Architecture matters: evaluating design choices for deep learning registration networks

Hanna Siebert[1,2], Lasse Hansen[1], Mattias P. Heinrich[1]

[1]Institute of Medical Informatics, University of Lübeck
[2]Graduate School for Computing in Medicine and Life Sciences, University of Lübeck
siebert@imi.uni-luebeck.de

Abstract. The variety of recently proposed deep learning models for deformable pairwise image registration leads to the question how beneficial certain architectural design considerations are for the registration performance. This paper aims to take a closer look at the impact of some basic network design choices, i.e. the number of feature channels, the number of convolutions per resolution level and the differences between partially independent processing streams for fixed and moving images and direct concatenation of input scans. Starting from a simple single-stream U-Net architecture, we investigate extensions and modifications and propose a model for 3D abdominal CT registration evaluated on data from the Learn2Reg challenge that outperforms the baseline network VoxelMorph used for comparison.

1 Introduction

Deformable image registration aims to align pairs of images or image volumes by predicting non-linear transformations that optimises an appearance or shape-based metric. Registration of medical images helps to analyse large image datasets for research purposes and plays an important role in clinical practice, including diagnostic tasks, image-guided interventions, and motion tracking [1]. Recent deep learning-based image registration methods [2,3,4,5,6] show the potential to outperform conventional methods in terms of improved registration speed and accuracy. However, the estimation of large deformations is still considered challenging. Meanwhile there is a large variety of publications presenting different deep learning networks for image registration offering multiple suggestions for the design of architectures consisting of different architectural modules.

We take up the idea of several registration networks which include an U-Net architecture to learn deformations [2,4]. The idea of not directly concatenating fixed and moving image before feature extraction is examined as well and has been used in [7] where fixed and moving images are analyzed in separate pipelines for affine registration or in the dual-stream registration network proposed in [8].

This paper aims to take a closer look at the impact of different architectural design ideas on the registration performance in order to finally propose an architecture for abdominal CT registration that combines the most convincing of the considerations examined. We compare our results to the simple baseline network for unsupervised pairwise image registration VoxelMorph [2].

2 Materials and Methods

We examine four different architectural designs for pairwise image registration, starting from a simple single-stream U-Net architecture, which is then extended and further modified. All considered architectures are visualised in Fig. 1.

Our investigations start with a registration model that is from its basic structure similar to VoxelMorph [2], but with one difference that it contains fewer skip connections. The model concatenates fixed and moving images at the beginning and uses sequences of convolution followed by instance normalisation and leaky ReLU. The resolution of the spatial dimensions is first successively decreased by strided convolutions to $\frac{1}{24}$ of the input image dimensions and then increased again by upconvolutions. The central part of the architecture consists of an U-Net like part using two skip connections [9]. Conversely to the reduction in resolution, the number of feature channels from the convolution layers is firstly

