International Journal of Computer Assisted Radiology and Surgery (2020) 15:269–276
https://doi.org/10.1007/s11548-019-02089-8

ORIGINAL ARTICLE

Multimodal 3D medical image registration guided by shape encoder–decoder networks

Max Blendowski[1] · **Nassim Bouteldja**[2] · **Mattias P. Heinrich**[1]

Received: 22 February 2019 / Accepted: 4 November 2019 / Published online: 18 November 2019

Abstract

Purpose Nonlinear multimodal image registration, for example, the fusion of computed tomography (CT) and magnetic resonance imaging (MRI), fundamentally depends on a definition of image similarity. Previous methods that derived modality-invariant representations focused on either global statistical grayscale relations or local structural similarity, both of which are prone to local optima. In contrast to most learning-based methods that rely on strong supervision of aligned multimodal image pairs, we aim to overcome this limitation for further practical use cases.

Methods We propose a new concept that exploits anatomical shape information and requires only segmentation labels for both modalities individually. First, a shape-constrained encoder–decoder segmentation network without skip connections is jointly trained on labeled CT and MRI inputs. Second, an iterative energy-based minimization scheme is introduced that relies on the capability of the network to generate intermediate nonlinear shape representations. This further eases the multimodal alignment in the case of large deformations.

Results Our novel approach robustly and accurately aligns 3D scans from the multimodal whole-heart segmentation dataset, outperforming classical unsupervised frameworks. Since both parts of our method rely on (stochastic) gradient optimization, it can be easily integrated in deep learning frameworks and executed on GPUs.

Conclusions We present an integrated approach for weakly supervised multimodal image registration. Achieving promising results due to the exploration of intermediate shape features as registration guidance encourages further research in this direction.

Keywords Multimodal fusion · Guided image registration · Encoder–decoder network · Nonlinear shape interpolation

Introduction

Fusion of data acquired from different modalities (or sensors) plays a very important role in remote sensing, medical imaging, and cross-modal learning. We present a new concept for aligning multimodal medical images by exploring the relations between shape encodings and spatial nonlinear transformations within convolutional autoencoders. The deformable registration of magnetic resonance imaging (MRI) and computer tomography (CT) scans has numerous clinically highly relevant applications, including radiotherapy, image-guided interventions, and multimodal diagnostics. In contrast to same-modality alignment, which is mainly complicated by anatomical deformations due to motion or disease progression, the fusion of multimodal medical scans is in addition highly challenging due to non-functional intensity mapping across CT and MRI and locally varying contrast patterns.

Max Blendowski and Nassim Bouteldja contributed equally to this work.

This work was funded by the German Research Foundation (DFG) under Grant Number 320997906.

✉ Max Blendowski
blendowski@imi.uni-luebeck.de

Nassim Bouteldja
nassim.bouteldja@lfb.rwth-aachen.de

Mattias P. Heinrich
heinrich@imi.uni-luebeck.de

[1] Institute of Medical Informatics, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany

[2] Institute of Imaging and Computer Vision, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany

In this work, we propose to learn a modality-independent mapping to a common shape space that enables us to separate the intensity matching and spatial alignment tasks. First, a joint convolutional encoder–decoder network without skip connections is trained using segmentation masks as supervision to learn a low-dimensional embedding that accurately represents anatomical shapes regardless of input domain/modality. Abolishing skip connections, we enforce the latent code to comprise all relevant information to adequately reconstruct the target shapes. Second, we propose to align the images of two unseen CT and MRI scans based on their reconstructed shapes—using gradient descent and a cross-entropy loss together with a regularization penalty to ensure the smoothness of the estimated nonlinear displacement field. Especially for large deformations, misalignment can occur when a local minima of the cost landscape is reached. Here, we smoothly interpolate realistic intermediate shapes from our learned space and can therefore make use of a divide-and-conquer strategy that concatenates small and thus easier registration steps.

After reviewing the related work in the next section, we describe our proposed method in detail in the "Methods" section. Due to the two-part nature of our approach, we perform experiments in the "Experiments and results" section to evaluate the quality of the learned nonlinear shape embedding on the one hand, as well as the robustness of the proposed iterative registration guidance on the other hand. Finally, we discuss our results in the "Discussion" section and give an outlook on future work to further improve our proposed approach.

Related work

Registration of image pairs usually relies on image metrics that assess how well corresponding structures are aligned [4]. In the monomodal case, similarity measures like the sum of squared differences (SSD) are often sufficient. In contrast, registering volumes from different domains, e.g. CT and MRI, requires more elaborate strategies. Classical approaches contain, e.g. information theoretic methods to compute similarities based on mutual information [13]. However, different modalities may result in deceptive statistical correlations for certain image patterns that do not correspond to real anatomical structures—leading to a physiologically implausible alignment [20].

