DIR-Lab-COPDgene, subjects were instructed to inhale with maximum effort (Table B.1, Appendix B), resulting in large pre-registration landmark errors of 22.4 mm (13.6-32.2) (Figure 5). In contrast, the training dataset (NLST) contains images from different time points, not respiratory phases, with

Table 2: Effect of different intensity sampling methods (static, dynamic, dynamic*) in the synthetic dataset on brain MR (T1-T1 and T1-T2) and lung CT registration. The best median (IQR) values, excluding the from-scratch model, are highlighted in **bold**. *p < 0.05

Lung CT (NLST)							
		Keypoint error (mm) ↓				SSIM	SDLogJ
Before reg.		6.9 (4.7-11.1)				0.31 (0.23-0.37)	-
From-scratch		2.2 (1.3-4.6)				0.55 (0.50-0.60)	0.33 (0.30-0.35)
Static		**2.9 (1.4-6.6)***				**0.67 (0.60-0.70)***	0.39 (0.35-0.42)
Dynamic		3.5 (2.0-7.5)				0.56 (0.53-0.61)	0.35 (0.32-0.36)
Dynamic*		4.8 (3.0-8.9)				0.49 (0.47-0.56)	**0.31 (0.30-0.32)***
Brain MRI$_{T1\text{-}T1}$ (OASIS)							
		Cerebral cortex	Brainstem	Lat. Ventricle	Hippocampus	SSIM	SDLogJ
Before reg.	HD$_{95}$ ↓	3.0 (3.0-3.3)	2.8 (2.2-3.2)	4.2 (3.3-6.2)	3.0 (2.7-3.7)	0.69 (0.68-0.70)	-
	ASSD ↓	1.1 (1.06-1.14)	1.15 (0.98-1.32)	1.4 (1.18-2.54)	1.22 (1.06-1.46)		
From-scratch	HD$_{95}$ ↓	1.4 (1.4-1.4)	1.0 (1.0-1.0)	1.0 (1.0-1.0)	1.2 (1.2-1.2)	0.83 (0.82-0.84)	0.47 (0.44-0.52)
	ASSD ↓	0.7 (0.6-0.7)	0.4 (0.3-0.4)	0.3 (0.3-0.3)	0.4 (0.4-0.4)		
Static	HD$_{95}$ ↓	**1.4 (1.4-1.6)***	1.4 (1.4-1.7)	**1.4 (1.2-2.2)***	**1.8 (1.8-2.1)***	**0.85 (0.83-0.86)***	0.33 (0.32-0.35)
	ASSD ↓	**0.7 (0.6-0.7)***	**0.5 (0.5-0.6)***	**0.5 (0.5-0.7)***	**0.6 (0.6-0.7)***		
Dynamic	HD$_{95}$ ↓	2.2 (2.2-2.3)	1.4 (1.4-1.7)	2.2 (1.8-3.9)	2.2 (1.9-2.6)	0.77 (0.75-0.78)	0.23 (0.23-0.24)
	ASSD ↓	0.9 (0.9-0.9)	0.6 (0.5-0.6)	0.8 (0.7-1.3)	0.8 (0.7-1.0)		
Dynamic*	HD$_{95}$ ↓	2.3 (2.2-2.6)	1.4 (1.4-1.7)	3.1 (2.6-4.8)	2.5 (2.1-3.0)	0.75 (0.74-0.76)	**0.18 (0.18-0.19)***
	ASSD ↓	0.9 (0.9-1.0)	0.6 (0.6-0.6)	1.0 (0.9-1.5)	1.0 (0.8-1.1)		
Brain MRI$_{T1\text{-}T2}$ (Paired3T)							
		Cerebral cortex	Brainstem	Lat. Ventricle	Hippocampus	SSIM	SDLogJ
Before reg.	HD$_{95}$ ↓	2.8 (2.6-3.1)	4.5 (4.0-5.9)	4.8 (4.1-5.9)	2.8 (2.7-3.7)	-	-
	ASSD ↓	1.0 (0.9-1.0)	1.7 (1.4-2.0)	1.5 (1.5-1.6)	1.0 (1.0-1.2)		
From-scratch	HD$_{95}$ ↓	2.1 (2.0-2.2)	4.1 (2.8-4.9)	2.1 (1.9-2.6)	1.8 (1.6-2.1)	-	0.32 (0.30-0.32)
	ASSD ↓	0.8 (0.8-0.8)	1.3 (1.0-1.5)	0.7 (0.6-0.7)	0.7 (0.7-0.8)		
Static	HD$_{95}$ ↓	4.8 (4.4-5.0)	11.7 (10.2-13.3)	14.7 (13.8-15.2)	7.1 (6.8-7.5)	-	0.48 (0.47-0.49)
	ASSD ↓	1.3 (1.3-1.3)	3.2 (2.7-3.7)	4.7 (4.3-4.8)	2.4 (2.2-2.5)		
Dynamic	HD$_{95}$ ↓	**2.3 (2.2-2.6)***	**3.6 (3.0-4.1)***	**3.6 (3.0-4.2)***	**2.1 (2.0-2.4)***	-	**0.20 (0.20-0.21)***
	ASSD ↓	**0.9 (0.8-0.9)***	**1.1 (0.9-1.2)***	**0.9 (0.8-1.0)***	**0.7 (0.7-0.8)***		
Dynamic*	HD$_{95}$ ↓	**2.7 (2.3-3.1)***	3.7 (3.2-5.1)	**4.1 (2.7-4.5)***	**2.1 (1.9-2.5)***	-	**0.17 (0.17-0.17)***
	ASSD ↓	**0.9 (0.8-0.9)***	1.3 (1.0-1.4)	1.1 (0.9-1.2)	**0.7 (0.7-0.9)***		

pre-registration landmark errors of 6.9 mm (4.7-11.1).

In the other out-of-distribution lung CT dataset, DIR-Lab, the fine-tuned model performed comparably to the from-scratch models with and without data augmentation, achieving landmark errors of 1.3 mm (0.9-1.9), 1.3 mm (0.9, 1.9), and 1.4 mm (0.9-2.0), respectively (p>0.1) (Figure 4a and Table D.2a, Appendix D.2). Interestingly, the fine-tuned model resulted in slightly lower keypoint errors in the (in-distribution) NLST dataset (p<0.05), with 92% of keypoints under 2 mm, compared to 85% and 88% for the from-scratch models with and without data augmentation. The pattern in the DIR-Lab landmarks' cumulative density before registration, visible at 2.5 mm intervals, is likely caused by the original slice thickness (Figure 4a).

The fine-tuned model showed small improvements compared to from-scratch models in brain MRI datasets. In IXI, it achieved marginally better HD and ASSD values than from-scratch training when averaged over the brain structures (p<0.05) (Table D.2b, Appendix D.2). However, the accuracy of most individual brain structures was comparable (Figure 4b). In the Paired3T dataset, the fine-tuned model had a slightly better accuracy than the from-scratch models for most brain structures, with HD$_{95}$ improvements between 6% and 14%. For example, it achieved median (IQR) HD$_{95}$ 1.8 mm (1.6-2.8) for the lateral ventricle, compared to 1.9 mm (1.7-3.1) and 2.0 mm (1.7-3.0) for the model trained from scratch with and without data augmentation (p<0.05).

A notable additional finding is that the pre-trained model produced higher SDLogJ values than the other models in lung registration (Table D.2, Appendix D.2), which is indicative of less smooth deformation fields. For example, the pre-trained model and the fine-tuned model had respective median (IQR) SDLogJ values of 0.37 (0.33-0.39) and

0.21 (0.21-0.24) ($p<0.05$) in the lung NLST dataset. In contrast, the pre-trained model produced lower SDLogJ values than the other models in the brain OASIS dataset (0.27 (0.27-0.28) vs. 0.57 (0.56-0.60) for fine-tuned, $p<0.05$), which were accompanied with lower SSIM values (0.78 (0.77-0.79) vs. 0.83 (0.83-0.84), $p<0.05$).

5. Discussion

This study investigates transfer learning for deep learning models in deformable image registration. We pre-trained a model on synthetic data and then fine-tuned it on target data, resulting in robust performance across different datasets. Fewer training epochs and less target domain data were required compared to training a model from scratch, as demonstrated in lung CT registration.

Synthetic images were generated using random geometric shapes and dynamic contrast, following Hoffmann et al. (2022), to promote generalization across anatomies and modalities. While random geometric shapes and dynamic contrast were important to achieving the most consistent registration across tasks, we found that different target tasks may benefit from different synthetic data. For example, static contrast achieved the best mono-modal brain registration. After pre-training on exclusively synthetic data, the model performed worse than iterative registration and from-scratch models and produced less smooth deformation fields in lung registration. These findings highlight the importance of model fine-tuning on target domain images, contrasting previous reports by Hoffmann et al. (2022) and He et al. (2022), who found competitive accuracies with synthetic training data alone. Two factors might explain the lower accuracy in our study: (1) Training was limited to 1000 epochs on synthetic data from 100 pre-set label maps, and (2) image corruptions were excluded to simplify data generation. However, additional analysis confirmed that these factors are not the leading causes of the reduced accuracy (Figures D.2 and D.3, Appendix D.1).

