Learning interpretable multi-modal features for alignment with supervised iterative descent

Max Blendowski BLENDOWSKI@IMI.UNI-LUEBECK.DE
Mattias P. Heinrich HEINRICH@IMI.UNI-LUEBECK.DE
Institute of Medical Informatics, University of Lübeck, DE

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 contrast to most previous approaches, our model does not require full deformation fields as supervision but rather only small incremental descent targets generated from organ labels during training. By mapping the complex appearance to a common feature space in which update steps of a first-order Taylor approximation (akin to a regularised Demons iteration) match the supervised descent direction, we can train a CNN-model that learns interpretable modality invariant features. Our experimental results demonstrate that these features can be plugged into conventional iterative optimisers and are more robust than state-of-the-art hand-crafted features for aligning MRI and CT images.

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

1. Introduction

Much recent research has aimed at improving the alignment of images by means of learning optical flow (deformable registration) (Dosovitskiy et al., 2015; Hu et al., 2018). Deep convolutional networks 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 lighting. However, medical image analysis in particular requires the comparison of related structures in different modalities. Here, the importance lies in obtaining interpretable cross-modal features that enable meaningful correspondences for evaluating changes across patients, to aid diagnosis and for biomarker discovery.

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 allows the usage of standard similarity metrics, e.g. the sum of squared differences. (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.

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 fully-convolutional encoder-decoder architectures, e.g the SVF-Net (Rohé et al., 2017) or (Balakrishnan et al., 2018). 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 datasets with (pseudo-)ground truth labels during training. We 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.

Overview and contributions: Our strategy is to integrate deep learning methods into the classical registration pipeline by learning expressive cross-modal features. Similar to (Xiong and De la Torre, 2013) with their supervised descent approach, we also learn features for a descent direction. However, their method is restricted to mono-modal data and population-based, thus inapt for one-to-one alignment problems. Based on 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 from scratch and give a detailed explanation of our approach in the following. 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 propose a new **SU**pervised **IT**erative de**S**cent algorithm (SUITS) that unrolls a conventional iterative optimisation of regularised B-spline transformations into a differentiable (recurrent) network. By employing the generally applicable constraints of regularised iterative alignment, our model requires only very few trainable weights for learning expressive multi-modal features and can be trained with small datasets under only weak-supervision of segmentation labels. Our experimental validation on multi-modal CT to MRI registration achieves encouraging improvements over hand-crafted features and serves as proof of concept.

2. Methods

In this section, we introduce our proposed supervised iterative descent algorithm (SUITS) for multi-modal image registration. Figure 1 shows the general structure of the method for both the training and inference phases. In addition, the modular interrelationships are indicated, which allow the learning of meaningful feature extracting networks and an iterative estimation of the displacement fields. First, the differentiable B-Spline Descent module essential for this approach is presented, before the entire process is explained in more detail.

2.1. B-Spline Descent module

The main source of inspiration for our work is the classical pipeline of feature-based iterative image registration. If the images f (fixed) and m (moving) are displayed in a common characteristic space