Alternative strategies transform both modalities into a shared space, where they become comparable. Self-similarity based modality-invariant local image representations have been successfully used in computer vision [16] and medical imaging [7]. Despite their convincing results, with the ongoing success of convolutional neural networks (CNNs) [12], there is currently a clear trend to learn expressive features instead of using handcrafted ones.

In the context of image registration, the learning of features is challenging for several reasons. Most importantly, large amounts of ground-truth data , that are normally a prerequisite for deep learning, are very scarce. In [15,19] aligned images are generated with "traditional" registration approaches, thus the learned transformations simply mimic the latter or serve as parameter initializations for the classical methods. Furthermore, various strategies emerged to simulate pseudo-ground truth deformations that could be employed for the training of CNNs, e.g. [3,17]—without guarantees to comply with real anatomical deformations. Using differentiable image sampling, as first proposed in [9], was employed in [2] to derive a feed-forward network that was trained with classical cost terms (similarity and displacement regularization). This idea was further extended in [8] to employ segmentation labels as supervision for multimodal alignment. Using shape information as prior for learning to segment images from a new modality without paired image data was recently proposed in [11].

We refer interested readers to [14] and [18] for a more detailed overview on classical medical image registration and segmentation-based registration methods in particular. The accuracy of segmentation-based registration methods is always limited by the structure-of-interest delineation quality. However, with the advent of deep learning techniques, this elemental segmentation step has improved dramatically and gained our attraction. Therefore, instead of relying on statistical shape models that could provide perfect 1-to-1 surface correspondences, but only under very costly computations, we want to exploit the advantages of a well-defined shape space to guide a segmentation-based registration process—detailed in the following.

Contributions

Our work aims to overcome certain limitations of previous work on deep learning-based multimodal image registration. Firstly, our method does not require aligned images or landmarks (as in [15]) or (synthetic) ground-truth deformation fields to be trained (cf [3]). It therefore avoids substantial problems of ambiguous correspondences and time-consuming training data generation. Similar to [8] and [11], we rely on weak-supervision through anatomical segmentation labels. New to our work is (1) the novel use of these weakly learned shape priors in a classical optimization driven registration framework and (2) the exploitation of an appropriately constrained shape space to perform meaningful intermediate interpolations of the two considered anatomies. With this approach, we are able to decouple the learning of modality-invariant similarity from known pairwise correspondences and employ well-known regularization cost functions and iterative optimization of consecutive deformations. While our method would be applicable to same-modality

alignment, the practical need for improving multimodal registration and fusion together with joint multimodal representation learning is of much greater importance.

Methods

Our proposed multimodal registration approach is based on two main ideas that we explain in more detail in the following: First, we assume that reasonable correspondences between images from such fundamentally different domains as CT and MRI scans can be more easily found, when aligning consistent segmentations of anatomical structures. In order to generate these segmentations in the first place, we make use of a convolutional encoder–decoder architecture. Second, suitably training such encoder–decoder networks enables to linearly interpolate codes between the shape embedding of both input images to yield smooth and realistic shape interpolations. Reconstructed shapes from these intermediate encodings guide the registration iteratively, instead of facing possibly large nonlinear deformations in the direct registration problem.

CAE for shape-constrained segmentation

Our approach is based on our convolutional autoencoder (CAE) architecture (see Fig. 1) previously published in [1]. Its two central features with regard to the low-dimensional shape representation are as follows: First, our network avoids any skip connections to enable interpretable shape representations. This is necessary and crucial to note for the subsequent registration guidance, since we aim to interpolate between two shapes only by moving through the embedding space. Second, we also proposed in [1] a novel joint training of the model using CT and MR images I_i, as well as segmentations S_i $(i = 1, \ldots, N)$ as alternating inputs to the same network enabling multimodal end-to-end training.

Our model follows a traditional CAE structure, i.e. based on its low-dimensional encoding, it aims at optimally reconstructing the input. The intermediate representation (shape space, see Fig. 1) is of low-dimensional nature to force the network to capture the most salient (rather global) features of the underlying anatomy. Our encoder E is of multimodal nature and projects different input domains (CT, MR, shapes) into a joint 1584 dimensional shape space resulting in very smooth shape predictions. Since multi-organ integer labels are converted to multi-channel one-hot encodings, while MRI and CT are single-channel inputs, the first layer of our network is the only one that differs for grayscale scans and segmentations.

Joint training and CE-optimization: We employ the optimization approach of our previous work [1] to train the model. Here, mini-batches of solely segmentations S_i or grayscale images I_i (MRI and/or CT) are inputted to the network in an alternating fashion. In the former case, the inputted shapes are encoded in the low-dimensional shape space (by E) and subsequently propagated through D for reconstruction. This traditional CAE structure is optimized by cross-entropy (CE) loss minimization between shape inputs and predictions.

