sensors

MDPI

Article

Learning a Metric for Multimodal Medical Image Registration without Supervision Based on Cycle Constraints

Hanna Siebert *, Lasse Hansen and Mattias P. Heinrich

Institute of Medical Informatics, Universität zu Lübeck, 23538 Lübeck, Germany; hansen@imi.uni-luebeck.de (L.H.); heinrich@imi.uni-luebeck.de (M.P.H.)
* Correspondence: siebert@imi.uni-luebeck.de

Abstract: Deep learning based medical image registration remains very difficult and often fails to improve over its classical counterparts where comprehensive supervision is not available, in particular for large transformations—including rigid alignment. The use of unsupervised, metric-based registration networks has become popular, but so far no universally applicable similarity metric is available for multimodal medical registration, requiring a trade-off between local contrast-invariant edge features or more global statistical metrics. In this work, we aim to improve over the use of handcrafted metric-based losses. We propose to use synthetic three-way (triangular) cycles that for each pair of images comprise two multimodal transformations to be estimated and one known synthetic monomodal transform. Additionally, we present a robust method for estimating large rigid transformations that is differentiable in end-to-end learning. By minimising the cycle discrepancy and adapting the synthetic transformation to be close to the real geometric difference of the image pairs during training, we successfully tackle intra-patient abdominal CT-MRI registration and reach performance on par with state-of-the-art metric-supervision and classic methods. Cyclic constraints enable the learning of cross-modality features that excel at accurate anatomical alignment of abdominal CT and MRI scans.

Keywords: image registration; cycle constraint; multimodal features; self-supervision; rigid alignment

Citation: Siebert, H.; Hansen, L.; Heinrich, M.P. Learning a Metric for Multimodal Medical Image Registration without Supervision Based on Cycle Constraints. *Sensors* **2022**, 22, 1107. https://doi.org/10.3390/s22031107

Academic Editors: Vahid Abolghasemi, Hossein Anisi and Saideh Ferdowsi

Received: 28 December 2021
Accepted: 27 January 2022
Published: 1 February 2022

1. Introduction

Medical image registration based on deep learning methods has gathered great interest over the last few years. Yet, certain challenges, especially in multimodal registration, need to be addressed for learning based approaches, as evident from the recent MICCAI challenge *Learn2Reg* [1]. In order to avoid an elaborate comprehensive annotation of all relevant anatomies and to avoid label bias, unsupervised, metric-based registration networks are widely used for intramodal deep learning based registration [2,3].

However, this poses an additional challenge for multimodal registration problems, as currently no universal metric has been developed and a trade-off has to be made between using local contrast-invariant edge features such as NGF, LCC, and MIND or more global statistical metrics like mutual information. Metric-based methods also entail the difficulty of tuning hyperparameters that balance similarity weights (ensuring similarity between fixed image and warped moving image) and regularisation weights (ensuring plausible deformations).

Ground truth deformations for direct supervision are only available when using synthetic deformation fields. The now very popular FlowNet [4] estimates deformation fields between pairs of input images from a synthetically generated dataset that has been obtained by applying affine transformations to images. However, for medical applications, synthetic deformations have been deployed for monomodal image registration [5–7]. Alternatively, label supervision that primarily maximises the alignment of known structures with expert annotations could be employed [2,8,9]. This leads to improved registration of anatomies

that are well represented, but can introduce a bias and deteriorating performance for unseen labels.

On the one hand, the focus of supervised approaches on a limited set of labelled structures may be particularly inadequate for diagnosis of a pathology that cannot be represented sufficiently in the training data. Using metric supervision, on the other hand, has little potential to improve upon classical algorithms that employ the same metric as similarity terms during optimisation. With efficient (parallelised) implementations, adequate runtimes of less than a minute have recently been achieved for classical algorithms.

Learning completely without metric or label supervision, self-supervision, would remedy the aforementioned problems and enable the development of completely new registration methods and multimodal feature descriptors without introducing annotation or engineering biases.

Self-supervision approaches have been used in medical and non-medical learning based image processing tasks. Recently, a self-supervised approach for learning pretext-invariant representations for object detection has outperformed supervised pre-training in [10]. By minimising a contrastive loss function, the authors construct image representations that are invariant to image patch pertubation, similar to the representation of transformed versions of the same image and differ from representations of other images. In [11], semantic features have been learned with self-supervision in order to recognise the rotation that has been applied to an image given four possible transformations as multiples of 90 degrees. The learned features have been useful for various visual perception tasks. For rigid registration between point clouds, an iterative self-supervised method has been proposed in [12]. Here, partial-to-partial registration problems have been addressed by learning geometric priors directly from data. The method comprises a keypoint detection module which identifies points that match in the input point clouds based on co-contextual information and aligns common keypoints. For monomodal medical image registration, in [13] spatial transformations between image pairs have been estimated in a self-supervised learning procedure. Therefore, an image-wise similarity metric between fixed and warped moving images is maximised in a multi-resolution framework while the deformation fields are regularised for smoothness.

