Medical Image Analysis (2020)

Contents lists available at ScienceDirect

Medical Image Analysis

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

Weakly-supervised learning of multi-modal features for regularised iterative descent in 3D image registration

Max Blendowski[a], Lasse Hansen[a], Mattias P. Heinrich[a]

[a]*Institute of Medical Informatics, University of Lübeck, Germany*

ARTICLE INFO

Article history:
Received 1 September 2019
Received in final form 26 June 2020
Accepted 4 August 2020

Keywords: Multi-Modal Features, Machine Learning, Image Registration

ABSTRACT

Methods for deep learning based medical image registration have only recently approached the quality of classical model-based image alignment. The dual challenge of both a very large trainable parameter space and often insufficient availability of expert supervised correspondence annotations has led to slower progress compared to other domains such as image segmentation. Yet, image registration could also more directly benefit from an iterative solution than segmentation. We therefore believe that significant improvements, in particular for multi-modal registration, can be achieved by disentangling appearance-based feature learning and deformation estimation. In this work, we examine an end-to-end trainable, weakly-supervised deep learning-based feature extraction approach that is able to to map the complex appearance to a common space. Our results on thoracoabdominal CT and MRI image registration show that the proposed method compares favourably well to state-of-the-art hand-crafted multi-modal features, Mutual Information-based approaches and fully-integrated CNN-based methods - and handles even the limitation of small and only weakly-labeled training data sets.

1. Introduction and motivation

Much recent research has aimed at improving the alignment of images by means of learning optical flow (deformable registration, e.g. Dosovitskiy et al. (2015); Hu et al. (2018)). Deep convolutional neural networks (DCNNs) for displacement field prediction share some similarities to conventional alignment strategies, but in general require a large amount of trainable parameters in a succession of convolution layers, which make the interpretation of learned features difficult. In addition, most previous work has focused on image sequences of the same modality with only subtle changes in appearance and contrast. However, medical image analysis in particular requires the comparison of related structures in different modalities. Multi-modal alignment for diagnosis or biomarker discovery and image-guided interventions is one of the clinically most relevant applications of image registration, e.g. for the registration of a dose-planning CT with an MRI required for organs-at-risk delineation and intra-operative ultrasound to pre-operative MRI fusion in neurosurgery (Heinrich, 2018). Here, the importance lies in obtaining interpretable cross-modal features that enable meaningful correspondences for evaluating changes due to deformations and positioning as well as across patients.

The structure of this paper is as follows: first, we give a brief overview of works that deal with multi-modal registration in order to present the different research directions in this field. Then, we examine the differences to related work and briefly outline the main ideas of the approach to be presented. More detailed information regarding the type of networks used and iterative registration schemes will be presented in the methods section of this work. For the evaluation of the presented ap-

*Corresponding author:
e-mail: blendowski@imi.uni-luebeck.de (Max Blendowski)

proaches, serve extensive, comparative experiments on 3D thoracoabdominal data, the results of which are subsequently discussed. Finally, we summarise the most relevant results once again and give an outlook on further research.

2. Related work

In general, most classical multi-modal image registration approaches make use of two core ideas: either they rely on a similarity measure that is able to handle multi-modal input or they try to find a mapping for both modalities to a shared space and use a monomodal metric. Based on information theoretic insights and as a representative of the first group, Maes et al. (1997) introduced mutual information as a similarity measure that does not require cross-modal features. Exemplary for the second group, inspired by the concept of self similarity proposed in Shechtman and Irani (2007), Heinrich et al. (2012) introduced the expressive, cross-modal MIND descriptor that extracts modality invariant neighbourhood information and enables the use of standard similarity metrics, e.g. the sum of squared differences. The concept has later been extended to compute a 12-dimensional contextual self-similarity descriptor for each voxel in Heinrich et al. (2013). However, these classical methods neglect the central paradigm of more recent CNN-based approaches, which is that problem-specific, data-driven learned features are often more accurate.

As an early example of this idea, Simonovsky et al. (2016) employ a Siamese network architecture to explicitly learn a similarity metric between two multi-modal brain images for registration. A drawback of their method is that training of the CNN weights is not performed in an end-to-end fashion, but uses a pseudo-classification task beforehand. Kim et al. (2017) present end-to-end trainable CNN-based self similarity features, however only trained on mono-modal input. Modality conversion has been employed for learning transferable representations with unpaired multi-modal CycleGANs (Tanner et al., 2018) and Mahapatra et al. (2018) use GANs for multimodal image registration. Furthermore, other early works of Wu et al. (2015), Miao et al. (2016) & Liao et al. (2017) use CNN regressors for the purpose of training. First steps towards a broader comparison on the capabilities of learned features and hand-crafted ones as input into an image registration pipeline are made in (Majumdar et al., 2018). Here, different CNN architectures are again pre-trained for segmentation as an auxiliary task and subsequently compared to hand-crafted features within a discrete optimization framework. With regard to the question to learn or not to learn features for deformable image registration, they find no unambiguous answer for not end-to-end trained features. Recently, Lee et al. (2019) follow an approach that tries to separate the structural representation and the parameter estimation of the transformation. However, while we use learnable weights exclusively in the representative part of our architecture, their representation-learning filters are connected with other parts of the network as well. The use in the spatial-transformer part of their architecture softens the desired separation again.

A variety of recent learning based registration methods has emerged that, in contrast to classical modular techniques, comprise the whole process to generate a displacement field from a given image pair in a fully integrated feed-forward step. Therefore, it is difficult to determine which parts are responsible for alignment or feature extraction in methods that resemble one-step fully-convolutional encoder-decoder architectures, e.g the SVF-Net (Rohé et al., 2017) or VoxelMorph (Balakrishnan et al., 2019). Furthermore, due to their high number of parameters, e.g. the FlowNet proposed by Dosovitskiy et al. (2015) or the label-driven, weakly supervised method of Hu et al. (2018) require large data sets with (pseudo-)ground truth labels during training. A limitation of one-step solutions is their limited ability to capture large deformations. This is evident from results on inhale-exhale CT lung registration (Eppenhof et al., 2019; Hering et al., 2019; de Vos et al., 2019), where large residual errors of 2-3 mm remain for non-iterative encoder-decoder networks while conventional methods reach less than 1mm (Rühaak et al., 2017). Among others the following three complementary ideas have been proposed to learn iterative update steps for registration to overcome these limitations. 1) Using regression forests, Gutierrez-Becker et al. (2017) learn guided update steps by recombining an ensemble of input features. Instead, we aim to learn these features in a data-driven manner. 2) In Ma et al. (2017) an agent-driven approach based on deep reinforcement learning rigidly aligns 2D projections of preoperatively acquired computed tomography scans with depth images. Due to computational demands their method so far is limited to two-dimensional data. 3) Xiong and De la Torre (2013) propose to mimic the output of traditional mathematical descent optimisers (Gauss-Newton, etc.) and learn directions based on feature images within their supervised descent approach. Their method, however, is restricted to mono-modal data and population-based, thus not directly applicable for one-to-one alignment problems. We explored a similar idea in our previous MIDL paper (Blendowski and Heinrich, 2019), where descent directions were learned through guidance from an auxiliary iterative alignment of signed distance maps.