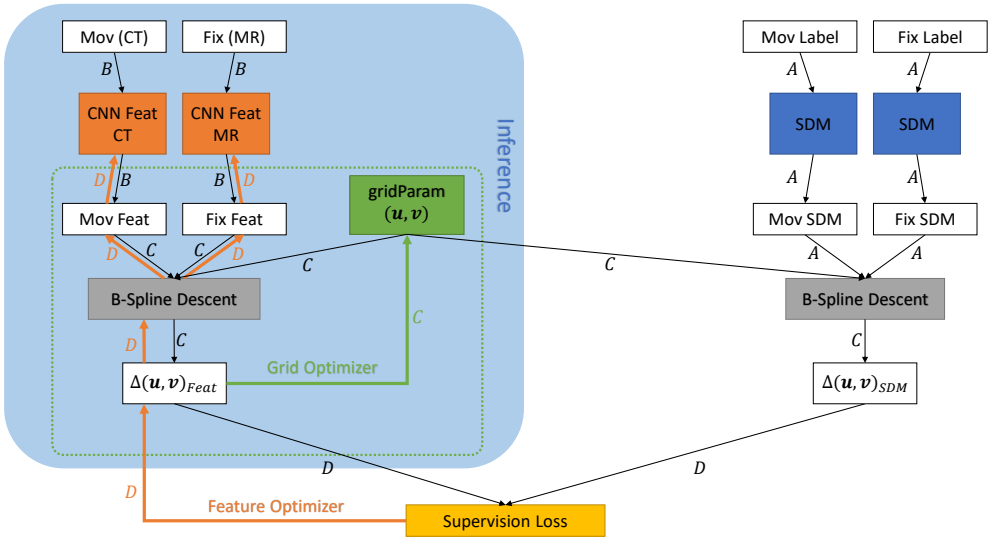

Figure 1: Method overview: During training our approach employs two Adam optimizers; *Grid Optimizer* (green) updates the displacement field parameters based on incremental steps $\Delta(\mathbf{u},\mathbf{v})_{Feat}$ per B-Spline Descent module iteration (*C*). *Feature Optimizer* (orange) adapts the feature CNNs' weights (one complete run *B-D*) guided by the supervision signal stemming from differences to the SDM (*A* - precomputed once) based incremental steps. During inference, only *Grid Optimizer* adjusts the displacement grid, operating on fixed features (MIND/CNNFeat).

by the selection of suitable features, it allows the use of optical flow methods, because the brightness consistency constraint, which results in a sum of squared differences metric, is approximately valid. Based on this assumption, we introduce a B-Spline descent module (grey B-Spline Descent Block in Figure 1). Multi-channel feature representations M and F (per pixel 8 channels for the CNN-based approach and 6 for MIND descriptors) of m and f as well as their current displacement parameters $(\mathbf{u},\mathbf{v})$ serve as input to compute incremental updates $\Delta(\mathbf{u},\mathbf{v})$ for the displacements as output. $(\mathbf{u},\mathbf{v})$ holds a two-dimensional displacement vector at every pixel and $\Delta(\mathbf{u},\mathbf{v})$ contains the gradients to update the displacements. We adopt a widely used energy term to compute $\Delta(\mathbf{u},\mathbf{v})$. To simplify their calculation, a linearization using a first order Taylor approximation is performed - only valid under the assumption of small displacement updates per step used to iteratively refine the warping result based on the initial moving image (Papenberg et al., 2006). Per *image channel c* and pixel position this leads to

$$E_c(\mathbf{u}_c(\mathbf{x}),\mathbf{v}_c(\mathbf{x})) = \frac{1}{2}\left(M_c(\mathbf{x}) + M_{c,\partial x}\cdot\mathbf{u}_c(\mathbf{x}) + M_{c,\partial y}\cdot\mathbf{v}_c(\mathbf{x}) - F_c(\mathbf{x})\right)^2 + \frac{\lambda}{2}(\mathbf{u}_c(\mathbf{x})+\mathbf{v}_c(\mathbf{x}))^2 \quad (1)$$

Here, $M_{c,\partial x/y}$ denote the partial moving image derivatives of channel c and $\frac{\lambda}{2}(\mathbf{u}_c(\mathbf{x})+\mathbf{v}_c(\mathbf{x}))^2$ penalizes big displacement updates. Taking the partial derivatives $\frac{\partial E_c(\mathbf{u}_c,\mathbf{v}_c)}{\partial \mathbf{u}_c(\mathbf{x})/\mathbf{v}_c(\mathbf{x})}$ to minimize this expression

and sorting the terms into a linear system of equations yields:

$$\begin{bmatrix} M_{c,\partial x}^2 + \lambda & M_{c,\partial x} M_{c,\partial y} \\ M_{c,\partial x} M_{c,\partial y} & M_{c,\partial y}^2 + \lambda \end{bmatrix} \begin{bmatrix} \mathbf{u}_c(\mathbf{x}) \\ \mathbf{v}_c(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} (F_c - M_c) M_{c,\partial x} \\ (F_c - M_c) M_{c,\partial y} \end{bmatrix} \tag{2}$$