In [14], cycle-consistency in time is used for learning visual correspondence from unlabelled video data for self-supervision. Their idea is to obtain supervision for correspondence by tracking backward and then forward, i.e., along a cycle in time, and use the inconsistency between the start and end points as the loss function. For image-to-image translation, a cycle-consistent adversarial network approach is introduced in [15]. The authors use a cycle consistency loss that induces the assumption that forward and backward translation should be bijective and inverse of each other. Another approach that addresses inconsistency is introduced in [16] for medical image registration. It uses information from a complete set of pairwise registrations, aggregates inconsistency, and minimizes the group-wise inconsistency of all pairwise image registrations by using a regularized least-squares algorithm. The idea to measure consistency via registration cycles for monomodal medical image data has been used in [17] that estimates forward and reverse transformation jointly in a non-deep-learning approach and [18] using registration circuits to correct registration errors. In [19], a monomodal unsupervised medical image registration method that trains deep neural network for deformable registration is presented using CNNs with cycle-consistency. This approach uses two registration networks that process the two input images as fixed and moving images inversely to each other and gives the deformed volumes to the networks again to re-deform the images to impose cycle-consistency.

Previous deep learning based registration work has often omitted the step of rigid or affine registration, despite its immense challenges due to often large initial misalignments. Image registration challenges such as [1] provide data that has been pre-aligned with help of non-deep-learning-based methods, whereas the challenge's image registration tasks are then often addressed with deep learning based methods. Rigid transformation is often the inital step before performing deformable image registration, and only few

works [20] investigate deep learning techniques for this step. As evident from the CuRIOUS challenge [21], so far no CNN approach was able to learn a rigid or affine mapping between multimodal scan pairs (MRI and ultrasound of neurosurgery) with an adequate robustness. Besides that, no label bias can occur with rigid alignment. Hence, a learning model for large linear transformations is of great importance.

Contributions

In order to avoid the difficulty of choosing a metric for multimodal image registration, we propose a completely new concept. For learning multimodal features for image registration, our learning method requires neither label supervision nor handcrafted metrics. It extends upon research that successfully learned monomodal alignment through synthetic deformations, but transforms this concept to multimodal tasks without resorting to complex modality synthesis.

The basic idea of our novel learning based approach is illustrated in Figure 1. It relies on geometric instead of metric supervision. In this work

- We introduce a cycle based approach including cycles that for each pair of CT and MRI scans comprise two multimodal transformations to be estimated and one known synthetic monomodal transformation.
- We restrict ourselves to rigid registration and aim to learn multimodal registration between CT and MRI without metric supervision by minimising the cycle discrepancy.
- We use a CNN for feature extraction with initially separate encoder blocks for each modality followed by shared weights within the last layers.
- We use a correlation layer without trainable weights and a differentiable least squares fitting procedure to find an optimal 3D rigid transformation.
- We created to the best of our knowledge the first annotated MRI/CT dataset with paired patient data that are made publicly available with manual segmentations for liver, spleen, left and right kidney.

Our extensive experimental validation on 3D rigid registration demonstrates the high accuracy that can be achieved and the simplicity of training such networks.

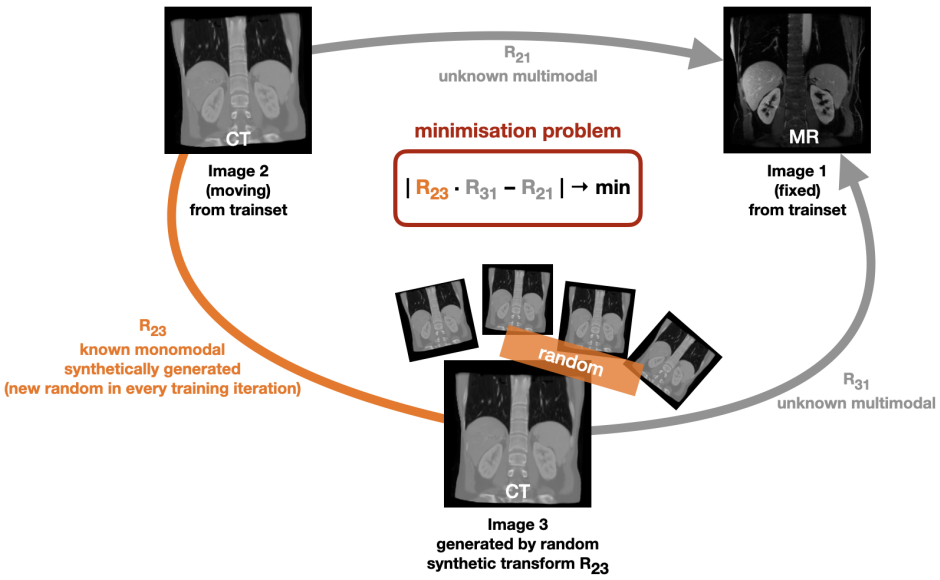


Figure 1. Our proposed self-supervised learning concept for multimodal image registration aiming to minimise a cycle discrepancy. In every training iteration, another (known) random transformation matrix R_{23} is used to generate a synthetic image. Like this, a cycle consisting of two unknown multimodal transformations (with the transformation matrices R_{21} and R_{31}) and a known monomodal transformation (with the transformation matrix R_{31}) is obtained, leading to the minimisation problem of $|R_{23} \cdot R_{31} - R_{21}| \rightarrow \min$ that is used for learning.