Here, we explore the idea of making the computation of a regularised optimisation step itself differentiable and take inspiration from recent work in computer vision, where Brachmann et al. (2017) developed DSAC (differentiable RANSAC) as a modular end-to-end trainable fitting approach, which effectively disentangles feature learning from regressing a transformation - but so far only for low-parametric homographies.

Contributions

In our conference submission at MIDL 2019 (Blendowski and Heinrich, 2019), we introduced our strategy to integrate deep learning methods into the classical registration pipeline by learning expressive cross-modal features. We guide the feature learning process by penalizing misleading displacement parameter updates based on label-driven weak-supervision as in Hu et al. (2018). In contrast to current work on weakly-supervised optical flow learning, we focus on the disentanglement of appearance and deformation (which has also seen interest in face analysis (Shu et al., 2018)). We thus still follow our

SUpervised **IT**erative de**S**cent algorithm (SUITS) with some modifications. By employing the generally applicable constraints of regularised iterative alignment, we emphasize that our models require less trainable weights ($\approx 10^5$) than fully integrated registration CNNs for learning expressive multi-modal features (typically ≥ 1016) and can be trained with small data sets under only weak-supervision of segmentation labels. Using a robust, iterative optimization framework, even large deformations can be captured that are so far hard to incorporate in one-step fully integrated CNN approaches. Furthermore, the modular nature of the approach allows the flexibility to replace e.g. the displacement update computation algorithm as in classical registration pipelines.

Whereas we limited ourselves to 2-dimensional thoracic data from CT and MRI in our primal work and examined only one iterative registration approach, in this work, we employ a different and well-studied iterative registration update scheme - a Gauss-Newton method introduced Brox et al. (2004) - to implement and examine our end-to-end feature learning strategy. Also differing from our original submission, where we penalised deviating parameter updates, we now compute a loss directly based on the warped moving ground truth labels during training. Our validation on 3-dimensional multi-modal thoracoabdominal unpaired CT to MR registration achieves superior results compared to hand-crafted features and well established multi-modal registration frameworks such as SimpleElastix (Marstal et al., 2016) or the more recent CNN-based VoxelMorph approach (Balakrishnan et al., 2019).

3. Method

In this section, we introduce our proposed supervised iterative descent algorithm (SUITS) for multi-modal image registration. Fig.1 shows the general structure of the method for both the training and inference phases. The modules indicate their interrelationships, which allow the disentangled learning of meaningful feature extracting networks and an iterative estimation of the displacement fields. First, the modular relationships of the entire process are explained, before we go into the details of maintaining differentiability during the iterative computation of displacment field updates.

*3.1. **SU**pervised **IT**erative De**S**cent (SUITS)*

When image pairs originate from the same modality, a variety of registration frameworks is available to align them with each other by iteratively updating the displacement grid parameters. However, our goal is to align multi-modal images that violate e.g. the brightness consistency constraint that is often assumed for mono-modal inputs. There are several unsupervised, one-step procedures that bring prior knowledge into the multimodal registration process by choosing the similarity metric accordingly. We agree with the importance of using prior knowledge, but alter the approach to rely on using it during feature extraction. Especially on CT and MRI data, the MIND descriptor (Heinrich et al., 2012) represents a state-of-the-art reference method for transforming input images into a common space using the concept of self-similarity. Based on these representations, iterative registration procedures for monomodal data can be applied whose metrics do not require any special choice. However, these descriptors do not follow the recent CNN paradigm of learning features in a data driven manner.

For this reason, we are investigating whether and to what extent feature CNNs, which were initially trained as MIND replicators as a form of prior knowledge, can be adapted data-driven. This raises the question of how to adapt these feature mappings.

In general, image registration tackles the problem of finding the transformation $\mathbf{u}$

$$\min_{\mathbf{u}} \mathcal{D}\Big(\mathcal{R}_{fix}(\mathbf{x}), \mathcal{R}_{mov}(\mathbf{x}+\mathbf{u})\Big) + C(\mathbf{u}) \tag{1}$$

that best aligns a given representation of a moving image $\mathcal{R}_{mov}$ at any image position $\mathbf{x}$ with a representation of a fixed image $\mathcal{R}_{fix}$ in terms of a distance measure $\mathcal{D}$ and under additional constraints C, such as plausible smoothness of the displacement field. Below, we give insights on how we are able to learn these shared representations.

Learning Feature CNNs: Our central idea is to use a form of weak supervision in order to achieve a meaningful gradient guidance for error backpropagation to adapt the MIND-pretrained feature CNNs. As depicted in Fig.1, during training we make use of an auxiliary image representation that is computed and processed in the right-hand stream: while using unpaired multi-modal inputs, for each individual image, organ segmentations, that are available during training, represent a simple form of a shared feature space.

Since we follow a classical iterative registration approach during test time (cf. the black arrows), within our framework, we calculate over several runs step-by-step displacement field updates (iteration cycles start and end in *displacement* $\mathbf{u}^k$). In the following we use dense displacement fields $\mathbf{u}$ to describe non-linear local deformations. For each run k, we compute the warped representation of the moving image according to current displacement parameters $\mathbf{u}^k$. We process the fixed and warped images by a feature-extracting CNN - FeatCNN - with Y-shaped architecture. These representations serve as inputs to the *Displacement Update Step Computation* module that outputs the displacement grid parameter updates $\Delta\mathbf{u}^{k-1}$ and for the sake of clarity, encapsulates the mathematical methodology for calculating the updates. During inference, as output of this last module, we simply add the updates multiplied by a step size factor γ to obtain the displacement parameters for the next iteration:

$$\mathbf{u}^k = \mathbf{u}^{k-1} + \gamma \cdot \Delta\mathbf{u}^{k-1}, \text{with } \mathbf{u}^0 = \mathbf{0} \tag{2}$$

Before we consider the gradient flow to adjust the FeatCNN weights, note the following: as in the classical registration setting, where handcrafted descriptors process the input images, the FeatCNN weights remain identical during the whole iterative alignment process at test time. Therefore, the network should learn features that are expressive and applicable during the *complete* iterative alignment process for a given image pair. This is differing from multi-stage regression approaches as proposed e.g. in Xiong and De la Torre (2013). To this end, we make use of the provided training organ labelings as weak su-