Finally, using the Sherman-Morrison-Woodbury formula a matrix inversion-free expression is used to efficiently compute displacement grid updates:

$$\begin{bmatrix} \mathbf{u}_c(\mathbf{x}) \\ \mathbf{v}_c(\mathbf{x}) \end{bmatrix} = \frac{1}{\lambda + M_{c,\partial x}^2 + M_{c,\partial y}^2} \cdot \begin{bmatrix} (F_c - M_c) M_{c,\partial x} \\ (F_c - M_c) M_{c,\partial y} \end{bmatrix} \tag{3}$$

Following the heuristic presented in (Guimond et al., 2002), we solve (2) using (3) independently per channel and average the individual, channel-wise solutions to obtain the displacement updates $\Delta(\mathbf{u}, \mathbf{v})$. This update step is basically the core of our approach. It is composed only of differentiable operations and therefore easily to integrate into backpropagation engines used for machine learning such as PyTorch.

In accordance with most image registration approaches, our module also allows the use of displacement fields with a more coarse grid spacing with respect to the actual image. In order to generate dense fields to warp the moving image, we employ third order cardinal B-Spline interpolations (Tustison and Avants, 2013). Due to their definition on a uniform spaced grid and their recursive formulation, interpolation between the knots equals a convolution operation with a smoothing kernel. Using differentiable upsampling followed by two average pooling layers, we can efficiently implement this interpolation scheme to generate dense displacement fields. While choosing $\lambda = (M - F)^2$ locally adapting following (Vercauteren et al., 2009), the incremental displacement updates often exhibit implausible strong local changes. On that account, an additional smoothness penalty that considers the deviation of $\Delta(\mathbf{u}, \mathbf{v})$ from a smoothed version of itself is introduced.

Overall, it is worth noting, that our B-Spline Descent module already outputs an update direction with $\Delta(\mathbf{u}, \mathbf{v})$ that can be backpropagated by off-the-shelf optimizers. Details on how to use the module in the larger context of our method follow below.

2.2. Supervised Iterative Descent (SUITS)

With the B-Spline Descent module at hand, image pairs fulfilling the brightness consistency constraint, should be able to be aligned with each other by iteratively updating the displacement grid parameters. However, our goal is to align multi-modal images that violate this constraint. To this end, we want to learn CNN-based feature mappings for both modalities from scratch - each representing a mapping to a shared representation. This raises the question of how to learn these feature mappings.

Learning Feature CNNs: Here, our idea is to use a form of weak supervision in order to achieve a meaningful gradient guidance for error backpropagation to train the feature CNNs. As depicted in Figure 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 there are organ segmentations available, which we transform to their signed distance maps (SDM) to represent a simple form of a shared feature space (A). We assume that inserting SDMs into the B-Spline Descent module will yield $\Delta(\mathbf{u}, \mathbf{v})_{SDM}$ as a sufficient guidance for a single descent step of $\Delta(\mathbf{u}, \mathbf{v})_{Feat}$ - even if the SDM representation is incomparable (directly) to the original image input.

In our case of unpaired image data, we depend only on this supervisory gradient signal that can be used to propagate update information to the feature CNN weights. An Adam optimizer (named *Feature Optimizer*), that keeps track of which operations the image pairs undergo until $\Delta(\mathbf{u},\mathbf{v})_{Feat}$ is generated, performs updates only on the weights of the feature CNNs (D). The loss signal for this backpropagation is simply the mean squared error $MSE(\Delta(\mathbf{u},\mathbf{v})_{Feat},\Delta(\mathbf{u},\mathbf{v})_{SDM})$.