Fine-tuning the model significantly improved performance, producing domain-specific models that remained robust to discrepancies between training and test data. The largest robustness gains over from-scratch training were observed in the DIR-Lab-COPDgene dataset (34-62%, $\text{NLST}_{train} \rightarrow$DIR-Lab-COPDgene) and Paired3T dataset (6-14%, $\text{OASIS}_{train} \rightarrow$Paired3T). Other studies have improved cross-dataset robustness using test-time instance optimization, for example, improving the landmark error from 4.1 to 2.1 mm in DIR-Lab-COPDgene (50%) (Wang et al., 2022) and the Dice from 0.79 to 0.82 mm in brain MRI data (3%) (Zhu et al., 2021; Mok et al., 2024). Our approach achieves comparable improvements without requiring additional computational time, making it better suited for time-sensitive applications.

However, robustness did not improve across all datasets. For example, DIR-Lab and IXI showed minimal improvement. One possible explanation is that the from-scratch models were already robust to differences between these datasets and the training set (disease pathology, scan type, scanner manufacturer, and imaging protocol), leaving limited room for improvement. The negligible impact of data augmentation in these datasets supports this interpretation. To compare transfer learning and data augmentation further, we conducted additional experiments involving simulated domain shifts (Appendix D.3). Interestingly, both approaches led to comparable improvements, even when the augmentation strategies were specifically tailored to the simulated shifts.

The fine-tuned model achieved registration accuracies close to state-of-the-art methods. In DIR-Lab, it reached a mean (SD) landmark error of 1.6 (1.3) mm, comparable to the errors of leading methods, which ranged from 1.1 (0.8) mm to 1.6 (1.6) mm (Hering et al., 2021; Mok and Chung, 2020; Hansen and Heinrich, 2021; Eppenhof and Pluim, 2019; Fu et al., 2020) (Table D.3, Appendix D.5). The DIR-Lab-COPDgene dataset posed a greater challenge due to large initial deformations caused by breath-hold CT scans taken at maximum inhalation. Our approach achieved an average error of 4.3 (median 2.3) mm, compared to the errors between 1.3 and 2.3 mm of other methods (Hansen and Heinrich, 2021; Heinrich and Hansen, 2022; Wang et al., 2022; Tian et al., 2024) (Table D.3, Appendix D.5). These competing methods used breath-hold CT data from COPD patients for training, better matching the test set, and used techniques more suited for large deformations, such as discretized displacements (Hansen and Heinrich, 2021), keypoint correspondences (Heinrich and Hansen, 2022), and cascaded networks with four levels and three resolutions (Tian et al., 2024). Our use of a simpler two-level cascaded U-Net may have limited performance on large deformation registration.

An additional observation was that the pre-trained model behaved differently across target domains, producing less smooth deformations in the lungs than in the brain. These findings may stem from the synthetic training deformations, which were non-diffeomorphic and did not fully represent the target data. Lung datasets involved larger deformations (up to 16 voxels in DIR-Lab-COPDgene versus 11 voxels in the synthetic data), while brain datasets involved smaller displacements (up to four voxels).

This study used a simple approach to investigate the feasibility of transfer learning in deformable image registration. A limitation is that we did not focus on finding the best possible transfer learning strategy. Further improvements may be expected from more advanced techniques such as learning rate warm-up, progressive layer unfreezing, or domain adaptation. These strategies are promising for

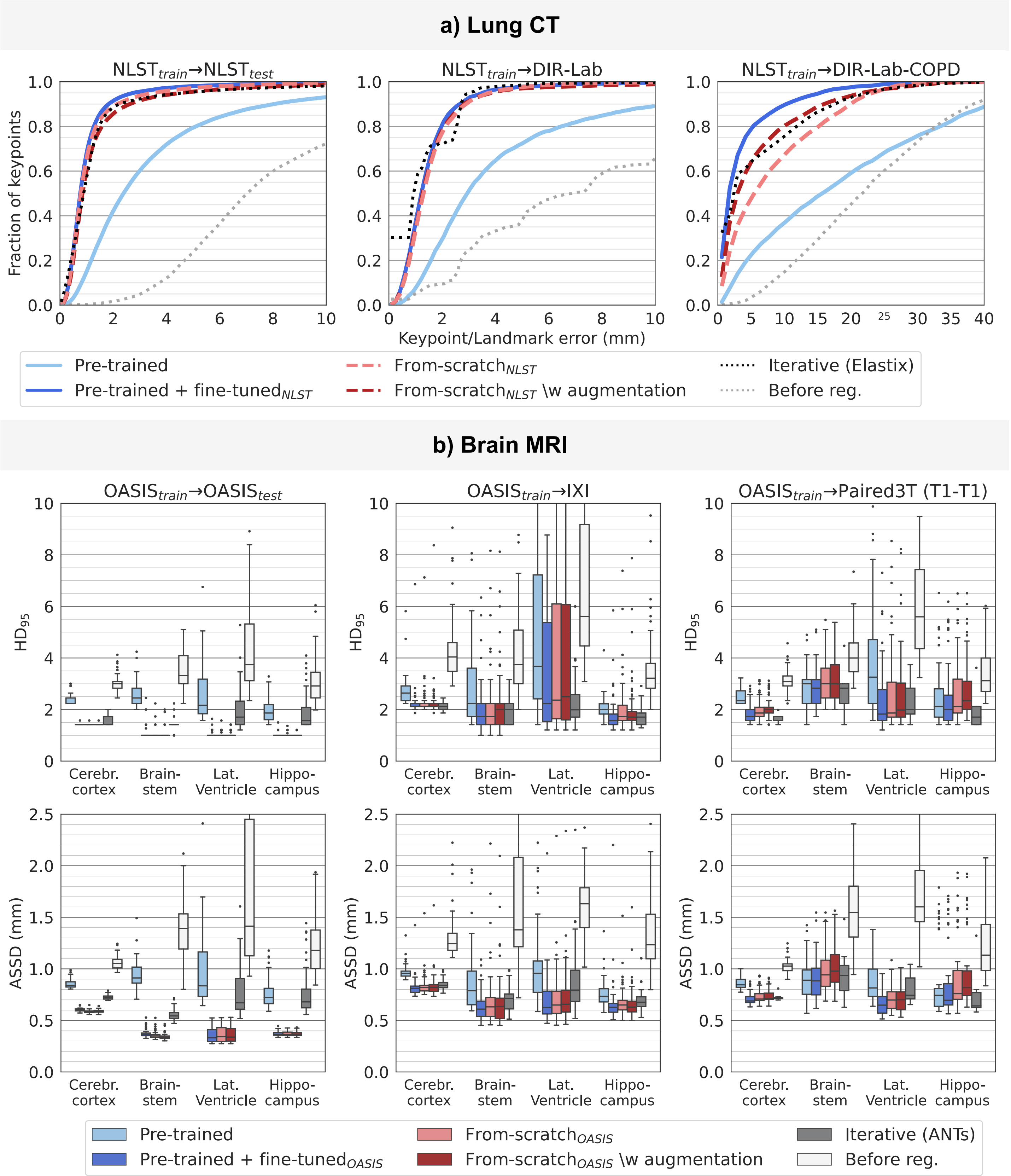

Figure 4: Model robustness across datasets. **a)** Registration accuracy across lung CT datasets (NLST, DIR-Lab, and DIR-Lab-COPDgene), shown as the cumulative density of points (keypoints or landmarks) vs. the registration error. **b)** Registration accuracy across brain MRI datasets (OASIS, IXI, and paired3T), shown as the HD_{95} and ASSD of brain structures.