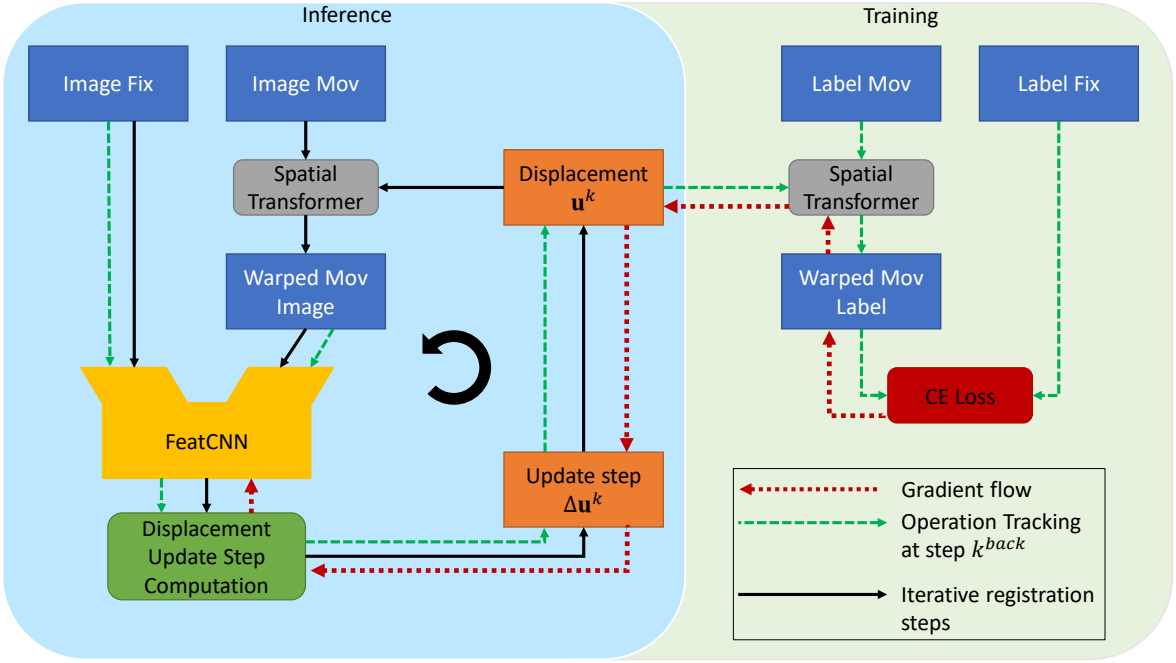

Fig. 1. General structure of our method. The iterative process during inference is depicted on the left hand side (light blue box). For every step, we warp the moving image according to the current displacement parameters $\mathbf{u}^k$ and generate the representations $\mathcal{R}_{mov/fix}$ of the moving and fixed image using the Y-shaped FeatCNN (yellow). Subsequently, based on these representations, the Displacement Update Step Computation module outputs $\Delta\mathbf{u}^k$ for the next iteration. To learn features that can robustly be employed during the whole iterative registration process, during training, we perform a randomly chosen number of k^{back-1} iterations using the provided label maps as representations from a shared space. At iteration k^{back} we track all operations using an autograd engine (dashed green arrows) and compute Δu^k based on the FeatCNN-based representations. During training, we perform a randomly chosen number of k^{back} iterations to learn features that can robustly be employed during the whole iterative registration process. At this iteration we track all operations using an autograd engine (dashed green arrows). We use the weakly-supervised label-driven loss based on the segmentation illustrated at the right hand side to generate a gradient flow backwards to the learnable FeatCNN weights (dotted red arrows).

pervision to guide the registration process. We apply the *Displacement Update Step Computation* for a randomly chosen number $k_{back} - 1$ of iterations directly on the label images, as they represent a monomodal and therefore valid input. Having reached this iteration, we warp the original grey value moving image according to the current displacement parameters $\mathbf{u}^{k_{back}-1}$.

For the next steps, we activate the operation tracking of the autograd engine and enable gradient flow (cf. the dashed green arrows). As a first step, the FeatCNN representations $\mathcal{R}_{mov}^{k_{back}-1}$ of the warped moving image and $\mathcal{R}_{fix}$ of the fixed image are computed. Subsequently, based on these representations, the *Displacement Update Step Computation* yields the update step $\Delta\mathbf{u}^{k_{back}-1}$ and finally - using (2) - $\mathbf{u}^{k_{back}}$. In order to establish a guidance for the FeatCNN, that has to map multi-modal input images to a shared representation, we make use of the weak supervision provided by multi-channel one-hot segmentation labels that are available during training. We generate the warped moving segmentation $\mathcal{S}_{mov}^{k_{back}}$ according to $\mathbf{u}^{k_{back}}$.

Finally, with this transformed label image at hand, we are able to generate a loss signal for the backpropagation process to adapt the FeatCNN weights (cf. the dotted red arrows). We employ a *cross entropy loss* to compute the guidance loss based on the label representation space

$$\mathcal{L}_{guide} = \mathcal{L}_{CE}(\mathcal{S}_{mov}^{k_{back}}, \mathcal{S}_{fix}) \tag{3}$$

Note, that $\mathcal{S}_{k_{back}}^{mov}$ is *not* a multi-channel one-hot label image - due to the trilinear interpolation within the *Spatial Transformer* usage to warp the original $\mathcal{S}_{mov}^{0}$ according to $\mathbf{u}^{k_{back}}$.

We present the training process also in pseudocode below in Algorithm 1.

3.2. Displacement Update Step Computation

In the previous section we have explained the general procedure, which enables us to train the parameters of our FeatCNN with the help of a gradient signal generated by weakly-supervised label-driven loss. Having these essentials at hand, we now go into details of the examined iterative registration method, which is encapsulated in Fig. 1 as *Displacement Update Step Computation*.

Compared to our initial work in Blendowski and Heinrich (2019), where we used a Demons algorithm based on a closed form solution with an adaptive update step-size affecting a denominator, we adopted an iterative registration method inspired by Brox et al. (2004) to compute the displacement parameter updates based on diffusion regularisation. Because this method is based on Gau Newton optimisation, it requires to solve a large, but sparse linear system of equations (LSE). To this end, we employ an *algebraic multigrid* (AMG) solver (Ruge and Stüben, 1987).

It is important to note, that the registration scheme needs to be implemented using only differentiable operations. Thus, during the backward step, we have to differentiate the solving of sparse linear systems itself in order to correctly modify the incoming gradient flow and pass it on to the learnable feature weights. In the following, we will first present the methodological basis of the registration method. Then we will describe the calculation of local gradients during the solution of a linear equation system.

Algorithm 1: Schematic overview of the proposed SUITS approach depicted in Fig.1.

Input: CT & MRI thoracoabdominal images + organ labels
Output: FeatCNN trained for Feature Extraction

Initialise FeatCNN;
Initialise Feature Optimiser & register the FeatCNN weights;
for $epx \leftarrow 0$ **to** $\#epochs$ **do**
  Set $\mathbf{u}^0 = \mathbf{0}$;
  Draw a new random batch of patient pairs;
  Randomly choose k_{back} from $[0, \#iterations]$;
  for $k \leftarrow 1$ **to** $k_{back} - 1$ **do**
    `// NOT Tracked by Feature Optimiser`
    Set fixed Label Image $\mathcal{S}_{fix}$ as fixed representation $\mathcal{R}_{fix}$;
    Set moving Label Image $\mathcal{S}_{mov}$ as moving representation $\mathcal{R}_{mov}$;
    Compute $\mathcal{R}_{mov}^{k-1}$ by warping the moving image $\mathcal{R}_{mov}$ with $\mathbf{u}^{k-1}$;
    Generate $\Delta\mathbf{u}^{k-1}$ using the *Displacement Update Step Computation*;
    Set $\mathbf{u}^k = \mathbf{u}^{k-1} + \gamma \cdot \Delta\mathbf{u}^{k-1}$
  end
  `// At step` k_{back} `tracked by Feature Optimiser`
  Compute $\mathcal{R}_{fix}$ using FeatCNN as fixed representation;
  Compute $\mathcal{R}_{mov}^{k_{back}-1}$ as moving representation by warping the moving greyvalue image with $\mathbf{u}^{k_{back}-1}$ and applying FeatCNN;
  Generate $\Delta\mathbf{u}^{k_{back}-1}$ using the *Displacement Update Step Computation* module;
  Set $\mathbf{u}^{k_{back}} = \mathbf{u}^{k_{back}-1} + \gamma \cdot \Delta\mathbf{u}^{k_{back}-1}$;
  Generate $\mathcal{S}_{k_{back}}^{mov}$ by warping the moving label image according to $\mathbf{u}^{back}$;
  Compute the cross entropy guidance loss $\mathcal{L}_{guide} = \mathcal{L}_{CE}(\mathcal{S}_{mov}^{k_{back}}, \arg\max \mathcal{S}_{fix})$ and backpropagate the error;
end

