# Deep Learning-Based Affine Medical Image Registration for Multimodal Minimal-Invasive Image-Guided Interventions - A Review and Comparative Study on Generalizability

Anika Strittmatter[a,b,*], Lothar R. Schad[a,b], Frank G. Zöllner[a,b]

[a] *Computer Assisted Clinical Medicine, Medical Faculty Mannheim, Heidelberg University, Theodor-Kutzer-Ufer 1-3, 68167 Mannheim, Germany*

[b] *Mannheim Institute for Intelligent Systems in Medicine, Medical Faculty Mannheim, Heidelberg University, Theodor-Kutzer-Ufer 1-3, 68167 Mannheim, Germany*

**Abstract**

Multimodal image registration is applied in medical image analysis as it allows the integration of complementary data from multiple imaging modalities. In recent years, various neural network-based approaches for medical image registration have been presented in papers, but due to the use of different datasets, a fair comparison is not possible. In this research 20 different neural networks for an affine registration of medical images were implemented. The networks' performance and the networks' generalizability to new datasets were evaluated using two multimodal datasets - a synthetic and a real patient dataset - of three-dimensional CT and MR images of the liver. The networks were compared using our own developed CNN as benchmark and a conventional affine registration with SimpleElastix as baseline. Five networks improved the pre-registration Dice coefficient of the synthetic dataset significantly ($p$-value $< 0.05$) and seven networks improved the pre-registration Dice coefficient of the patient dataset significantly and are therefore able to generalize to new datasets. Many different machine learning-based methods have been proposed for affine multimodal medical image registration, but few are generalizable to new data and applications. It is therefore necessary

*Corresponding author

*Email address:* `Anika.Strittmatter@medma.uni-heidelberg.de` (Anika Strittmatter)

to conduct further research in order to develop medical image registration techniques that can be applied more widely.

*Keywords:* Registration, Affine, Deep Learning, Neural Networks, Medical images, Multimodal data

## 1. Introduction

Patients with oligometastics disease have a poor survival probability when employing standard therapeutic options due to local lesion heterogeneity. However, using tailored techniques, such as minimally invasive image-guided therapies that incorporate tumor heterogeneity, greatly increases survival (Qiu et al., 2016; Ruers et al., 2017). Such interventions are commonly performed using computer tomography (CT) imaging while tumor heterogeneity can be derived using MRI. Also, visibility of certain lesions in CT is not always given but can be visualized in MRI.

The research campus "Mannheim Molecular Intervention Environment" ($M^2$OLIE) is working on optimizing the treatment of oligo metastatic liver cancer, especially image-guided procedures. For this purpose, three-dimensional computed tomography (CT) and magnetic resonance (MR) images are acquired pre- and postoperatively, while three-dimensional cone beam computed tomography (CBCT) is applied intraoperatively. To facilitate oncological interventions as well as to monitor the progress of treatment, the images acquired by the different modalities are merged into a multimodal dataset, thus combining all available morphological and functional information regarding the tumour (Waldkirch, 2020). Therefore, registered images of CT and MRI is warranted.

Typically, conventional methods are used for medical image registration. The transformation parameters are iteratively updated to optimise a similarity metric. However, conventional methods are very time-consuming, as the registration of each new image is performed iteratively again. Registration methods based on deep learning address this issue and are used to speed up the registration process. The network is trained once. To register new images, a forward pass through the network is performed, eliminating the need for time-consuming iterations (Balakrishnan et al., 2019).

In recent years, several supervised, semi-supervised or unsupervised neural networks for medical image registration with a wide variety of architectures have been published. These publications are also collected in reviews

(Chen et al., 2021a), (Fu et al., 2020). However, as different datasets were used to train and evaluate the networks, a fair comparison is difficult to achieve and to the best of our knowledge has yet not been performed. This paper addresses this point. The aim of this paper is to analyse and assess the performance of several neural networks for multimodal affine image registration and to evaluate the generalizability of these networks to new datasets. For this purpose, 20 different neural networks were implemented and compared using two multimodal datasets - a synthetic and a patient dataset - of three-dimensional CT and MR images of the liver. The paper is organised as follows: Section 2 outlines the existing image registration methods, section 3 describes the framework used to compare the network architectures and section 4 presents the experimental results.

## 2. Background

In image registration, a transformation is calculated to optimally align two or more images. In this process, the source (moving) image is registered to the reference (fixed) image. The calculated transformation is applied to the moving image, thus creating the moved image. An affine transformation consists of the following individual transformations: translation, rotation, scaling and shearing. The affine transformation is represented for three-dimensional images in a 4x4 matrix with 12 parameters (see equation 1). A combination of the single transformations is achieved by matrix multiplication (Chumchob and Chen, 2009). The right column (parameters $a_{14}, a_{24}, a_{34}$) of the transformation matrix contains the translation parameters. The remaining nine parameters represent the combined values for rotation, scaling and shearing (Waldkirch, 2020).

$$T_{affine} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

Conventional registration methods apply a mathematical optimization model that is solved individually for each image pair. In an iterative process, a function is maximized that describes the similarity of fixed and moved images. This is computationally intensive and slow due to the iterative process. Thus, affine registration with state of the art algorithms such as SimpleElastix (Marstal et al., 2016) or NiftyReg (Modat et al., 2014) takes several seconds to a minute on the CPU. Some of the conventional algorithms also have GPU implementations that shorten the duration of a single registration (a few seconds), but then a GPU is needed for each registration process (Fluck et al., 2011).

When using neural networks, on the other hand, the time-consuming optimization for each new image pair is omitted. A neural network takes two n-dimensional images as input and outputs the transformation parameters. The weights of the neural network are optimized by training. Thus a global function is optimized once, the optimization for each individual test image pair is omitted. For registration, the learned function is evaluated, which saves a lot of time on the CPU as well as on the GPU compared to conventional methods. The networks are trained supervised, unsupervised or semi-supervised (weakly-supervised). Supervised methods require ground truth, which usually does not exist, especially in medical image registration. Instead, images are used as ground truth, which were registered manually or by conventional methods. This has the drawback that the trained networks are limited by the accuracy of the used ground truth. The networks can therefore only deliver results with a maximum accuracy that corresponds to the accuracy of the ground truth. Synthetically generated transformations can also be used. However, it is disputable to what extent these transformations correspond to real transformations.

Unsupervised learning, on the other hand, does not require ground truth. During training, the moved image, i.e. the moving image transformed by the computed transformation, is compared with the fixed image. If anatomical segmentations are available for the training data set, the training can be performed semi-supervised (weakly-supervised). In addition to the fixed and moving image, segmentations of the fixed and moving image are also passed to the neural network during training. During the evaluation of the networks, i.e. for the registration of new image pairs, no segmentations are needed. In addition to the image-based losses, segmentation driven loss functions, such as Dice loss, are used, which can improve the registration result compared to unsupervised learning (Balakrishnan et al., 2019).

## 3. Methods

We implemented 20 neural networks for affine registration of medical images, which we found on IEEE Xplore, PubMed, Web of Science and ScienceDirect with the keywords affine, registration, deep learning, artificial neural network, medical images. Generative Adversarial Networks (GAN) are not taken into account, since an additional loss (adversarial loss) is applied during training by using a second network (discriminator). Generative Adversarial Networks can therefore not be fairly compared with the other networks, which were trained with only one loss (Dice loss).

### *3.1. Network Architectures*

We implemented and compared the architectures for affine registration of 17 papers, one abstract and 2 theses. The architectures can be broadly grouped into different categories, as shown in Figure 1. Seventeen architectures can be categorised as Convolutional Neural Networks (CNN). The CNNs can be further subdivided into regression networks (16 papers), for example the encoder part of the U-Net followed by dense layers - as well as U-Net-like networks (1 thesis), where the affine parameters are formed at the end of the network by a Global Average Pooling layer. Regression networks can be further subdivided into convolution-based (12 papers) and Siamese network-based (3 papers and 1 abstract). Two papers use transformer, the Swin transformer (Chen et al., 2022) and an adaptation of the Vision transformer (Mok and Chung, 2022). One network architecture (Hasenstab et al., 2019) is assigned to the category "other network architectures", a neural network for affine registration is used which contains only one single dense layer with 12 neurons.