When CT and MR images are inputs, we found a CE-based optimization for grayscale input encoding into the shape space to be superior over directly regressing image encodings to their corresponding shape encodings by minimizing the ℓ_1-distances $||E(I_i) - E(S_i)||_1$ (as in [10]). In fact, we fix the decoder D and propagate input images through E as well as D, and minimize the CE-loss between predictions and ground-truth labels $CE\{D(E(I_i)), S_i\}$ to improve image embedding quality. Despite potential vanishing gradient issues, this improves the image embedding due to the following advantages: firstly, the embedding is optimized for the optimal shape code in the current shape space instead of its (suboptimal) shape encoding. Secondly, CE has constantly shown its superiority over the ℓ_1-loss for classification tasks by providing more helpful gradients for optimization, and thirdly, E is solely trained on CE-loss-based updates that makes it redundant to find a proper weighting between ℓ_1- and CE-loss-updates instead, thus improving the stability during training.

Hence, E is trained to improve the reconstruction quality of segmentations, and simultaneously learns to transfer shape as well as multimodal image features into a common shape space trying to yield an equal representation of each domain. In contrast to most previous work, we let E provide about three times as many convolutional layers (and therefore abstractational depth) as D, since D is only optimized for reconstruction quality of segmentations and E for extracting domain-invariant, high-level features, which requires enough preceding nonlinear transformations to reveal common representations of shape features. Interestingly, we found that five convolutional layers suffice for D to map from the shape space into the segmentation domain with a high representation ability.

Iteratively guided registration

Having successfully trained our shape CAE, the subsequent registration approach is illustrated in Fig. 2. Given a pair of images $(\mathcal{F}, \mathcal{M})$, where the moving image $\mathcal{M}$ should be aligned with the fixed image $\mathcal{F}$, we formulate this problem as

$$\underset{\varphi}{\operatorname{argmin}} \ \mathcal{D}(\mathbf{S}_{\mathcal{F}}, \varphi \circ \mathbf{S}_{\mathcal{M}}) + \alpha \mathcal{R}(\varphi) \tag{1}$$

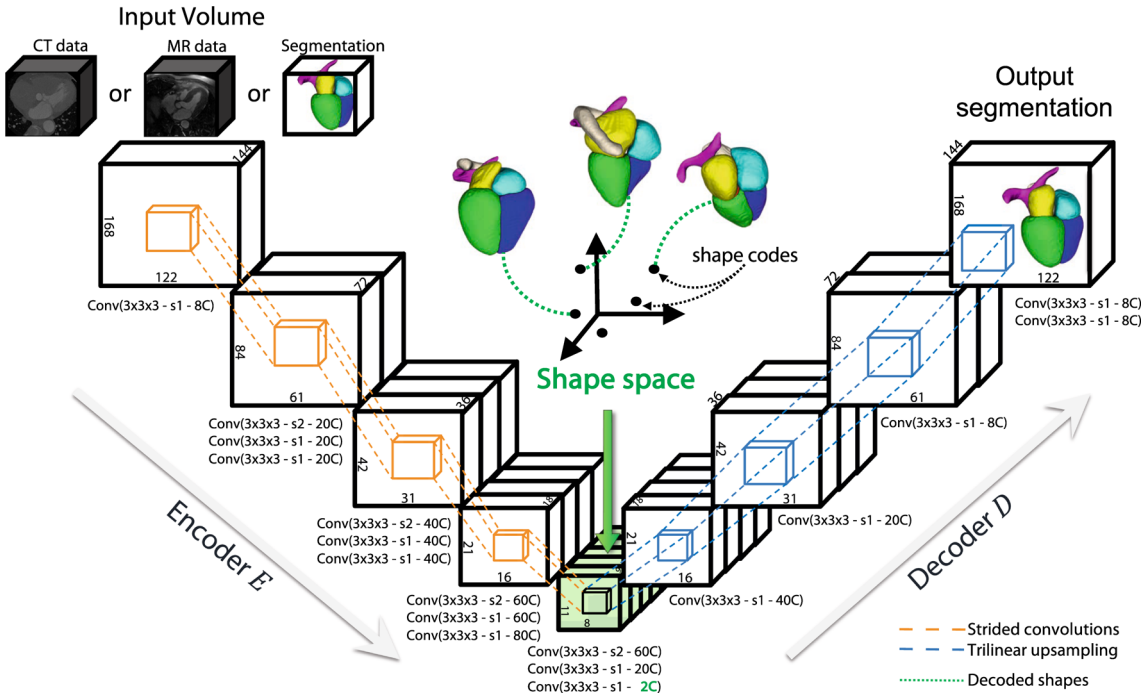

Fig. 1 Block diagram of our proposed all-convolutional model providing 624K trainable parameters. The abbreviation "conv($3 \times 3 \times 3$-s1-10C)" stands for a convolutional layer with $3 \times 3 \times 3$ kernel size, $1 \times 1 \times 1$ striding and 10 output channels. E projects its input into the $2 \cdot 8 \cdot 9 \cdot 11 = 1584$-dimensional shape space. The low-dimensional shape code is then propagated through D for segmentation synthesis. Note that E provides three times as more convolutional layers as D