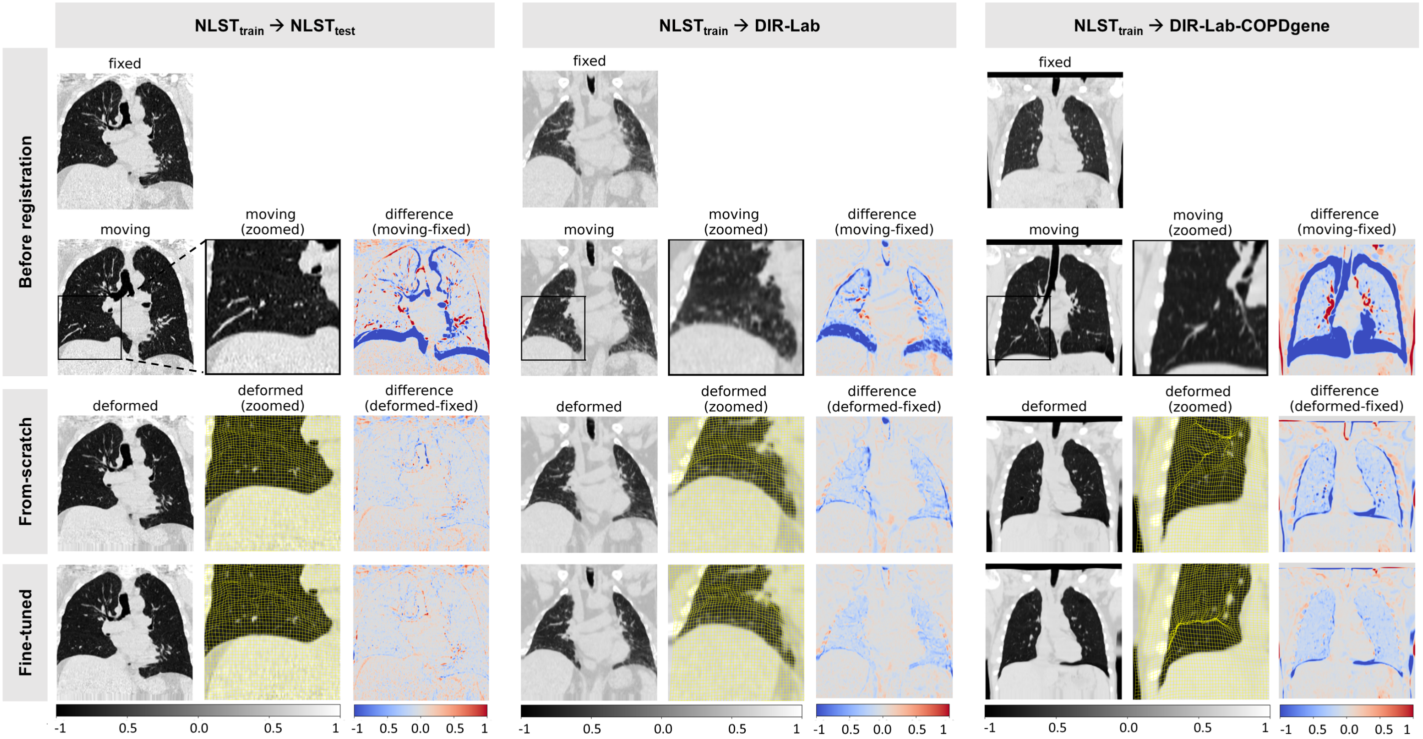

Figure 5: Registration of three example cases in the NLST, DIR-Lab, and DIR-Lab-COPDgene test sets. The first two rows display the fixed and moving images and an image showing their differences. After registration with the from-scratch and fine-tuned models, the third and fourth rows show the deformed moving images and their differences from the fixed image. The yellow grid indicates the coronal plane deformations.

addressing issues such as catastrophic forgetting. Furthermore, our method relied on weak supervision using Dice loss, which limits its applicability to settings with target domain labels. Additional experiments in Appendix D.4 illustrate that the underlying concept extends to unsupervised settings without labels. However, they only serve as proof of concept, as the model was not optimized for this setting.

An alternative to transfer learning based on synthetic data is pre-training on acquired images. Tian et al. (2024) demonstrated this with UniGradICON, trained on over 5,000 images spanning multiple modalities and anatomies. Fine-tuning the model with target datasets achieved state-of-the-art performance. For reference, we report UniGradICON's published accuracy on the DIR-Lab-COPDgene dataset, on which it outperformed our method (Table D.3, Appendix D.5). However, it is difficult to directly compare performances because different network architectures were used. Like UniGradICON, our work highlights the potential of transfer learning for deep learning-based image registration. Synthetic data offers a key advantage: it eliminates the need for large, labeled datasets of acquired images. Our approach required only a small fraction of target domain data for fine-tuning, making it valuable in clinical scenarios where data are scarce or unrepresentative, such as radiotherapy for rare tumors or cancer types.

6. Conclusion

This paper presents a transfer learning approach based on pre-training with synthetic data to improve the robustness of deep learning-based image registration. Our approach remained robust to a broader range of domain shifts than data augmentation and reduces the need for large and diverse training datasets from multiple centers. Transfer learning promises to improve robustness to discrepancies between training and real-world data in deep learning-based deformable image registration.

Acknowledgments

The Irene Curie Fellowship program and the Eindhoven Artificial Intelligence Systems Institute supported this work.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treating human subjects.

Conflicts of Interest

We declare we do not have conflicts of interest.

Data availability

This work used open-source datasets obtained from `https://learn2reg.grand-challenge.org/learn2reg-2023/` and `https://brain-development.org/ixi-dataset/`.

References

IXI Dataset, 2024. URL `https://brain-development.org/ixi-dataset/`. Accessed: 2024-09-20.

Denise R. Aberle, Amanda M. Adams, Christine D. Berg, William C. Black, Jonathan D. Clapp, and others (National Lung Screening Trial Research Team). Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011.

Vincent Arsigny, Olivier Commowick, Xavier Pennec, editor="Larsen Rasmus Ayache, Nicholas", Mads Nielsen, and Jon Sporring. A Log-Euclidean framework for statistics on diffeomorphisms. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 924–931. Springer Berlin Heidelberg, 2006.

Brian B. Avants, Charles L. Epstein, Murray Grossman, and James C. Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, 2008.

Brian B. Avants, Nick Tustison, and Gang Song. Advanced normalization tools (ants). *Insight j*, 2(365):1–35, 2009.

Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Voxelmorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019.

Edward Castillo, Richard Castillo, Josue Martinez, Maithili Shenoy, and Thomas Guerrero. Four-dimensional deformable image registration using trajectory modeling. *Physics in Medicine & Biology*, 55(1):305, 2009a.

Richard Castillo, Edward Castillo, Rudy Guerra, Valen E. Johnson, Travis McPhail, et al. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. *Physics in Medicine & Biology*, 54(7):1849, 2009b.

Richard Castillo, Edward Castillo, David Fuentes, Moiz Ahmad, Abbie M. Wood, et al. A reference dataset for deformable image registration spatial accuracy evaluation using the COPDgene study archive. *Physics in Medicine & Biology*, 58(9):2861, 2013.

Monzer Chehab, Brian E. Kouri, Michael J. Miller, and Aradhana M. Venkatesan. Image fusion technology in interventional radiology. *Techniques in Vascular and Interventional Radiology*, 26(3):100915, 2023.

Xiaoyang Chen, Liangqiong Qu, Yifang Xie, Sahar Ahmad, and Pew-Thian Yap. A paired dataset of T1-and T2-weighted MRI at 3 Tesla and 7 Tesla. *Scientific Data*, 10 (1):489, 2023.

Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, et al. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65 (5):545–563, 2021.

Jorge Corral Acero, Ernesto Zacur, Hao Xu, Rina Ariga, Alfonso Bueno-Orovio, et al. SMOD - data augmentation based on statistical models of deformation to enhance segmentation in 2D Cine Cardiac MRI. In Yves Coudière, Valéry Ozenne, Edward Vigmond, and Nejib Zemzemi, editors, *Functional Imaging and Modeling of the Heart*, volume 11504, pages 361–369, 2019.

Bob D. de Vos, Hessam Sokooti, Marius Staring, and Ivana Išgum. Machine learning in image registration. In Alejandro F. Frangi, Jerry L. Prince, and Milan Sonka, editors, *Medical Image Analysis*, The MICCAI Society book Series, pages 501–515. Academic Press, 2024.

Koen A.J. Eppenhof and Josien P.W. Pluim. Pulmonary CT registration through supervised learning with convolutional neural networks. *IEEE Transactions on Medical Imaging*, 38(5):1097–1105, 2019.

Enzo Ferrante, Ozan Oktay, Ben Glocker, and Diego H. Milone. On the adaptability of unsupervised CNN-based deformable image registration to unseen image domains. In Yinghuan Shi, Heung-Il Suk, and Mingxia Liu, editors, *International Workshop on Machine Learning in Medical Imaging (MLMI)*, pages 294–302. Springer, 2018.

Bruce Fischl. Freesurfer. *NeuroImage*, 62(2):774–781, 2012.

Yabo Fu, Yang Lei, Tonghe Wang, Kristin Higgins, Jeffrey D. Bradley, et al. LungRegNet: An unsupervised deformable image registration method for 4D-CT lung. *Medical Physics*, 47(4):1763–1774, 2020.

Shaoya Guan, Tianmiao Wang, Kai Sun, and Cai Meng. Transfer learning for nonrigid 2D/3D cardiovascular images registration. *IEEE Journal of Biomedical and Health Informatics*, 25(9):3300–3309, 2021.

