Learning Deformable Registration of Medical Images with Anatomical Constraints

Lucas Mansilla, Diego H. Milone, Enzo Ferrante

Research institute for signals, systems and computational intelligence, sinc(i)
FICH-UNL, CONICET
Santa Fe, Argentina

Abstract

Deformable image registration is a fundamental problem in the field of medical image analysis. During the last years, we have witnessed the advent of deep learning-based image registration methods which achieve state-of-the-art performance, and drastically reduce the required computational time. However, little work has been done regarding how can we encourage our models to produce not only accurate, but also anatomically plausible results, which is still an open question in the field. In this work, we argue that incorporating anatomical priors in the form of global constraints into the learning process of these models, will further improve their performance and boost the realism of the warped images after registration. We learn global non-linear representations of image anatomy using segmentation masks, and employ them to constraint the registration process. The proposed AC-RegNet architecture is evaluated in the context of chest X-ray image registration using three different datasets, where the high anatomical variability makes the task extremely challenging. Our experiments show that the proposed anatomically constrained registration model produces more realistic and accurate results than state-of-the-art methods, demonstrating the potential of this approach.

Keywords: medical image registration, convolutional neural networks, x-ray image analysis

1. Introduction

Deformable image registration is one of the pillar problems in the field of medical image analysis. Disease diagnosis, therapy planning, surgical and radiotherapy procedures are a few examples where image registration plays a crucial role. In the medical context, the problem consists in aligning two or more images coming from different patients, modalities, moments or view points. Such alignment is achieved by means of a deformation field, that warps the so called *source* image, aligning it with the corresponding *target* image. Traditionally, image registration has been formulated as an optimization problem, where the objective function represents a similarity measure, which indicates how well the source image matches the target.

Recently, the latest advances in machine learning allow us to conceive image registration under an entirely different paradigm. In particular, deep convolutional neural networks (CNN) have proved to outperform all existent strategies in other fundamental tasks of computer vision, like image segmentation (Long et al., 2015) and classification (Krizhevsky et al., 2012). During the last years, we have witnessed the advent of deep learning-based image registration methods (Li and Fan, 2018; Yang et al., 2017; Rohé et al., 2017; Sokooti et al., 2017; de Vos et al., 2017, 2019; Balakrishnan et al., 2018; Dalca et al., 2018; Estienne et al., 2019), which achieve state-of-the-art performance, and drastically reduce the required computational time. These works have made a fundamental contribution by setting novel architectures for CNN-based deformable image registration (following supervised, unsupervised and semi-supervised training approaches). However, little work has been done regarding how can we encourage our models to produce not only accurate, but also anatomically plausible results, which is still an open question in the image registration community.

Image registration is employed in several medical tasks, e.g. when segmenting organs at risk in radiotherapy planning (Oh and Kim, 2017), co-registering MRI brain images prior to neuro-morphometry analysis (Gaser, 2016) or fusing

pre and post-operative images to perform minimally invasive surgeries (Ferrante and Paragios, 2017)). In these applications, producing anatomically plausible results is of paramount importance to guarantee the correctness of such medical and analytical procedures. Most of the existing registration methods impose smoothness constraints to the deformation fields or incorporate pixel-level losses into the objective functions to encourage anatomical plausibility. However, smooth deformation fields do not guarantee realistic results: in some cases, like moving organs for example, we require sharper deformations in the organ boundaries in order to preserve the anatomy. Regarding current pixel-based metrics (like Dice coefficient or Cross Entropy, for example), they do not consider the complete global context and therefore do not necessarily correlate with higher anatomical plausibility. The challenge we address is how CNN-based image registration models can produce accurate and anatomically plausible results after registration, that is, realistic results. Objectively quantifying the degree of realism of a medical image and incorporating it into the learning process of deep learning-based registration models is an extremely complex task. Here we propose to learn such measure from anatomical segmentation masks.

In this work, we argue that incorporating priors in the form of global anatomical constraints (Oktay et al., 2018) into the learning process of deep learning-based registration models, will further improve the accuracy of the results and boost the realism of the warped images after registration. We address the question about how we can incorporate anatomical priors into deep learning-based image registration methods in order to obtain more realistic results. In that sense, our contributions are four-fold: (i) we extend, for the first time, the concept of anatomically constrained neural networks (Oktay et al., 2018) to the image registration problem, (ii) we perform a deeper study of the complementarity between global and local loss functions defined over segmentation masks, (iii) we introduce the novel anatomically constrained registration network (AC-RegNet) architecture and validate it in the challenging task of X-ray chest image registration, comparing its performance with state-of-the-art existing methods and (iv) we showcase several application scenarios for AC-RegNet

in the context of X-ray chest image analysis including multi-atlas segmentation, automatic quality control and pathology classification, where anatomical plausibility is highly relevant, especially when performing pathology classification using anatomical segmentation masks. In this case, the masks are obtained through registration-based label propagation. Thus, the anatomical plausibility of the deformed segmentation masks is crucial for the classification task. We evaluate the proposed method in the context of X-ray chest imaging using three different datasets, including an interesting study about the behaviour of the global anatomical constraints when compared with a local metric. We show that the proposed method encourages the registration models to warp images in the space of anatomically plausible solutions while, at the same time, increasing the accuracy of the results.

2. Related works

Inspired by Horn and Schunck. (1980) and Lucas and Kanade (1981), the research communities of computer vision and medical imaging have made major efforts towards developing more accurate and efficient registration methods. Since then, deformable image registration has been modelled in multiple ways (see Sotiras et al. (2013) for a comprehensive description), most of them posing image registration as an optimization problem, which in its general form can be formulated as

$$\hat{\mathcal{T}} = \arg\min_{\mathcal{T}} \mathcal{M}(I \circ \mathcal{T}, J) + \mathcal{R}(\mathcal{T}), \tag{1}$$

where I is the source (moving) image, J is the target (fixed) image, $\mathcal{T}$ parameterizes a spatial transformation that maps each point of the image I to J, $\mathcal{M}$ corresponds to the criterion of (dis)similarity that quantifies the quality of the alignment between the warped source image $I \circ \mathcal{T}$ and the target image J, and $\mathcal{R}$ corresponds to the regularization term that imposes geometric constraints on the solution. In deformable image registration, the spatial transformation $\mathcal{T}$ is characterized by a deformation field, which represents the pixel displacements. The optimal transformation $\hat{\mathcal{T}}$ aligning I with J is computed by solving this

minimization problem. Classical non-learning based methods, like the one implemented in the popular toolbox SimpleElastix (Marstal et al., 2016), which will be used as baseline in this work, compute the optimal transformation by iteratively exploring the space of potential transformation parameters. In this case, a B-spline transformation is used to parameterize the deformation field, whose parameters are adjusted by minimizing the differences between the images. The main advantage of these methods is that they do not require a training process. Therefore, they can easily adapt to unseen images, being robust enough to be used in different image modalities and organs. However, when dealing with deformable registration, these algorithms are computationally expensive due to the high dimensionality of the parameter space, making the image registration process highly time consuming.