3.2.1. AMG-Diffusion

In this work, we examine an approach using a diffusion regulariser inspired by Brox et al. (2004) that extends the seminal work of Horn and Schunck (1981) by an additional penalty. We are interested in finding the minimal $\mathbf{u}$ for the energy equation

$$\frac{1}{2}\Big\| \frac{\partial}{\partial x}\mathcal{R}_{mov}(\mathbf{x})\cdot\mathbf{u}_x + \frac{\partial}{\partial y}\mathcal{R}_{mov}(\mathbf{x})\cdot\mathbf{u}_y + \frac{\partial}{\partial z}\mathcal{R}_{mov}(\mathbf{x})\cdot\mathbf{u}_z + \mathcal{R}_{mov}(\mathbf{x}) - \mathcal{R}_{fix}(\mathbf{x})\Big\|_2^2 + \frac{\lambda}{2}\|\nabla\mathbf{u}\|_2^2 = \min_{\mathbf{u}} E(\mathbf{u}) \tag{4}$$

- where under the assumption of small steps a first-order Taylor approximation is legit to linearise the term $\mathcal{R}_{mov}(\mathbf{x}+\mathbf{u}_{x/y/z})$ using the gradient image of dimension $x/y/z$ given by $\frac{\partial}{\partial x/y/z}\mathcal{R}_{mov}$. Minimizing this energy term yields

$$\frac{\partial E(\mathbf{u}_{x/y/z})}{\partial \mathbf{u}_{x/y/z}} = \Big(\frac{\partial}{\partial x}\mathcal{R}_{mov}(\mathbf{x})\cdot\mathbf{u}_x + \frac{\partial}{\partial y}\mathcal{R}_{mov}(\mathbf{x})\cdot\mathbf{u}_y + \frac{\partial}{\partial z}\mathcal{R}_{mov}(\mathbf{x})\cdot\mathbf{u}_z + \mathcal{R}_{mov}(\mathbf{x}) - \mathcal{R}_{fix}(\mathbf{x})\Big)\cdot\frac{\partial}{\partial x/y/z}\mathcal{R}_{mov}(\mathbf{x}) - \lambda\Delta\mathbf{u}_{x/y/z} \stackrel{!}{=} 0 \tag{5}$$

where $-\Delta\mathbf{u}_{x/y/z} = \mathbf{L}\mathbf{u}_{x/y/z}$, with $\mathbf{L}$ being the graph Laplacian operator on the image grid. In a local 6-neighborhood $\mathcal{N}_{\mathbf{x}}^6$ around position $\mathbf{x}$, the Laplacian $\Delta\mathbf{u}_{x/y/z}(\mathbf{x})$ can be approximated by $\sum_{l\in\mathcal{N}_{\mathbf{x}}^6}\mathbf{u}_{x/y/z}(l) - 6\cdot\mathbf{u}_{x/y/z}(\mathbf{x})$.

Since we follow the iterative strategy to start with $\mathbf{u}^0 = \mathbf{0}$ and warp the moving representation $\mathcal{R}_{mov}^k$ according to the current displacements $\mathbf{u}^k = \mathbf{u}^{k-1} + \Delta\mathbf{u}^{k-1}$, we have to compute the update $\Delta\mathbf{u}^{k-1}$. We therefore have to solve the following sparse linear system of equations for $\mathbf{z}$ (with $\odot$ being an elementwise multiplication & $\mathcal{R}_{mov/fix}$ being diagonal matrices)

$$\mathbf{A}\cdot\mathbf{z} = \begin{bmatrix} -\Big(\mathcal{R}_{mov}^{k-1} - \mathcal{R}_{fix}\Big)\odot\frac{\partial}{\partial x}\mathcal{R}_{mov}^{k-1} - \lambda\mathbf{L}\mathbf{u}_x^{k-1} \\ -\Big(\mathcal{R}_{mov}^{k-1} - \mathcal{R}_{fix}\Big)\odot\frac{\partial}{\partial y}\mathcal{R}_{mov}^{k-1} - \lambda\mathbf{L}\mathbf{u}_y^{k-1} \\ -\Big(\mathcal{R}_{mov}^{k-1} - \mathcal{R}_{fix}\Big)\odot\frac{\partial}{\partial z}\mathcal{R}_{mov}^{k-1} - \lambda\mathbf{L}\mathbf{u}_z^{k-1} \end{bmatrix} \tag{6}$$

where $\mathbf{A}$ is composed as the following block matrix:

$$\begin{bmatrix} \Big(\frac{\partial}{\partial x}\mathcal{R}_{mov}^{k-1}\Big)^2 + \lambda\cdot\mathbf{L} & \frac{\partial}{\partial x}\mathcal{R}_{mov}^{k-1}\frac{\partial}{\partial y}\mathcal{R}_{mov}^{k-1} & \frac{\partial}{\partial x}\mathcal{R}_{mov}^{k-1}\frac{\partial}{\partial z}\mathcal{R}_{mov}^{k-1} \\ \frac{\partial}{\partial y}\mathcal{R}_{mov}^{k-1}\frac{\partial}{\partial x}\mathcal{R}_{mov}^{k-1} & \Big(\frac{\partial}{\partial y}\mathcal{R}_{mov}^{k-1}\Big)^2 + \lambda\cdot\mathbf{L} & \frac{\partial}{\partial y}\mathcal{R}_{mov}^{k-1}\frac{\partial}{\partial z}\mathcal{R}_{mov}^{k-1} \\ \frac{\partial}{\partial z}\mathcal{R}_{mov}^{k-1}\frac{\partial}{\partial x}\mathcal{R}_{mov}^{k-1} & \frac{\partial}{\partial z}\mathcal{R}_{mov}^{k-1}\frac{\partial}{\partial y}\mathcal{R}_{mov}^{k-1} & \Big(\frac{\partial}{\partial z}\mathcal{R}_{mov}^{k-1}\Big)^2 + \lambda\cdot\mathbf{L} \end{bmatrix}$$

and $\mathbf{z} = \Delta\mathbf{u}^{k-1} = \Big[\mathbf{u}_x, \mathbf{u}_y, \mathbf{u}_z\Big]^T$ denotes the current displacement updates (not to be confused with the Laplacian here and therefore named $\mathbf{z}$). According to (Brox et al., 2004), the right hand side terms $-\lambda\mathbf{L}\mathbf{u}_{x/y/z}^{k-1}$ emerge from regularizing the sum of the previous displacements and their update direction: $\lambda\|\nabla(\mathbf{u}_{x/y/z}^{k-1} + \Delta\mathbf{u}_{x/y/z}^{k-1})\|_2^2$. Therefore, (6) directly follows from (5) by separating and rearranging the variable terms belonging to $\Delta\mathbf{u}_{x/y/z}^{k-1}$ in the three resulting equations to be minimized. Because there is no closed solution for this explicit LSE, we make use of an *algebraic multigrid solver* approach to solve the given problem efficiently. In addition to its fast convergence, it is particularly interesting in our context that it inherently uses a multiscale strategy.