2. Materials and Methods

We introduce a learning concept for multimodal image registration that learns without metric supervision. Therefore, we propose a method to learn with the help of a self-supervised learning procedure using three-way cycles. For our registration models, the architectural design consists of modules for feature extraction, correlation, and registration. Implementation details, open source code and trained models can be found at github.com/multimodallearning/learning_without_metric (accessed on 8 March 2021).

2.1. Self-Supervised Learning Strategy

Our deep learning based method learns multimodal registration without using metric supervision. Instead, it is based on geometric self-supervision by minimising the cycle discrepancy created through a cycle consisting of two multimodal transformation and one monomodal transformation. The basic cycle idea is illustrated in Figure 1: Initially, a fixed image (Image 1) and a moving image (Image 2) exist. The transformation R_{21} is unknown and is to be learned by our method. In each training iteration, we randomly deform the moving image (Image 2) by applying a known random transformation R_{23} and hereby obtain a synthetic image (Image 3). By bringing the individual transformations into a cycle, the minimisation problem of

$$|R_{23} \cdot R_{31} - R_{21}| \rightarrow \min \tag{1}$$

can be derived. We chose to minimise the discrepancy as given in Equation (1) instead of minimising the difference between the transformation combination $R_{23} \cdot R_{31} \cdot R_{12}$ and the identity transformation Id with $|R_{23} \cdot R_{31} \cdot R_{12} - \mathrm{Id}| \rightarrow \min$ in order to avoid that our method only learns identity warping. For optimisation, we use the mean squared error loss function to minimise the cycle discrepancy between the two flow fields generated by the transformation matrices R_{21} and $R_{23,31} = R_{23} \cdot R_{31}$.

As we restrict our model to rigid registration, we create the synthetic transformations R_{23} by randomly initialising rigid transformation matrices with values that are assumed to be realistic from an anatomical point of view.

The advantages of our learning concept are manifold. First, in comparison to supervising the learning with a known similarity metric and regularisation term, the need for balancing a weighting term is removed and the method is applicable to new datasets without domain knowledge. Second, it enables multimodal learning, which is not feasible using synthetic deformations in conjunction with image-based loss terms (cf. [6]). Third, it avoids the use of domain discriminators as used, e.g., in the CycleGAN approach [15,22], which usually requires a large set of training scans with comparable contrast in each modality and may be sensitive to hyper-parameter choices.

On first sight, it might seem daring to use such a weak guidance. While it is clear that once suitable features are learned the loss term enables convergence, since the cycle constraint is fulfilled. Yet to initiate training towards improved features, we primarily rely on the power of randomness (by drawing multiple large synthetic deformations) and explorative learning. In addition, the architecture contains a number of stabilising elements: a patch-based correlation layer computation, outlier rejection and least squares fitting, that are described in detail below in Sections 2.3 and 2.4.

2.2. Training Pipeline

We apply our self-supervised learning strategy in the training procedure by going through the same steps in each training iteration as visualised in Figure 2: First, a random transformation matrix R_{23} is generated and applied on the moving image in order to obtain the synthetic image. Then, moving and fixed image are passed through feature extraction, correlation layer and transformation computation module to obtain the transformation matrix R_{21}. The same step is also performed for fixed and synthetic image to obtain R_{31}. After this, R_{23} and R_{31} are combined to obtain $R_{23,31}$. Finally, the mean squared error of the

deformations calculated with help of R_{21} and $R_{23,31}$ is determined. The individual modules for this training pipeline are described in more detail in the following Sections 2.3 and 2.4.

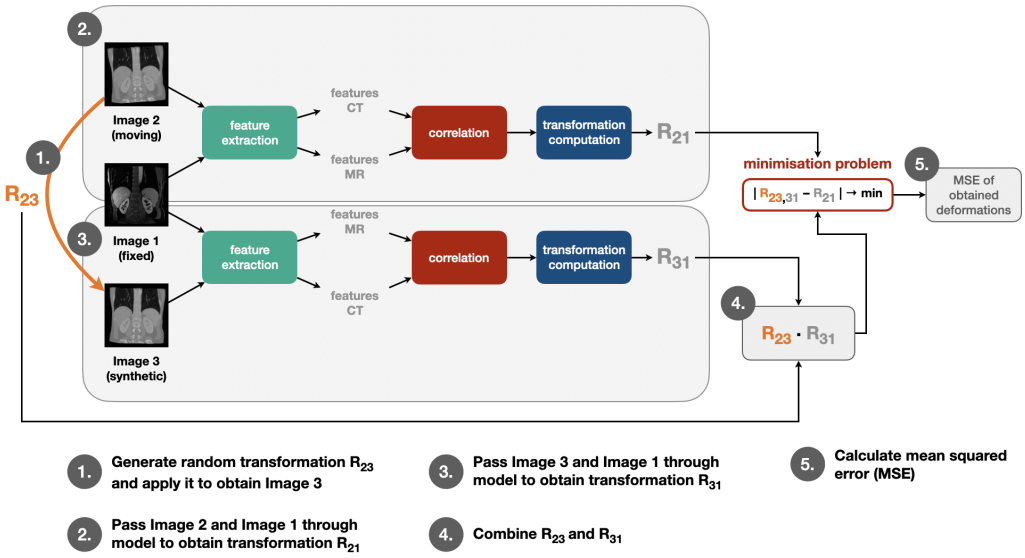


Figure 2. Pipeline to train our registration model: A random transformation matrix R_{23} is generated and used to obtain the synthetic image. The pair of moving and fixed image as well as the pair of synthetic and fixed image are passed through feature extraction, correlation layer and transformation computation module (see following Sections 2.3 and 2.4) to obtain the transformation matrices R_{21} and R_{31}. Then, R_{23} and R_{31} are combined to obtain $R_{23,31}$. As a final step, the mean squared error (MSE) of the deformations calculated with help of R_{21} and $R_{23,31}$ is determined.