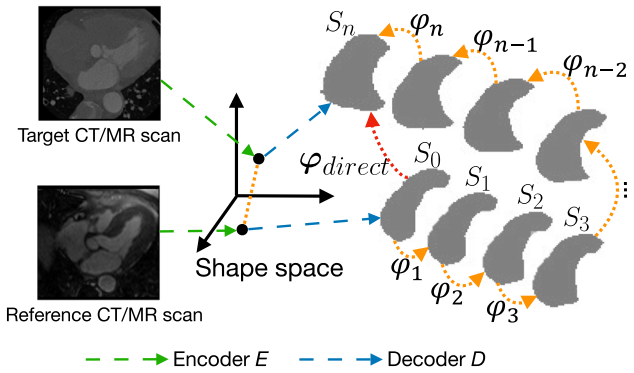

Fig. 2 Iteratively guided registration: we compute CAE encodings $E(\mathcal{M})$ and $E(\mathcal{F})$ for moving ($\mathcal{M}$) and fixed ($\mathcal{F}$) image. We then generate n linearly interpolated encodings to reconstruct $\mathbf{S}_0, \ldots, \mathbf{S}_n$. Instead of directly looking for a possibly large transformation to align $\mathbf{S}_0$ with $\mathbf{S}_n$ via φ_{direct}, we iteratively compute each φ_i between $\mathbf{S}_i$ and $\mathbf{S}_{i-1}$

i.e. we are looking for a transformation φ, that minimizes a distance measure $\mathcal{D}$ between CAE-generated organ labelings and an additional regularization term $\mathcal{R}$, e.g. responsible for smooth deformation fields. Our hypothesis is that we can linearly interpolate between both encodings in this space to yield $n - 1$ smooth shape-interpolated versions evolving between the CAE-generated segmentations $\mathbf{S}_{\mathcal{F}/\mathcal{M}} = D(E(\mathcal{F}/\mathcal{M}))$ by evaluating

$$\mathbf{S}_\lambda = D\left(E(\mathcal{M}) - \frac{\lambda}{n} \cdot (E(\mathcal{M}) - E(\mathcal{F}))\right) \tag{2}$$

with $\lambda \in \{0, \ldots, n\}$, such that $\mathbf{S}_0 = \mathbf{S}_\mathcal{M}$ and $\mathbf{S}_n = \mathbf{S}_\mathcal{F}$. We aim to iteratively guide the registration process between the given moving and fixed images when iteratively aligning their organ labelings.

Consequently, we break down the complex registration problem of finding an optimal transformation φ_{direct} into multiple, and due to smaller deformations, much easier ones:

$$\varphi_{\text{direct}} \approx \varphi_n \circ \varphi_{n-1} \circ \cdots \circ \varphi_2 \circ \varphi_1 \tag{3}$$

The number of interpolation steps n controls the deformation magnitude per iteration. Note that the CAE outputs $\mathbf{S}$ are of dimension $c \times x \times y \times z$, where $c = \#labels$. We extract voxelwise labels L_k by selecting the argmax of $\mathbf{S}_k$ along the first dimension. This is necessary, since we minimize the cross-entropy loss—here acting as distance measure $\mathcal{D}$—between target labels L_k and the warped label map $\varphi \circ \mathbf{S}_{k-1}$ per voxel to obtain the transformation φ_k between two shape interpolations $\mathbf{S}_{k-1}$ and $\mathbf{S}_k$. In order to additionally generate anatomically plausible deformations, we penalize abrupt local changes by summing the squared differences between the deformation and a smoothed version of itself and also favor small transformations by

adding the length of all displacement vectors to our loss: i.e. $\mathcal{R} = \sum_{\mathbf{x}\in\Omega} \|\varphi^{\mathbf{x}} - \varphi^{\mathbf{x}}_{\text{smooth}}\|_2^2 + \|\varphi^{\mathbf{x}}\|_2^2$.

By only employing differentiable loss terms, we can employ *gradient-descent* schemes to iteratively estimate the transformation φ_k that best aligns $\mathbf{S}_{k-1}$ to $\mathbf{S}_k$. Note that we employ an Adam optimizer at this point, that updates the displacements whose gradients we track using the autograd engine of the PyTorch framework. In order to restrict the number of parameters for our transformation model, we use a coarse grid of control points. For every gridpoint g, a three-dimensional displacement vector d^g needs to be estimated, that in combination with its positional identity id^g forms the transformation $\varphi_k^g = id^g + d^g$ at this position. To yield a dense transformation for every image voxel, we use trilinear upsampling, also a differentiable operation.