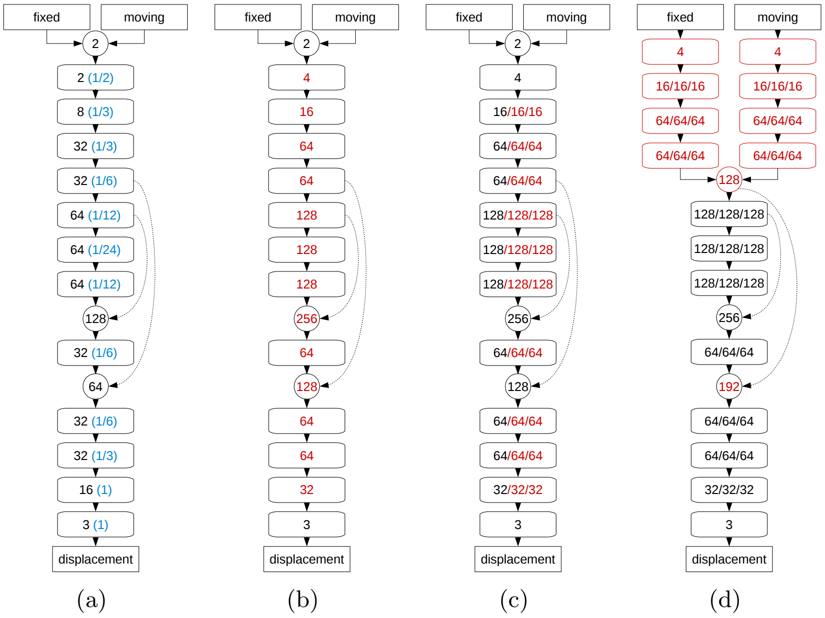

Fig. 1. Overview of the different considered architectures with given number of feature channels being output from the convolutions (rounded rectangles) and concatenations (circles). The initial architecture (a) is modified so that first the number of feature channels (b) and then the number of convolutions (c) is increased. The last modification (d) results in a two-stream architecture that starts with separate encoder blocks for fixed and moving image. Modifications to the previous architecture are marked in red respectively. The corresponding output resolutions are indicated in blue within the visualisation of model (a) and apply to all of the models.

increased up to 64 and then decreased (see Fig. 1 (a)) until the output yields 3 feature channels that correspond to the 3 displacement dimensions. As the displacements are considered to be within the the range of $[-1, 1]$, the last convolution layer is followed by a *tanh* activation function and the obtained output is used for warping with the moving input image.

The first modification we make to this initial architecture is to double the number of feature channels (quadrupling the parameters) of all convolution layers of the network (see Fig. 1 (b)). We then extend the number of Convolution-InstanceNorm-ReLU sequences per resolution level to three (Fig. 1 (c)). Finally, we propose a two-stream architecture with separate encoder blocks for fixed and moving image and their concatenated output as input for the U-Net part of the architecture (Fig. 1 (d)). Different from [8], which introduces a continuous dual-stream architecture, we concatenate the two streams within the encoder part of the network at a spatial resolution of $\frac{1}{6}$ of the input dimensions. As we use monomodal data for our experiments, the weights are shared between the two encoders of this two-stream architecture. In Fig. 2, our final image registration approach is illustrated.

We train our models using a loss function which ensures similarity of fixed and warped moving image and smooth deformation fields. Modality independent neighbourhood descriptors (MIND) with self-similar context (SSC) [10] are extracted from fixed and warped moving image and the mean squared error between them is calculated. Additionally, we apply diffusion regularisation to achieve smooth and plausible deformation fields. For our proposed two-stream

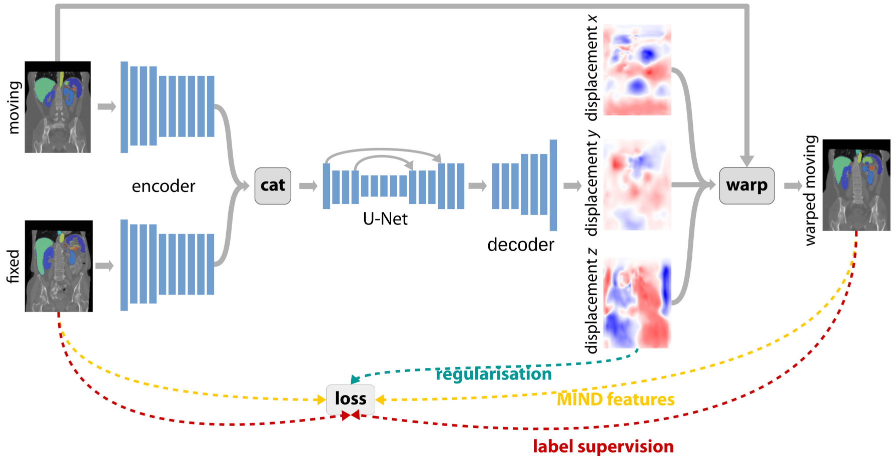

Fig. 2. Our model for pairwise image registration with label supervision: Fixed and moving images are given into separate encoder blocks for the extraction of features that are then concatenated and passed to an U-Net and following decoder block for the estimation of displacements. The obtained displacement fields are used to warp the moving image. The loss function is designed so that the warped moving image and labels resemble the fixed image and labels (similarity of MIND features and label supervision) and furthermore the deformation fields are smooth (diffusion regularisation).

architecture, we furthermore investigate the benefits of label supervision by further extending the loss function by computing the mean squared error between fixed an warped moving one-hot encoded label maps (background excluded) weighted inversely proportional to the square root of the class frequency.