2.3. Architecture

The architecture used for our registration method comprises three main components for feature extraction, correlation, and transformation computation.

We chose to use a CNN for feature extraction with initially separate encoder blocks for each modality and shared weights within the last few layers. These features are subsequently fed into the correlation layer, which has no trainable weights and whose output could be directly converted into displacement probabilities. Our method employs a robust and differentiable least squares fitting to find an optimal 3D rigid transformation subject to outlier rejection. Figure 3 visualises the procedure for feature extraction, correlation, and computation of the rigid transformation matrix that is used for registration.

For our feature extraction CNN, we use a Y-shaped architecture (cf. Figure 3) [9] starting with a separate network part for each of the two modalities (ModalityNet), which takes the respective input and passes it through two sequences with a structure of $2\times$.

- (Strided) 3D convolution with a kernel size of three and padding of one;
- 3D instance normalisation;
- leaky ReLU.

The two convolutions of the first sequence are non-strided and output eight feature channels. The first convolution of the second sequence has a stride of two and doubles the number of feature channels to 16, whereas the second convolution of the second sequence is non-strided and keeps the number of 16 feature channels. Whereas the size of the input dimensions are preserved within the first convolution sequence, the strided convolution within the second sequence leads to a halving of each feature map dimension. The output of the ModalityNets are passed into a final module with shared weights (SharedNet), which finalises the feature extraction by applying two sequences of the same structure as used for the separate ModalityNets. Here, the first sequence comprises non-strided convolutions that output 16 feature channels while keeping the spatial dimensions as output by the ModalityNets. The first convolution of the second sequence has a stride of two leading to

another halving of the spatial dimension's sizes and doubles the number of feature channels to 32. The second convolution of the second sequence is non-strided and keeps the number of 32 feature channels. The output of the SharedNet is given to a $1 \times 1 \times 1$-convolution providing the final number of 16 feature channels followed by a Sigmoid activation function. As we use correlation and transformation estimation techniques without trainable weights, our model only comprises 80k trainable parameters within the feature extraction part.

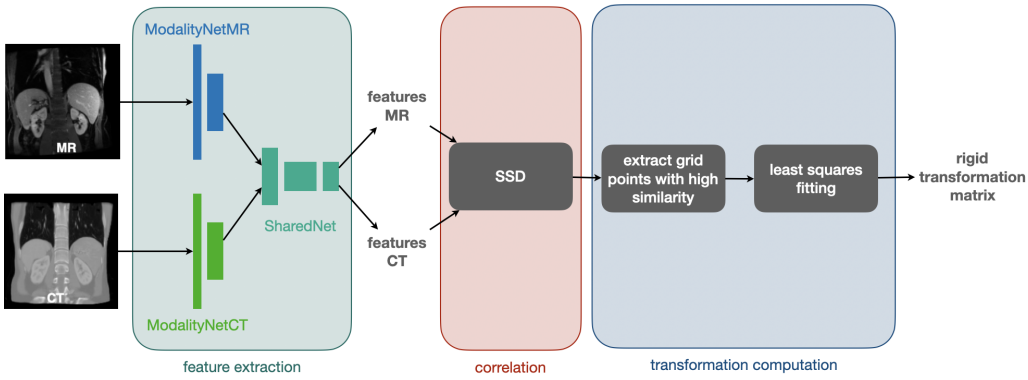


Figure 3. The process of feature extraction, correlation, and computation of the rigid transformation matrix: A CNN is used for feature extraction starting with a separate network part for each modality (ModalitiyNetMR and ModalityNetCT) followed by a module with shared weights (SharedNet). The obtained features are correlated by calculating patch-wise the sum of squared differences (SSD). Subsequently, grid points with high similarity are extracted and used to define point-wise correspondences to calculate the rigid transformation matrix with a least squares fitting.

2.4. Correlation and Transformation Computation

As suggested in previous research [3,4], the use of a dense correlation layer that explores a large number of discretised displacements at once is employed to capture larger deformations robustly. This way the learned features are used to define a sum of squared differences cost function akin to metric learning [23].

Similar to [24], which operates directly on input image pairs and uses normalised cross correlations (NCC), we use a block-matching technique to find correspondences between the fixed features and a set of transformed moving features. We correlate the obtained features by calculating patch-wise the sum of squared differences (SSD) and extract points with high similarity of a coarse grid with a spacing of 12 voxel. The extracted grid points are used to define point-wise correspondences to calculate the rigid transformation matrix with a robust (trimmed) least squares fitting procedure.