In some papers, parts of the networks' implementation were not disclosed in detail, especially parameter settings were not reported making the reproducibility difficult. For example, in (Zhao et al., 2020) the activation function was not specified and in (de Silva et al., 2020) the kernel size of the convolutions was missing. Assumptions were made for the parameters, which were not specified in the paper or in published source code. Settings were chosen that are commonly used in affine registration network architectures, for example, the activation function LeakyReLU and a kernel size of 3x3x3. In four papers two-dimensional datasets were used. This paper evaluates the networks using three-dimensional data. For this reason, all networks have been adapted to three dimensions. Only the paper (Zeng et al., 2020) uses a two-dimensional affine network to register three-dimensional data. For comparison with the architectures of the other papers, the two-dimensional original network and the three-dimensional adaptation were used.

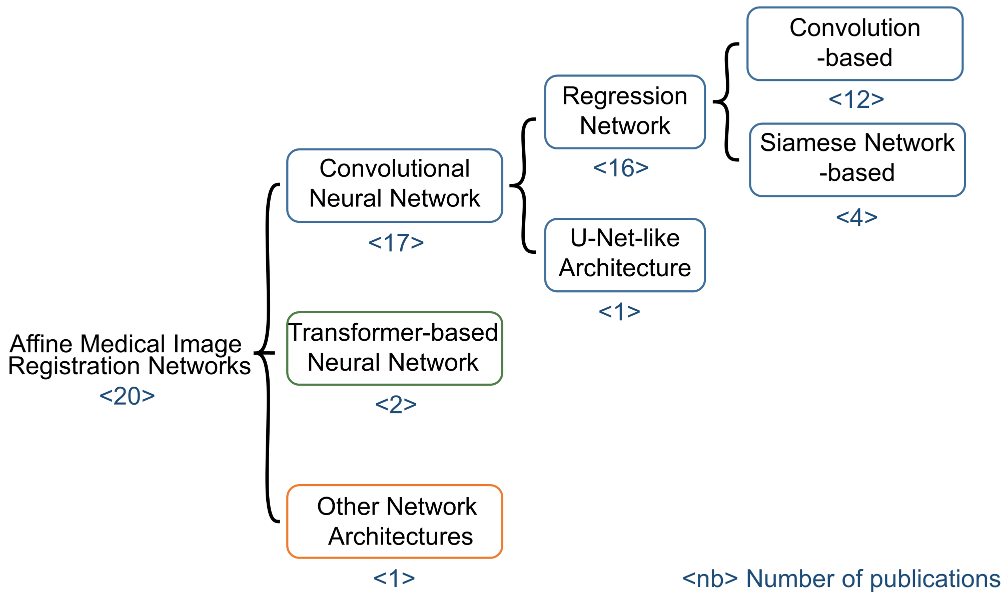


Figure 1: Categories of the implemented architectures

### *3.2. Benchmark Neural Network*

As benchmark, we used our own developed CNN (Figure 2), a ResNet-18 encoder and a U-Net-like decoder with one convolution per convolutional block without skip connections. The decoder consists of six three-dimensional convolutions with a kernel size of 3x3x3, a stride of 1 and padding same. The first five convolutions and activation functions (LeakyReLU) of the decoder are followed by upsampling. Global Average Pooling is then applied to generate the 12 parameters needed for a three-dimensional affine transformation. A spatial transformer then calculates the moved image from the moving image and the affine transformation matrix (Jaderberg et al., 2015).

### *3.3. Baseline method*

As baseline, i.e. as state-of-the-art, we applied the Elastix affine registration algorithm implemented in the SimpleElastix Toolbox using default parameters for an affine transformation. Mutual Information (MI) is used as objective function.

### *3.4. Training Setting*

The training of the implemented neural networks is performed with semi-supervised learning according to the VoxelMorph framework (Balakrishnan

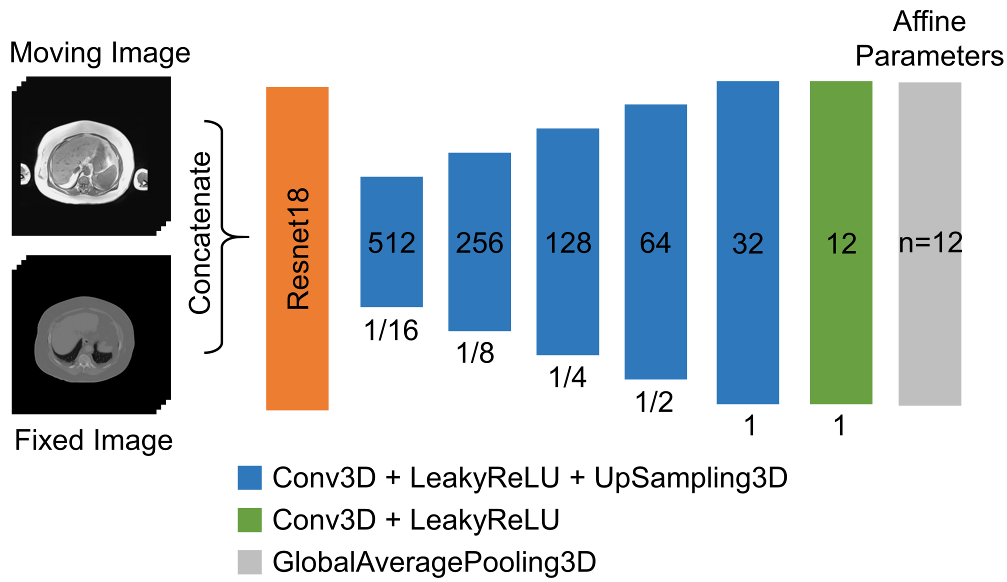


Figure 2: Architecture of the benchmark CNN for affine registration

et al., 2019). During training, in addition to the fixed and moving images, the segmentations of the fixed and moving images are given as input. We trained the networks using Dice loss. The overlap of the segmentation of the moved image and the segmentation of the fixed image is calculated. We performed five-fold cross-validation and trained the networks for 100 epochs with a batch size of one, stopping after five epochs with no improvement in validation loss. The network's weights with the highest accuracy on the validation data are saved. The Adam optimiser was used with a learning rate of 1e-4. We resized the input of the network, i.e. the fixed image, the moving image and the segmentation masks to 256x256x64 voxels. The images were used in half resolution due to memory limitations of the GPUs. For evaluation, the resulting affine transformation was applied to the full resolution images. During preprocessing, the image intensity was normalised to the range [0,1]. No augmentation was applied. Finetuning of the networks was done with the same setup as the training.

All experiments were performed in Python using Tensorflow (Abadi et al., 2016) on a workstation equipped with a NVIDIA RTX A4000 with 16 GB GPU memory, a NVIDIA RTX A5000 with 24 GB GPU memory and a NVIDIA RTX A6000 with 48 GB GPU memory.

## 4. Results

### *4.1. Datasets*

To investigate the performance of the neural networks and the networks' generalizability to new datasets we used two datasets.

The networks were trained and evaluated with a synthetic dataset of 112 CT and MR T1-weighted three-dimensional images of the liver. These volumes were generated from images of the XCAT (Segars et al., 2010) phantom using CycleGAN (Bauer et al., 2021). Each CT volume is reconstructed to a 512x512 matrix with 82-124 slices and a resolution of 1x1x2mm$^3$. The MR volumes were created with a matrix of 330x450 and 54-83 slices to match a resolution of 1x1x3mm$^3$. This dataset contains anatomical segmentation masks for the liver. The CT and MR T1-weighted images are available in two breathing states: inhaled and exhaled. In the scope of this paper, the MR images are registered affinely to the CT images, the inhaled state to the exhaled state and vice versa.