Iterative Image Alignment: In order to iteratively align the input images, we employ a second Adam optimizer (named *Grid Optimizer*). The green dotted box in Figure 1 encloses the region where it is active. Based on their alignment due to the displacements $(\mathbf{u},\mathbf{v})$ derived from the current (learned) feature representations of two images (B), the B-Spline Descent module outputs an descent step $\Delta(\mathbf{u},\mathbf{v})_{Feat}$ (C). This is fed into the *Grid Optimizer* as gradient value for the parameters $(\mathbf{u},\mathbf{v})$ when it performs its update step. Note that these parameters $(\mathbf{u},\mathbf{v})$ are shared for both the CNN- and the SDM-stream of our method. Also, they are only indirectly influenced by the SDM stream through the supervision signal during training when updating the feature CNNs' weights (D) (this means we never directly use an SDM gradient for spatial alignment). Since the displacement grid parameters $(\mathbf{u},\mathbf{v})$ and their momentum act as memory encoding the current alignment state when adjusting the feature weights, our approach can be interpreted as a recurrent method. Nevertheless, to generate robust features that have to remain fixed during inference, it is important to consider image pairs of different alignment stages throughout training. We achieve this by dividing a given number of total iterations per image pair into p subintervals and randomly choosing the pair to be presented at each step. Thereby within one mini-batch image pairs with different degrees of alignment (number of current update steps) will be used at the same time. Here, we differ from multi-stage regression approaches as proposed e.g. in (Xiong and De la Torre, 2013), since we want our learned features (that are fixed during inference) to be expressive and applicable during the *complete* iterative alignment process for a given image pair.

Our method allows to use multiscale strategies by a stepwise refinement of the grid spacing: starting from a coarse control point grid and performing a fixed number of incremental displacement updates, the displacement field parameters $(\mathbf{u},\mathbf{v})$ of the next stage with a smaller spacing are initialized by upscaled versions of their predecessors.

During inference, the CNN feature representations of an unseen image pair - now without the need of additional annotations - will be iteratively aligned for a fixed number of iterations at different gridscales (indicated by the blue box). Finally, their resulting displacement parameters $(\mathbf{u},\mathbf{v})$ (green block) can be used to warp the moving towards the fixed image. For further algorithmic details refer to Algorithm 1 in Appendix A. [1]

3. Experiments & Results

To verify the applicability of our approach, we perform multi-modal image registrations on 2D coronal slices of unpaired abdominal CT and MRI scans from the VISCERAL dataset (Jimenez-del Toro et al., 2016) with a similar slice thickness. As additional information to train our method, we use the provided label maps for the following structures: liver, spleen, kidneys and psoas major muscles. Dice scores that our approach achieves with fixed learned features during testing serve as our quality measure. We resample these images to an isotropic pixelsize of 1.5mm^2 and compensate only for through-plane transformations of the 3D volumes using the deeds-SSC approach of (Heinrich et al., 2013) and leaving all non-rigid in-plane deformations, which results in a large

1. We plan to make our code publicly under https://github.com/multimodallearning/midl19_suits.

initial misalignment with Dice overlap of 0.44.Subsequently, we crop them without any guidance to dimensions of 320x312. Figure 2 shows example slices from this dataset along with signed distance maps for different organ structures. First, we conducted an unsupervised experiment for monomodal

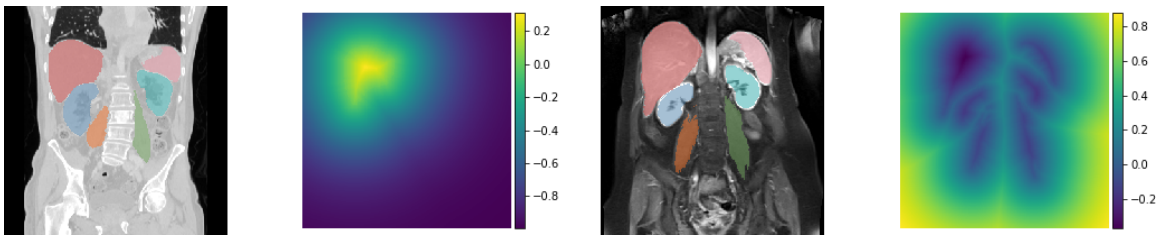

Figure 2: Exemplary Abdominal Scans: (left to right) CT image with provided expert segmentations; scaled signed distance map (SDM) of the liver (CT); MR image with provided expert segmentations; SDM of the background (MR)

CT registration on image intensities to demonstrate the suitability of our B-Spline Descent module for general purposes. Here, initial mean Dice scores increased from 0.44 to 0.69.