Lasse Hansen and Mattias P. Heinrich. GraphRegNet: Deep graph regularisation networks on sparse keypoints for dense registration of 3D lung CTs. *IEEE Transactions on Medical Imaging*, 40(9):2246–2257, 2021.

Yuanbo He, Aoyu Wang, Shuai Li, Yikang Yang, and Aimin Hao. Nonfinite-modality data augmentation for brain image registration. *Computers in Biology and Medicine*, 147:105780, 2022.

Mattias P. Heinrich and Lasse Hansen. VoxelMorph++: Going beyond the cranial vault with keypoint supervision and multi-channel instance optimisation. In Alessa Hering, Julia Schnabel, Miaomiao Zhang, Enzo Ferrante, Mattias Heinrich, and Daniel Rueckert, editors, *Biomedical Image Registration (WBIR)*, page 85–95. Springer-Verlag, 2022.

Mattias P. Heinrich, Heinz Handels, and Ivor J. A. Simpson. Estimating large lung motion in COPD patients by symmetric regularised correspondence fields. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 338–345. Springer, 2015.

Alessa Hering, Stephanie Häger, Jan Moltz, Nikolas Lessmann, Stefan Heldmann, et al. CNN-based lung CT registration with multiple anatomical constraints. *Medical Image Analysis*, 72:102139, 2021.

Malte Hoffmann, Benjamin Billot, Douglas N. Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V. Dalca. SynthMorph: Learning contrast-invariant registration without acquired images. *IEEE Transactions on Medical Imaging*, 41(3):543–558, 2022.

Andrew Hoopes, Malte Hoffmann, Bruce Fischl, John Guttag, and Adrian V. Dalca. Hypermorph: Amortized hyperparameter learning for image registration. In *Information Processing in Medical Imaging (IPMI): 27th International Conference*, pages 3–17. Springer, 2021.

Fabian Isensee, Paul F. Jaeger, Simon A.A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

Zhuoran Jiang, Fang-Fang Yin, Yun Ge, and Lei Ren. A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration. *Physics in Medicine & Biology*, 65 (1):015011, 2020.

Jaeyong Kang and Jeonghwan Gwak. Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. *IEEE Access*, 7:26440–26447, 2019.

Michael D Ketcha, Tharindu De Silva, Runze Han, Ali Uneri, Sebastian Vogt, et al. Learning-based deformable image registration: effect of statistical mismatch between train and test images. *Journal of Medical Imaging*, 6(4): 044008–044008, 2019.

Stefan Klein, Marius Staring, Keelin Murphy, Max A. Viergever, and Josien P.W. Pluim. Elastix: A toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging*, 29(1):196–205, 2009.

Iris D. Kolenbrander, Matteo Maspero, Allard A. Hendriksen, Ryan Pollitt, Jochem R. N. van der Voort van Zyp, et al. Deep-learning-based joint rigid and deformable contour propagation for magnetic resonance imaging-guided prostate radiotherapy. *Medical Physics*, 51(4):2367–2377, 2024.

Learn2Reg. Task1: CT Lung Registration (NLST) [Data set]. `https://learn2reg.grand-challenge.org/learn2reg-2023/`, 2023. Accessed: 2023-06-16.

Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro R.A.S. Bassi, Yijia Shi, et al. AbdomenAtlas: a large-scale, detailed-annotated, multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, 97:103285, 2024.

Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, et al. Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9): 1498–1507, 2007.

Tony C.W. Mok and Albert Chung. Large deformation diffeomorphic image registration with Laplacian pyramid networks. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 211–221. Springer, 2020.

Tony C.W. Mok, Zi Li, Yingda Xia, Jiawen Yao, Ling Zhang, et al. Deformable medical image registration under distribution shifts with neural instance optimization. In Xiaohuan Cao, Xuanang Xu, Islem Rekik, Zhiming Cui, and Xi Ouyang, editors, *International Workshop on Machine Learning in Medical Imaging (MLMI)*, pages 126–136. Springer, 2024.

Hugo Oliveira and Jefersson dos Santos. Deep transfer learning for segmentation of anatomical structures in chest radiographs. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 204–211, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M.

Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

Marius Staring, Stefan Klein, Johan H.C. Reiber, Wiro J. Niessen, and Berend C. Stoel. Pulmonary image registration with Elastix using a standard intensity-based algorithm. *Medical Image Analysis for the Clinic: A Grand Challenge*, pages 73–79, 2010.

Muhammed Talo, Ulas Baran Baloglu, Özal Yıldırım, and U Rajendra Acharya. Application of deep transfer learning for automated brain abnormality classification using MR images. *Cognitive Systems Research*, 54:176–188, 2019.

National Lung Screening Trial Research Team. Data from the National Lung Screening Trial (NLST) [Data set]. The Cancer Imaging Archive, `https://doi.org/10.7937/TCIA.HMQ8-J677`, 2013.

Maarten L. Terpstra, Matteo Maspero, Tom Bruijnen, Joost J.C. Verhoeff, Jan J.W. Lagendijk, and Cornelis A.T. van den Berg. Real-time 3D motion estimation from undersampled MRI using multi-resolution neural networks. *Medical Physics*, 48(11):6597–6613, 2021.

Lin Tian, Hastings Greer, Roland Kwitt, François-Xavier Vialard, Raúl San José Estépar, et al. UniGradICON: a foundation model for medical image registration. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, et al., editors, *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 749–760. Springer Nature Switzerland, 2024.

Di Wang, Yue Pan, Oguz C. Durumeric, Joseph M. Reinhardt, Eric A. Hoffman, et al. PLOSL: Population learning followed by one shot learning pulmonary image registration using tissue volume preserving and vesselness constraints. *Medical Image Analysis*, 79:102434, 2022.

Yan Wang, Zixuan Feng, Liping Song, Xiangbin Liu, and Shuai Liu. Multiclassification of endoscopic colonoscopy images based on deep transfer learning. *Computational and Mathematical Methods in Medicine*, 2021 (1):2485934, 2021.

Wentao Zhu, Yufang Huang, Daguang Xu, Zhen Qian, Wei Fan, et al. Test-time training for deformable multi-scale image registration. In *IEEE International Conference on Robotics and Automation*, pages 13618–13625, 2021.

Appendix A. Label map generation

3D label maps were generated as part of the synthetic pre-training data pipeline using the following procedure:

- Random geometry: The label maps are generated from a Gaussian image. To obtain the Gaussian image, we first form a set of 26 (J) 3D grids of size $s_L \times s_L \times s_L$, which contain values sampled from a Gaussian distribution $\mathcal{N}(0, 1)$. The grids are then upsampled to full-resolution images by linear interpolation, deformed using a DVF denoted as ϕ_2, and concatenated to obtain the Gaussian image (size $192 \times 192 \times 192 \times J$). The label map is obtained from this image by assigning the channel j with the highest value for each pixel. The small, medium and large geometric shapes were created by setting s_L to 12, 6, and 3 and b_{ϕ_1} to 10, 50, and 50, respectively.

  The DVF denoted as ϕ_2 is obtained similarly to ϕ_F and ϕ_M (in the main manuscript) but from a different starting grid. It starts as a $6 \times 6 \times 6 \times 3$ grid with values sampled from a Gaussian distribution $\mathcal{N}(0, \sigma^2_{\phi_2})$ with σ_{ϕ_2} sampled uniformly between 0 and 1. The grid is upsampled to a full-resolution image through linear interpolation, resulting in a stationary velocity vector field, which is integrated through scaling and squaring (step=5) to produce the DVF.

- Rectangles: The label maps are generated from 26 (J) 3D grids of size $8 \times 8 \times 8$ with gray values sampled between 0 and 1 at 1000 random pixel locations. The grids are upsampled to full-size (192^3) images and concatenated into a 26-channel image, from which a label map is obtained by assigning the channel with the highest value.

- Spheres: The label maps are generated by drawing 2000 j-valued spheres per label $j \in [1, 26]$ in a full-size (192^3) image. Each sphere's location is randomly assigned, and its radius is randomly sampled between 15 and 25 pixels.

Appendix B. Imaging parameters

This study evaluates the registration accuracy across datasets of the same modality and anatomical region but acquired as part of different studies. The imaging parameters of each dataset are detailed in Table B.1.

Appendix C. Statistical deformation model

We generated synthetic training subjects on the fly (Figure C.1), applying synthetic deformations to the original images. The synthetic deformations are created using a statistical model of deformations Corral Acero et al. (2019), which captures the variations in the deformations between images, representing breathing motion in lung CT and inter-patient deformations in brain MRI. We follow three steps to obtain the model and construct synthetic deformations:

Step 1. Iterative registration: We iteratively registered the fixed and moving images of a subset of 50 training subjects (from NLST$_{train}$ or from OASIS$_{train}$) to obtain a set of representative deformations. The registration was performed in SimpleElastix (SimpleITK Python extension) with B-spline registration, optimizing the NCC similarity metric and a bending energy term (weight factor=1.0) via stochastic gradient descent in a multi-resolution scheme of four resolutions. The optimization ran 256 iterations in each resolution, and the final B-spline parameter grid is $23 \times 23 \times 23 \times 3$.