The second dataset consists of CT and MR T1-weighted patient images of the liver. These were acquired on 39 patients as part of the M$^2$OLIE project. This dataset contains 47 CT images and 73 MR T1-weighted images (in-phase and opposed-phase) in transversal orientation. The in-plane resolution of the CT scans is 0.78x0.78 ($\pm$ 0.12) mm$^2$ and the slice thickness is 2.04 ($\pm$ 0.76) mm. The CT volumes were made with a 512x512 matrix and 25-132 slices. The MR scans have an in-plane resolution of 1.27 x 1.27 ($\pm$ 0.17) mm$^2$ and slice thickness of 3.77 ($\pm$ 0.46) mm. The matrices used for the MR scans range from 190-440x288-640 voxels with 52-120 slices. Anatomical segmentation masks of the liver were created by a neural network. A three-dimensional U-Net was used. To perform liver segmentation on CT data, the network was trained with the 3D-IRCADb-01 dataset (Soler et al., 2010), for liver segmentation on MR data, the network was trained with the MR images of the Combined Healthy Abdominal Organ Segmentation (CHAOS) dataset (Kavur et al., 2019). Afterwards, the networks were finetuned with 20 segmentations of the M$^2$OLIE dataset, which were created manually in the Medical Imaging Interaction Toolkit (MITK) (German Cancer Research Center (DKFZ) Division of Medical Image Computing). In the scope of this paper, the MR images are registered affinely to the CT images.

For both datasets, we used trilinear interpolation to resample all volumes to an identical voxel spacing of 1x1x2 mm$^3$. The volumes were then cropped and zero-padded to a size of 512x512x128 voxels.

### *4.2. Evaluation Metrics*

We evaluated the registration's performance using three metrics. Normalised Mutual Information (NMI) was applied to quantify structural differences of the fixed image (F) and the moved image (M$\circ\phi$, moving image warped by the deformation field). The NMI is defined as:

$$NMI(F, M \circ \phi) = \frac{2 \cdot I(F; M \circ \phi)}{H(F) + H(M \circ \phi)} \tag{2}$$

with

$$I(F; M \circ \phi) = H(F) + H(M \circ \phi) - H(F, M \circ \phi) \tag{3}$$

where I is the Mutual Information and H is the entropy.

In addition, we measured the overlap of the segmentations of the moved image ($S_M \circ \phi$) and the fixed image ($S_F$) using the Dice coefficient and the Hausdorff distance (HD).

The Dice coefficent is given by:

$$DICE(S_F, S_M \circ \phi) = 2 \cdot \frac{|S_F \cap (S_M \circ \phi)|}{|S_F| + |S_M \circ \phi|} \tag{4}$$

The Hausdorff distance is defined as:

$$H(S_F, S_M \circ \phi) = max(h(S_F, S_M \circ \phi), h(S_M \circ \phi, S_F)) \tag{5}$$

with

$$h(S_F, S_M \circ \phi) = \underset{s_F \epsilon S_F}{max} \underset{(s_M \circ \phi) \epsilon (S_M \circ \phi)}{min} ||s_F - (s_M \circ \phi)|| \tag{6}$$

We compared the metrics' mean and standard deviation (SD) across the neural networks described in the papers, using our CNN as a benchmark and the conventional affine registration method as a baseline (Table 1).

In addition, we used a paired t-test (if normally distributed) or a Wilcoxon signed-rank test (if not normally distributed) to examine the significance of the improvement in NMI, Dice coefficient, and Hausdorff distance by registration compared with the unregistered data. The null hypothesis states that registration by the different implemented methods does not improve the three specified evaluation metrics. This hypothesis is rejected if $p < 0.05$.

Table 1: The Normalized Mutual Information (NMI), Dice coefficient (%) and Hausdorff distance (HD, in mm) results (mean ± SD) on synthetic Dataset and real patient Dataset for the implemented networks, the benchmark (our CNN) and the baseline (SimpleElastix affine). NMI: higher value is better (maximum is 1), Dice coefficient: higher value is better (maximum is 100%), Hausdorff distance: lower value is better (minimum is 0). Marked in green: improvement compared to unregistered data, with the bold numbers indicating significant improvement ($p$-value $< 0.05$) to the unregistered data

| Method | Synthetic Dataset | | | Patient Dataset evaluated | | | Patient Dataset finetuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | NMI | DICE (%) | HD (mm) | NMI | DICE (%) | HD (mm) | NMI | DICE (%) | HD (mm) |
| Before registration | 0.246 ± 0.024 | 66 ± 4 | 23 ± 3 | 0.128 ± 0.042 | 50 ± 18 | 50 ± 27 | 0.128 ± 0.042 | 50 ± 18 | 50 ± 27 |
| SimpleElastix affine | **0.271 ± 0.021** | **73 ± 5** | **18 ± 3** | **0.225 ± 0.067** | **60 ± 19** | **36 ± 18** | **0.225 ± 0.067** | **60 ± 19** | **36 ± 18** |
| Benchmark (Our CNN) | 0.218 ± 0.025 | **82 ± 7** | **19 ± 7** | 0.122 ± 0.042 | 50 ± 19 | **49 ± 27** | 0.132 ± 0.041 | **57 ± 18** | **48 ± 28** |
| Hu et al., 2018 | 0.153 ± 0.042 | 55 ± 10 | 56 ± 26 | 0.104 ± 0.037 | 37 ± 18 | 82 ± 42 | 0.077 ± 0.035 | 34 ± 13 | 112 ± 56 |
| Gu et al., 2019 | 0.227 ± 0.021 | **81 ± 4** | **16 ± 5** | 0.125 ± 0.043 | **52 ± 19** | **48 ± 27** | **0.144 ± 0.041** | **63 ± 18** | **45 ± 29** |
| Shen et al., 2019 | 0.210 ± 0.021 | 66 ± 8 | 26 ± 3 | 0.122 ± 0.043 | 47 ± 19 | 51 ± 27 | 0.121 ± 0.042 | 47 ± 19 | 51 ± 27 |
| Zhao et al., 2020 | 0.229 ± 0.020 | **81 ± 4** | **15 ± 4** | 0.121 ± 0.041 | 51 ± 18 | 51 ± 28 | **0.139 ± 0.042** | **62 ± 18** | **47 ± 29** |
| Luo et al., 2020 | 0.004 ± 0.002 | 0 ± 1 | 132 ± 17 | 0.004 ± 0.005 | 1 ± 3 | 139 ± 22 | 0.002 ± 0.005 | 1 ± 3 | 141 ± 20 |
| Tang et al., 2020 | 0.109 ± 0.031 | 44 ± 14 | 83 ± 45 | 0.112 ± 0.060 | 38 ± 17 | 83 ± 40 | 0.095 ± 0.042 | 25 ± 20 | 82 ± 35 |
| Zeng et al., 2020 2D | 0.013 ± 0.015 | 0 ± 1 | 134 ± 29 | 0.008 ± 0.008 | 1 ± 3 | 138 ± 27 | 0.018 ± 0.015 | 2 ± 4 | 136 ± 25 |
| Zeng et al., 2020 3D | 0.011 ± 0.009 | 1 ± 1 | 148 ± 47 | 0.011 ± 0.010 | 3 ± 4 | 126 ± 25 | 0.018 ± 0.017 | 2 ± 4 | 149 ± 44 |
| Chen et al., 2021b | 0.226 ± 0.036 | 63 ± 9 | 29 ± 11 | 0.127 ± 0.050 | 45 ± 21 | 55 ± 30 | **0.132 ± 0.051** | 45 ± 21 | 55 ± 31 |
| Gao et al., 2021 | 0.157 ± 0.024 | 54 ± 11 | 42 ± 9 | 0.116 ± 0.042 | 40 ± 19 | 57 ± 26 | 0.125 ± 0.047 | 45 ± 19 | 54 ± 27 |
| Roelofs, 2021 | 0.230 ± 0.023 | 68 ± 7 | 28 ± 4 | 0.127 ± 0.041 | 50 ± 18 | 50 ± 28 | **0.137 ± 0.050** | **51 ± 18** | 50 ± 28 |
| Shao et al., 2021 | 0.020 ± 0.016 | 3 ± 7 | 124 ± 19 | 0.016 ± 0.016 | 5 ± 11 | 120 ± 28 | 0.005 ± 0.007 | 1 ± 3 | 142 ± 30 |
| Zhu et al., 2021 | 0.210 ± 0.025 | 64 ± 5 | 29 ± 4 | 0.125 ± 0.043 | 46 ± 14 | 66 ± 28 | **0.132 ± 0.044** | 50 ± 18 | 53 ± 27 |
| Chee and Wu, 2018 | 0.174 ± 0.050 | 51 ± 13 | 57 ± 23 | 0.110 ± 0.038 | 41 ± 16 | 67 ± 30 | 0.125 ± 0.041 | 45 ± 17 | 56 ± 26 |
| de Vos et al., 2019 | 0.220 ± 0.019 | **78 ± 4** | 23 ± 6 | 0.126 ± 0.042 | 51 ± 18 | 51 ± 27 | **0.140 ± 0.040** | **60 ± 17** | **48 ± 30** |
| de Silva et al., 2020 | 0.244 ± 0.023 | 66 ± 4 | 24 ± 3 | 0.126 ± 0.042 | 50 ± 18 | 50 ± 27 | **0.134 ± 0.044** | **52 ± 17** | **49 ± 28** |
| Venkata et al., 2022 | 0.193 ± 0.031 | 51 ± 14 | 51 ± 16 | 0.114 ± 0.043 | 42 ± 16 | 62 ± 27 | 0.123 ± 0.042 | 45 ± 16 | 57 ± 27 |
| Waldkirch, 2020 | 0.226 ± 0.021 | **81 ± 4** | **18 ± 6** | 0.123 ± 0.041 | **51 ± 19** | 49 ± 28 | **0.141 ± 0.040** | **57 ± 18** | **48 ± 29** |
| Chen et al., 2022 | 0.224 ± 0.022 | **74 ± 11** | **18 ± 7** | 0.123 ± 0.042 | 47 ± 17 | 53 ± 28 | 0.128 ± 0.043 | **52 ± 18** | **48 ± 27** |
| Mok and Chung, 2022 | 0.241 ± 0.023 | 67 ± 5 | 23 ± 3 | 0.128 ± 0.042 | 50 ± 18 | 50 ± 27 | 0.128 ± 0.042 | 50 ± 18 | **50 ± 27** |
| Hasenstab et al., 2019 | 0.000 ± 0.000 | 0 ± 0 | 361 ± 44 | 0.001 ± 0.005 | 1 ± 3 | 143 ± 25 | 0.001 ± 0.005 | 1 ± 3 | 143 ± 25 |