To sum up, the *Displacement Update Step Computation* outputs a transformation parameter update $\Delta\mathbf{u}^{k-1}$ per iteration by using AMG to solve Eq.(6).

3.2.2. Local gradients of LSE solvers

Having explained the methodological foundations of our iterative registration approach, the following section deals with the backpropagation of the gradient through the LSE solver.

Since we implement our method within a typical autograd engine of a deep learning framework, every module needs to pass local gradients to its inputs with respect to gradients associated with its outputs.

In principle, the solution of Eq.(6) with its form $\mathbf{A}\cdot\mathbf{z} = [...] = \mathbf{b}$ is given by the expression $\mathbf{z} = \mathbf{A}^{-1}\cdot\mathbf{b}$. However, the explicit calculation of the inverse of a sparse matrix is of prohibitive computational demand. Thus, the following known local gradients for a matrix-vector-multiplication are not straightforward to apply in our setting:

MATRIX-VECTOR-MULTIPLICATION

$$\begin{aligned} &\text{forward: } \mathbf{M}\cdot\mathbf{v} = \mathbf{w},\ \mathbf{M}\in\mathbb{R}^{nxn}, \mathbf{v},\mathbf{w}\in\mathbb{R}^n \\ &\text{backward: } grad@\mathbf{v} = \mathbf{M}^T\cdot grad@\mathbf{w} \\ &grad@\mathbf{M} = \begin{bmatrix} grad@\mathbf{w}_1\cdot[v_1,...,v_n] \\ \vdots \\ grad@\mathbf{w}_n\cdot[v_1,...,v_n] \end{bmatrix} \end{aligned} \tag{7}$$

The notation $grad@\mathbf{y}$ refers to the local gradient of a variable $\mathbf{y}$ with respect to the chain rule terms that have to be tracked and computed during the backward pass within the autograd engine following (Hecht-Nielsen, 1992). Considering Eq.(6), we are interested in determining the gradient for the right hand side $\mathbf{b}$ during the backward pass. Preliminary experiments have shown that the calculation of the gradient of the feature difference images $\Big(\mathcal{R}_{mov}^{k-1} - \mathcal{R}_{fix}\Big)$ has the most direct influence on the network parameters. The influence of the memory-intensive calculation for the matrix gradients was negligible.

In order to compute $grad@\mathbf{b}$, we start from an expression involving the inverse system matrix and show a way to circumvent its explicit computation. With $\mathbf{z} = \mathbf{A}^{-1}\cdot\mathbf{b}$, we know from Eq.(7) that $grad@\mathbf{b} = \Big(\mathbf{A}^{-1}\Big)^T\cdot grad@\mathbf{z}$. Multiplying both sides with $\Big(\Big(\mathbf{A}^{-1}\Big)^T\Big)^{-1} = \mathbf{A}^T$ yields

$$\mathbf{A}^T\cdot grad@\mathbf{b} = grad@\mathbf{z}$$

This equation however can readily be solved using the *algebraic multigrid solver* again during the backward pass:

$$grad@\mathbf{b} = amg\_solve\Big(grad@\mathbf{z}, \mathbf{A}^T\Big) \tag{8}$$

Before we summarise the necessary steps of the LSE solver layer within an autograd framework in Algorithm 2, the

$grad@\mathbf{A}$ computation has to be defined for completeness:

$$grad@\mathbf{A} = -1 \cdot \begin{bmatrix} grad@\mathbf{b}_1 \cdot [z_1, ..., z_n] \\ \vdots \\ grad@\mathbf{b}_n \cdot [z_1, ..., z_n] \end{bmatrix} \tag{9}$$

Algorithm 2: Pseudocode for the LSE solver as autograd layer.

`LSESolverForward(`$\mathbf{A}, \mathbf{b}$`):`
  Compute $\mathbf{x} = amg\_solve(\mathbf{b}, \mathbf{A})$;
  Save the tensors $\mathbf{A}$, $\mathbf{b}$ & $\mathbf{x}$ for the backward step;
  return $\mathbf{x}$;
`LSESolverBackward(`$grad@\mathbf{x}$`):`
  Load the saved tensors $\mathbf{A}$, $\mathbf{b}$ & $\mathbf{x}$;
  Compute $grad@\mathbf{b} = amg\_solve(grad@\mathbf{x}, \mathbf{A}^T)$;
  Compute $grad@\mathbf{A} = -1 \cdot \begin{bmatrix} grad@\mathbf{b}_1 \cdot [x_1, ..., x_n] \\ \vdots \\ grad@\mathbf{b}_n \cdot [x_1, ..., x_n] \end{bmatrix}$
  return $grad@\mathbf{A}$, $grad@\mathbf{b}$

4. Experiments

Before we begin to describe the various experiments carried out in the context of this contribution, the setting of our MIDL submission (Blendowski and Heinrich, 2019) is briefly outlined to serve as contrast to the proposed method in this work. Originally we dealt with the multi-modal registration of thoracic CT and MRI data, but only on 2-dimensional data in the sense of a proof-of-concept.

In this work, we use the 3-dimensional thoracoabdominal volume scans of the VISCERAL dataset for oncological staging purposes (Jimenez-del Toro et al., 2016). We make use of the provided unpaired gold corpus training data sets and choose a subset consisting of 20 contrast enhanced CT scans and 20 MRI scans. For each of these patients the same manually annotated ground truth segmentations from medical experts for 6 different structures - namely liver, spleen, left & right psoas major muscles and left & right kidneys - are available that we will use in our proposed methods for the weakly-supervised label-driven guidance. In short, the study population comprises roughly the same number of images from male and female patients (62 male, 69 female) with an average patient age of 59.9 years(± 9.79 years standard deviation). We resampled all images to a 2.0 mm^3 isotropic voxelspacing (coronal: 138, sagittal: 187, axial: 192 voxels) and used the z-transformation per image to normalise the inputs. We pre-registered all scans using the multi-modal, multi-level affine method based on block-matching and trimmed-least squares as described in Heinrich (2018), which achieved state-of-the-art performance in a recent MICCAI challenge (Xiao et al., 2019) (MRI to US brain shift correction). All pairwise transformations were computed and an unbiased groupwise mean was obained using matrix logarithms (see Modat et al. (2014) for details). Using these steps a robust initial affine alignment could be achieved.

For all experiments, we have divided the data set into 15 training patients and 5 test patients per modality, resulting in 225 possible pairs for inter-patient registrations during training and 25 in each test run per fold to evaluate our results. During all registrations, we set the CT images as *moving* images that are to be aligned with the *fixed* MRI volumes. We implemented our experiments using the PyTorch framework running on a Nvidia RTX 2070 GPU.

We roughly subdivide our experiments into 2 categories: Firstly, we distinguish *baseline* experiments using comparative methods from related work. The second group investigates the registration performance of the proposed framework with and without data-driven adapted FeatCNN weights.