Existing CNN-based image registration methods are usually classified as supervised or unsupervised, depending on whether or not they use ground truth deformation fields to compute the loss function during training. Inspired by the original FlowNet for vector flow estimation (Dosovitskiy et al., 2015), supervised CNN-based image registration methods (Yang et al., 2017; Rohé et al., 2017; Sokooti et al., 2017) posed image registration as a regression problem. Given a pair of source and target images, they aim at regressing a deformation field that matches the ground-truth. One of the advantages of these methods is its independence with respect to image modalities: given a training dataset with pairs of images and their corresponding ground-truth deformations, it learns to map images to deformation fields without using any kind of similarity measure to compare them. However, getting such good datasets is a difficult task and makes these approaches impractical.

On the contrary, unsupervised CNN-based medical image registration (Li and Fan, 2018; de Vos et al., 2017, 2019; Balakrishnan et al., 2018; Dalca et al., 2018; Stergios et al., 2018; Balakrishnan et al., 2019) does not require ground-truth deformation fields. Instead, these methods (and the original CNN-based unsupervised optical flow estimation method (Ren et al., 2017)) solve the registration process by minimizing a loss function based on the (dis)similarity $\mathcal{M}$

between the deformed source image and the target. They use a differentiable warping module similar that used in spatial transformers (Jaderberg et al., 2015), to warp the source image during the forward-pass, and allow the gradients flow back during backpropagation. In such way, the model is trained to produce deformation fields that minimize the similarity-based loss function. At test time, a single forward pass will return the deformation field. In this work, we will follow this strategy to construct a baseline architecture (referred RegNet throughout this text) that will serve as baseline when evaluating the impact of the proposed anatomically constrained registration method.

Various approaches were envisioned in the literature to improve the accuracy and realism of the registration methods by incorporating prior information (about image modalities, anatomy and structure) into the registration process. Two of the most common strategies are knowledge-based transformations, where the information is encoded within the deformation model (Wouters et al., 2006; Glocker et al., 2009; Del Palomar et al., 2008) and segmentation-aware strategies, which directly incorporate segmentation priors to the registration process. The main disadvantage of knowleadge-based transformations is that they are highly domain specific, especially in methods like Del Palomar et al. (2008), which employ biomechanical models of specific organs to predict real deformations. In this work, we focus on the second alternative. Several non-deep learning based approaches (Shakeri et al., 2016; Ferrante et al., 2017, 2018a) were proposed to take advantage of such segmentations in the context of discrete graph-based image registration (Paragios et al., 2016). In Shakeri et al. (2016), probabilistic priors are incorporated to the registration process through a new term in the energy function of the proposed discrete formulation. In Ferrante et al. (2017, 2018a) segmentation masks are used to perform weak supervision when learning to aggregate standard similarity measures for image registration. However, these methods are orders of magnitude much slower than deep learning based image registration models. The first multi-modal CNN-based image registration method proposed in Hu et al. (2018), incorporates segmentation masks into the loss function of a weakly supervised approach to guide the learning

process. They use a pixel-level similarity measure defined on the segmentation masks, that makes it possible to register images independently of their modality. A similar pixel-level measure based on the Dice coefficient was incorporated in the VoxelMorph (Balakrishnan et al., 2019) and U-ResNet (Estienne et al., 2019) frameworks, and used in tandem with a standard intensity based loss. Still, local pixel-level loss functions do not consider the global context and might produce similar values for anatomically plausible and non-plausible segmentations. In this work, we build on top of these ideas by regularizing the learning process using a global and non-linear representation of the underlying anatomy. We show that this global term is complementary to existent pixel-level loss functions. Moreover, in the context of X-ray chest image registration, we improve the performance of existent registration methods by a significant margin while producing more realistic images after deformation.

3. Learning deformable image registration with anatomical constraints

In this section, we provide a brief description of the existing basic CNN architecture used to perform unsupervised image registration, which constitutes the building blocks for the proposed model. We then discuss how can we learn compact and non-linear representations of the image anatomy using denoising autoencoders (DAE) (Vincent et al., 2010), and how these representations can be introduced in the loss function to act as an anatomical regularizers, encouraging the learnt model to produce anatomically plausible images after deformation (see Figure 1 for an overview of the proposed novel architecture).

3.1. Basic architecture

The basic CNN architecture for image registration is composed of two main modules. The first one (referred as VectorCNN in Figure 1) follows an encoder-decoder structure similar to that of U-Net (Ronneberger et al., 2015). Given a pair of source image I and target image J as input, VectorCNN predicts a deformation field $\mathcal{T} = \text{VectorCNN}(I, J; \Theta)$ where $\mathcal{T} : \mathcal{R}^n \rightarrow \mathcal{R}^n$ (n beign

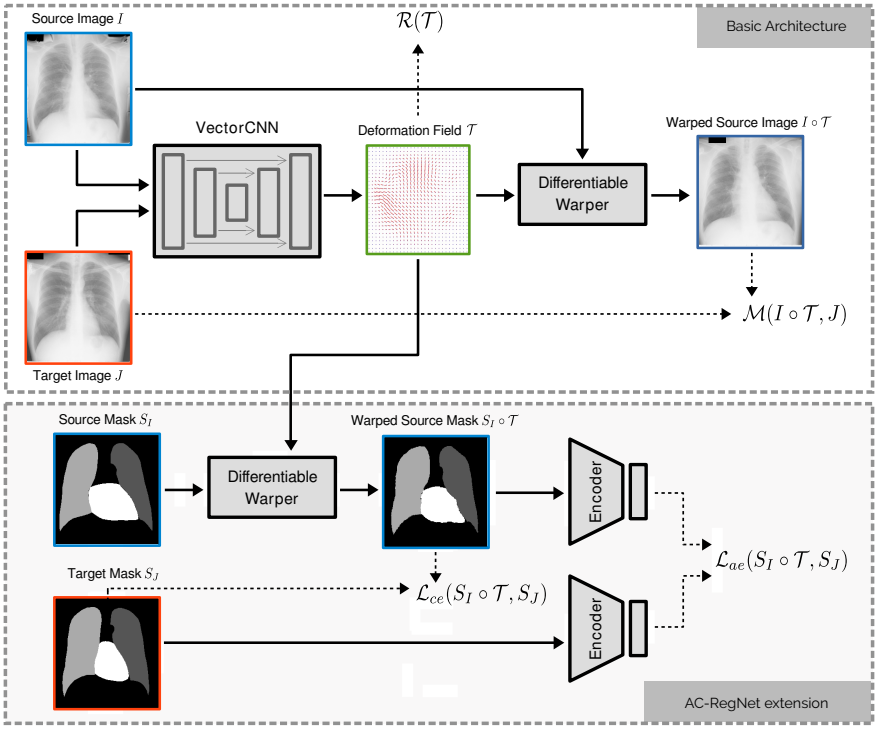

Figure 1: Architecture of the proposed AC-RegNet image registration model. We combine a basic CNN architecture for image registration (upper box) with local ($\mathcal{L}_{ce}$) and a global ($\mathcal{L}_{ae}$) loss functions based on anatomical segmentations S_I, S_J (lower box). We train the network VectorCNN to produce deformation fields $\mathcal{T}$ which minimize the differences between source (I) and target (J) images while, at the same time, ensuring anatomical correspondence after registration.

the image dimensionality) and Θ corresponds to the network parameters learnt during training. Images I and J are concatenated and fed to the network as a single multi-channel image. VectorCNN processes the two images through a series of convolutional and pooling layers, and outputs a 2-channel filter map representing the 2D deformation field (3-channels in case we are dealing with 3D images). The second component is a differentiable warping module similar to that used in spatial transformer networks (Jaderberg et al., 2015), that uses $\mathcal{T}$ to deform the source image I, producing a warped image $I \circ \mathcal{T}$.