### 4.3. Synthetic Dataset

We trained and evaluated the networks with the synthetic dataset (column "Synthetic Dataset" in Table 1).

By using the conventional registration method SimpleElastix, the NMI, Dice coefficient as well as the Hausdorff distance could be improved compared to the unregistered images.

In contrast, no implemented neural network could increase the NMI of the unregistered images (pre-registration NMI). Five out of 20 networks implemented from papers improved the Dice coefficient significantly ($p$-value $< 0.05$) and four networks were able to reduce the Hausdorff distance significantly. Four networks from papers improved the two segmentation-based metrics Dice coefficient and Hausdorff distance significantly simultaneously.

The best results in terms of NMI were obtained with the neural network of de Silva et al. (2020), but registration with this network did not increase the NMI compared to the unregistered data. The network also did not succeed in improving the segmentation-based metrics Dice coefficient and Hausdorff distance.

The neural network of Gu et al. (2019) achieved the highest Dice coefficient. Our benchmark network yielded a similar Dice coefficient. The networks of Zhao et al. (2020) and Waldkirch (2020) also achieved good Dice coefficient values, which were above 80%.

The worst results for all three metrics were obtained by the neural network of Hasenstab et al. (2019). The network failed to register the data at all. The image intensities of the moved image and its segmentation are set completely to the background value.

Examples of registration results can be seen in Figure 3. Before registration, the abdomen in the CT and MR images is not well aligned. After registration with the neural network of Tang et al. (2020) (a poorly performing network), the structural similarity and overlap of the anatomical segmentations has worsened (Figure 3, second row). An inaccurate affine transformation matrix was predicted by the neural network, so the moved image was rotated incorrectly.

An example of a good registration is shown in Figure 3 in the third row. This is the registration result of the best performing network from papers in terms of the Dice coefficient (Gu et al., 2019). The abdomen of the moved image and the fixed image align better than before registration, and the overlap of the liver segmentations also improved. The registration result from our own developed CNN (Figure 3 last row) further increased the overlap of the anatomical segmentations compared to the best performing network from the papers. However, it achieved a lower structural similarity than the network of Gu et al. (2019). This can be seen, for example, in the outline of the abdomen (Figure 3, fourth column). The abdomens align better in the registration result of the network of Gu et al. (2019). The moved image of our own developed CNN is shifted a little too far down.

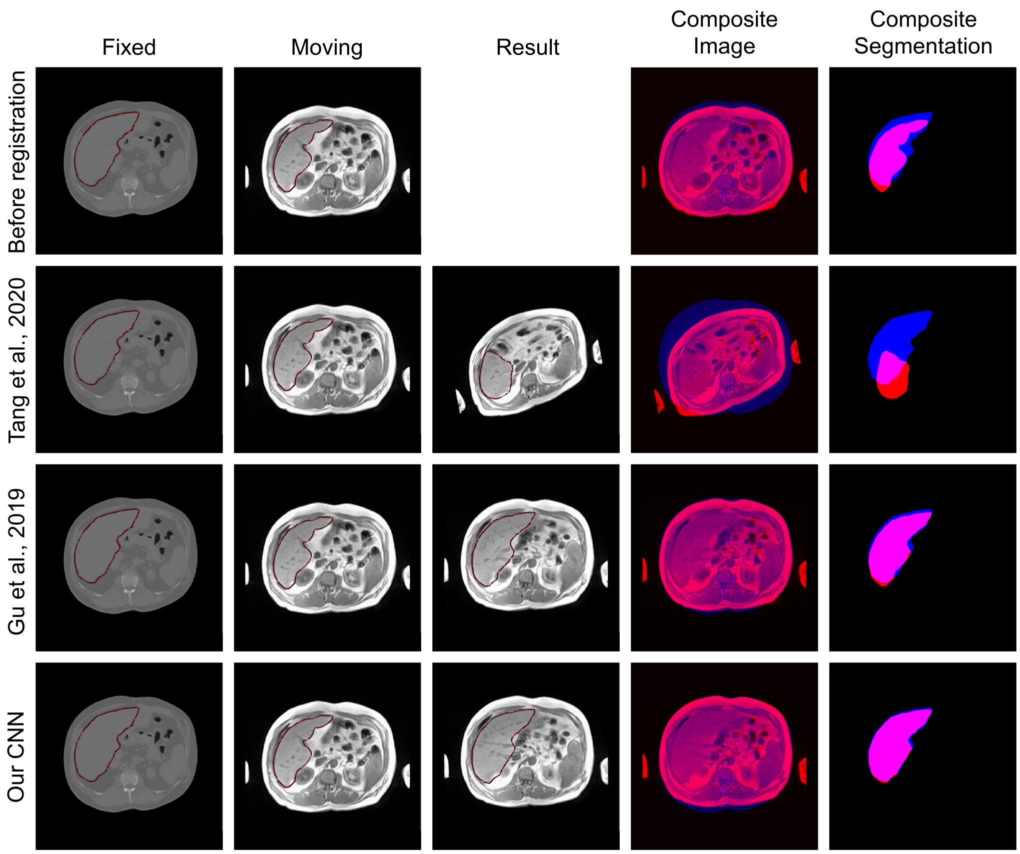


Figure 3: Example results of an affine registration of two volumes from the synthetic dataset. The images show central slices of the axial plane. The fixed images, moving images and result (moved) images are overlaid by the liver segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red. The second row illustrates an example of a poor registration result. Row three and four represent good registration results.