Step 2. Dimensionality reduction: The resulting B-spline grids were used to build a model capturing 80% of the deformations' variations. The 50 B-spline parameter grids were flattened into 50 arrays and organized as columns in a matrix, obtaining three matrices M_i (size $23^3 \times 50$) for three directions i ($i \in x, y, z$). We calculated the average B-spline parameters ($\overline{M_i}$) for each matrix along the second dimension. We also reduced each matrix's second dimension, through principal component analysis (PCA), to a set of Eigenvectors U_i (setting the number of principal components to capture 80% of the variations). The Eigenvectors were scaled with the Eigenvalues.

Step 3. Deformation construction: For each direction i ($i \in x, y, z$), the scaled Eigenvectors were combined with the average B-spline parameters and a random component x to construct the synthetic deformation parametrized by B-splines, the $\phi_{bspline,i}$:

$$\phi_{bspline,i} = \overline{M_i} + U_i \cdot x \tag{4}$$

Here, x is a matrix with values sampled from the uniform distribution $\mathcal{U}(-c, +c)$ with c sampled between 4000 and 6000 for lung deformations and 2000 and 4000 for brain deformations. Finally, the full-resolution DVF, ϕ_{M2F}, was obtained by upsampling $\phi_{bspline,i}$ through third-order B-Spline interpolation and concatenating the directions.

Diffeomorphic properties: While B-spline parameterization promotes smooth DVFs, it does not guarantee invertibility (i.e., diffeomorphism). For lung registration, the synthetic DVFs showed foldings between 0% and 0.3%, and SDLogJ values ranging between 0.14 and 1.1 (with a 95th percentile of 0.38). For brain registration, the folding ranged between 0.0% and 0.025%, while SDLogJ values ranged between 0.13 and 0.40. These indicate that the DVFs were predominantly smooth.

Appendix D. Additional findings

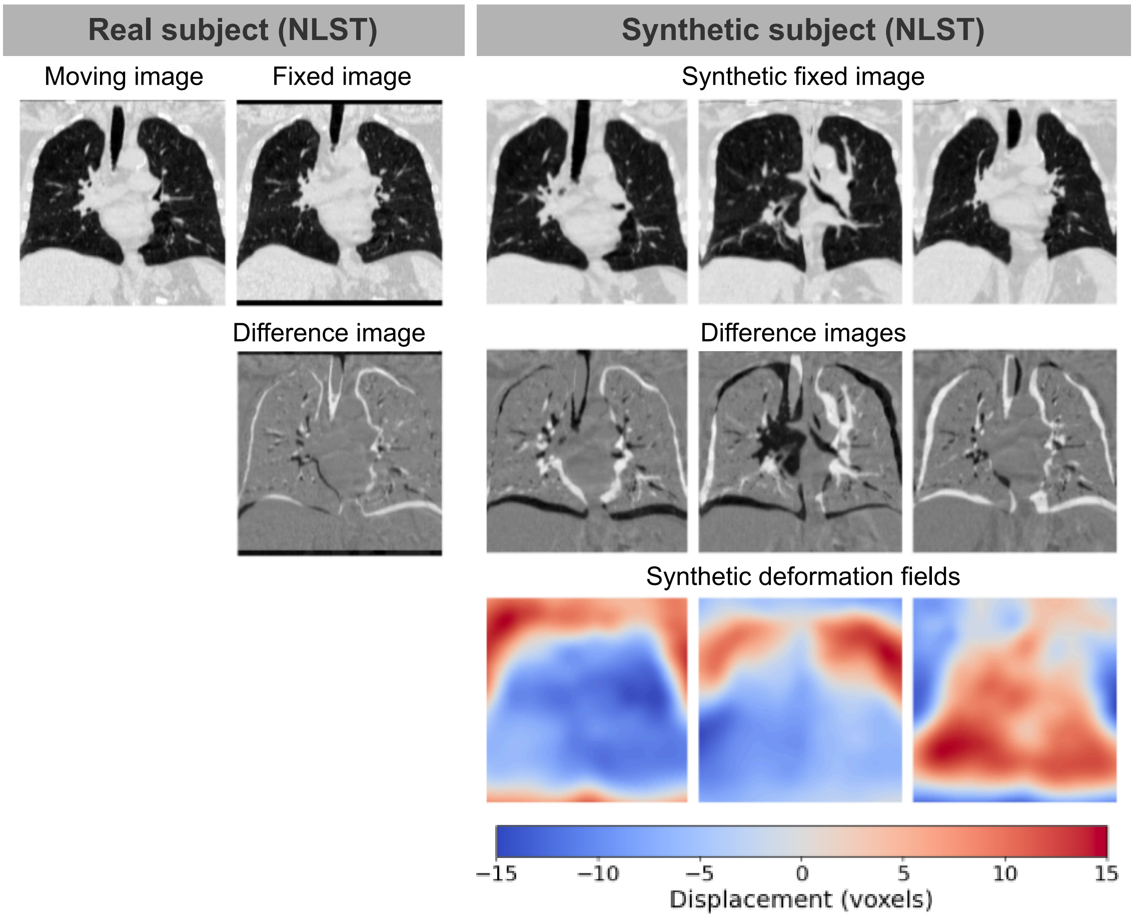

Figure C.1: Example synthetic lung CTs generated with the statical deformation model

D.1 Synthetic data: the impact of image corruption removal and the number of label maps

We studied the impact of image corruptions (synthetic bias fields, contrast augmentation, and simulated partial volume effects) on registration accuracy. A U-Net was trained in 200 epochs on synthetic images with and without these corruptions and evaluated on brain MRI and lung CT validation sets. None of the corruptions improved the registration consistently (Figure D.2).

We also assessed the impact of the number of label maps in synthetic data by training a cascaded U-Net on 100 label maps (1000 epochs) and 1000 label maps (100 and 200 epochs). These setups achieved comparable registration accuracies on brain MRI and lung CT (Figure D.3). A slight improvement in lung CT registration was observed with 1000 label maps and extended training. The comparable accuracy for 100 and 1000 label maps may be explained by the on-the-fly deformation during training, resulting in unique fixed and moving label maps, and the on-the-fly generation of gray-value images. This approach ensures that no training image pair is identical even with only 100 synthetic label maps.

D.2 Robustness: All evaluation metrics

Table D.2 presents all evaluation metrics, including the Structural Similarity Index (SSIM), folding, and the standard deviation of the logarithm of the Jacobian determinant (SDLogJ).

D.3 Robustness against introduced variations in CT

This section assesses the robustness of the fine-tuned lung CT registration model under controlled intensity and dose modifications in the NLST test set images.

Experimental set-up: The images in the test set are modified using intensity and dose perturbations. The fine-tuned lung CT registration model is evaluated on these modified test sets while comparing it with training from scratch and data augmentation, i.e., contrast and noise augmentations that mimic intensity and dose variations (section 2.4.1 in the main manuscript).

Intensity variations involve shifting the HU window by clipping the image intensities between -1100 and 100 HU. Dose variations involve simulating ultralow-dose CTs from the original low-dose CTs by adding noise patch-wise to the original CTs. To do this, we first extract the noise power spectrum (NPS) from a homogeneous patch (size

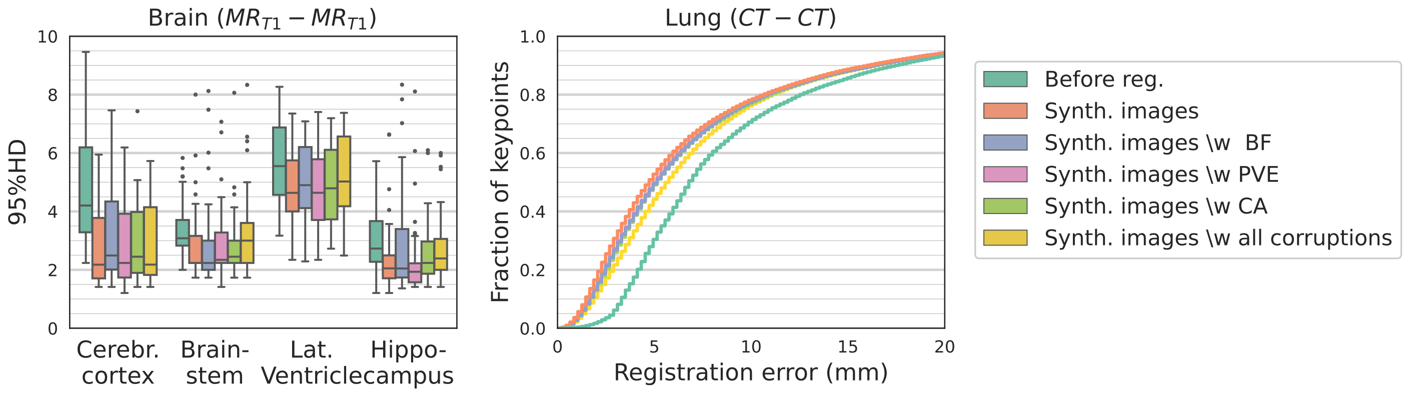

Figure D.2: Effect of image corruptions in the synthetic dataset: bias field (BF) artifacts, contrast augmentation (CA), and partial volume effects (PVE). The lung CT registration accuracy is shown as the cumulative density of keypoints vs. the registration error