At the beginning of the training process, VectorCNN will produce random deformation fields $\mathcal{T}$. During training, the parameters Θ are adjusted so that the warped source image $I \circ \mathcal{T}$ minimizes the (dis)similarity criterion $\mathcal{M}$ with the target image J, in the same spirit that classic registration methods. In this work, we use the negative normalized cross correlation (NCC) to quantify image alignment. NCC has been previously used in the context of CNN-based registration (Balakrishnan et al., 2018) and is a common choice when dealing with monomodal registration. We also consider a simple regularization term $\mathcal{R}(\mathcal{T})$ imposing smoothness to the deformation field by computing the total variation of the field (Li and Fan, 2018; Ferrante et al., 2018b). The basic loss function is therefore defined as

$$\mathcal{L}(I, J, \mathcal{T}) = \mathcal{M}(I \circ \mathcal{T}, J) + \lambda_r \mathcal{R}(\mathcal{T}), \tag{2}$$

where λ_r is a weighting factor for the total variation-based regularization term.

3.2. Segmentation-aware local loss functions

In order to augment the anatomical context provided to the network, we consider a simple initial strategy to include anatomical segmentations into the loss function by combining the aforementioned intensity-based loss $\mathcal{M}(I \circ \mathcal{T}, J)$, with a segmentation aware loss $\mathcal{L}_{ce}(S_I \circ \mathcal{T}, S_J)$. This new loss quantifies the alignment between a target anatomical segmentation mask S_J and a warped version of a source segmentation mask S_I. The size of each segmentation mask is the same as that of the corresponding image, and the mask is formed by the elements (or pixels) $s_k \in \mathcal{C}$, where $\mathcal{C}$ is the set of classes. $\mathcal{L}_{ce}(S_I \circ \mathcal{T}, S_J)$ is implemented as the classical categorical cross-entropy defined at the pixel level on the one-hot encoded versions of S_I and S_J. The segmentation-aware local loss function is thus defined as

$$\mathcal{L}(I, J, S_I, S_J, \mathcal{T}) = \mathcal{M}(I \circ \mathcal{T}, J) + \lambda_r \mathcal{R}(\mathcal{T}) + \lambda_{ce} \mathcal{L}_{ce}(S_I \circ \mathcal{T}, S_J), \tag{3}$$

where λ_{ce} is a weighting factor for the additional term $\mathcal{L}_{ce}$. Note that segmentation masks S_I, S_J are only required during training time to compute the loss

function. At test time, a single pair of images will be fed into the network to produce a deformation field and no segmentation masks are required.

3.3. Auto-encoding global anatomical priors

The local loss function $\mathcal{L}_{ce}(S_I \circ \mathcal{T}, S_J)$ defined in the previous section looks at pixel level predictions; therefore, it does not guarantee a good matching at the global scale between the deformed source and target anatomical masks. The segmentation masks used in this work represent anatomical structures like lungs and heart. This is different from, for example, segmentation masks corresponding to pathological structures or lesions (like brain tumors or skin lesions), which are highly irregular both in terms of shape and topology. Even if anatomical structures present high variability among different patients, human organs maintain a high degree of regularity that we exploit to constraint the registration process. We are interested in designing a loss function to analyze anatomical masks at a global scale, taking into account the anatomical plausibility of the deformed source mask when comparing it with the target mask. Since $\mathcal{L}_{ce}(S_I \circ \mathcal{T}, S_J)$ operates at the pixel level, the back-propagated gradients are parametrized only by pixel-wise individual probability terms and thus provide little global context (Oktay et al., 2018).

We learn a lower-dimensional representation of the anatomical segmentations using denoising autoencoders (DAE) (Vincent et al., 2010). Autoencoders are neural networks designed to learn a mapping from the input space X to a novel, lower-dimensional representation h, that retains significant information about the input. These neural networks usually follow an encoder-decoder architecture, where the encoding $h = \text{enc}(X)$ is extracted from an intermediate fully connected layer. This encoding contains significant information to decode the original input through a decoding phase $X \simeq \text{dec}(\text{enc}(X))$.

The model is trained to minimize the reconstruction error of the input masks, what results in maximizing a lower bound on the mutual information between the input X and learnt representation h (Vincent et al., 2010). In other words, the network is forced to store significant information (useful to reconstruct the

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. Mansilla, D. H. Milone & E. Ferrante; "Learning deformable registration of medical images with anatomical constraints"
Neural Networks, Vol. 124, pp. 269-279, 2020.

original anatomical masks) into the learnt representation. A DAE considers noisy versions of the segmentation masks as input, and is trained to reconstruct clean versions of the corrupted input. This denoising effect, together with the bottleneck imposed by the encoder-decoder architecture, leads the model towards learning a manifold that captures the main variations in the data and maps similar segmentation masks into regions which are close in the manifold.

We train the DAE so that it minimizes the categorical cross-entropy defined over the one-hot encodings of our multi-organ anatomical masks. The noisy input segmentation masks for the DAE were constructed taking the clean segmentation masks and swapping the border pixels of the anatomical structures with the label of its left neighbor with a probability of 0.1. Therefore, the learnt representation concentrates significant information about global anatomical properties of the organs, such as shape and topology, which can be used to introduce global anatomical priors in the learned registration model. In Section 5.1 we perform an experiment to better reflect the advantages of such representation when used as loss function, and discuss why it is beneficial to combine it with standard pixel-level losses.