In order to assess the results of our proposed *multi-modal* approach, we use the non-trainable MIND descriptor (Heinrich et al., 2012) within our B-Spline Descent scheme as comparison that has been shown to provide robust and expressive modality-independent features for image registration. We also conduct experiments with a state-of-the-art algorithm (SimpleElastix) for multi-modal images (metric: mutual information, 4 level multi-resolution alignment with affine pre-registration) as additional baseline.

In total, we use 10 slices per modality, which corresponds to 100 possible registration pairs. To train and evaluate our approach, we split our patient set in 7 randomly drawn training slices per modality (49 pairs) and evaluate the registration accuracy on the remaining 9 pairs. We repeat this procedure 10 times and provide the mean Dice scores per organ structure for both the MIND baseline as well as for our learned features.

As feature extractors, we use two identical architectures for both modalities: a simple feed-forward design consisting of 7 convolutional blocks each. Using appropriate padding to maintain the input size, we encode the characteristics of the convolutional layers as tuples containing (kernel_size, #output_channels). The number of input channels of the first layer equals 1, since it processes the original images. The following layers choose their number of input channels according to the output channels of their predecessor. Thus, the sequence (7,4)-(7,6)-(5,6)-(5,8)-(5-8)-(3,8)-(3-8) unambiguously defines our feature learning networks. Except from the last convolution that also learns its bias values, they are followed by Group normalization blocks (2 groups) and $tanh$-activation functions. During training, the SDMs are used to supervise the gradient signal. They are computed for the background as well as for every organ based on the expert segmentations and stored channelwise. Since the raw values of the SDMs exhibit large variations depending on different sizes or positions for each organ, we normalize their range to $[-1,1]$ by applying a $tanh(\alpha \cdot x)$ function, where the scaling factor α is set to 0.01 (see Figure 2 for examples). Due to this, we ensure similar value ranges for the SDM feature maps and the CNN feature outputs.

Taking the training schedule for robust feature learning into account as described above (preventing the network from seeing only already similarly aligned images), we update the feature CNNs' weights with *Feature Optimizer* after every 5th iteration of *Grid Optimizer*. *Feature Optimizer* uses an initial learning rate of 0.001 for its parameter updates, while using 0.005 for *Grid Optimizer*. In total, we use 3 grid spacing scales, refining the control point spacing from 20, over 10 to every 7th pixel position. At each stage we compute 300 updates $\Delta(\mathbf{u},\mathbf{v})$ for the displacement grid parameters. The additional penalty, that considers deviations of $\Delta(\mathbf{u},\mathbf{v})$ from its smoothed version, is given more weighting in the finest stage with 0.025 compared to 0.0125 at the first ones. We use two image pairs as one batch. Finally, we only backpropagate updates from regions around organ borders with *Feature Optimizer* to learn CNN weights, i.e. where $abs(SDM_{f/m}(\mathbf{x})) < 0.1$, because only in this regions the supervision can be expected to be informative.

At test time, we maintain all parameters as described above, except that our Feature CNNs are fixed now, i.e. only *Grid Optimizer* performs its iterations. Also, when extracting features with the MIND descriptor instead of our trained CNN features, we keep these parameters. There is no longer a need for the SDM stream as depicted in Figure 1, since it is only used for the supervisory signal during training.

Results: Qualitative results are illustrated in Figure 3 for CT to MR registrations. Here, for both descriptor extraction methods results are depicted. The top row shows the initial images overlayed with their respective segmentations, followed by the fixed image overlayed with segmentations of the moving image that have been warped according to the displacement fields generated based on MIND- and CNN-based features (MIND-Dice: 0.63, CNN-Dice: 0.74), respectively. The bottom row first illustrates the CNN-based displacements by its effects on a grid image after all stages and iterations with the B-Spline Descent module. The following checkerboard visualization gives an intuition of the initial organ alignment (Mean Dice Score 0.40). Also, for both feature types, the checkerboard images depict the organ aligment after registration, where the approach achieves Dice scores of 0.65 based on MIND-features and 0.75 with the learned features for this exemplary image pair.