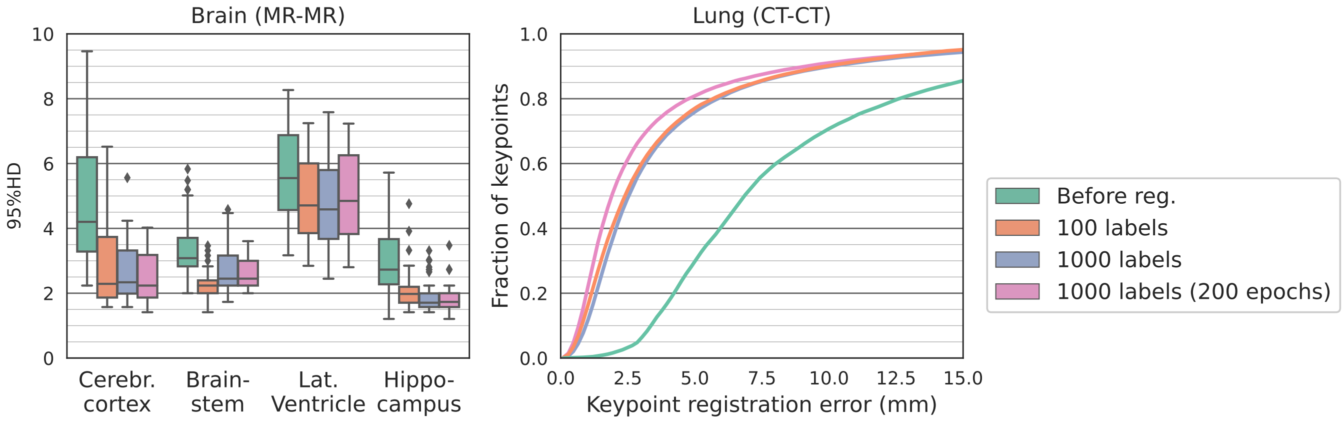

Figure D.3: Effect of 100 or 1000 label maps in the synthetic data on brain MR and lung CT registration.

$32 \times 32 \times 32$) in an original CT. We then use this NPS to filter noise sampled from a Gaussian distribution ($\mathcal{N}(0,1)$) through multiplication in the frequency domain. For each patch (size $32 \times 32 \times 32$) in the original CT (patch$_{\text{orig}}$), the new ultra-low dose CT patch (patch$_{\text{new}}$) is obtained as follows:

$$\text{patch}_{\text{new}} = \text{patch}_{\text{orig}} + \mathcal{F}^{-1}\left\{\sqrt{\text{NPS}} \cdot \mathcal{F}\{\text{patch}_{\text{noise}}\}\right\} * \sigma * 5.0 \tag{5}$$

Here, $\mathcal{F}$ indicates the fast Fourier Transform, patch$_{\text{noise}}$ is a patch with sampled noise, and σ is the standard deviation of the image intensities in the original CT.

Results: The fine-tuned lung CT registration model registers the shifted HU-window and ultra-low-dose datasets slightly better than the model trained from scratch and achieves similar registrations as data augmentation (Figure D.4). For example, it resulted in 87.5% of keypoints with errors below 2 mm on the ultra-low-dose dataset, compared to the 77.3% of the from-scratch model and 87.9% of the model trained with data augmentation.

D.4 Unsupervised fine-tuning

This section evaluates the feasibility of transfer learning with unsupervised fine-tuning in settings without target domain labels, focusing on lung CT registration.

Experimental set-up: We start with the pre-trained cascaded U-Net, trained using weakly supervised learning on synthetic data, and then fine-tune it using unsupervised learning on the NLST$_{\text{train}}$ dataset. The fine-tuning procedure follows the same setup as the weakly supervised fine-tuning described in Section 2.2.2 of the main manuscript, with the only change being the Dice term in the objective function, which is replaced with an intensity-based image similarity term. The similarity term is the normalized mutual information (NMI) computed at three resolution levels to capture both local and global image similarity:

$$\mathcal{L}_{\text{intensity-based}}(F, M) = \sum_{s=0}^{s=2} \frac{1}{2^s} \cdot \mathcal{L}_{\text{NMI}}\left(F^{(s)}, M^{(s)}\right) \tag{6}$$

Here, $F^{(s)}$ and $M^{(s)}$ are the fixed and moving images, respectively, downsampled by a factor 2^s using average pooling. Although the normalized cross-correlation would be a logical choice for the intensity-based loss term, our preliminary experiments resulted in unstable training.

We compare the fine-tuned model to unsupervised models trained from scratch, with and without data augmentation (Section 2.4.1), as well as to the pre-trained model without fine-tuning and the model fine-tuned with weak supervision.

Results: The fine-tuned model achieved the most consistent registration performance across lung CT datasets, showing improvements in landmark errors over models trained from scratch (Figure D.5). In the NLST dataset, it achieved a median (interquartile range, IQR) landmark error of 1.9 mm (IQR: 1.1–3.7), compared to 2.7 mm (1.4–5.4) and 2.2 mm (1.1–4.4) ($p<0.05$) for from-scratch models with and without data augmentation, respectively. Similarly, in the DIR-Lab dataset, the fine-tuned model achieved an error of 2.1 mm (1.3–4.0), while the from-scratch models reported 3.5 mm (1.8–6.7) with augmentation and 3.1 mm (1.7–6.3) without ($p<0.05$).

These findings illustrate that transfer learning extends to unsupervised settings, suggesting broader potential in settings without target domain labels. In addition, they suggest that data augmentation did not improve registration accuracy and even slightly degraded it. Finally, unsupervised fine-tuning underperformed compared to weakly supervised fine-tuning, potentially due to the method not being optimized for this setting (e.g., learning rate and regularization weight).

D.5 Comparison of existing works on DIR-Lab

The DIR-Lab datasets (DIR-Lab-4DCT and DIR-Lab-COPDgene) are widely used as benchmarks for evaluating the registration accuracy. Numerous studies have documented the registration errors for each subject. Table D.3 showcases these reported results from various deep learning-based registration methods alongside our findings.

Table B.1: The datasets' imaging parameters. NR: Not reported

Parameter	**Dataset**		
	NLST	**DIR-Lab**	**DIR-Lab-COPDgene**
Subjects (N)	210	10	10
Train/val/test (N image pairs)	170/20/20	0/0/10	0/0/10
Pathology	Subjects at high risk for lung cancer	Esophageal or lung cancer	Chronic obstructive pulmonary disease (COPD)
Imaging site	33 different U.S. medical centers	Houston (USA)	NR
Scanner type	Various	General Electric (GE) Discovery ST PET/CT scanner	GE VCT 64-slice scanner
Manufacturer	Various	GE HealthCare Technologies, Waukesha, WI	GE HealthCare Technologies, Waukesha, WI
Scan period	August 2002-April 2004	NR	NR
Subject positioning	NR	Supine	Supine
Breathing instructions	NR	Resting tidal breathing	Normal expiration and maximum effort full inspiration
4D binning	NA; longitudinal low-dose CT scans	Phase binning	NA; Breath-hold CT
Table pitch (mm)	Various		1,375
Tube voltage (kV)	Various	NR	120
Tube current (mA)	Various	NR	400 (inhale); 100 (exhale)
Effective exposure (mAs/pitch)	Various		200 (inhale); 50 (exhale)
Axial plane voxel spacing (mm)	Various	0.97-1.16	0.59-0.74
Slice thickness (mm)	Various	2.5	2.5
Image dimensions (voxels)	Various	256-512 x 256-512 x 94-136	512 x 512 x 102-135