4.1. Baseline experiments

SimpleElastix-MI baseline: In order to assess the registration results of our proposed methods, we use SimpleElastix (Marstal et al., 2016) as a robust and well-studied baseline. In the context of this work, it will serve as a representative of classical, multi-modal registration algorithms that employ mutual information as distance measure based on information theoretical insights.

We followed the standard non-rigid registration setting, resulting in a 4 level multi-resolution alignment hierarchy with only non-linear quadratic B-spline transformations since the data has already been preregistered affine. As mentioned before, the framework employed mutual information as distance measure during all steps of this *SimpleElastix-MI* baseline experiment. With hindsight to the FeatCNN experiments described below, we choose the same control point spacing of every 4th voxel for a fair comparison.

Voxelmorph: Introduced in Balakrishnan et al. (2019), the *Voxelmorph* approach implements a fully-integrated, unsupervised CNN-based registration method. This method is based on the optimization of an UNet-like architecture, that predicts a dense displacement field directly from the concatenation of the moving and fixed image. The loss term is composed of a similarity metric and a regularization term. The advantage of this fully-integrated approach is that during inference, only one forward pass of the image pair through the network outputs the transformation parameters. However, there is no clear disentanglement of feature extracting layers and parts of the CNN that are responsible for the transformation parameter estimation. Moreover, as the loss term is comparable to the objective function of conventional iterative registration frameworks, this approach may not outperform its corresponding non-learned counterpart. We performed the experiments with the mentioned train & test splits using the publicly available implementation of *Voxelmorph*[1]. To adapt the *Voxelmorph* framework for the difficulties of multi-modal registration we employ an MSE loss on the MIND features of the warped moving and fixed image. Note, that this is in contrast to our proposed approach, where the CNN is pretrained on predicting MIND-like features and weights can be further optimised on the specific registration

[1] `https://github.com/voxelmorph/voxelmorph`

Table 1. Architectural details of the Y-shaped FeatCNN.

Conv-Layer	1	2	3	4	5
$chan_{in}$	1	16	32	32	32
$chan_{out}$	16	32	32	32	12
dilation	2	1	1	1	1
stride	1	1	2	1	1
pad	4	1	1	1	1
kernelsize	5	3	3	3	3
instance norm	yes	yes	yes	yes	no
activation	ReLU	ReLU	ReLU	ReLU	Sigmoid
shared weights	no	no	yes	yes	yes

task. We adopt the diffusion regulariser on the spatial gradients of the displacement field (weighted with a factor of 0.1 in the combined loss term) and train the network for a total 15000 epochs with an initial learning rate of 0.0001 using the Adam optimiser.

4.2. FeatCNN experiments

FeatCNN architecture: Having described all baseline experiments, we now provide details on the FeatCNN architecture. Taking into account the multimodal nature of our input data, we use a Y-shaped feed-forward architecture. In total, the network consists of 5 layers. The last 3 layers are shared, whereas the first two layers per input stream are adapted to the specific input modalities.

Table 1 contains detailed information regarding the parameter settings of each layer. Overall, this FeatCNN architecture has ≈ 155k trainable parameters. To implement our idea to examine whether and to what extend MIND features as a form of prior knowledge can be adapted in a data driven manner, we first have to train the FeatCNN as MIND replicators. In order to do so, we use the whole gold corpus of the VISCERAL data. For 5000 iterations, we train our Y-shaped architecture to generate the same 12-channel feature output as the handcrafted MIND descriptor on both CT and MRI scans. We train the network weights using an Adam optimiser with an initial learning rate of 0.001 in combination with an $L1$ loss. We use this *pre-trained MIND* FeatCNN in both following experiments.

Note that since the FeatCNN downsamples our input images by a factor of 4, we have to upsample the resulting displacement field updates $\Delta\mathbf{u}^k$ to the original image dimensions for both following approaches using our proposed iterative registration framework. This trilinear upsampling postprocessed by an averaging filter with kernelsize 5 to generate the dense displacement field implements a quadratic B-spline transformation model with control points at every 4th voxel. We empirically chose the step-size parameter $\gamma = 1$.

Pre-trained MIND & iterative diffusion regularization: In order to provide insights on the performance of the iterative registration framework itself, we perform the following experiment. Without any further learning, we freeze the pre-trained FeatCNN and register all 25 test pairs. For 15 iterations each, we update the displacement field and set the regularization parameter $\lambda = 10$.

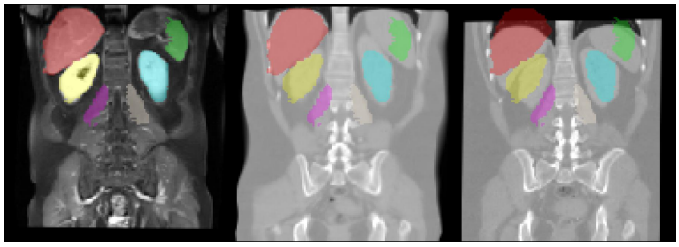

Fig. 2. Exemplar inter-patient registration result. Using our proposed *SUITS* approach, we register a moving CT image (right) to the fixed MR image (left). The resulting warped CT image is shown in between. The ground truth organ segmentations of the fixed MR image are displayed as overlay for all images to illustrate the achieved organ alignment before and after registration.

SUITS end-to-end with pre-training (proposed method): This experiment aims at answering the question if the whole end-to-end feature adaption process provides benefits compared to the replication of hand-crafted descriptors like MIND. As well as the *pre-trained MIND* experiment, it also allows the comparison between a methodology that brings in prior knowledge in the form of features and e.g. *Voxelmorph* that use it in its distance metric. At test time we follow the same procedure as described in the previous section.

To adapt the FeatCNN weights during the training, our strategy is as follows: We use an Adam optimiser with an initial learning rate of $1e-5$ and train with a batch size of 3 image pairs for 250 training iterations. For every iteration, we randomly draw a number $k_{back} \in [0, 15]$ to train expressive features for the *whole* iterative registration process (as described in Sec.3.1). Furthermore, we relax the regularization parameter and set $\lambda = 5$. In doing so, we allow for slightly larger deformations during the k_{back}-iteration and therefore for a more informative cross entropy loss based on the label image differences.

5. Results and discussion

Fig. 3 and Fig. 4 provide on overview over all conducted experiments within this work. Dice segmentation overlap scores over the whole test set are depicted for all 6 foreground structures as well as 95% Hausdorff distance values. Their according numerical values are also listed in Table 2.

As expected, *SimpleElastix* robustly performs registrations on all 20 inter-patient registration pairs (using the publicly available CPU-only implementation: ≈2 min per pair) and forms a solid baseline that all other methods have to compete against. Initial mean Dice scores of 41.3% without any registration increase to 45.9%. Whereas the mean initial 95% Hausdorff distance of 36.4mm remained unchanged for this method.

The registration performance of *VoxelMorph* as a non-iterative CNN-based approach also yields convincing results of an increased Dice score to 47.4% for this fairly difficult multimodal problem. Also, the 95% Hausdorff distance slightly decreased to 35.1mm.