More quantitative results can be found in Table 1. Here, the mean after 10 test iterations (a total of 90 registration pairs) using the CNN based features slightly outperforms the MIND features. While the MIND experiments achieves better results for the psoas muscles, the trained features (Mean Dice Score 0.72) yield better results aligning larger organ structures, such as the liver or the spleen. Moreover, it compares favourably well against SimpleElastix (Mean Dice Score 0.70) as representative of elaborate, classical algorithms.

4. Discussion & Conclusion

In our work, we developed a new approach to integrate CNN-based multi-modal features into the classical image registration pipeline. We showed that even with unpaired images and only a weak supervision by few organ labels during training, expressive image representations in a shared space are generated - readily employable in iterative multi-stage alignment frameworks. While especially for smaller structures handcrafted MIND features perform competitively, due to the increased receptive field of a multi-layer CNN, larger structures can more easily be aligned. Overall, our method enables to replace the need for dense correspondences by costly to generate displacement fields with organ label maps as the only required additional information during training.

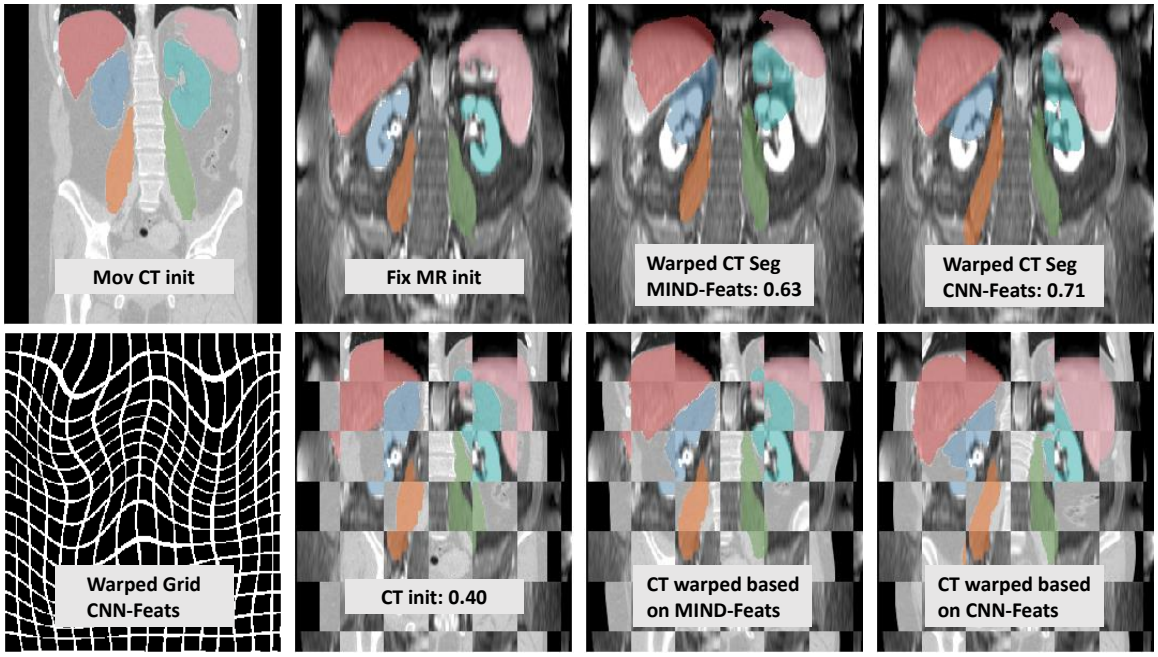

Figure 3: Exemplary Results: (Top row) CT & MRI scans overlayed with respective segmentations and illustration of the MRI scan overlayed with the warped segmentations based on MIND (Dice 0.63) and learned CNN features (Dice 0.71). (Bottom row) An accordingly to the CNN feature based displacement field warped grid; checkerboard illustrations of the initial organ misalignments (Dice: 0.40) and after warping with the respective displacement fields.