### *4.4. Patient Dataset*

We evaluated the networks trained on the synthetic dataset with the patient dataset (column "Patient Dataset evaluated" in Table 1). No implemented network was able to increase the pre-registration NMI. Only two networks could improve the pre-registration Dice coefficient significantly ($p$-value $< 0.05$) and one network was able to reduce the Hausdorff distance significantly. Comparable to the synthetic dataset, the network of Gu et al. (2019) yielded the highest Dice coefficient for the patient dataset.

We then finetuned the networks trained on the synthetic dataset with the patient dataset (column "Patient Dataset finetuned" in Table 1). Seven finetuned networks could increase the pre-registration Dice coefficient of the patient dataset significantly ($p$-value $< 0.05$) and seven networks improved the pre-registration Hausdorff distance significantly. Once again, the network of Gu et al. (2019) yielded the highest Dice coefficient. It even achieved a higher Dice coefficient than our own developed network.

Examples of registration results of the patient dataset are shown in Figure 4. Before registration (first row in Figure 4), the abdomen is not well aligned in the CT and MR images. This can be observed distinctly in the composite image and composite segmentation (second to last and last column). After registration with the neural network of Tang et al. (2020) (a poor performing network), the structural similarity and the overlap of the anatomical segmentations decreased (Figure 3, second row). As already observed for the network of Tang et al. (2020) trained with the synthetic dataset, the network of Tang et al. (2020) finetuned with the patient dataset computed a wrong affine transformation, so that the moved image was rotated incorrectly. The NMI, the Dice coefficient and the Hausdorff distance deteriorated compared to the unregistered images (see Table 1, row "Tang et al., 2020").

The third row in Figure 4 shows the registration result of the network with the best performance in terms of the Dice coefficient (Gu et al., 2019). The overlap of the liver segmentations improved compared to the unregistered data. The registration result of our CNN (Figure 4 last row) is worse than the registration result of the network of Gu et al. (2019). Our own developed CNN could only slightly improve the the structural similarity and the overlap of the segmentation of fixed and moved image compared to the unregistered images.

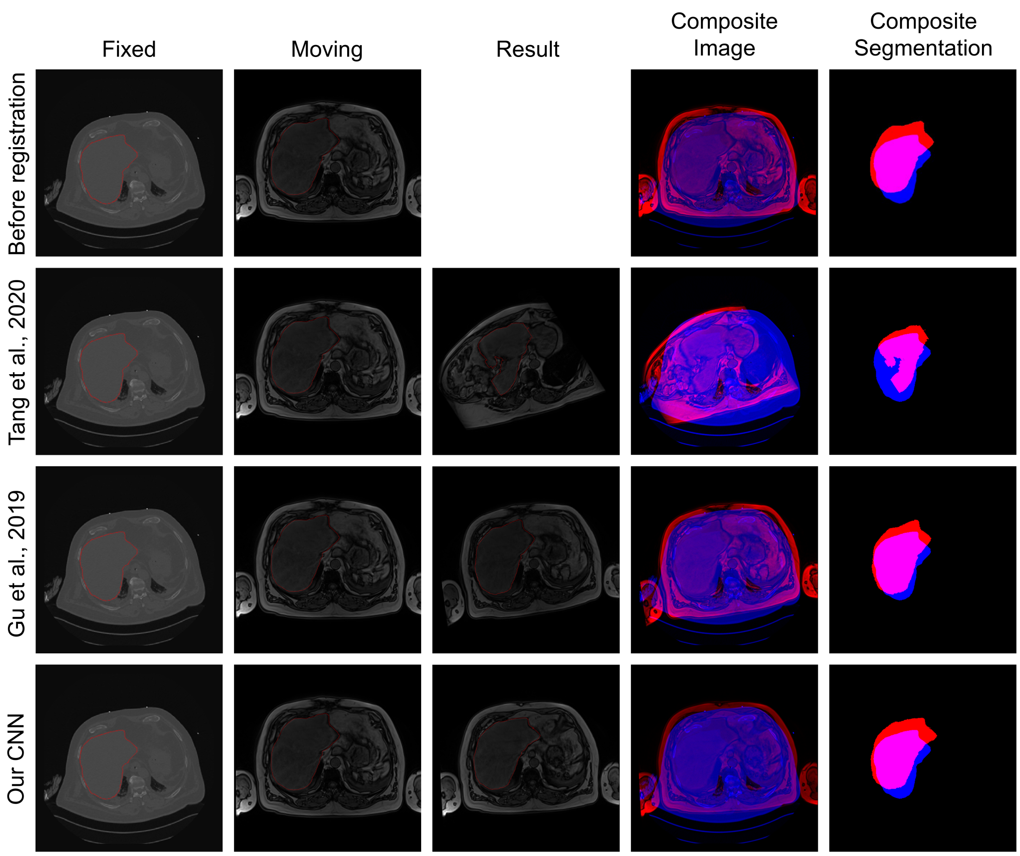


Figure 4: Example results of an affine registration of two volumes from the patient dataset. The images show central slices of the axial plane. The fixed images, moving images and result (moved) images are overlaid by the liver segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red. The second row illustrates an example of a poor registration result. Row three and four represent good registration results.

## 5. Discussion

We implemented 20 different machine learning-based image registration methods and evaluated them with respect to an affine multimodal registration of CT and MR images of the liver. Furthermore, we implemented a framework to test and evaluate these approaches systematically. Our results show that although each network was trained on the synthetic data and patient data only few could generalize, i.e. adapt to the new data and produce good registration results compared to a conventional affine registration algorithm and an inhouse developed deep learning approach.

To allow a fair comparison, we tried to reproduce the algorithms as best as possible, however, in some papers, not all parameters for the neural networks' implementation were described. This is detrimental to the reproducibility and assumptions were made, that may limit possible better registration results.

The use of augmentation, as well as testing different normalisation techniques, learning rates and loss functions could improve the registration results. Furthermore, the batch size influences the training and affects the accuracy of the network. In the experiments in this paper, a batch size of one was used due to the memory limitation of the GPUs.

### *5.1. Synthetic Dataset versus Patient Dataset*

Five networks trained with the synthetic dataset could improve the pre-registration Dice coefficient of the synthetic dataset significantly ($p$-value $< 0.05$). When evaluated with the patient dataset, only two networks could improve the pre-registration Dice coefficient significantly. This is due to the fact that the networks have not been finetuned with the patient data. The synthetic data simulates breathing, but does not consider the fact that in real patient data the patient was positioned differently in the CT and MR scanner. In the patient dataset, there is not only a deformation of the organs due to respiration, but also due to a different positioning of the patient in the two scanners. Finetuning of the networks with the real patient dataset was necessary here in order to also take the positioning-related deformation into account. Seven finetuned networks then improved the pre-registration Dice coefficient of the patient dataset significantly ($p$-value $< 0.05$).

Compared with the registration results of the synthetic dataset, the registration results of the patient dataset had a higher standard deviation (Table 1). This is due to the fact that in the synthetic dataset the respiration - i.e.

the deformation - was modeled similarly in all models. There are no outliers here. However, in the patient dataset, the respiration and the additional deformation due to the different positioning of the patient in the CT and MR scanner was not similar for the different patients. For example, there were seven cases in which the patient was not lying on his back, but rather in a more sideways position in the CT scanner, whereas he was lying on his back in the MR scanner. For these CT-MR image pairs, the registration result was thus worse than for image pairs in which the patient was positioned similarly in the CT and MR scanner.