Experiments and results

According to the split of our method into two steps, we also perform two experiments for the respective parts: First, we examine the proposed encoder–decoder network for shape-constrained segmentation. Subsequently, we perform registrations with a varying number of shape interpolations. Additionally, as comparison, we report registration results obtained with an end-to-end CNN-based approach [8] using their publicly available code, as well as results achieved with SSC deeds [5] as a representative of classical registration frameworks.

We evaluate our approach on the MM-WHS training dataset which consists of 40 multimodality whole-heart images (20 cardiac CT/CTA and MRI) covering whole-heart substructures of different patients each. As preprocessing, our pipeline starts with data resampling into isotropic voxel sizes of $1.5 \times 1.5 \times 1.5\,\text{mm}^3$. We then crop bounding boxes with sizes of $144 \times 122 \times 168$ around the region of interest. The pipeline ends with applying a zero mean unit variance transformation on the cropped grayscale patches. In order to thoroughly validate our method, we employ a fourfold cross-validation (15 CT/MRI for training, 5 for testing).

CAE shape reconstruction

First, we need to examine the robustness of our proposed segmentation approach. While state-of-the-art segmentation methods employ U-Net architectures, we omit skip connections. Although expecting this to cause inferior performance, this step is crucial in order to interpolate between shape encodings to generate intermediate organ labelings as registration guidance.

For the *CAE* experiment (as conducted in [1]), we trained our model on random mini-batches of size 3 containing either CT and/or MR data, or solely segmentations, in an alternating order for 1000 epochs. We use the Xavier method to initialize the parameters of the model and optimize them with Adam. We have empirically chosen the hyper-parameters as follows: The learning rate starts with 0.002 and is reduced by a factor of 0.9 every 30th epoch. Besides, every convolutional layer is followed by batch normalization and a LeakyReLU activation function ($\alpha = 0.1$)—except for a softmax function as final output layer which together with the negative log likelihood loss constitutes the cross-entropy loss on shape reconstructions. Furthermore, we use affine transformations for data augmentation and additionally apply weight decay with a weighting of 10^{-5} to avoid over-fitting. For comparison, we also train the same architecture incorporating skip connections in the *U-Net* experiment only on CT and/or MR grayscale input and evidently without decoder fixation, to judge their effect with regard to the resulting segmentations. To measure the segmentation accuracy, we report the mean Dice–Sørensen coefficient obtained by our fourfold cross-validation experiment.

As expected, *U-Nets* with their skip connections yield average Dice scores of 0.87 (CT) and 0.84 (MR), outperforming the *CAE* variant with 0.84 (CT) and 0.79 (MR). However, as illustrated in Fig. 3, the CAE-generated segmentations can guide iterative registration steps since they still exhibit strong similarities with the ground truth.

Iteratively guided registration

Based on the trained CAE, we subsequently aim to analyze our proposed iterative registration approach. As clarified in the "Methods" section, we argue that a more plausible transformation can be found when guiding the registration process by intermediate shape representations. We therefore conduct experiments with regard to an increasing number of interpolated shapes $\mathbf{S}_i$ in between $\mathbf{S}_\mathcal{M} = \mathbf{S}_0$ and $\mathbf{S}_\mathcal{F} = \mathbf{S}_n$. We use the same fourfold cross-validation splits as in the preceding shape reconstruction experiment. We register five MR images as moving images $\mathcal{M}$ to fixed CT images $\mathcal{F}$ per fold (25 pairs in total) and gradually increase the number of composed transformations φ from $n = 1$ over $n = \{3, 5, 8\}$ to $n = 15$. Because the images contain large anatomical variabilities (unpaired patients) and originate from different domains, their registration is very challenging. For every transformation φ_i again we use Adam with a learning rate of 0.01 and optimize for 50 epochs (# found empirically: ensures converging of $\varphi \circ \mathbf{S}_{k-1}$ and $\mathbf{S}_k$). With placing our control grid points at every 8th image voxel and setting $\alpha = 0.01$ as for the additional deformation constraint $\mathcal{R}$, we regularize the transformations φ.

Figure 4 illustrates the obtained Dice values for each organ label (true CTSeg to warped true MRSeg) when using different numbers of intermediate shape representations as registration guidance. As baselines, we show the initial Dice