Table 1: Organ Dice Scores Listings: Compared to the initial overlaps, both feature extraction approaches achieve better alignments. While the MIND-based registrations perform competitively especially on fine structures (Psoas muscles), the CNN-Feature-based alignments yield favourable results in particular for large structrues (liver, spleen).

	Organs						
Experiment	Liver	Spleen	LKidney	RKidney	LPsoas	RPsoas	∅
Initial Overlap	0.56	0.37	0.52	0.55	0.53	0.65	0.53
SimpleElastix	0.75	**0.68**	0.58	**0.72**	0.68	**0.76**	0.70
MIND Descriptor	0.67	0.45	0.70	0.69	**0.72**	0.75	0.66
Feature CNNs	**0.83**	0.64	**0.74**	0.68	**0.72**	0.73	**0.72**

For future work, this proof of concept encourages to investigate our method in several ways. First, with an extension to 3D, our approach can be examined on challenging 3D datasets with higher demands on memory and computational restrictions. Also, an evaluation of effects depending on

architectural design choices regarding the feature generating networks clearly is of interest. Finally, studying the effects of other possible supervision signals, e.g. combining SDM and MIND features as auxiliary guidance, will provide further insights.

To conclude with, our experiments support the assumption, that disentangling appearance-based feature learning and deformation estimation - as practised in traditional, well-studied iterative approaches - can provide an alternative to parameter-intense end-to-end CNN registration methods.

Acknowledgments

This work was supported by the German Research Foundation (DFG) under grant number 320997906 (HE 7364/2-1). We would like to thank the reviewers for their many insightful comments and suggestions helping to improve our paper. We gratefully acknowledge the support of the NVIDIA Corporation with their GPU donations for this research.

References

Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: A learning framework for deformable medical image registration. *arXiv preprint arXiv:1809.05231*, 2018.

Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, 2017.

Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.

Alexandre Guimond, Charles RG Guttmann, Simon K Warfield, and C-F Westin. Deformable registration of dt-mri data based on transformation invariant tensor characteristics. In *Proceedings IEEE International Symposium on Biomedical Imaging*, pages 761–764. IEEE, 2002.

Benjamin Gutierrez-Becker, Diana Mateus, Loic Peter, and Nassir Navab. Guiding multimodal registration with learned optimization updates. *Medical image analysis*, 41:2–17, 2017.

Mattias P Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V Gleeson, Michael Brady, and Julia A Schnabel. Mind: Modality independent neighbourhood descriptor for multimodal deformable registration. *Medical image analysis*, 16(7):1423–1435, 2012.

Mattias P Heinrich, Mark Jenkinson, Michael Brady, and Julia A Schnabel. Mrf-based deformable registration and ventilation estimation of lung ct. *IEEE transactions on medical imaging*, 32(7): 1239–1248, 2013.

Yipeng Hu, Marc Modat, Eli Gibson, Wenqi Li, Nooshin Ghavami, Ester Bonmati, Guotai Wang, Steven Bandula, Caroline M Moore, Mark Emberton, et al. Weakly-supervised convolutional neural networks for multimodal image registration. *Medical image analysis*, 49:1–13, 2018.

Oscar Jimenez-del Toro, Henning Müller, Markus Krenn, Katharina Gruenberg, Abdel Aziz Taha, Marianne Winterstein, Ivan Eggel, Antonio Foncubierta-Rodríguez, Orcun Goksel, András Jakab, et al. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks. *IEEE transactions on medical imaging*, 35(11):2459–2475, 2016.

Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *Proc. IEEE Conf. Comp. Vision Patt. Recog*, page 8, 2017.

Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997.

Dwarikanath Mahapatra, Bhavna Antony, Suman Sedai, and Rahil Garnavi. Deformable medical image registration using generative adversarial networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1449–1453. IEEE, 2018.

Nils Papenberg, Andrés Bruhn, Thomas Brox, Stephan Didas, and Joachim Weickert. Highly accurate optic flow computation with theoretically justified warping. *International Journal of Computer Vision*, 67(2):141–158, 2006.

Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. Svf-net: Learning deformable image registration using shape matching. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 266–274. Springer, 2017.

Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

Zhixin Shu, Mihir Sahasrabudhe, Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. *arXiv preprint arXiv:1806.06503*, 2018.

Christine Tanner, Firat Ozdemir, Romy Profanter, Valeriy Vishnevsky, Ender Konukoglu, and Orcun Goksel. Generative adversarial networks for mr-ct deformable image registration. *arXiv preprint arXiv:1807.07349*, 2018.

Nicholas James Tustison and Brian Avants. Explicit b-spline regularization in diffeomorphic image registration. *Frontiers in neuroinformatics*, 7:39, 2013.

Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, 2009.

Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

Appendix A. Algorithmic Details

Algorithm 1: Schematic overview of the Training procedure

Input: CT & MR images + organ labels
Output: CNNs trained for Feature Extraction
Initialize FEATURE CNNS;
Initialize FEATURE OPTIMIZER & register the CNNS WEIGHTS;
Initialize GRID OPTIMIZER & register the displacement parameters $(\mathbf{u},\mathbf{v})$;
Generate PAIRPRESENTATIONSCHEME; // different alignment stages
Compute Fixed Signed Distance Maps M_{SDM} & F_{SDM}; // cf. As in Figure 1
for #*grid_scales* **do**
  while *batch_pairs* **in** PAIRPRESENTATIONSCHEME **do**
    // Tracked by FEATURE OPTIMIZER
    Compute $M_{feat} = \text{CNN}_{CT}(m)$ & $F_{feat} = \text{CNN}_{MRI}(f)$; // cf. Bs
    // NOT Tracked by FEATURE OPTIMIZER
    for #*grid_iters* **do**
      // Perform several displacement grid update steps
      Compute GridUpdate $\Delta(\mathbf{u},\mathbf{v})_{Feat} = \text{BSTModule}(M_{Feat}, F_{Feat}, (\mathbf{u},\mathbf{v}))$; // cf. Cs
      Use GRID OPTIMIZER to update $(\mathbf{u},\mathbf{v})$ by $\Delta(\mathbf{u},\mathbf{v})_{Feat}$;
    end
    // Tracked by FEATURE OPTIMIZER
    Compute $\Delta(\mathbf{u},\mathbf{v})_{Feat} = \text{BSTModule}(M_{Feat}, F_{Feat}, (\mathbf{u},\mathbf{v}))$;
    Compute $\Delta(\mathbf{u},\mathbf{v})_{SDM} = \text{BSTModule}(M_{SDM}, F_{SDM}, (\mathbf{u},\mathbf{v}))$;
    Compute $\text{MSE}(\Delta(\mathbf{u},\mathbf{v})_{Feat}, \Delta(\mathbf{u},\mathbf{v})_{SDM})$ as loss; // cf. Ds
    Use FEATURE OPTIMIZER to update the CNN WEIGHTS
  end
end

Algorithm 2: Schematic overview of the Pairwise Registration during Inference

Input: CT & MR image pairs; CNNs trained by Alg. 1
Output: Warped Moving Image, Displacement Parameters $(\mathbf{u},\mathbf{v})$
Initialize GRID OPTIMIZER & register the displacement parameters $(\mathbf{u},\mathbf{v})$;
for #*grid_scales* **do**
  Compute $M_{feat} = \text{CNN}_{CT}(m)$ & $F_{feat} = \text{CNN}_{MRI}(f)$; // cf. Bs
  for #*iters_per_scale* **do**
    Compute GridUpdate $\Delta(\mathbf{u},\mathbf{v})_{Feat} = \text{BSTModule}(M_{Feat}, F_{Feat}, (\mathbf{u},\mathbf{v}))$; // cf. Cs
    Use GRID OPTIMIZER to update $(\mathbf{u},\mathbf{v})$ by $\Delta(\mathbf{u},\mathbf{v})_{Feat}$;
  end
end
Warp m according to $(\mathbf{u},\mathbf{v})$;
return warpedMovingImage, $(\mathbf{u},\mathbf{v})$