Parameter	**Dataset**		
	OASIS	**IXI**	**Paired3T**
Subjects (N)	414	78	10
Train/val/test (N image pairs)	312/50/50	0/0/75	0/0/45
Pathology	Healthy adults and adults diagnosed with mild to moderate Alzheimer's disease	Healthy adults	Healthy adults
Imaging site	Washington University	Various hospitals in Londen: Guy's Hospital, Hammersmith Hospital,and Institue of Psychiatry	School of Medicine of the University of North Carolina at Chapel Hill
Scanner type	1.5 Tesla	1.5 and 3 Tesla	3 Tesla Magnetom Prisma
Manufacturer	Siemens	Philips and GE	Siemens
Axial plane voxel spacing (mm)	1	0.94	0.65
Slice thickness (mm)	1.25	1.20	0.65
Image dimensions (voxels)	256x256x128	130 150×256×256	256x304x308

Table D.2: All evaluation metrics for lung CT and brain MRI registration. The HD$_{95}$ and ASSD values are averaged over all brain structures (cerebral cortex, brainstem, lateral ventricle, hippocampus).

a) Lung CT

NLST$_{train}$ → NLST$_{test}$

	Landmark error (mm) ↓	**SSIM ↑**	**Folding (%) ↓**	**SdLogJ ↓**
Pre-trained	2.4 (1.4-4.4)	0.61 (0.54-0.64)	0.00 (0.00-0.00)	0.37 (0.33-0.39)
Pre-trained + fine-tuned$_{\text{NLST}}$	**0.8 (0.6-1.2)***	0.60 (0.56-0.69)	0.00 (0.00-0.00)	**0.21 (0.21-0.24)***
From-scratch$_{\text{NLST}}$	**0.8 (0.6-1.2)***	0.59 (0.53-0.68)	0.00 (0.00-0.00)	**0.21 (0.19-0.24)***
From-scratch$_{\text{NLST}}$ \w augmentation	0.9 (0.6-1.4)	0.59 (0.56-0.69)	0.00 (0.00-0.00)	**0.22 (0.21-0.25)***

NLST$_{train}$ → DIR-Lab

Pre-trained	3.0 (1.9-5.5)	0.85 (0.82-0.87)	0.00 (0.00-0.00)	0.29 (0.29-0.31)
Pre-trained + fine-tuned$_{\text{NLST}}$	**1.3 (0.9-1.9)***	0.83 (0.77-0.85)	0.00 (0.00-0.00)	0.16 (0.15-0.17)
From-scratch$_{\text{NLST}}$	**1.4 (0.9-2.0)***	0.81 (0.77-0.85)	0.00 (0.00-0.00)	**0.14 (0.13-0.15)***
From-scratch$_{\text{NLST}}$ \w augmentation	**1.3 (0.9-1.9)***	0.83 (0.78-0.86)	0.00 (0.00-0.00)	0.17 (0.16-0.18)

NLST$_{train}$ → DIR-Lab-COPDgene

Pre-trained	16.3 (6.3-30.1)	0.51 (0.48-0.54)	0.00 (0.00-0.00)	0.62 (0.45-0.68)
Pre-trained + fine-tuned$_{\text{NLST}}$	**2.3 (1.3-4.7)***	0.54 (0.50-0.62)	0.00 (0.00-0.00)	**0.29 (0.23-0.32)***
From-scratch$_{\text{NLST}}$	6.0 (2.5-13.3)	0.50 (0.46-0.55)	0.00 (0.00-0.00)	**0.26 (0.24-0.29)***
From-scratch$_{\text{NLST}}$ \w augmentation	3.5 (1.8-8.5)	0.55 (0.51-0.60)	0.00 (0.00-0.00)	**0.29 (0.23-0.32)***

b) Brain MRI

OASIS$_{train}$ → OASIS$_{test}$

	HD$_{95}$ ↓	**ASSD ↓**	**SSIM ↑**	**Folding (%) ↓**	**SdLogJ ↓**
Pre-trained	2.2 (2.1-2.7)	0.82 (0.78-0.96)	0.78 (0.77-0.79)	**0.00 (0.00-0.00)***	**0.27 (0.27-0.28)***
Pre-trained + fine-tuned$_{\text{OASIS}}$	**1.1 (1.1-1.2)***	0.42 (0.41-0.44)	**0.83 (0.83-0.84)***	0.04 (0.04-0.04)	0.57 (0.56-0.60)
From-scratch$_{\text{OASIS}}$	**1.1 (1.1-1.2)***	0.42 (0.40-0.44)	**0.84 (0.83-0.84)***	0.04 (0.03-0.04)	0.57 (0.55-0.61)
From-scratch$_{\text{OASIS}}$ \w augmentation	**1.1 (1.1-1.2)***	**0.41 (0.40-0.43)***	**0.84 (0.83-0.84)***	0.04 (0.03-0.04)	0.57 (0.55-0.59)

OASIS$_{train}$ → IXI

Pre-trained	3.0 (2.4-3.6)	0.85 (0.81-0.96)	0.86 (0.85-0.89)	0.00 (0.00-0.00)	0.26 (0.25-0.27)
Pre-trained + fine-tuned$_{\text{OASIS}}$	**2.0 (1.8-3.0)***	**0.67 (0.65-0.71)***	**0.89 (0.89-0.91)***	0.00 (0.00-0.00)	0.24 (0.21-0.30)
From-scratch$_{\text{OASIS}}$	2.2 (1.9-3.1)	0.69 (0.66-0.74)	**0.89 (0.88-0.91)***	0.00 (0.00-0.00)	**0.22 (0.20-0.23)***
From-scratch$_{\text{OASIS}}$ \w augmentation	2.2 (1.9-2.9)	0.69 (0.65-0.74)	**0.89 (0.88-0.91)***	0.00 (0.00-0.00)	**0.22 (0.19-0.26)***

Pre-trained	2.8 (2.3-3.5)	0.82 (0.77-0.96)	0.82 (0.81-0.83)	0.00 (0.00-0.00)	0.27 (0.27-0.28)
Pre-trained + fine-tuned$_{\text{OASIS}}$	**2.0 (1.9-2.8)***	**0.74 (0.69-0.86)***	**0.86 (0.85-0.87)***	0.00 (0.00-0.00)	0.28 (0.27-0.30)
From-scratch$_{\text{OASIS}}$	2.4 (2.1-3.3)	0.79 (0.74-0.89)	**0.85 (0.84-0.86)***	0.00 (0.00-0.00)	**0.25 (0.25-0.26)***
From-scratch$_{\text{OASIS}}$ \w augmentation	2.4 (2.1-3.3)	0.80 (0.76-0.91)	**0.86 (0.85-0.87)***	0.00 (0.00-0.00)	**0.25 (0.24-0.26)***