### *5.2. Individual Networks*

For the synthetic dataset, our own developed CNN yielded similar results as the best performing network from papers in terms of the Dice coefficient (Gu et al., 2019). For the patient dataset our own developed CNN - finetuned on the patient dataset - improved the pre-registration Dice coefficient significantly, but the network did not give as good results as the best performing network from papers in terms of the Dice coefficient (Gu et al., 2019). We developed our own CNN specifically for the synthetic dataset. Hence, this network architecture delivered good results for the synthetic dataset. If the network was then trained with other datasets (e.g. the patient dataset), the accuracy of the result decreased depending on the generalizability of the network.

In (Hu et al., 2018) DownBlocks consisting of convolutional layers and residual blocks are applied. The network achieved good registration results when applied to the dataset used in (Hu et al., 2018) (T2-weighted MR volumes and transrectal ultrasound images of prostate cancer patients). However, the results presented in this paper (as seen in Table 1) indicate that the network developed by Hu et al. (2018) lacks generalizability when applied to new datasets. Therefore, further adjustments to the network architecture, such as the number of DownBlocks, are necessary in order to achieve good registration results on new datasets as well.

The network of Gu et al. (2019) uses a U-Net-like encoder part and and a fully-connected layer to create the 12 parameters for the affine transformation. This is a structure that occurs in several other papers in a similar form ((Zhao et al., 2020), (Zeng et al., 2020), (Gao et al., 2021)), as it achieves a high accuracy. The network of Gu et al. (2019) yielded the best registration results, i.e. the best Dice coefficient, in all 3 cases: 1) network trained and evaluated on synthetic dataset; 2) network trained on synthetic dataset and evaluated with patient dataset; 3) network trained on synthetic dataset and finetuned and evaluated with patient dataset. This shows that the network has a high generalizability to new datasets. A high degree of generalizability was already ensured by Gu et al. (2019) during the development of the network by testing and evaluating the developed framework using four publicly available datasets (Gu et al., 2019).

An affine network structure with multiple steps is applied in (Shen et al., 2019). Comparable to a recurrent network, the parameters of the affine network from all steps are identical. The number of steps of the network influences the performance of the network and could lead to poor generalizability of the entire network. In addition to modifying the individual layers in the network, the number of steps could also be adjusted in order to improve the generalizability of the network and its performance on new datasets.

Zhao et al. (2020) used a network architecture similar to the one used by Gu et al. (2019), but with a different number of filters for each convolutional layer and a different number of convolutional layers per convolutional block. This parameter variation can change the representational power of the network, enabling it to learn more complex features. On the other hand, this can also increase the risk of overfitting. Zhao et al. (2020) and Gu et al. (2019) found the right balance between these factors and their networks thus achieved good registration results even for new datasets.

In (Luo et al., 2020) a coarse and a fine registration stage are used. The coarse scale calculates anatomical segmentations and a coarse transformation (translation and scaling), which is used to reduce large global image displacements. The fine scale corrects smaller affine image displacements. However, no implementation details of the U-Net were given in (Luo et al., 2020). Only the fine stage was described properly and could be implemented. If the coarse stage is added, the registration result for the synthetic dataset and the patient dataset would probably be improved.

Tang et al. (2020) used dropout layers as final layers of their neural network. Dropout layers are applied to prevent overfitting, which is particularly important for small datasets such as our synthetic dataset and patient dataset. On the other hand, the use of dropout layers or a too high dropout rate can cause the network to lose information. That occurred while training the network of Tang et al. (2020) with the synthetic dataset and the patient dataset and has led to a poor registration result.

The adapted three-dimensional network of Zeng et al. (2020) yielded better registration results than the original two-dimensional version. However, both networks worsened the NMI, Dice coefficient and Hausdorff distance compared to the unregistered images, despite using an architecture resembling the best performing network in terms of Dice coefficient (Gu et al., 2019). This could be due to the number of filters for each convolutional layer being too high. The first convolution in the network of Zeng et al. (2020) already contains 128 filters. The network of Gu et al. (2019) starts with only 4 filters in the first convolutional layer. Especially for small datasets, it might happen that for large networks with many parameters to be learned, there is not enough training data available to support the complexity of the network and a model with fewer parameters performs better.

Cross-stitch units are applied in (Chen et al., 2021b) to learn the best combination of the outputs of the previous layers. However, the registration results in Table 1 show that the cross-stitch units might be inadequate for small datasets and a simpler architecture without cross-stitch units could achieve better registration results. To overcome this problem, caused by small training datasets, Chen et al. (2021b) used 6,000 pairs of synthetic CT scans (2,400 pairs for training) as initialisation of the three-dimensional network before training the network on a small dataset (Hering et al., 2020b,a) consisting of 60 three-dimensional CT volumes (Chen et al., 2021b).

(Gao et al., 2021) also applied a network with an architecture similar to the best performing network in terms of Dice coefficient (Gu et al., 2019), using a different number of filters for each convolutional layer and a different number of convolutional blocks. However, these parameters are set in the network of Gao et al. (2021) in such a way that the network yielded good registration results for the dataset used in their work, but did not generalize well and achieved poor registration results for the datasets used in this paper. In order to achieve a higher generalizability, the parameters aforementioned have to be adjusted.

In his master's thesis, Roelofs (2021) evaluated different ResNet architectures for affine registration of medical images. The best results were obtained with ResNet-18, therefore ResNet-18 was examined in this paper. The skip connections used in ResNets overcome the vanishing gradient problem and considerably improve the overall performance of neural networks comprising a substantial number of layers. For this reason, the network used in (Roelofs, 2021) (ResNet-18) was able to improve the Dice coefficient of the synthetic dataset and the patient dataset (trained on the synthetic dataset as well as finetuned with the patient dataset).

The affine network of Shao et al. (2021) consists of feature extraction, matching and affine regression network. A correlation matrix was used as matching part, which has the advantage over concatenating or subtracting the CNN features that it improves the generalizability and performance of the network (Rocco et al., 2018). The network provided good registration results for the data used in (Shao et al., 2021) (MR images and histopathology images of the prostate). However, the entire network architecture did not generalize well and thus produced poor registration results for the datasets used in this paper. To improve generalizability, the architecture of the feature extraction network and the affine regression network could be further modified.

In (Zhu et al., 2021) a joint affine and deformable network is used. In the implementation of the network for this work, only the affine part was used. However, the affine and deformable subnetworks are not independent of each other but are linked through parameter sharing of the affine network and the downsample part of the deformable network. Training the joint affine and deformable network therefore also affects the predicted affine transformation matrix and could improve the affine registration result.

Chee and Wu (2018) used an encoder which, in addition to convolutional layers and pooling layers, also uses dense blocks. In dense blocks, skip connections are applied between convolutional layers to combine features from different levels. Chee and Wu (2018) evaluated their network on a single dataset (three-dimensional MR images of the human brain). For this dataset, the network achieved good registration results. However, poor results were obtained for new datasets due to weak generalizability. The network architecture, for example the number of dense blocks, could be further adjusted to improve generalizability to new datasets.

As first part of their network, de Vos et al. (2019) used two distinct pipelines for fixed and moving image to enable inputting image pairs into the network that do not have the same volume size without prior resizing. Weight sharing is applied between the two pipelines. This results in a less complex model with a reduced number of parameters and enhances its generalizability. The network also achieved good registration results for new datasets and increased the Dice coefficient significantly for the synthetic dataset and the patient dataset (Table 1).