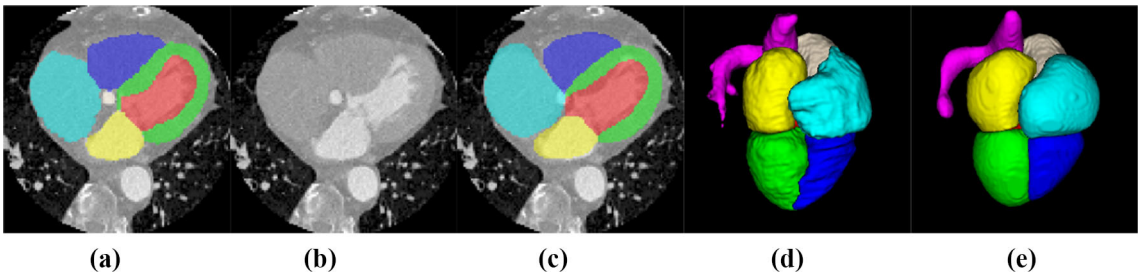

Fig. 3 Illustrating *CAE*-based segmentations: **a** expert segmentation of the axial CT slice in (**b**); **c** *CAE*-generated labeling; 3D renderings of the ground truth (**d**) and the corresponding *CAE*-result (**e**)

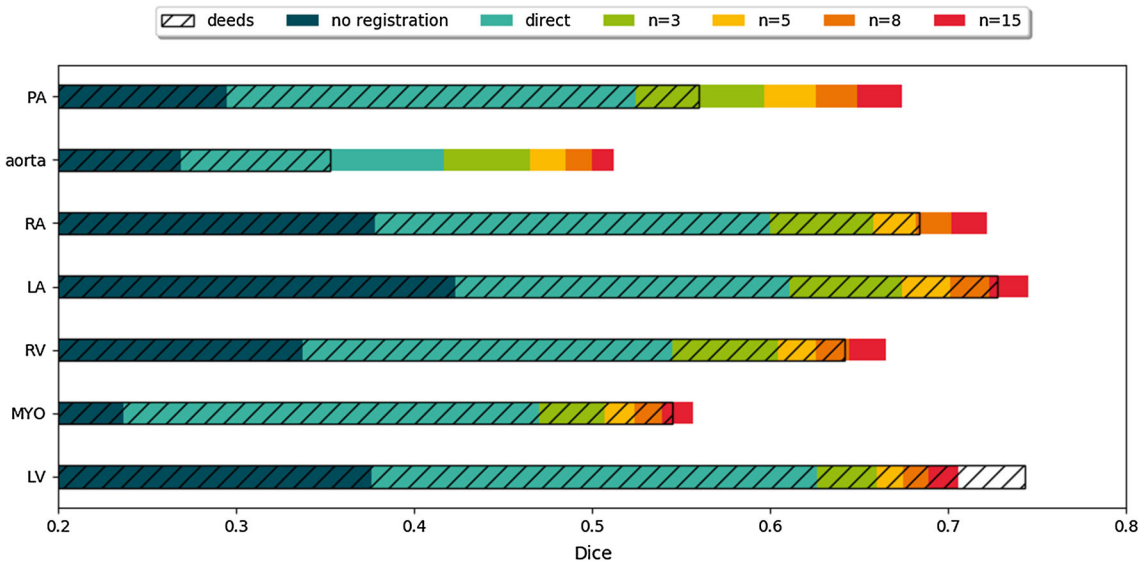

Fig. 4 Registration results for 20 MR to CT pairs: average Dice scores between ground-truth CT segmentations and warped ground truth MR segmentations with increasing number n of composed transformations $\varphi_n \circ \cdots \circ \varphi_1$. $n = 15$ (red) is achieving the best results and outperforms a direct registration ($n = 1$, light blue) by $+11.65\%$. Striped bars indicate the *SSC deeds* [5] results

values when transferring the true segmentations without any registration (dark blue) and Dice scores obtained with a classical image registration framework, that was specifically designed for multimodal MR-CT alignment (SSC deeds, [5,6], striped bars). Furthermore, we conducted experiments with the publicly available code for [8]. Compared to the initial average Dice score of 33% without registration, only a minimal improvement to 35% could be achieved. With regard to our approach, composing $n = 15$ (red) transformations φ leads to a rise of 11.65% (∅ Dice: 65.27%; ∅ stddev field Jacobian: 0.3994, indicating volume changes; % rate Jacobian $< 0 : 0.001$, indicating foldings) compared to a direct registration ($n = 1$, light blue, ∅ Dice: 53.62%, ∅ stddev field Jacobian: 0.2210, % rate Jacobian < 0: 0.001). Using the Wilcoxon rank sum test, this increase in dice scores is statistically significant ($p = 7.98e{-}4$). Although getting gradually smaller with a growing number of iterations, Dice scores steadily increase as indicated by the stacked, correspondingly colored horizontal barplots. Table 1 summarizes the results.

In Fig. 5, we exemplarily show the same axial slice of a patient before and after registration.

The bottom row with its checkerboard representations visualizes the alignment of heart structures after registering both volumes with our proposed approach. Only exhibiting slight unevenness for label transitions at checker borders in the middle after registration, this visual inspection confirms the enhanced alignment of foreground structures—which our employed cross-entropy loss function in conjunction with the Adam optimizer targets for. While these image contents undergo larger deformations, most background parts remain nearly unaltered.