3.4. AC-RegNet: Learning deformable image registration with anatomical constraints

The novel anatomically constrained registration network (AC-RegNet) architecture is depicted in Figure 1. We combine the basic CNN architecture for image registration described in Section 3.1 with the local segmentation-aware loss function $\mathcal{L}_{ce}$ (Section 3.2) and a new loss term based on the learnt anatomical representations (Section 3.3). This term will encourage global agreement between deformed source and target segmentation masks, ultimately resulting in more realistic and anatomically plausible images after warping. The new term $\mathcal{L}_{ae}$ is defined as the squared Euclidean distance between the codes h generated from the deformed source $S_I \circ \mathcal{T}$ and the corresponding target segmentation mask S_J as:

$$\mathcal{L}_{ae}(S_I \circ \mathcal{T}, S_J) = \|enc(S_I \circ \mathcal{T}) - enc(S_J)\|_2^2 . \tag{4}$$

Note that both, the Euclidean norm and enc(X) are differentiable operations and therefore $\mathcal{L}_{ae}$ is a differentiable loss. The final loss function for our AC-RegNet model considering both, local and global constraints, is given by:

$$\begin{aligned}\mathcal{L}(I, J, S_I, S_J, \mathcal{T}) &= \mathcal{M}(I \circ \mathcal{T}, J) + \lambda_r \mathcal{R}(\mathcal{T}) + \lambda_{ce}\mathcal{L}_{ce}(S_I \circ \mathcal{T}, S_J) + \\ &\quad \lambda_{ae}\mathcal{L}_{ae}(S_I \circ \mathcal{T}, S_J).\end{aligned} \quad (5)$$

The influence of the new term $\mathcal{L}_{ae}$ in the loss function is controlled with a weighting factor λ_{ae}. The main difference between $\mathcal{L}_{ae}$ and the pixel-level $\mathcal{L}_{ce}$ is that the first one acts at a global scale, better reflecting agreement in terms of anatomical shape variations. A deeper study about the complementarity of both losses is provided in Section 5.1.

3.4.1. Training the AC-RegNet model

The training is organized in two stages. First, we train the autoencoder to learn a global and lower-dimensional representation of the anatomical structures using the segmentation masks. Second, we train the AC-RegNet model, by learning the parameters Θ of the VectorCNN that will produce the deformation field $\mathcal{T}$, considering the loss function defined in (5). In this second stage, the parameters of the encoder model used to produce the codes $h = \text{enc}(S)$ are fixed. We highlight the fact that segmentation masks are used during training (of both, the autoencoder and AC-RegNet models) but, at test time, we only require the pair of images to be registered. This is possible since the CNN which produces the deformation field given the pair of images (referred as "Basic Architecture" in Figure 1) does not use the segmentation masks. Instead, segmentation masks representing anatomical structures of interest in the corresponding images are used during training to enforce anatomical correspondences between the target and deformed source through the loss function defined in (5). Once the AC-RegNet model is trained, the registration process of a given pair of images is performed by calculating the deformation field with VectorCNN and deforming the source image with the Warper module (see upper block in Figure 1), without the need of segmentations. The anatomical constraints are therefore introduced

in the model during training.

4. Data and Experimental Setup

4.1. Image dataset

The proposed registration model is evaluated in the context of inter-subject 2D chest X-ray image registration. Performing such task for different patients is challenging, since the anatomical variability between two different subjects can be really high. In our experiments, we use three image databases: the Japanese Society of Radiological Technology (JSRT) database (Shiraishi et al., 2000), the Montgomery County X-ray database and the Shenzhen Hospital X-ray database (Candemir et al., 2013; Jaeger et al., 2013). These last two databases were created by the U.S. National Library of Medicine in collaboration with the Department of Health and Human Services, Montgomery County, Maryland, USA and the Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China, respectively.

JSRT is a public database containing 247 PA chest X-ray images with and without lung nodules (154 nodule and 93 non-nodule images) of 2048x2048 pixels and a spacing of 0.175 mm/pixel. The Montgomery set contains 138 PA X-ray images with and without manifestations of tuberculosis (80 normal and 58 pathological images) of 4020x4892 or 4892x4020 pixels and a spacing of 0.0875 mm/pixel. The Shenzhen set contains 662 X-ray images with and without manifestations of tuberculosis (326 normal and 336 pathological images) in different sizes. Spacing is not provided, so we report results in pixel space when computing distance based measures like Hausdorff distance, which measures the maximum distance between segmentation contours[1]. JSRT provides manual lung and heart segmentations for each image. Manual lung segmentations are available for Montgomery and Shenzhen sets. These segmentation

[1]We used the Hausdorff distance implementation available in MedPy package.

masks will be used to learn the lower-dimensional representations and introduce anatomical context to the registration problem.

The images and segmentations of the Montgomery and Shenzhen sets were preprocessed in order to obtain square images in the same spatial resolution. In each dataset, an image was taken as a reference image and resized by filling its shortest side with background color to make it square. Then, all the images of each dataset were registered against this image, taken as a reference image, through a similarity transform using SimpleElastix (Marstal et al., 2016)[2], finally obtaining images of 4892x4892 pixels in the Montgomery set and 3000x3000 pixels in the Shenzhen set.

4.2. Experimental setting

We randomly divided the images of each dataset in 60% training, 20% validation and 20% test. In the training stage, we sample random pairs of images from the training fold and built mini-batches of size 32. For testing, we sample $2 \times N$ random pairs of images from the test fold, where N represents the number of images in that fold of the dataset.

In order to evaluate the performance of image registration algorithms we employ three metrics commonly used in the literature, which quantify the agreement between the warped source segmentation after registration and the target masks: (i) Dice Similarity Coefficient (DSC), which measures the overlapping between the segmentations (Dice, 1945), (ii) Hausdorff Distance (HD), maximum distance between segmentation contours, and (iii) Average Symmetric Surface Distance (ASSD), computed as the average distance between the segmentation contours. DSC varies between 0 and 1, with 1 indicating a total correspondence between segmentations. HD and ASSD measure the distance between contours in milimeters, and lower values indicate better performance.

[2]The configuration files used to run Elastix can be found online at `https://github.com/lucasmansilla/ACRN_Chest_X-ray_IA/tree/master/acregnet/config/JSRT/elastix`

4.3. Implementation details

The proposed models were implemented in TensorFlow and trained with Adam optimizer considering learning rate of 10^{-3} and default TensorFlow values for the remaining optimization meta-parameters[3]. Models were trained until convergence and the weighting factors for the loss functions were chosen trough grid search using the validation fold, resulting in $\lambda_r = 5 \times 10^{-5}$, $\lambda_{ce} = 1$ and $\lambda_{ae} = 10^{-1}$. A detailed description of the CNN architectures is provided in Appendix A.

5. Results and Discussion

5.1. Understanding the anatomical constraints

We perform a first experiment to compare the behaviour of the proposed global loss function ($\mathcal{L}_{ae}$) with the standard pixel-level loss ($\mathcal{L}_{ce}$), when comparing anatomically plausible and non-plausible segmentation masks. We take 20 random segmentation masks from our dataset, and generate 4 modified versions of each one by changing a constant number of pixels (120 pixels in our example) from the original segmentation (see images (a), (b), (c) and (d) in Figure 2). While the segmentation mask (a) corresponds to an anatomically plausible version of the original mask (we just erode the mask by changing to background 120 pixels in the lungs and heart border), the other versions correspond to anatomically non-plausible masks (we remove blocks of 120 pixels representing complete parts of the lungs or heart). Remind that, in all these cases, a fixed number of pixels was changed. We then compute both losses $\mathcal{L}_{ae}$, $\mathcal{L}_{ce}$ and compare the reference segmentation with its modified versions. As expected, the local $\mathcal{L}_{ce}$ remained constant for the 4 cases, regardless of the place where the pixels are modified, since the number of non-agreeing pixels was 120