In (de Silva et al., 2020) a similar network structure is applied as in (Shao et al., 2021), i.e. feature extraction, matching and regression network. Both networks are based on the network architecture of Rocco et al. (2018). But Shao et al. (2021) applied a ResNet-101 as feature extractor, while de Silva et al. (2020) used a VGG-16 network. This difference in the feature extractor might have resulted in the network of Shao et al. (2021) achieving a poor registration result, while the network of de Silva et al. (2020) finetuned for the patient dataset significantly improved the registration of the patient dataset for the three investigated metrics.

Venkata et al. (2022) used a modified network architecture of (Chee and Wu, 2018). In (Chee and Wu, 2018), an encoder is used whose first part includes convolutional layers with two-dimensional filters and the second part of the encoder applies three-dimensional filters. This structure is used to handle that the volume of a three-dimensional medical image is usually not cube-shaped, but has a lesser depth than height and width (Chee and Wu, 2018). Venkata et al. (2022), on the other hand, used three-dimensional filters for all convolutional layers. As shown in Table 1, this change in the network architecture had no significant effect on the Dice coefficient of the registration result for the synthetic dataset and the patient dataset.

Waldkirch (2020) used a U-Net architecture and a global average pooling layer. The U-Net architecture has demonstrated high accuracy in biomedical image segmentation (Ronneberger et al., 2015). The U-Net's high generalizability also allows it to be applied to other image-related tasks, e.g. it is already often used for deformable image registration, and achieved good affine registration results for the datasets used in this paper.

In (Chen et al., 2022) the swin transformer is used, which combines transformer and convolutions. Transformers provide the advantage over convolutions that the receptive field of the transformer is as large as the image, i.e. it is not restricted to a small area as in convolutions. Spatial relationships in the two images can therefore be better investigated and understood (Chen et al., 2022). The network of (Chen et al., 2022) (trained on the synthetic dataset as well as finetuned with the patient dataset) was therefore able to improve the spatial alignment (based on the segmentation-based metrics) of the synthetic dataset and the patient dataset significantly.

The C2FViT transformer of Mok and Chung (2022) is a modification of the Vision Transformer (Dosovitskiy et al., 2020). Vision Transformer already yielded good results in the area of image classification, but has the drawback that training with a large dataset is necessary (Dosovitskiy et al., 2020). However, our in-house synthetic dataset and patient dataset are relatively small. This led to a poor performance of the network. For larger datasets, the network will probably achieve better registration results.

The neural network of Hasenstab et al. (2019) achieved the worst registration results. This is due to the fact that the network consists of only one single dense layer with 12 parameters. Convolutions or transformers were used for all other implemented networks. Convolutions have the advantage over dense layers in that they have fewer parameters to be learned during training, and the kernels are translationally-invariant and can recognise local patterns. Applying convolutions will lead to an improvement of the efficiency and performance of the network (Yamashita et al., 2018).

### *5.3. Individual Metrics*

Some networks improved only one of the two segmentation-based metrics (Dice coefficient or Hausdorff distance), not both at the same time. The reason for this may be that the Hausdorff distance is more sensitive to outliers than the Dice coefficient.

It is also important to note that the value of the evaluation metric is biased towards the loss function used during the optimisation. As baseline SimpleElastix was applied using the default settings for an affine transformation. The similarity metric Mutual Information was optimised, as there is no Dice coefficient similarity metric in SimpleElastix. The Normalized Mutual Information of the data registered with SimpleElastix was therefore relatively high. The neural networks, on the other hand, were trained with the Dice loss. The Dice coefficient of the data registered with a neural network trained with Dice loss was higher than if the neural networks were trained with the mutual information loss. The registration results of the baseline and the neural networks therefore cannot be compared directly in terms of value of the evaluation metrics. The baseline was used to show that affine registration of the datasets is basically possible.

By optical evaluation, we verified that the structural similarity of a network improved as the Dice coefficient increased through registration. However, this was not always directly visible in the NMI, which decreased for some networks after registration. One reason for this could be that the NMI is calculated globally, i.e. over the entire image. It is therefore not able to take spatial information into account. The intensity distribution is calculated over the entire image, the distance of the voxels with the same intensities is not considered. A locally calculated NMI would therefore be suitable here. This way, only the voxels that are close to each other would be taken into account (Guo, 2019).

As demonstrated with the synthetic dataset in Figure 3, structural similarity could improve, i.e. overlap of the liver increased, but the the overlap of the abdomens in total deteriorated. In the experiments of this paper, only affine registration was used, which is limited in its transformation. Due to respiration, movement of the patient and a different positioning of the patient in the CT and MR scanner, the abdomen is transformed deformable. In this case, an affine registration is not sufficient to spatially align both the entire abdomen in total and the liver at the same time. For this purpose, a deformable transformation should be used to further improve the spatial alignment of the images. Future work will focus on further transformations, e.g. deformable registration.

## 6. Conclusion

Although various machine learning algorithms for affine multimodal image registration are proposed, we could demonstrate that few can generalize to new data and applications. More work should be invested in developing more general applicable image registration techniques to be able to be used in minimal-invasive image guided interventions to support precision medicine in cancer treatment.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This research project is part of the Research Campus $M^2$OLIE and funded by the German Federal Ministry of Education and Research (BMBF) within the Framework "Forschungscampus: public-private partnership for Innovations" under the funding code 13GW0388A.
This project was supported by the German Federal Ministry of Education and Research (BMBF) under the funding code 01KU2102, under the frame of ERA PerMed.

## References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I.J., Harp, A., Irving, G., Isard, M., Jia, Y., Józefowicz, R., Kaiser, L., Kudlur, M., . . . Zheng, X., 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. CoRR abs/1603.04467. URL: hiip://arxiv.org/abs/1603.04467, arXiv:1603.04467.

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. doi:10.1109/tmi.2019.2897538.

Bauer, D.F., Russ, T., Waldkirch, B.I., Tönnes, C., Segars, W.P., Schad, L.R., Zöllner, F.G., Golla, A.K., 2021. Generation of annotated multimodal ground truth datasets for abdominal medical image registration. International Journal of Computer Assisted Radiology and Surgery 16, 1277–1285. doi:10.1007/s11548-021-02372-7.

Chee, E., Wu, Z., 2018. AIRNet: Self-Supervised Affine Registration for 3D Medical Images using Neural Networks. CoRR abs/1810.02583. URL: hiip://arxiv.org/abs/1810.02583, arXiv:1810.02583.

Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y., 2022. TransMorph: Transformer for unsupervised medical image registration. Medical Image Analysis 82, 102615. doi:hiips://doi.org/10.1016/j.media.2022.102615.

Chen, X., Dianz-Pinto, A., Ravikumar, N., Frangi, A.F., 2021a. Deep learning in medical image registration. Prog. Biomed. Eng. 3, 012003. doi:10.1088/2516-1091/abd37c.

Chen, X., Meng, Y., Zhao, Y., Williams, R., Vallabhaneni, S.R., Zheng, Y., 2021b. Learning Unsupervised Parameter-Specific Affine Transformation for Medical Images Registration. pp. 24–34. doi:10.1007/978-3-030-87202-1_3.

Chumchob, N., Chen, K., 2009. A robust affine image registration method. International Journal of Numerical Analysis and Modeling 6, 311–334.

German Cancer Research Center (DKFZ) Division of Medical Image Computing, . Medical Imaging Interaction Toolkit (MITK). URL: hiips://wwwmitkorg/wiki/The _Medical_Imaging_Interaction_Toolkit_(MITK). v2021.02.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. CoRR abs/2010.11929. URL: `hiips://arxiv.org/abs/2010.11929`, `arXiv:2010.11929`.

Fluck, O., Vetter, C., Wein, W., Kamen, A., Preim, B., Westermann, R., 2011. A survey of medical image registration on graphics hardware. Computer Methods and Programs in Biomedicine 104, e45–e57. doi:hiips://doi.org/10.1016/j.cmpb.2010.10.009.

Fu, Y., Lei, Y., Wang, T., Curran, W.J., Liu, T., Yang, X., 2020. Deep learning in medical image registration: a review. Phys. Med. Biol. 65, 20TR01. doi:10.1088/1361-6560/ab843e.