For our experiments we use the Learn2Reg challenge [11] dataset containing 30 abdominal CT scans with thirteen manually labeled abdominal organs, including spleen 🟦, right kidney 🟦, left kidney 🟦, gall bladder 🟦, esophagus 🟦, liver 🟩, stomach 🟩, aorta 🟨, inferior vena cava 🟨, portal and splenic vein 🟧, pancreas 🟧, left adrenal gland 🟥, and right adrenal gland 🟫 [12]. The data has been linearly pre-registered and we re-sample the data to dimensions of $144 \times 112 \times 192$ to reduce computational complexity. The dataset is split into 20 training cases and 10 test cases. For evaluation we consider all possible pairwise combinations of the test cases (leading to 45 unique pairs). We train our networks using Adam and a learning rate of 0.001 (0.0001 for the baseline network of VoxelMorph) for $50,000$ iterations. Diffusion regularisation is weighted in such a way that the standard deviation of the Jacobian determinant stays below 1.0 on the training set and for label supervision we chose a weighting of $\lambda_{ls} = 2$.

3 Results

In Tab. 1, we report the average Dice overlap and properties of the Jacobian determinant as well as the inference time on GPU and the number of trainable parameters. Comparing the registration performance, the first model examined (1-stream (a), unsupervised) was only able to achieve a gain in Dice overlap of 2.5 % points (compared to the initially overlap of 25.15 %) and yields a worse performing network compared to the VoxelMorph with its higher number of skip connections and lower number of parameters. The model with an increased number of feature channels (1-stream (b), unsupervised) led to an improvement of about 2 % points compared to the first model. Increasing the number of convolution-normalisation-activation blocks per resolution level from one to three (1-stream (c), unsupervised) increased the Dice overlap by another 2 % points leading to a score similar to VoxelMorph, whereas the deformation field estimated by VoxelMorph is less smooth. With our unsupervised 2-stream model (d), we were able to achieve an average Dice overlap of 35.39 %, outperforming VoxelMorph by nearly 4% points. When training with label supervision, this score could be further improved to 43.85 %, which is competitive to many other approaches (cf. learn2reg.grand-challenge.org).

Tab. 2 shows Dice scores for the different considered label classes pointing out that the registration methods showed the best improvement of Dice overlap for comparably large and medium-sized organs. For these organs, also label supervision during training showed the highest improvement of registration performance compared to unsupervised training. These findings are also exemplary visually confirmed by Fig. 3.

Table 1. Evaluation results using Dice scores and Jacobian determinant, average inference time on GPU (Quadro P6000), and number of parameters. Dice scores are averaged over all thirteen label classes (background excluded). Initial refers to average values before registration. The quality of the deformation field is evaluated through the Jacobian determinant. Small standard deviations indicate smooth deformation fields and values below 0 indicate singularities, i.e. foldings.

	avg Dice [%]	std $det(J)$	$det(J)$ < 0 [%]	inf. time/ pair [ms]	param. count
initial	25.14 ± 12.85	-	-	-	-
VoxelMorph unsuperv.	31.70 ± 13.75	0.5853	3.61	117.58	396451
1-stream (a), unsuperv.	27.85 ± 12.56	0.4096	0.89	44.09	746467
1-stream (b), unsuperv.	29.78 ± 12.60	0.4600	1.10	50.17	2985251
1-stream (c), unsuperv.	31.72 ± 13.01	0.4184	0.96	75.09	6814499
2-stream (d), unsuperv.	**35.39 ± 14.05**	0.4681	1.38	102.32	7449123
2-stream (d), superv.	**43.85 ± 11.33**	0.5012	1.37	102.32	7449123

Table 2. Dice overlap in % for the 13 different label classes (see text for colour coding).