For the correlation layer, we choose a set of $11 \times 11 \times 11$ discrete displacements with a capture range of approx. 40 voxel in the original volumes. After calculating the sum-of-squared-differences cost volume, we sort the obtained SSD costs and reject the 50% of the displacement choices that entail the highest similarity costs. We apply the Softmax function on the remaining displacement choices to obtain differentiable soft correspondences. While we use this differentiable approach to estimate regularised transformations within a framework that comprises trainable CNN parameters, the learned features could also be used for other optimisation frameworks [9].

The displacement candidates output by the Softmax function are added to the coarse moving grid points. In a least squares fitting procedurce comprising five iterations, the final rigid transformation matrix that serves for transformation of the moving image is determined. The best-fitting rigid transformation can be found by computing the singular value decomposition $S = U\Sigma V^T$ with the matrix $S = X^T Y^T$ (X: centered fixed grid points x_i, Y:

centered moving grid points with added displacement candidates y_i) and the orthogonal matrices U and V obtained by the singular value decomposition. This leads to the rotation

$$Q = V \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & \ldots & & \\ & & & 1 & \\ & & & & \det(VU^T) \end{pmatrix} U^T \tag{2}$$

and the translation

$$t = \bar{y} - Q\bar{x} \tag{3}$$

with $\bar{x}$ being the mean values for fixed grid points and $\bar{y}$ the mean moving grid points with added displacement candidates.

This way, the rigid transformation matrices R_{21} and R_{31} are determined. To combine the synthetic transformation R_{23} and the predicted transformation R_{31} matrix multiplication is used yielding $R_{23,31}$. The transformation matrices R_{21} and $R_{23,31}$ are used to compute the affine grids that are then given to the MSE loss function during training and the affine grid computed by R_{21} is used for warping during inference to align the moving image to the fixed image.

This approach has the advantage of being very compact with only 80k parameters ensuring memory efficiency and fast convergence of training. The multimodal features learned by our model are generally usable for image alignment and can be given to various optimisation methods for image registration once trained with our method.

3. Experiments and Results

Our experiments are performed on 16 paired abdominal CT and MR scans from collections of The Cancer Imaging Archive (TCIA) project [25–28]. We have manually created labels for four abdominal organs (liver, spleen, left kidney, right kidney), which we use for the evaluation of our methods. Apart from a withheld test set, they are publicly released for other researchers to train and compare their multimodal registration models. The pre-processing comprises reorientation, resampling to an isotropic resolution of 2 mm and cropping/padding to volume dimensions of $192 \times 160 \times 192$.

To increase the number of training and testing pairs and model realistic variations in initial misalignment we augment the scans with 8 random rigid transformations each that on average reflect the same Dice overlap (of approx. 43%) as the raw data. All models are trained for 100 epochs with a mini-batch size of 4 in less than 45 min each using ≈8 GByte GPU memory.

The weights of the CNN used for feature extraction (FeatCNN) are trained for 100 epochs using the Adam optimiser with an initial learning rate of 0.001 and an cosine annealing scheduling.

3.1. Comparison of Training Strategies

We compare three different strategies to train our FeatCNN in a two-fold cross-validation:

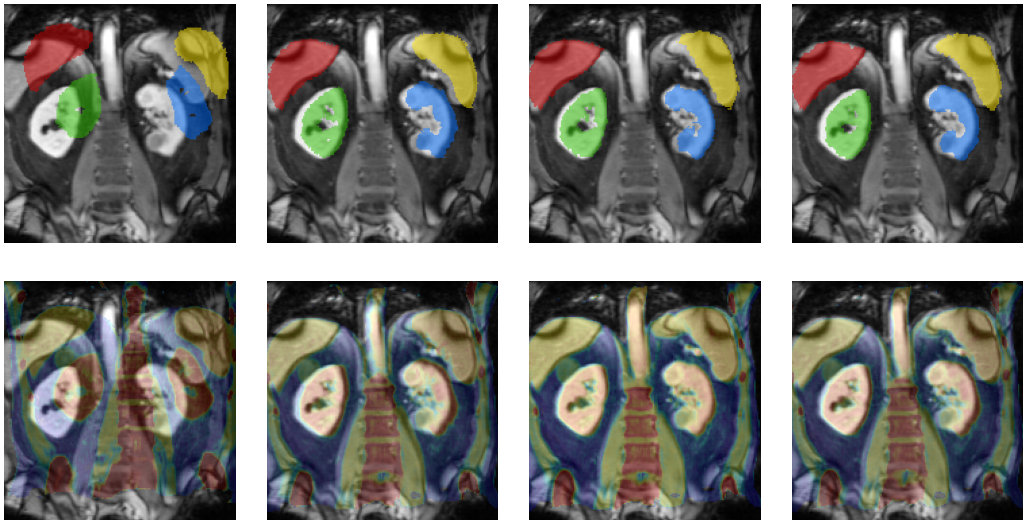
1. FeatCNN + Cycle Discrepancy (ours): Our proposed self-supervised cycle learning strategy;
2. FeatCNN + MI Loss: Learning with metric-supervision using Mutual Information (MI) as implemented by [29];
3. FeatCNN + NCC^2 Loss: Learning with metric-supervision using squared local normalised cross correlations (NCC^2) [24,30];
4. FeatCNN + Label Loss: Supervised learning with label supervision.