[3]Our code is available at `https://github.com/lucasmansilla/ACRN_Chest_X-ray_IA`. Requirements and instructions are listed at `https://github.com/lucasmansilla/ACRN_Chest_X-ray_IA/tree/master/CLI_application/acregnet`

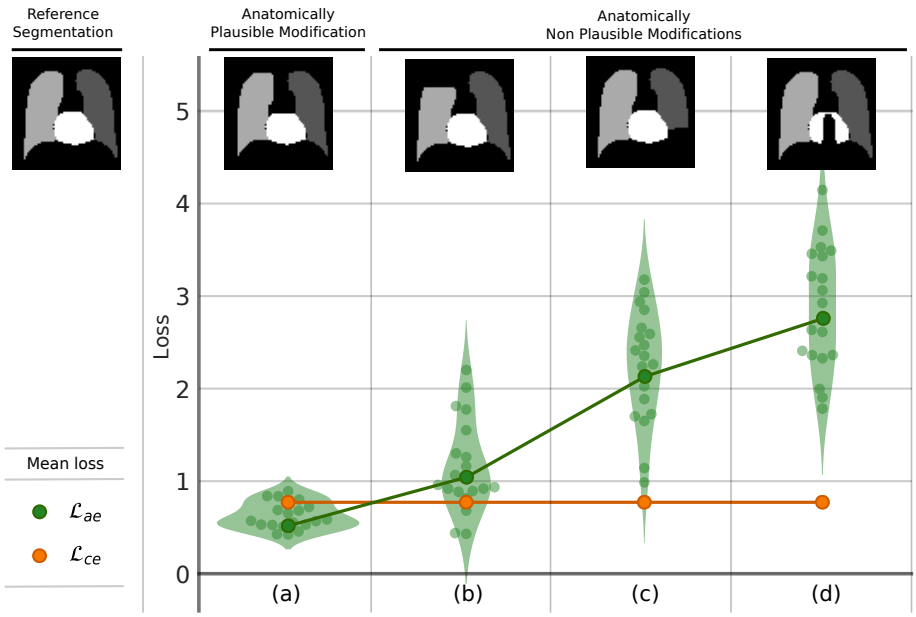

Figure 2: Comparison between the local loss $\mathcal{L}_{ce}$ defined at the pixel level, and the global loss $\mathcal{L}_{ae}$ based on the learnt representation, when comparing 20 random reference segmentations from JSRT dataset with its modified versions. The modified masks were obtained by manually setting 120 foreground pixels to background forming anatomically plausible (a) and non-plausible (b,c,d) versions. The figure shows that $\mathcal{L}_{ae}$ encodes complementary information with respect to $\mathcal{L}_{ce}$. Note that while the local $\mathcal{L}_{ce}$ remains constant, the global $\mathcal{L}_{ae}$ is much lower for the anatomically plausible cases. This indicates that the proposed $\mathcal{L}_{ae}$ loss can discriminate different degrees of anatomical plausibility in cases where the standard pixel level functions indicate equivalent accuracy.

for all of them. However, when observing the behaviour of the global loss $\mathcal{L}_{ae}$, it returned a much lower value for the anatomically plausible case than for the non-plausible cases. Figure 2 shows the loss value for the 20 modified random masks following the same tendency: while the local $\mathcal{L}_{ce}$ remained constant in all cases, the global $\mathcal{L}_{ae}$ returned higher values for the non-plausible masks.

This confirms our intuition about how $\mathcal{L}_{ae}$ encodes complementary information with respect to $\mathcal{L}_{ce}$. In the next section, we will see how this sensitivity to anatomical differences at the global scale can be exploited to improve the

Dataset	Method	Metric		
		DSC	HD	ASSD
JSRT	**AC-RegNet**	**0.943 (0.020)**	**17.973 (7.356)**	**3.340 (1.210)**
	AE-RegNet	0.934 (0.021)	19.464 (8.277)	3.846 (1.320)
	CE-RegNet	0.925 (0.025)	21.973 (8.966)	4.466 (1.553)
	RegNet	0.809 (0.085)	42.177 (19.751)	11.229 (5.035)
	SimpleElastix	0.846 (0.087)	35.713 (18.180)	9.028 (5.050)
Montgomery	**AC-RegNet**	**0.953 (0.017)**	**14.963 (7.910)**	**2.645 (0.957)**
	AE-RegNet	0.947 (0.019)	16.880 (8.621)	2.981 (1.167)
	CE-RegNet	0.929 (0.027)	33.425 (22.813)	4.349 (1.945)
	RegNet	0.869 (0.052)	45.152 (35.702)	8.078 (5.002)
	SimpleElastix	0.879 (0.073)	42.504 (27.480)	7.136 (5.130)
Shenzhen	**AC-RegNet**	**0.931 (0.027)**	**277.386 (182.207)**	**31.738 (15.891)**
	AE-RegNet	0.924 (0.032)	285.549 (179.823)	34.452 (18.259)
	CE-RegNet	0.908 (0.039)	325.958 (201.213)	42.845 (23.560)
	RegNet	0.830 (0.073)	410.012 (225.783)	73.758 (35.849)
	SimpleElastix	0.883 (0.058)	353.562 (217.423)	51.978 (30.299)

Table 1: Mean and standard deviation of Dice Similarity Coefficient (DSC), Hausdorff Distance (HD) and Average Symmetric Surface Distance (ASSD) along all classes (left/right lung and heart) from JSRT, Montgomery and Shenzhen datasets. HD and ASSD for JSRT and Montgomery are expressed in milimeters, while Shenzen is expressed in pixels. Differences among the distributions for all pairs of method are statistically significant according to a paired Wilcoxon test considering Bonferroni correction. Note that AC-RegNet achieves the best results, presenting the highest DSC and lowest HD and ASSD.

accuracy of our registration algorithm.

5.2. Model comparison

The proposed AC-RegNet model was compared with two initial baselines: SimpleElastix (Marstal et al., 2016) and the baseline RegNet described in Section 3.1, which do not consider segmentation-aware loss functions during training. SimpleElastix is a classic medical image registration toolbox, considered