With the fixed *pre-trained MIND & iterative diffusion regularization* outperforming both approaches by reaching a Dice

Table 2. Registration results (mean of Dice segmentation scores and 95th Haussdorf distance (in mm)) for all experiments on 25 pairwise multi-modal inter-patient registration pairs. The *p*-values are calculated with the signed Wilcoxon rank-sum test and show the statistical significance of our proposed method.

	liver		spleen		l pmm		r pmm		l kidney		r kidney		mean		std		*p*-val	
	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95
Initial	.47	60.9	.35	49.1	.30	46.5	.43	23.9	.48	19.9	.45	18.1	.41	36.4	.19	27.1	$1.3\cdot10^{-5}$	$1.7\cdot10^{-5}$
SimpleElastix	.57	60.0	.45	46.4	.35	47.2	.50	25.2	.45	21.1	.43	18.8	.46	36.5	.20	26.6	$2.9\cdot10^{-5}$	$3.6\cdot10^{-4}$
Voxelmorph	**.59**	58.8	**.46**	**43.6**	.35	46.7	.49	22.9	.46	21.1	.50	17.7	.48	35.1	.17	25.8	$2.7\cdot10^{-3}$	0.03
pre-trained MIND	.51	59.0	.39	48.7	.39	43.7	.58	20.9	.52	20.8	.52	16.5	.48	34.9	.22	27.4	$4.6\cdot10^{-5}$	$9.0\cdot10^{-5}$
SUITS (ours)	.55	**57.8**	.44	46.5	**.41**	**42.4**	**.59**	**20.6**	**.55**	**19.7**	**.54**	**16.0**	**.51**	**33.9**	.21	26.7	-	-

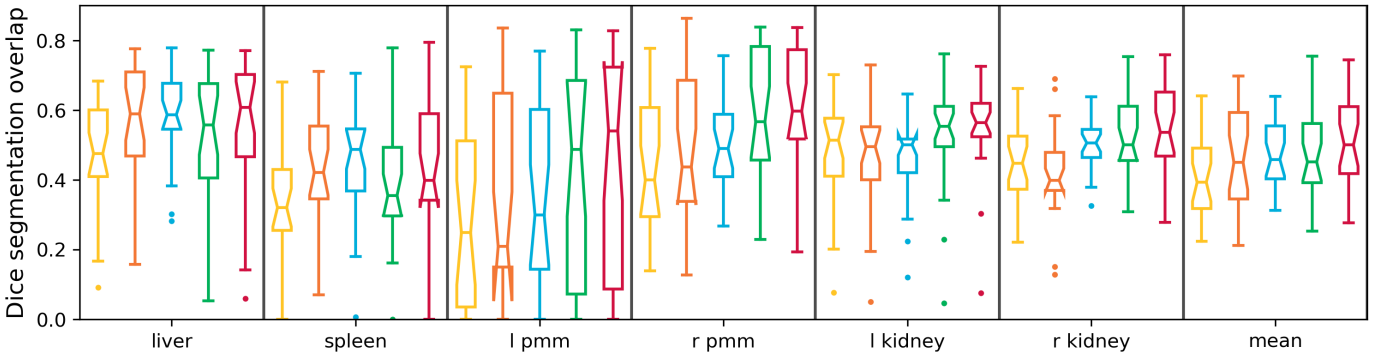

Fig. 3. Box plot illustration of Dice segmentation scores for all experiments on 25 pairwise multi-modal inter-patient registration pairs and per organ mean values for liver, spleen, left & right psoas major muscles (l & r pmm) and left & right kidneys. Following colour coding is used to distinguish the experiments: *initial Dice values* , *SimpleElastix* , *VoxelMorph* , *pre-trained MIND & iterative diffusion regularization* and *SUITS end-to-end with pre-training* (proposed method) .

score of 48.4% and a 95% Hausdorff distance of 34.9mm, the functionality of our iterative diffusion regularised registration method can be determined as being successful (using the publicly available GPU implementation: ≈120ms per pair).

Using *SUITS* - our proposed end-to-end trainable framework with pre-training (implementing an AMG solver following (Bertalan et al., 2014) into the GPU-PyTorch pipeline: ≈30s/15s per training/inference pair) - in combination with the iterative diffusion regularised registration approach finally achieves the highest Dice score of 51.3% and also the lowest 95% Hausdorff distance of 33.8mm. An exemplar inter-patient registration result is depicted in Fig. 2. Finally, by applying the Wilcoxon ranksum test (see Tab.2) to the average organ Dice scores and 95% Hausdorff distance values per registration pair, the accuracy gains of our proposed *SUITS* approach are proven to be statistically significant in comparison to all baseline methods. To sum up our experimental findings, we found that end-to-end training of FeatCNN descriptors specifically for iterative alignment schemes in a multi-modal setting is possible and furthermore also beneficial considering the increasing registration accuracy.

6. Conclusion

In this work, we have presented a framework that allows to disentangle feature learning and registration - even in the presence of multi-modal image data. Our proposed *SUITS* approach outperforms both, the state-of-the-art classical *SimpleElastix* framework (Marstal et al., 2016) using mutual information and a multi-level hierarchy combined with linear and non-rigid registration steps and also *VoxelMorph* as a more recent fully-integrated CNN-based approach (Balakrishnan et al., 2019).

Our method is applicable even with only very limited training data, thus providing an interesting alternative for further multi-modal problem settings.

Compared to Blendowski and Heinrich (2018) - where monomodal COPD lung data are registered using descriptors trained on an auxiliary retrieval task under regularizing binarisation constraints - or Simonovsky et al. (2016) - where a distance metric for T1-T2 brain image registration is learned on classifying patch pairs as pseudo task -, we have demonstrated that using a priori knowledge and specifically adapting feature extractors in an end-to-end fashion for the iterative registration framework in which they are employed, firstly, is indeed possible and, secondly, leads to better registration accuracy.

In contrast to the fully-integrated *VoxelMorph* approach and the method introduced in Lee et al. (2019) - where there is no clear disentanglement of feature learning layers and transformation parameter prediction parts of the network - our features are clearly separated from the iterative registration scheme. For future work, the use of second order optimization algorithms to update the transformation parameters could provide a worthwhile research direction as well as the examination of a possible label bias regarding the registration accuracies in fore- and background.

Overall, we demonstrated a method to adapt pre-trained net-

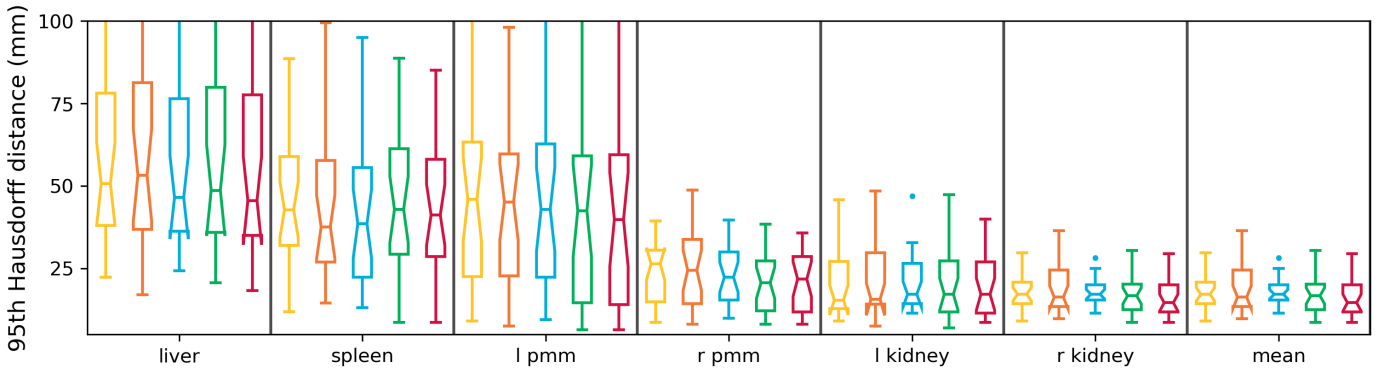

Fig. 4. Box plot illustration of the 95th Hausdorff distance for all experiments on 25 pairwise multi-modal inter-patient registration pairs and per organ mean values for liver, spleen, left & right psoas major muscles (l & r pmm) and left & right kidneys. Following colour coding is used to distinguish the experiments: *initial HD95 values* , *SimpleElastix* , *VoxelMorph* , *pre-trained MIND & iterative diffusion regularization* and *SUITS end-to-end with pre-training* (proposed method) .

works for feature extraction using a weakly-supervised label-driven approach in a multi-modal registration setting on limited training data.