All methods share the same settings for the correlation layer and a trimmed least square transform fitting (with five iterations and 50% outlier rejection). Hyperparameters were determined on a single validation scan (#15) for cyclic training and left unaltered for all other experiments. The same trainable FeatCNN comprising the layers as described in Section 2.3

is used to train with our Cycle Discrepancy Loss, MI, NCC^2, and Label Loss. For correlation, we chose to extract corresponding grid points within a grid with a spacing of 12 voxels and use patches with a radius of 2 to patch-wise calculate the SSD. We use a displacement radius of 4 and discretise the set of displacements possibilities for the correlation layer with a displacement step (resp. voxel spacing) of 5. To adjust the smoothness of the soft correspondences, the costs obtained by SSD computation are multiplied by a factor of 150 when given to the Softmax function. As the soft-correspondences are needed for differentiability only during training, we increase this factor to 750 for inference.

For our cycle discrepancy method, we create the synthetic transformation matrices R_{23} by randomly initialising them with values that are assumed to be realistic from an anatomical point of view. Therefore, the maximum rotation is $\pm 23^\circ$ and the maximum translation ± 42 voxel (which equals 84 mm for our experiments) in every image dimension.

The results demonstrate a clear advantage of our proposed self-supervised learning procedure with an average Dice of 72.3% compared to the state-of-the-art MI metric loss with 68.14% and NCC^2 Loss with 68.1%, which is suitable for multimodal registration due to its computation involving small local neighbourhoods [24] (see Table 1 for qualitative and Figure 4 for quantitative results). This result comes close to the theoretical upper bound of our model trained with full label supervision with 79.55%.



(**a**) (**b**) (**c**) (**d**)

Figure 4. Qualitative results of our proposed cycle discrepancy approach FeatCNN + Cycle Discrepancy (**c**). We visualise the comparison to initial (**a**) before warping as well as to the methods FeatCNN + MI Loss (**b**) and FeatCNN + Label Loss (**d**) (coronal slices). The **top row** shows the fixed MRI and (warped) moving labels. The **bottom row** visualizes the (warped) moving CT and a jet colourmap overlay of the fixed MRI scan.

Table 1. Results for our cross-validation experiments: Dice scores listed by anatomical structures of our 3D experiments using FeatCNN for feature extraction and MI Loss, NCC^2 Loss, Label Loss or our Cycle Discrepancy for training.

	Liver	**Spleen**	**Lkidney**	**Rkidney**	**Mean**
initial	59.32 ± 14.03	36.90 ± 19.49	36.59 ± 19.53	37.02 ± 22.08	42.46 ± 18.78
FeatCNN + MI Loss	75.07 ± 9.38	63.17 ± 22.13	69.86 ± 26.34	64.46 ± 29.45	68.14 ± 21.92
FeatCNN + NCC^2 Loss	75.08 ± 12.22	61.09 ± 23.69	72.19 ± 27.51	64.04 ± 31.76	68.10 ± 23.80
FeatCNN + Cycle Discrepancy	77.95 ± 8.16	69.89 ± 16.00	70.18 ± 24.34	71.85 ± 34.40	72.30 ± 20.75
FeatCNN + Label Loss	81.24 ± 8.75	73.84 ± 18.32	83.15 ± 26.62	79.97 ± 33.59	79.55 ± 21.82

3.2. Comparison of Inference Strategies and Increased Trainset

To further enhance our method, we extend it by a two-level warping approach during inference. Therefore, we present our model the input moving and fixed image to warp the moving image and then apply our model to the resulting warped moving image and the fixed image again. For both warping steps, we set a displacement radius of 7 voxel and a grid spacing of 8 voxel. For the first warping step, we use a displacement discretisation of 4 voxel and refine this hyperparameter to 2 voxel for the second warping step.

Moreover, as our dataset is quite small and our method does not require labels, when considering an application scenario where a number of MR/CT scan pairs have to be aligned offline, a fine-tuning of the networks on this test data would be feasible. Therefore, we aim to further increase the performance of our method with training on all available paired CT and MR scans without splitting the dataset.

In Table 2 we compare the results of single-level and two-level warping as well as the cross-validation results and the results when training on the whole available image data. We compare the results achieved by our method with the results achieved using the rigid image registration tool *reg_aladin* of NiftyReg [24] applied to the image pairs used without the symmetric version and one registration level.

Table 2. Results for our experiments comparing single-level and two-level warping approach as well as cross-validation and training without withheld data: Dice scores listed by anatomical structures of our experiments using Cycle Discrepancy for training.