Gao, X., Van Houtte, J., Chen, Z., Zheng, G., 2021. DeepASDM: a Deep Learning Framework for Affine and Deformable Image Registration Incorporating a Statistical Deformation Model, in: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 1–4. doi:10.1109/BHI50953.2021.9508553.

Gu, D., Liu, G., Tian, J., Zhan, Q., 2019. Two-Stage Unsupervised Learning Method for Affine and Deformable Medical Image Registration, in: 2019 IEEE International Conference on Image Processing (ICIP), pp. 1332–1336. doi:10.1109/ICIP.2019.8803794.

Guo, C.K., 2019. Multi-Modal Image Registration with Unsupervised Deep Learning. Master's thesis. Massachusetts Institute of Technology. United States.

Hasenstab, K.A., Cunha, G.M., Higaki, A., Ichikawa, S., Wang, K., Delgado, T., Brunsing, R.L., Schlein, A., Bittencourt, L.K., Schwartzman, A., Fowler, K.J., Hsiao, A., Sirlin, C.B., 2019. Fully automated convolutional neural network-based affine algorithm improves liver registration and lesion co-localization on hepatobiliary phase T1-weighted MR images. Eur Radiol Exp. 3(1):43. doi:10.1186/s41747-019-0120-7.

Hering, A., Murphy, K., van Ginneken, B., 2020a. Learn2Reg Challenge: CT Lung Registration - Test

Data. URL: hiip://doi.org/10.5281/zenodo.4048761, doi:10.5281/zenodo.4048761.

Hering, A., Murphy, K., van Ginneken, B., 2020b. Learn2Reg Challenge: CT Lung Registration - Training Data. URL: hiip://doi.org/10.5281/zenodo.3835682, doi:10.5281/zenodo.3835682.

Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, E., Bandula, S., Moore, C.M., Emberton, M., Ourselin, S., Noble, J.A., Barratt, D.C., Vercauteren, T., 2018. Weakly-supervised convolutional neural networks for multimodal image registration. Medical Image Analysis 49, 1–13. doi:10.1016/j.media.2018.07.002.

Jaderberg, M., Simonyan, K., Zisserman, A., kavukcuoglu, k., 2015. Spatial Transformer Networks, in: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.

Kavur, A.E., Selver, M.A., Dicle, O., Barış, M., Gezer, N.S., 2019. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data. URL: hiips://doi.org/10.5281/zenodo.3362844, doi:10.5281/zenodo.3362844.

Luo, G., Chen, X., Shi, F., Peng, Y., Xiang, D., Chen, Q., Xu, X., Zhu, W., Fan, Y., 2020. Multimodal affine registration for ICGA and MCSL fundus images of high myopia. Biomedical Optics Express 11. doi:10.1364/BOE.393178.

Marstal, K., Berendsen, F., Staring, M., Klein, S., 2016. SimpleElastix: A User-Friendly, Multi-lingual Library for Medical Image Registration, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 574–582. doi:10.1109/CVPRW.2016.78.

Modat, M., Cash, D., Daga, P., Winston, G., Duncan, J., Ourselin, S., 2014. Global image registration using a symmetric block-matching approach. Journal of medical imaging (Bellingham, Wash.) 1, 024003. doi:10.1117/1.JMI.1.2.024003.

Mok, T.C.W., Chung, A.C.S., 2022. Affine Medical Image Registration with Coarse-to-Fine Vision Transformer. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 20803–20812.

Qiu, H., Katz, A.W., Milano, M.T., 2016. Oligometastases to the liver: predicting outcomes based upon radiation sensitivity. J Thorac Dis (10):E1384-E1386. doi:10.21037/jtd.2016.10.88.

Rocco, I., Arandjelović, R., Sivic, J., 2018. Convolutional neural network architecture for geometric matching. IEEE Transactions on Pattern Analysis and Machine Intelligence , 2553–2567.

Roelofs, T.J.T., 2021. Deep Learning-Based Affine and Deformable 3D Medical Image Registration. Master's thesis. Aalto University. Espoo, Finnland.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham. pp. 234–241.

Ruers, T., Van Coevorden, F., Punt, C., Pierie, J., Borel-Rinkes, I., Ledermann, J., Poston, G., Bechstein, W., Lentz, M., Mauer, M., Folprecht, G., Van Cutsem, E., Ducreux, M., Nordlinger, B., 2017. Local Treatment of Unresectable Colorectal Liver Metastases: Results of a Randomized Phase II Trial. J Natl Cancer Inst 109(9):djx015. doi:10.1093/jnci/djx015.

Segars, W., Sturgeon, G., Mendonca, S., Grimes, J., Tsui, B., 2010. 4D XCAT phantom for multimodality imaging research. Medical physics 37, 4902–15. doi:10.1118/1.3480985.

Shao, W., Banh, L., Kunder, C.A., Fan, R.E., Soerensen, S.J., Wang, J.B., Teslovich, N.C., Madhuripan, N., Jawahar, A., Ghanouni, P., Brooks, J.D., Sonn, G.A., Rusu, M., 2021. ProsRegNet: A deep learning framework for registration of MRI and histopathology images of the prostate. Medical Image Analysis 68, 101919. doi:hiips://doi.org/10.1016/j.media.2020.101919.

Shen, Z., Han, X., Xu, Z., Niethammer, M., 2019. Networks for Joint Affine and Non-Parametric Image Registration, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 4219–4228. doi:10.1109/CVPR.2019.00435.

de Silva, T., Chew, E.Y., Hotaling, N., Cukras, C.A., 2020. Deep-Learning based Multi-Modal Retinal Image Registration for Longitudinal Analysis of Patients with Age-related Macular Degeneration. Biomedical Optics Express 12. doi:10.1364/BOE.408573.

Soler, L., Hostettler, A., Agnus, V., Charnoz, A., Fasquel, J., Moreau, J., Osswald, A., Bouhadjar, M., Marescaux, J., 2010. 3D image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database. IRCAD, Strasbourg, France, Tech. Rep .

Tang, K., Li, Z., Tian, L., Wang, L., Zhu, Y., 2020. ADMIR–Affine and Deformable Medical Image Registration for Drug-Addicted Brain Images. IEEE Access 8, 70960–70968. doi:10.1109/ACCESS.2020.2986829.

Venkata, S.P., Duffy, B.A., Datta, K., 2022. An unsupervised deep learning method for affine registration of multi-contrast brain MR images.

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. doi:10.1016/j.media.2018.11.010.

Waldkirch, B.I., 2020. Methods for three-dimensional Registration of Multimodal Abdominal Image Data. Ph.D. thesis. Ruprecht Karl University of Heidelberg.

Yamashita, R., Nishio, M., Do, R., Togashi, K., 2018. Convolutional neural networks: an overview and application in radiology. Insights into Imaging 9. doi:10.1007/s13244-018-0639-9.

Zeng, Q., Fu, Y., Tian, Z., Lei, Y., Zhang, Y., Wang, T., Mao, H., Liu, T., Curran, W., Jani, A., Patel, P., Yang, X., 2020. Label-driven MRI-US registration using weakly-supervised learning for MRI-guided prostate radiotherapy. Physics in Medicine & Biology 65. doi:10.1088/1361-6560/ab8cd6.

Zhao, S., Lau, T., Luo, J., Chang, E.I.C., Xu, Y., 2020. Unsupervised 3D End-to-End Medical Image Registration With Volume Tweening Network. IEEE Journal of Biomedical and Health Informatics 24, 1394–1404. doi:10.1109/jbhi.2019.2951024.

Zhu, Z., Cao, Y., Chenchen, Q., Rao, Y., Di, L., Dou, Q., Ni, D., Wang, Y., 2021. Joint affine and deformable three-dimensional networks for brain MRI registration. Medical Physics 48. doi:10.1002/mp.14674.