Table 1 Results of evaluated approaches. *label-reg* [8] only minimally improves NO_REG, while our approach with $n = 15$ outperforms classical *SSC deeds* [5]

Method	NO_REG	*label-reg* [8]	Ours $n = 1$	*SSC deeds* [5]	Ours $n = 15$
Dice	0.331	0.352	0.526	0.608	**0.653**

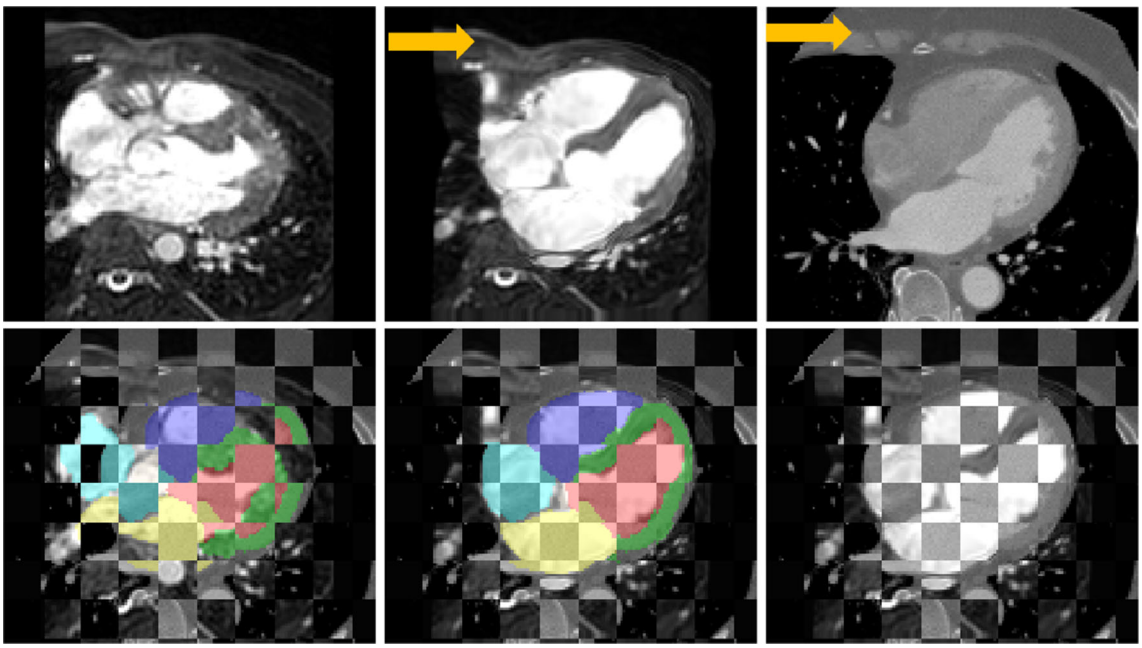

Fig. 5 CT-MR registration pair. Top f.l.t.r.: axial initial MR slice $\mathcal{M}$; same slice registered $\varphi_{15} \circ \cdots \circ \varphi_1 \circ \mathcal{M}$; corresponding CT slice $\mathcal{F}$. Yellow arrows indicate misaligned body borders in background in contrast to well aligned foreground structures. Bottom f.l.t.r.: $\mathcal{F}$ and $\mathcal{M}$ checkerboard images before/after registration with overlayed foreground labels

Discussion

In general, the resulting registrations are convincing in comparison with previous work and when considering the prevailing challenges of multimodal image alignment.

The CAE used to obtain shape embeddings, handles input data from different domains and still generates a compact and smooth shape encoding space. Thus, it enables us to generate realistic intermediate shapes between input CT and MR images for the iterative registration guidance task. However, omitting skip connections in its design results in an anticipated drop in segmentation accuracy with effects on the subsequent registration, because the CAE segmentation quality forms an upper bound regarding the expected registration alignment. Continuing experiments could search for ways to compensate the loss of spatial information when ommiting skip connections.

Our second experiment, confirms our hypothesis that a stepwise concatenation of small transformations achieves superior results compared to a direct estimation of possibly large deformations. Accuracy improvements with increasing numbers of intermediate steps clearly indicate that points along the interpolation path do not mislead the registration due to implausible shape transformations, thus the learned space itself is reasonably smooth. While outperforming the *SSC deeds* [5] as a baseline using $n = 15$ steps regarding Dice scores, only small deformations in the background occur. Although this is enforced by our regularizer, there is an obvious misalignment of body boundaries as shown in Fig. 5 when comparing the patients' chests and it will require further improvements. One potential solution would be to employ spatially more informative signed distance maps instead of only discrete label maps. Alternatively, increased supervision with the introduction of more classes or anatomical landmarks (cf. [8]) could improve the robustness of the shape embedding learned by the CAE.