	Liver	**Spleen**	**Lkidney**	**Rkidney**	**Mean**
initial	59.32 ± 14.03	36.90 ± 19.49	36.59 ± 19.53	37.02 ± 22.08	42.46 ± 18.78
cross-validation 1 warp	77.95 ± 8.16	69.89 ± 16.00	70.18 ± 24.34	71.85 ± 34.40	72.30 ± 20.75
cross-validation 2 warps	80.71 ± 9.33	72.12 ± 17.08	79.33 ± 26.06	74.65 ± 36.91	76.68 ± 22.34
trained without withheld data 1 warp	81.04 ± 8.22	71.11 ± 18.03	76.27 ± 24.25	76.49 ± 32.64	76.23 ±20.88
trained without withheld data 2 warps	81.85 ± 0.58	76.77 ± 13.64	79.81 ± 24.52	80.17 ± 34.65	79.65 ± 20.25
NiftyReg *reg_aladin*	83.97 ± 6.19	76.55 ± 12.00	79.83 ± 7.12	79.26 ± 37.55	79.90 ± 15.15

Introducing a second warping step increased our cross-validation results by more than 4% points. When training without a withheld testset, we achieved further improvements by another 3% points. These results are on a par with the results of state-of-the-art classic method NiftyReg- *reg_aladin*.

4. Discussion

In this work, we presented a completely new concept for multimodal feature learning with application to 3D image registration without supervision of labels or handcrafted metrics. We introduced a new supervision strategy that is based on synthetic random transformations (two across modality and one within) that form a triangular cycle. Minimising the two multimodal transformations in such a cycle constraint avoids singular solutions (predicting identity transforms) and enables the learning of large rigid deformations. Through explorative learning, we are able to successfully train modality independent feature extractors that enable highly accurate and fast multimodal medical image alignment by minimising a cycle discrepancy in training. We also created the first public multimodal 3D MRI/CT abdominal dataset with manual segmentations for validation. To the best of our knowledge our work is also the first deep learning model for robustly estimating large misalignments of multimodal scans.

Despite the very promising results, there are a number of potential extensions that could further improve our concepts. The idea of incremental learning and predicting more useful synthetic transformations to improve detail alignment could be considered and has already shown potential in preliminary 2D experiments.

While the gap between training and test accuracy is relatively small due to the robust architectural design, further fine-tuning would be applicable at test time (since no supervision is required) with moderate computational effort. Combining hand-crafted domain knowledge with self-supervised learning might further boost accuracy. Similarly, domain adaptation through adversarial training could be incorporated to explicitly model the differences of modalities. While the gap between training and test accuracy is relatively small due to the robust architectural design, further fine-tuning would be applicable at test time (since no supervision is required) with moderate computational effort.

5. Conclusions

With our method, we were able to improve over the use of handcrafted metric-based losses by using synthetic three-way cycles. By minimising the cycle discrepancy, we are able to learn multimodal registration between CT and MRI without metric supervision. We created a robust method to estimate large rigid transformations that is differentiable in end-to-end learning. Our method is able to successfully perform intra-patient abdominal CT-MRI registration that outperforms state-of-the-art metric-supervision.

Author Contributions: Conceptualization, M.P.H., L.H. and H.S.; methodology, M.P.H., L.H. and H.S.; software, M.P.H., L.H. and H.S.; validation, M.P.H., L.H. and H.S.; formal analysis, M.P.H., L.H. and H.S.; investigation, M.P.H., L.H. and H.S.; resources, M.P.H.; data curation, M.P.H. and H.S.; writing—original draft preparation, H.S.; writing—review and editing, M.P.H., L.H. and H.S.; visualization, H.S.; supervision, M.P.H.; project administration, M.P.H.; funding acquisition, M.P.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Federal Ministry of Education and Research (BMBF) grant number 16DHBQP052.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Our experiments are performed on abdominal CT and MR scans from collections of The Cancer Imaging Archive (TCIA) project [25–28]. Material is available under TCIA Data Usage Policy and Creative Commons Attribution 3.0 Unported License. Material has been modified for direct usage in registration and deep learning algorithms: We have reorientated the