Acknowledgments

This work was supported by the German Research Foundation (DFG) under grant number 320997906 (HE 7364/2-1). We gratefully acknowledge the support of the NVIDIA Corporation with their GPU donations for this research.

References

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging 38, 1788–1800.

Bertalan, T.S., Islam, A.W., Sidje, R.B., Carlson, E.S., 2014. Openmg: a new multigrid implementation in python. Numerical Linear Algebra with Applications 21, 685–700.

Blendowski, M., Heinrich, M.P., 2018. Combining mrf-based deformable registration and deep binary 3d-cnn descriptors for large lung motion estimation in copd patients. International journal of computer assisted radiology and surgery , 1–10.

Blendowski, M., Heinrich, M.P., 2019. Learning interpretable multi-modal features for alignment with supervised iterative descent, in: International Conference on Medical Imaging with Deep Learning, pp. 73–83.

Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C., 2017. Dsac-differentiable ransac for camera localization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Brox, T., Bruhn, A., Papenberg, N., Weickert, J., 2004. High accuracy optical flow estimation based on a theory for warping, in: European conference on computer vision, Springer. pp. 25–36.

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: Proceedings of the IEEE International Conference on Computer Vision, pp. 2758–2766.

Eppenhof, K.A., Lafarge, M.W., Veta, M., Pluim, J.P., 2019. Progressively trained convolutional neural networks for deformable image registration. IEEE transactions on medical imaging .

Gutierrez-Becker, B., Mateus, D., Peter, L., Navab, N., 2017. Guiding multimodal registration with learned optimization updates. Medical image analysis 41, 2–17.

Hecht-Nielsen, R., 1992. Theory of the backpropagation neural network, in: Neural networks for perception. Elsevier, pp. 65–93.

Heinrich, M.P., 2018. Intra-operative ultrasound to mri fusion with a public multimodal discrete registration tool, in: Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation. Springer, pp. 159–164.

Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.A., 2012. Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. Medical image analysis 16, 1423–1435.

Heinrich, M.P., Jenkinson, M., Papież, B.W., Brady, M., Schnabel, J.A., 2013. Towards realtime multimodal fusion for image-guided interventions using self-similarities, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 187–194.

Hering, A., van Ginneken, B., Heldmann, S., 2019. mlvirnet: Multilevel variational image registration network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 257–265.

Horn, B.K., Schunck, B.G., 1981. 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.

Kim, S., Min, D., Ham, B., Jeon, S., Lin, S., Sohn, K., 2017. Fcss: Fully convolutional self-similarity for dense semantic correspondence, in: Proc. IEEE Conf. Comp. Vision Patt. Recog, p. 8.

Lee, M.C., Oktay, O., Schuh, A., Schaap, M., Glocker, B., 2019. Image-and-spatial transformer networks for structure-guided image registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 337–345.

Liao, R., Miao, S., de Tournemire, P., Grbic, S., Kamen, A., Mansi, T., Comaniciu, D., 2017. An artificial agent for robust image registration, in: Thirty-First AAAI Conference on Artificial Intelligence.

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

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P., 1997. Multimodality image registration by maximization of mutual information. IEEE transactions on Medical Imaging 16, 187–198.

Mahapatra, D., Antony, B., Sedai, S., Garnavi, R., 2018. Deformable medical image registration using generative adversarial networks, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE. pp. 1449–1453.

Majumdar, A., Mehta, R., Sivaswamy, J., 2018. To learn or not to learn features for deformable registration?, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications. Springer, pp. 52–60.

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 Conference on Computer Vision and Pattern Recognition Workshops, pp. 134–142.

Miao, S., Wang, Z.J., Zheng, Y., Liao, R., 2016. Real-time 2d/3d registration via cnn regression, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1430–1434.

Modat, M., Cash, D.M., Daga, P., Winston, G.P., Duncan, J.S., Ourselin, S., 2014. Global image registration using a symmetric block-matching approach. Journal of Medical Imaging 1, 024003.

Rohé, M.M., Datar, M., Heimann, T., Sermesant, M., Pennec, X., 2017. Svf-net: Learning deformable image registration using shape matching, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 266–274.

Ruge, J.W., Stüben, K., 1987. Algebraic multigrid, in: Multigrid methods. SIAM, pp. 73–130.

Rühaak, J., Polzin, T., Heldmann, S., Simpson, I.J., Handels, H., Modersitzki, J., Heinrich, M.P., 2017. Estimation of large motion in lung ct by integrating regularized keypoint correspondences into dense deformable registration. IEEE transactions on medical imaging 36, 1746–1757.

Shechtman, E., Irani, M., 2007. Matching local self-similarities across images and videos, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE. pp. 1–8.

Shu, Z., Sahasrabudhe, M., Alp Guler, R., Samaras, D., Paragios, N., Kokkinos, I., 2018. Deforming autoencoders: Unsupervised disentangling of shape and appearance, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 650–665.

Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N., Komodakis, N., 2016. A deep metric for multimodal registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 10–18.

Tanner, C., Ozdemir, F., Profanter, R., Vishnevsky, V., Konukoglu, E., Goksel, O., 2018. Generative adversarial networks for mr-ct deformable image registration. arXiv preprint arXiv:1807.07349 .

Jimenez-del Toro, O., Müller, H., Krenn, M., Gruenberg, K., Taha, A.A., Winterstein, M., Eggel, I., Foncubierta-Rodríguez, A., Goksel, O., Jakab, A., et al., 2016. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks. IEEE transactions on medical imaging 35, 2459–2475.

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.

Wu, G., Kim, M., Wang, Q., Munsell, B.C., Shen, D., 2015. Scalable high-performance image registration framework by unsupervised deep feature representations learning. IEEE Transactions on Biomedical Engineering 63, 1505–1516.

Xiao, Y., Rivaz, H., Chabanas, M., Fortin, M., Machado, I., Ou, Y., Heinrich, M.P., Schnabel, J.A., Zhong, X., Maier, A., et al., 2019. Evaluation of mri to ultrasound registration methods for brain shift correction: The curious2018 challenge. IEEE Transactions on Medical Imaging .

Xiong, X., De la Torre, F., 2013. Supervised descent method and its applications to face alignment, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 532–539.