	🟦	🟦	🟦	🟦	🟩	🟩	🟩	🟨	🟨	🟧	🟧	🟥	🟫
initial	42	34	35	2	23	62	24	33	36	5	15	8	9
VM unsuperv.	53	45	45	3	28	72	29	47	44	7	17	12	10
1-stream (c), unsuperv.	54	41	44	4	29	73	31	44	45	8	16	13	10
2-stream (d), unsuperv.	60	49	50	4	28	78	35	50	46	11	19	17	12
2-stream (d), superv.	73	69	71	7	37	83	47	59	52	12	26	20	15

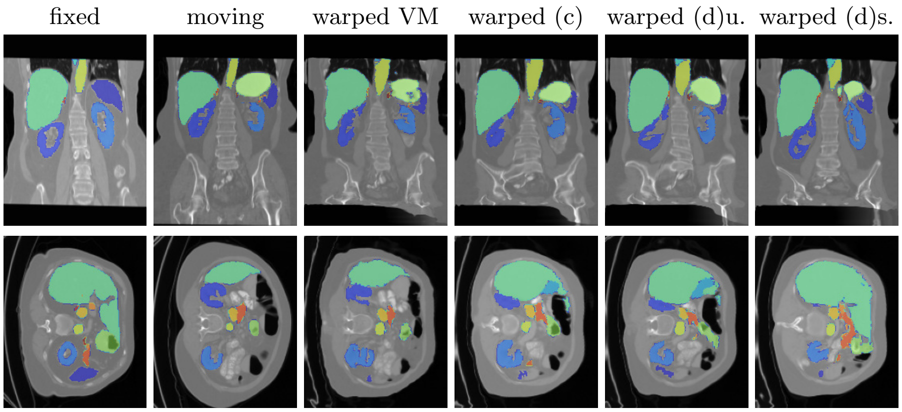

Fig. 3. Result visualization (coronal/axial) of different architectures (VoxelMorph (VM), 1-stream (c), unsupervised 2-stream (d)u., supervised 2-stream (d)s.) for one test pair: fixed and moving image and warped moving images output from the different models.

4 Discussion

We investigated several architectures for deep-learning based deformable image registration. Besides the expected observations that increased numbers of feature channels and convolution-normalisation-activation sequences led to improved registration results, we found out that concatenating the features extracted by separate encoder blocks for moving and fixed image achieved better results than directly concatenating the input images. With this two-stream architecture, we were able to outperform the simple baseline network for unsupervised pairwise image registration VoxelMorph. Due to the fact that we performed our experiments on a labeled dataset, we could further show that - starting from the initial untrained case - including label supervision when training our model led to a further substantial increase of Dice overlap of 8 % points compared to unsupervised training.

References

1. Hill DL, Batchelor PG, Holden M, et al. Medical image registration. Physics in medicine & biology. 2001;46(3):R1.
2. Balakrishnan G, Zhao A, Sabuncu MR, et al. Voxelmorph: a learning framework for deformable medical image registration. IEEE Trans Med Imag. 2019;38(8):1788–1800.
3. Mok TC, Chung AC. Large Deformation Diffeomorphic Image Registration with Laplacian Pyramid Networks. In: Proc. MICCAI. Springer; 2020. p. 211–221.
4. Eppenhof KAJ, Pluim JPW. Pulmonary CT Registration Through Supervised Learning With Convolutional Neural Networks. IEEE Trans Med Imag. 2019;38(5):1097–1105.
5. Heinrich MP. Closing the gap between deep and conventional image registration using probabilistic dense displacement networks. In: Proc. MICCAI. Springer; 2019. p. 50–58.
6. Eppenhof KAJ, Lafarge MW, Veta M, et al. Progressively Trained Convolutional Neural Networks for Deformable Image Registration. IEEE Trans Med Imag. 2020;39(5):1594–1604.
7. de Vos BD, Berendsen FF, Viergever MA, et al. A deep learning framework for unsupervised affine and deformable image registration. Med Image Anal. 2019;52:128–143.
8. Hu X, Kang M, Huang W, et al. Dual-Stream Pyramid Registration Network. In: Proc. MICCAI. Springer; 2019. p. 382–390.
9. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Proc. MICCAI. Springer; 2015. p. 234–241.
10. Heinrich MP, Jenkinson M, Papiez BW, et al. Towards realtime multimodal fusion for image-guided interventions using self-similarities. In: Proc. MICCAI. Springer; 2013. p. 187–194.
11. Hansen L, Hering A, Heinrich MP, et al.. Learn2Reg: 2020 MICCAI registration challenge; 2020. https://learn2reg.grand-challenge.org.
12. Xu Z, Lee CP, Heinrich MP, et al. Evaluation of six registration methods for the human abdomen on clinically acquired CT. IEEE Trans Biomed Eng. 2016;63(8):1563–1572.