To conclude our discussion, it is worth mentioning that the idea of iteratively guiding image registration by interpolating intermediate shapes not only enhances the results. Beyond that, it introduces an opportunity to indirectly measure the plausibility of the learned shape space—which is hardly possible so far—, by checking whether better results could be

achieved, when incorporating multi-step registrations. To this end, our method could be transferred to CAE-based shape modeling as a latent space evaluation tool.

Conclusion

In our work, we introduced an integrated approach for iteratively guided multimodal image registration in the context of medical volume data. Jointly learning shared features in a single, end-to-end trainable deep encoder–decoder model without skip connections results in good accuracies for multi-label CT and MRI whole-heart segmentations, while simultaneously restricting the underlying shape representation to be compact. The latter puts us in a position to interpolate between segmentation labels. This enables us to iteratively compute and concatenate small transformations—showing superior performance (65.27%, $n = 15$) when being evaluated on a challenging registration task compared to a single-step approach (52.62%, $n = 1$) as well as a well-known classical unsupervised multimodal registration tool (60.8%) that was designed specifically for multimodal MR-CT alignment. Our method can be trained without the need for strong supervision and requires no labels during inference. To conclude, our promising results encourage future work toward the exploration of intermediate shape features as registration guidance and the examination of multi-scale strategies.

Compliance with ethical standards

Conflict of interest The authors declare that they have no relevant conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Statement of informed consent was not applicable since the manuscript does not contain any participants' data.

References

1. Bouteldja N, Merhof D, Ehrhardt J, Heinrich MP (2019) Deep multi-modal encoder-decoder networks for shape constrained segmentation and joint representation learning. In: Bildverarbeitung für die Medizin 2019 in Lübeck, pp 23–28
2. de Vos BD, Berendsen FF, Viergever MA, Staring M, Išgum I (2017) End-to-end unsupervised deformable image registration with a convolutional neural network. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, Berlin, pp 204–212
3. 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
4. Hajnal JV, Hill DL (2001) Medical image registration. CRC Press, Boca Raton
5. Heinrich MP, Jenkinson M, Papież BW, Brady M, Schnabel JA (2013) Towards realtime multimodal fusion for image-guided interventions using self-similarities. In: International conference on medical image computing and computer-assisted intervention. Springer, Berlin, pp 187–194
6. Heinrich MP, Maier O, Handels H (2015) Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities. VISCERAL@ ISBI 2015 VISCERAL Anatomy3 Organ Segmentation Challenge
7. Heinrich MP, Jenkinson M, Bhushan M, Matin T, Gleeson FV, Brady M, Schnabel JA (2012) Mind: modality independent neighbourhood descriptor for multi-modal deformable registration. Med Image Anal 16(7):1423–1435
8. Hu Y, Modat M, Gibson E, Li W, Ghavami N, Bonmati E, Wang G, Bandula S, Moore CM, Emberton M, Ourselin S, Noble JA, Barrat DC, Vercauteren T (2018) Weakly-supervised convolutional neural networks for multimodal image registration. Med Image Anal 49:1–13
9. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K (2015) Spatial transformer networks. In: Advances in neural information processing systems, pp 2017–2025
10. Jetley S, Sapienza M, Golodetz S, Torr PHS (2016) Straight to shapes: real-time detection of encoded shapes. CoRR arXiv:1611.07932
11. Joyce T, Chartsias A, Tsaftaris SA (2018) Deep multi-class segmentation without ground-truth labels. In: Medical imaging with deep learning
12. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436
13. Maes F, Collignon A, Vandermeulen D, Marchal G, Suetens P (1997) Multimodality image registration by maximization of mutual information. IEEE Trans Med Imaging 16(2):187–198
14. Maintz JA, Viergever MA (1998) A survey of medical image registration. Med Image Anal 2(1):1–36
15. Rohé MM, 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, Berlin, pp 266–274
16. Shechtman E, Irani M (2007) Matching local self-similarities across images and videos. In: IEEE conference on computer vision and pattern recognition, 2007. CVPR'07. IEEE, pp 1–8
17. Sokooti H, de Vos B, Berendsen F, Lelieveldt BP, Išgum I, Staring M (2017) Nonrigid image registration using multi-scale 3d convolutional neural networks. In: International conference on medical image computing and computer-assisted intervention. Springer, Berlin, pp 232–239
18. Sotiras A, Davatzikos C, Paragios N (2013) Deformable medical image registration: a survey. IEEE Trans Med Imaging 32(7):1153
19. Yang X, Kwitt R, Styner M, Niethammer M (2017) Quicksilver: fast predictive image registration-a deep learning approach. NeuroImage 158:378–396
20. Zöllei L, Fisher JW, Wells WM (2003) A unified statistical and information theoretic framework for multi-modal image registration. In: Biennial international conference on information processing in medical imaging. Springer, Berlin, pp 366–377