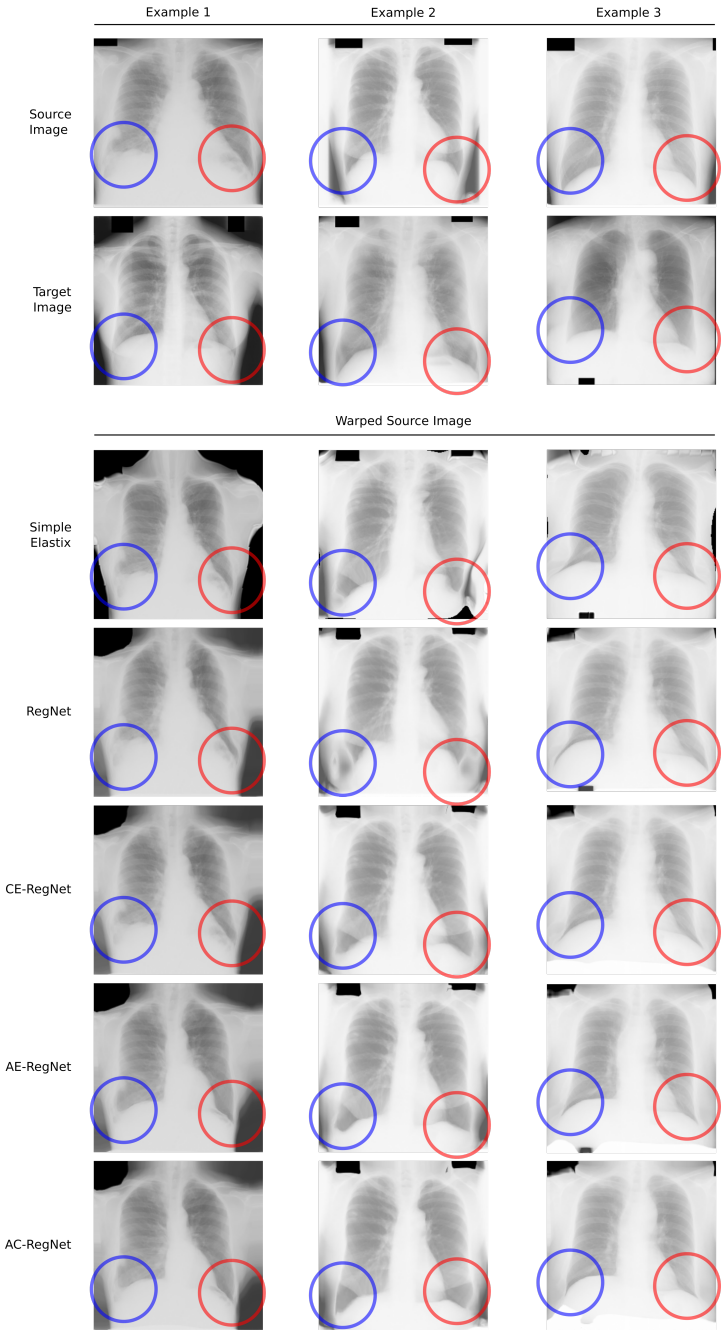

Figure 3: Visualization of the results after registering a pair of images. The blue and red circles highlight two of the areas of the lung anatomy that is better preserved by the AC-RegNet, when compared with the basic RegNet and the other segmentation-aware models. Note, for example, that unlike AC-RegNet, other methods such as SimpleElastix (row 3) and RegNet (row 4), produce images where lungs are cut or stretched down.

state-of-the-art and listed as one of the most popular software packages implementing iterative image registration during the last 20 years (see Viergever et al. (2016)). It has been recently used as baseline method not only for chest X-ray image registration (Ferrante et al., 2018b) but also in a variety of applications (see for example de Vos et al. (2019) or a complete list of publications in the Elastix website[4]).

We also include two segmentation-aware models, one considering only the local $\mathcal{L}_{ce}$ loss (referred as CE-RegNet for Cross Entropy) and another one considering only the global anatomical loss $\mathcal{L}_{ae}$ (referred as AE-RegNet for Auto-Encoder). The proposed model AC-RegNet considers a combination of both losses, as described in (5). The quantitative results reported in Table 1 show that our results are comparable to the state-of-the-art (Larrazabal et al., 2019; Candemir and Antani, 2019). Statistical significance of the results between each pair of methods was verified by performing a Wilcoxon signed-rank test considering Bonferroni correction for each evaluation metric. In particular, when comparing the two most accurate methods (AC-Regnet and AE-RegNet) we obtained $p < 0.005$ in JSRT, $p < 0.05$ in Montgomery and $p < 0.005$ in Shenzhen dataset. Table 1 shows that all segmentation-aware strategies (AC-RegNet, AE-RegNet and CE-RegNet) outperform baseline models (SimpleElastix and RegNet) by a significant margin, giving the highest DSC and lowest HD and ASSD values in each dataset. This already indicates that providing an anatomical context to the network helps to improve performance. Moreover, using the combined local and global metrics (AC-RegNet) yields better performance than the individual cases. Figure 3 illustrates the regularization effect produced by the AC-RegNet when compared with the other models. Note that registering images with AC-RegNet ensures a better anatomical correspondence between the organs, preventing unrealistic shapes in the lung areas as it happens with methods such as SimpleElastix and RegNet (where lungs are cut or stretched

[4]List of publications using Elastix and their corresponding parameter files: `http://elastix.bigr.nl/wiki/index.php/Parameter_file_database`.

down). These results confirm our previous study about the complementarity of both loss functions (see Section 5.1), and the importance of considering global shape information on top of pixel-level descriptors to obtain more anatomically plausible results.

5.3. Applications to X-ray image analysis

In this section we aim at highlighting the potential of AC-RegNet in a variety of medical image analysis tasks. We show three different applications of the proposed method in X-ray images: (i) multi-atlas image segmentation, (ii) reverse classification accuracy (RCA) estimation (Valindria et al., 2017) and (iii) representation learning for pathology classification. We use the well known NIH Chest-XRay14 dataset (Wang et al., 2017) that includes 112,120 chest X-ray images labeled with 14 common thorax diseases according to an automatic natural language processing (NLP) analysis of the radiology reports.

Multi-atlas image segmentation: Anatomical segmentations are useful when performing disease classification and population analysis. The Chest-XRay14 is one of the largest medical datasets publicly available. However, it does not include anatomical segmentations. We used the AC-RegNet model to implement a multi-atlas segmentation model (Iglesias and Sabuncu, 2015) and produce anatomical masks of lung and heart for all the images, which we are making publicly available[5]. We follow a simple multi-atlas segmentation strategy (Mansilla and Ferrante, 2018): given a target image, we take the 5 most similar images from the JSRT dataset (those which maximize the normalized cross correlation with that image) and apply AC-RegNet to register all of them to the target image space. We then transfer the JSRT segmentation labels by applying the resulting deformation field and fuse them using a simple majority voting

[5]The resulting anatomical segmentation masks together with their corresponding RCA coefficient that estimates the quality of the segmentation can be downloaded from: `https://github.com/lucasmansilla/NIH_chest_xray14_segmentations`

mechanism.

We believe that these segmentations are a valuable by-product contribution of our work, which may be used by the medical imaging community to perform further analysis based on the Chest-XRay14 dataset. We conducted automatic quality control to estimate the accuracy of the segmentation using RCA as described in the following paragraph.