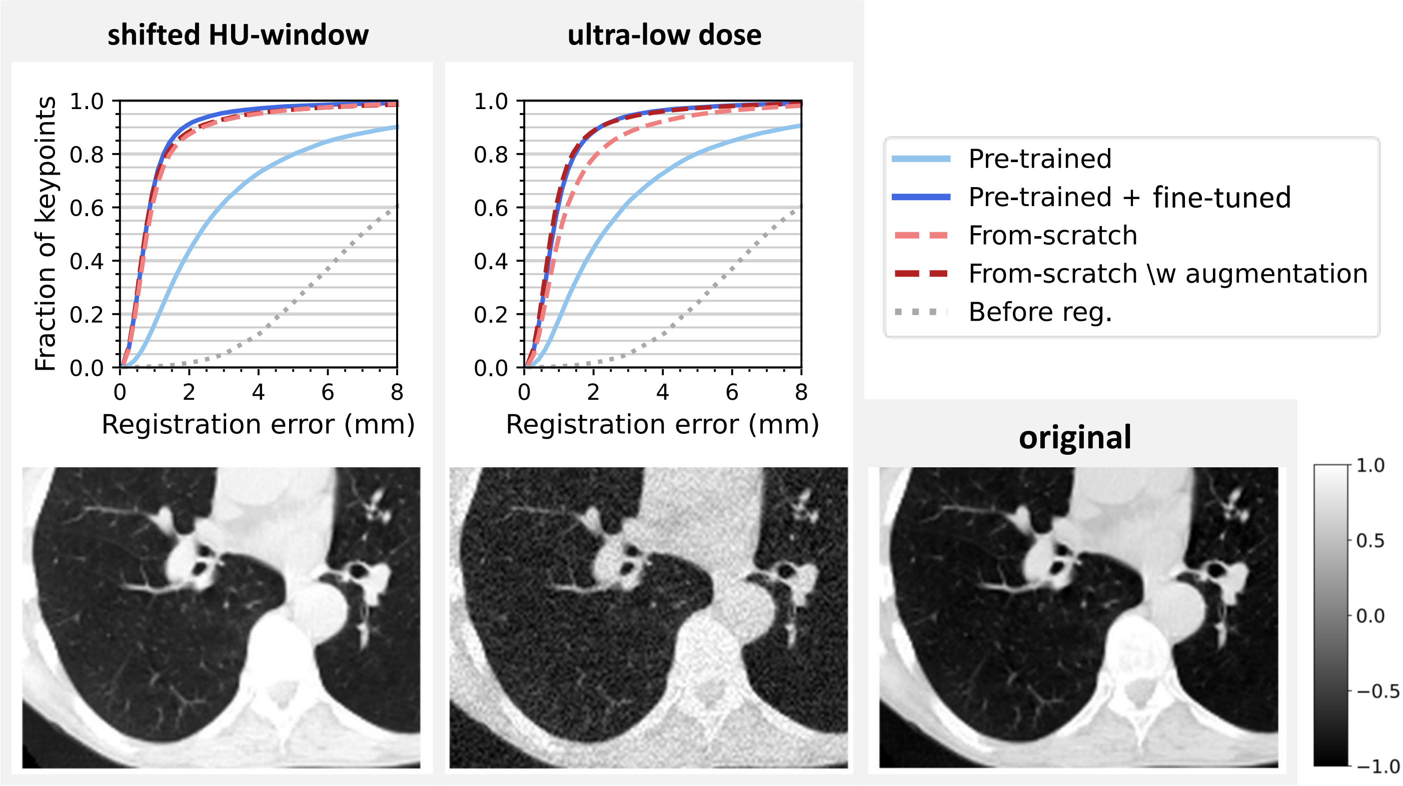

Figure D.4: Model robustness under intensity discrepancies: shifted HU window and ultra-low-dose. The first row shows the registration accuracy as the cumulative density of keypoints vs. the registration error, and the second row contains examples of the intensity variations applied.

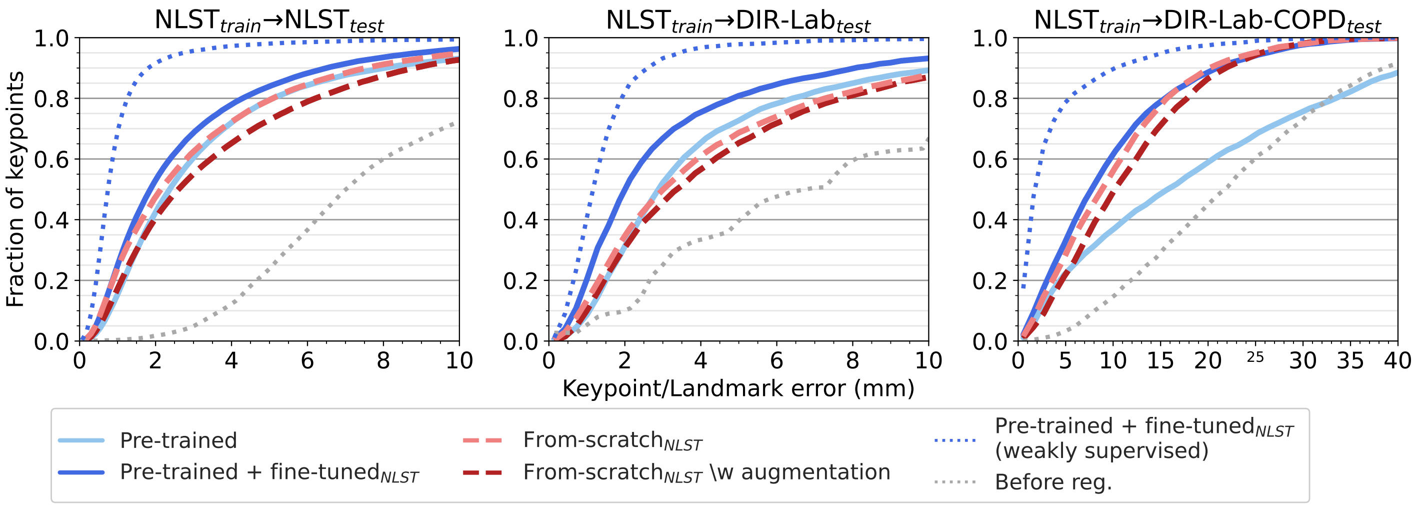

Figure D.5: The unsupervised model's robustness across lung CT datasets, shown as the cumulative density of points (keypoints or landmarks) vs. the registration error.

Table D.3: Mean ± standard deviation (SD) of the landmark registration error (in mm) of several deep-learning-based registration methods evaluated on the DIR-Lab datasets (4DCT and COPDgene). The SD is calculated across all landmarks from all subjects. Note that most methods, as opposed to ours, were trained on 4DCT or breath-hold CTs from COPD patients (e.g., from the COPDgene study archive data), representing an in-distribution performance evaluation. *Model is fine-tuned to a single image pair at test time (also called instance optimization or one-shot learning)

	Before reg.	**Eppenhof and Pluim 2019**	**LungRegNet 2020**	**Hering et al. 2021**	**LapIRN* 2024**	**VM++ * 2022**	**uniGradICON 2024**	**GraphregNet 2021**	**PLOSL* 2022**	**Adapted Model (ours)**
4DCT-01	3.89 ± 2.78	-	0.98 ± 0.54	0 .99 ± 0 .47	0.99 ± 0.46	-	-	0.86	1.12 ± 0.48	1.24 ± 0.60
4DCT-02	4.34 ± 3.90	1.24 ± 0.61	0.98 ± 0.52	0.98 ± 0.46	0.97 ± 0.47	-	-	0.90	1.06 ± 0.48	1.15 ± 0.58
4DCT-03	6.94 ± 4.05	-	1.14 ± 0.64	1.11 ± 0.61	1.10 ± 0.61	-	-	1.06	1.23 ± 0.64	1.26 ± 0.64
4DCT-04	9.83 ± 4.85	1.70 ± 1.00	1.39 ± 0.99	1.37 ± 1.03	1.33 ± 0.95	-	-	1.45	1.49 ± 0.96	1.72 ± 1.04
4DCT-05	7.48 ± 5.50	-	1.43 ± 1.31	1.32 ± 1.36	1.34 ± 1.21	-	-	1.60	1.61 ± 1.24	1.58 ± 1.34
4DCT-06	10.89 ± 6.96	-	2.26 ± 2.93	1.15 ± 1.12	1.16 ± 0.66	-	-	1.59	1.45 ± 0.79	1.79 ± 1.31
4DCT-07	11.03 ± 7.42	-	1.42 ± 1.16	1.05 ± 0.81	1.16 ± 0.62	-	-	1.74	1.40 ± 0.78	2.13 ± 2.16
4DCT-08	14.99 ± 9.00	-	3.13 ± 3.77	1.22 ± 1.44	1.19 ± 0.96	-	-	1.46	1.41 ± 1.08	1.96 ± 2.05
4DCT-09	7.92 ± 3.97	1.61 ± 0.82	1.27 ± 0.94	1.11 ± 0.66	1.16 ± 0.65	-	-	1.58	1.39 ± 0.72	1.66 ± 1.05
4DCT-10	7.30 ± 6.34	-	1.93 ± 3.06	1.05 ± 0.72	1.08 ± 0.58	-	-	1.71	1.33 ± 0.72	1.47 ± 1.02
Mean	8.46 ± 6.58	1.52 ± 0.85	1.59 ± 1.58	1.14 ± 0.76	1.15 ± 0.71	1.33	-	1.39 ± 1.29	1.35 ± 0.79	1.59 ± 1.33
COPD-01	26.33 ± 11.44	-	-	-	-	-	-	1.38	1.36 ± 0.78	5.00 ± 6.07
COPD-02	21.79 ± 6.47	-	-	-	-	-	-	2.09	2.06 ± 1.90	4.36 ± 4.17
COPD-03	12.64 ± 6.40	-	-	-	-	-	-	1.22	1.36 ± 0.73	1.55 ± 0.90
COPD-04	29.58 ± 12.95	-	-	-	-	-	-	1.58	1.75 ± 0.98	6.22 ± 6.31
COPD-05	30.08 ± 13.36	-	-	-	-	-	-	1.37	1.50 ± 0.81	7.78 ± 8.51
COPD-06	28.46 ± 9.17	-	-	-	-	-	-	1.10	1.45 ± 0.90	3.53 ± 3.83
COPD-07	21.60 ± 7.74	-	-	-	-	-	-	1.19	1.31 ± 0.76	2.59 ± 2.26
COPD-08	26.46 ± 13.24	-	-	-	-	-	-	1.19	1.54 ± 1.01	5.08 ± 5.80
COPD-09	14.86 ± 9.82	-	-	-	-	-	-	0.99	1.21 ± 0.77	2.95 ± 3.88
COPD-10	21.81 ± 10.51	-	-	-	-	-	-	1.38	1.79 ± 1.16	3.74 ± 2.81
Mean	23.36 ± 11.87	-	-	-	-	2.16	1.93	1.34 ± 1.44	1.53 ± 0.98	4.28 ± 5.33