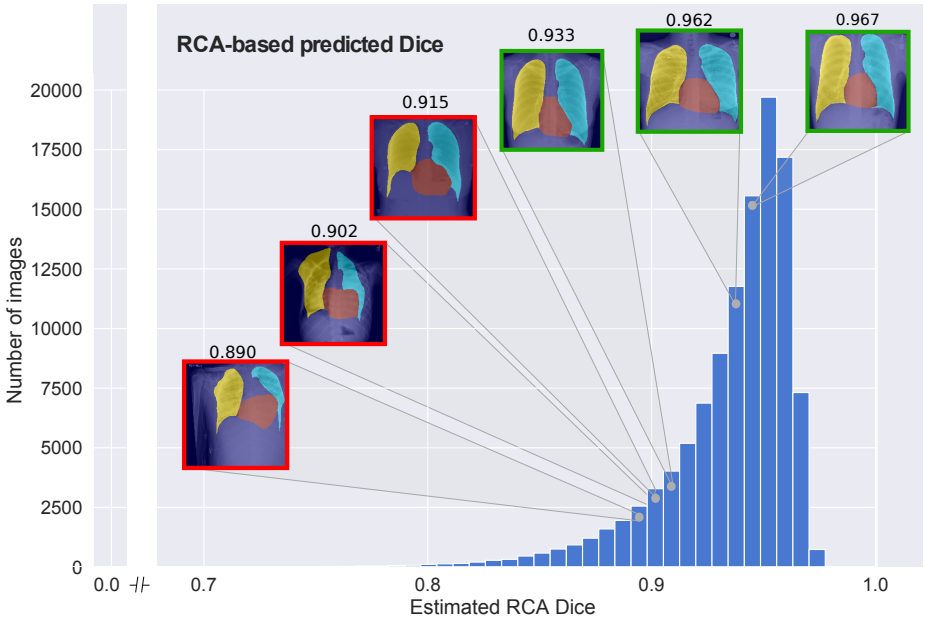

Figure 4: Histogram of estimated Dice coefficients for the resulting Chest-XRay14 segmentations. After visual inspection, we manually set a threshold of 0.92 to perform quality control and decide whether a segmentation meets (in green) or not (in red) the minimum quality standards. We include some visual examples, which represent segmentation masks of various qualities obtained with the propose method, together with their corresponding RCA-based predicted Dice.

Reverse classification accuracy (RCA) estimation: RCA is a framework for predicting the performance of a segmentation method on unseen data, first introduced in Valindria et al. (2017). RCA takes the predicted segmentation from a new image to "train" a reverse classifier (reverse in the sense that it is

trained using a prediction), which is evaluated on a set of reference images with available ground truth. Such reverse classifier may take different forms, ranging from a random forest classifier to a single-atlas segmentation method based on image registration. The hypothesis is that if the prediction is correct, then the RCA classifier trained with that predicted segmentation will perform well in the reference images (those with ground truth). For a more detailed description of RCA see Valindria et al. (2017).

Here we use AC-RegNet to implement RCA based on image registration, estimating the Dice coefficient of the Chest-XRay14 anatomical masks considering JSRT as reference images with ground-truth annotations. In such way, we provide an estimated quality index associated with every segmentation mask, that can be used to define if a given segmentation is to be trusted or not. After visual inspection, we set that threshold in 0.92. Figure 4 shows the histogram of RCA Dice coefficients together with some visual examples of segmentation masks below and above the minimum quality threshold.

Cardiomegaly classification: Cardiomegaly refers to an enlarged heart seen on any imaging test and can be diagnosed based on the cardiothoracic ratio (CR) (Dimopoulos et al., 2013). Since CR can be computed using the boundaries derived from heart and lung masks (Li et al., 2019), the anatomical segmentations should provide enough information to distinguish between healthy and pathological cases. We evaluated the discriminative power of the segmentations which meet the minimum quality requirements (RCA Dice > 0.92) in the task of cardiomegaly vs healthy control classification. After quality control, we kept 87,870 from the original 112,120 images, out of which 2019 were labeled with cardiomegaly. We also sampled 2019 healthy patients with RCA Dice > 0.92 to create a balanced dataset of 4,038 images including control and pathological.

We perform 20-fold cross validation on the aforementioned dataset training a support vector machine (SVM) (Cortes and Vapnik, 1995) with two alternative inputs based on the segmentation masks. As a first alternative, we applied principal component analysis (PCA) to reduce the dimensionality of a vectorized

version of the segmentation masks[6], keeping the 32 principal components and using them as features to train the SVM. Second, we employ the 32-dimensional representation learnt by the same autoencoder used to impose anatomical constraints to the AC-RegNet model. We trained the SVM using these two alternative representations and obtained accuracies of 0.77 and 0.79 respectively. This suggests that the learnt representation encodes useful information that can be exploited in other medical imaging scenarios.

6. Conclusions

In this paper, we introduced a new method to regularize CNN-based deformable image registration by considering global anatomical priors in the form of segmentation masks. Our method learns a non-linear and compact representation of the anatomy associated with medical images, and uses it to constraint the training process of standard CNN-based image registration architectures. Our method outperforms baseline approaches using 3 different datasets of X-ray images . What is more, we have used the to showcase different application scenarios where the proposed method can be used beyond standard image registration tasks, to perform multi-atlas segmentation, reverse classification accuracy estimation and pathology classification.

We provide a comprehensive evaluation of the AC-RegNet model in a challenging problem like chest X-ray image registration, including quantitative and qualitative results in three different datasets (namely JSRT, Montgomery and Shenzhen). We also showcase three different application scenarios in the context of X-ray image analysis using the NIH Chest X-Ray dataset (composed of more than 100,000 images), where the proposed AC-RegNet was used to perform image segmentation, quality control and pathology detection. In conclusion, we found that the proposed global loss function encodes significant information about the anatomical plausibility of a deformed segmentation mask, which

[6]We employ the Scikit-learn (`https://scikit-learn.org/`) implementation of PCA and SVM in our experiments.

complements existent local losses defined at the pixel level. When used in tandem with intensity-based metrics and local losses defined on the segmentation masks, the global loss introduces additional constraints during training that encourage the registration model to produce more anatomically plausible images after deformation.

The proposed model was applied in the context of 2D image registration, but extending it to 3D images is straightforward. In the future, we plan to validate our model in the context of brain 3D image registration, where anatomical structures can be clearly identified and used to constraint the training process. Moreover, as suggested in Hu et al. (2018), CNN-based image registration methods considering segmentation masks can help to alleviate the challenging task of multi-modal registration. We plan to explore how AC-RegNet can be used to develop fast, reliable and realistic image registration methods for multi-modal scenarios.

7. Acknowledgments

EF is beneficiary of an AXA Research Fund grant. The authors gratefully acknowledge NVIDIA Corporation with the donation of the GPUs used for this research, and the support of UNL (CAID-PIC-50220140100084LI and 2016-082) and ANPCyT (PICT 2014-2627, 2018-03907, 2018-3384).

Appendix A. Detailed architectures

References

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2018. An unsupervised learning model for deformable medical image registration, in: IEEE CVPR Proceedings, pp. 9252–9260.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. Voxelmorph: a learning framework for deformable medical image registration. IEEE TMI .

Candemir, S., Antani, S., 2019. A review on lung boundary detection in chest x-rays. IJCARS 14, 563–576.

Candemir, S., Jaeger, S., Palaniappan, K., Musco, J.P., Singh, R.K., Xue, Z., Karargyris, A., Antani, S., Thoma, G., McDonald, C.J., 2013. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE TMI 33, 577–590.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20, 273–297.

Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R., 2018. Unsupervised learning for fast probabilistic diffeomorphic registration. arXiv preprint arXiv:1805.04605 .

Del Palomar, A.P., Calvo, B., Herrero, J., López, J., Doblaré, M., 2008. A finite element model to accurately predict real deformations of the breast. Medical Engineering & Physics 30, 1089–1097.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Dimopoulos, K., Giannakoulas, G., Bendayan, I., Liodakis, E., Petraco, R., Diller, G.P., Piepoli, M.F., Swan, L., Mullen, M., Best, N., et al., 2013. Cardiothoracic ratio from postero-anterior chest radiographs: a simple, reproducible and independent marker of disease severity and outcome in adults with congenital heart disease. International journal of cardiology 166, 453–457.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T., 2015. Flownet: Learning optical flow with convolutional networks, in: IEEE ICCV Proceedings, pp. 2758–2766.

Estienne, T., Vakalopoulou, M., Christodoulidis, S., Battistela, E., Lerousseau, M., Carre, A., Klausner, G., Sun, R., Robert, C., Mougiakakou, S., et al.,

2019. U-resnet: Ultimate coupling of registration and segmentation with deep nets, in: MICCAI, Springer. pp. 310–319.

Ferrante, E., Dokania, P.K., Marini, R., Paragios, N., 2017. Deformable registration through learning of context-specific metric aggregation, in: MLMI, Springer. pp. 256–265.

Ferrante, E., Dokania, P.K., Silva, R.M., Paragios, N., 2018a. Weakly-supervised learning of metric aggregations for deformable image registration. J-BHI .

Ferrante, E., Oktay, O., Glocker, B., Milone, D.H., 2018b. On the adaptability of unsupervised cnn-based deformable image registration to unseen image domains, in: MLMI, Springer. pp. 294–302.

Ferrante, E., Paragios, N., 2017. Slice-to-volume medical image registration: A survey. Medical Image Analysis 39, 101–123.

Gaser, C., 2016. Structural mri: Morphometry, in: Neuroeconomics. Springer, pp. 399–409.

Glocker, B., Komodakis, N., Navab, N., Tziritas, G., Paragios, N., 2009. Dense Registration with Deformation Priors, in: IPMI, pp. 540–551.

Horn, B.K., Schunck., B.G., 1980. Determining Optical Flow. Artificial Intelligence 17, 185–203.

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

Iglesias, J.E., Sabuncu, M.R., 2015. Multi-atlas segmentation of biomedical images: a survey. Medical Image Analysis 24, 205–219.

Jaderberg, M., et al., 2015. Spatial transformer networks, in: NIPS, pp. 2017–2025.

Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R.K., Antani, S., et al., 2013. Automatic tuberculosis screening using chest radiographs. IEEE TMI 33, 233–245.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: NIPS, pp. 1097–1105.

Larrazabal, A.J., Martinez, C., Ferrante, E., 2019. Anatomical priors for image segmentation via post-processing with denoising autoencoders. In Proceedings of MICCAI 2019 .

Li, H., Fan, Y., 2018. Non-rigid image registration using self-supervised fully convolutional networks without training data, in: ISBI, IEEE. pp. 1075–1078.

Li, Z., Hou, Z., Chen, C., Hao, Z., An, Y., Liang, S., Lu, B., 2019. Automatic cardiothoracic ratio calculation with deep learning. IEEE Access 7, 37749–37756.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: IEEE CVPR Proceedings, pp. 3431–3440.

Lucas, B.D., Kanade, T., 1981. An Iterative Image Registration Technique with an Application to Stereo Vision. Imaging 130, 674–679. doi:10.1109/HPDC.2004.1323531.

Mansilla, L., Ferrante, E., 2018. Segmentación multi-atlas de imágenes médicas con selección de atlas inteligente y control de calidad automático, in: XXIV Congreso Argentino de Ciencias de la Computación (La Plata, 2018)., pp. 371–380.

Marstal, K., Berendsen, F., Staring, M., Klein, S., 2016. Simpleelastix: A user-friendly, multi-lingual library for medical image registration, in: Proceedings of the IEEE CVPR Workshops, pp. 134–142.

Oh, S., Kim, S., 2017. Deformable image registration in radiation therapy. Radiation oncology journal 35, 101.

Oktay, O., Ferrante, E., et al., 2018. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. IEEE TMI 37, 384–395.

Paragios, N., et al., 2016. (hyper)-graphical models in biomedical image analysis. Medical Image Analysis 33, 102 – 106.

Ren, Z., Yan, J., Ni, B., Liu, B., Yang, X., Zha, H., 2017. Unsupervised deep learning for optical flow estimation., in: AAAI, p. 7.

Rohé, M.M., Datar, M., Heimann, T., Sermesant, M., Pennec, X., 2017. Svf-net: Learning deformable image registration using shape matching, in: MICCAI, Springer. pp. 266–274.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: MICCAI, Springer. pp. 234–241.

Shakeri, M., Ferrante, E., Tsogkas, S., Lippe, S., Kadoury, S., Kokkinos, I., Paragios, N., 2016. Prior-based coregistration and cosegmentation, in: MICCAI, Springer. pp. 529–537.

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.i., Matsui, M., Fujita, H., Kodera, Y., Doi, K., 2000. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. American Journal of Roentgenology 174, 71–74.

Sokooti, H., et al., 2017. Nonrigid image registration using multi-scale 3d convolutional neural networks, in: MICCAI, Springer. pp. 232–239.

Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable medical image registration: A survey. IEEE TMI 32, 1153–1190.

Stergios, C., Mihir, S., Maria, V., Guillaume, C., Marie-Pierre, R., Stavroula, M., Nikos, P., 2018. Linear and deformable image registration with 3d convolutional neural networks, in: Image Analysis for Moving Organ, Breast, and Thoracic Images. Springer, pp. 13–22.

Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B., 2017. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. IEEE TMI 36, 1597–1606.

Viergever, M.A., Maintz, J.A., Klein, S., Murphy, K., Staring, M., Pluim, J.P., 2016. A survey of medical image registration – under review. Medical Image Analysis 33, 140–144. URL: http://www.sciencedirect.com/science/article/pii/S1361841516301074, doi:https://doi.org/10.1016/j.media.2016.06.030. 20th anniversary of the Medical Image Analysis journal (MedIA).

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. JMLR 11, 3371–3408.

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. Medical Image Analysis 52, 128–143.

de Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, I., 2017. End-to-end unsupervised deformable image registration with a convolutional neural network, in: DLMIA, ML-CDS. Springer, pp. 204–212.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: IEEE CVPR Proceedings, pp. 2097–2106.

Wouters, J., D'Agostino, E., Maes, F., Vandermeulen, D., Suetens, P., 2006. Non-rigid brain image registration using a statistical deformation model, in: Medical Imaging 2006: Image Processing, International Society for Optics and Photonics. p. 614411.

Yang, X., et al., 2017. Quicksilver: Fast predictive image registration–a deep learning approach. NeuroImage 158, 378–396.