data, resampled it to an isotropic resolution of 2 mm, and used cropping and padding to achieve voxel dimensions of 192 × 160 × 192. We have also manually created segmentations for liver, spleen, left kidney, and right kidney. The results shown here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/ (accessed on 7 January 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hering, A.; Hansen, L.; Mok, T.C.W.; Chung, A.C.S.; Siebert, H.; Häger, S.; Lange, A.; Kuckertz, S.; Heldmann, S.; Shao, W.; et al. Learn2Reg: Comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *arXiv* **2021**, arXiv:2112.04489.
2. Balakrishnan, G.; Zhao, A.; Sabuncu, M.R.; Guttag, J.; Dalca, A.V. An unsupervised learning model for deformable medical image registration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9252–9260.
3. Heinrich, M.P. Closing the gap between deep and conventional image registration using probabilistic dense displacement networks. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 50–58.
4. Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; Brox, T. Flownet: Learning optical flow with convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2758–2766.
5. Eppenhof, K.A.; Pluim, J.P. Pulmonary CT registration through supervised learning with convolutional neural networks. *IEEE Trans. Med. Imaging* **2018**, *38*, 1097–1105. [CrossRef]
6. Eppenhof, K.A.; Lafarge, M.W.; Moeskops, P.; Veta, M.; Pluim, J.P. Deformable image registration using convolutional neural networks. In Proceedings of the Medical Imaging 2018: Image Processing, Houston, TX, USA, 10–15 February 2018; International Society for Optics and Photonics: Bellingham, WA, USA, Volume 10574, p. 105740S.
7. Krebs, J.; Mansi, T.; Delingette, H.; Zhang, L.; Ghesu, F.C.; Miao, S.; Maier, A.K.; Ayache, N.; Liao, R.; Kamen, A. Robust non-rigid registration through agent-based action learning. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Quebec City, QC, Canada, 11–13 September 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 344–352.
8. Hu, Y.; Modat, M.; Gibson, E.; Li, W.; Ghavami, N.; Bonmati, E.; Wang, G.; Bandula, S.; Moore, C.M.; Emberton, M.; et al. Weakly-supervised convolutional neural networks for multimodal image registration. *Med. Image Anal.* **2018**, *49*, 1–13. [CrossRef] [PubMed]
9. Blendowski, M.; Hansen, L.; Heinrich, M.P. Weakly-supervised learning of multi-modal features for regularised iterative descent in 3D image registration. *Med. Image Anal.* **2021**, *67*, 101822. [CrossRef]
10. Misra, I.; van der Maaten, L. Self-supervised learning of pretext-invariant representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6707–6717.
11. Komodakis, N.; Gidaris, S. Unsupervised representation learning by predicting image rotations. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
12. Wang, Y.; Solomon, J.M. PRNet: Self-Supervised Learning for Partial-to-Partial Registration. In *Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Vancouver, BC, Canada, 2019; Volume 32.
13. Li, H.; Fan, Y. Non-rigid image registration using self-supervised fully convolutional networks without training data. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 1075–1078.
14. Wang, X.; Jabri, A.; Efros, A.A. Learning correspondence from the cycle-consistency of time. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2566–2576.
15. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
16. Gass, T.; Székely, G.; Goksel, O. Consistency-based rectification of nonrigid registrations. *J. Med. Imaging* **2015**, *2*, 014005. [CrossRef] [PubMed]
17. Christensen, G.E.; Johnson, H.J. Consistent image registration. *IEEE Trans. Med. Imaging* **2001**, *20*, 568–582. [CrossRef] [PubMed]
18. Datteri, R.D.; Dawant, B.M. Automatic detection of the magnitude and spatial location of error in non-rigid registration. In Proceedings of the International Workshop on Biomedical Image Registration, Nashville, TN, USA, 7–8 July 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 21–30.
19. Kim, B.; Kim, J.; Lee, J.G.; Kim, D.H.; Park, S.H.; Ye, J.C. Unsupervised deformable image registration using cycle-consistent cnn. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 166–174.
20. de Vos, B.D.; Berendsen, F.F.; Viergever, M.A.; Sokooti, H.; Staring, M.; Išgum, I. A deep learning framework for unsupervised affine and deformable image registration. *Med. Image Anal.* **2019**, *52*, 128–143. [CrossRef] [PubMed]

21. Xiao, Y.; Rivaz, H.; Chabanas, M.; Fortin, M.; Machado, I.; Ou, Y.; Heinrich, M.P.; Schnabel, J.A.; Zhong, X.; Maier, A.; et al. Evaluation of MRI to ultrasound registration methods for brain shift correction: The CuRIOUS2018 challenge. *IEEE Trans. Med. Imaging* **2019**, *39*, 777–786. [CrossRef] [PubMed]
22. Xu, Z.; Luo, J.; Yan, J.; Pulya, R.; Li, X.; Wells, W.; Jagadeesan, J. Adversarial uni-and multi-modal stream networks for multimodal image registration. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 222–232.
23. Simonovsky, M.; Gutiérrez-Becker, B.; Mateus, D.; Navab, N.; Komodakis, N. A deep metric for multimodal registration. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, 17–21 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 10–18.
24. Modat, M.; Cash, D.M.; Daga, P.; Winston, G.P.; Duncan, J.S.; Ourselin, S. Global image registration using a symmetric block-matching approach. *J. Med. Imaging* **2014**, *1*, 024003. [CrossRef] [PubMed]
25. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *J. Digit. Imaging* **2013**, *26*, 1045–1057. [CrossRef] [PubMed]
26. Akin, O.; Elnajjar, P.; Heller, M.; Jarosz, R.; Erickson, B.; Kirk, S.; Filippini, J. Radiology data from the cancer genome atlas kidney renal clear cell carcinoma [TCGA-KIRC] collection. *Cancer Imaging Arch.* **2016**. [CrossRef]
27. Linehan, M.; Gautam, R.; Kirk, S.; Lee, Y.; Roche, C.; Bonaccio, E.; Jarosz, R. Radiology data from the cancer genome atlas cervical kidney renal papillary cell carcinoma [KIRP] collection. *Cancer Imaging Arch.* **2016**. [CrossRef]
28. Erickson, B.; Kirk, S.; Lee, Y.; Bathe, O.; Kearns, M.; Gerdes, C.; Lemmerman, J. Radiology Data from The Cancer Genome Atlas Liver Hepatocellular Carcinoma [TCGA-LIHC] collection. *Cancer Imaging Arch.* **2016**. [CrossRef]
29. Sandkühler, R.; Jud, C.; Andermatt, S.; Cattin, P.C. AirLab: Autograd image registration laboratory. *arXiv* **2018**, arXiv:1806.09907.
30. Balakrishnan, G.; Zhao, A.; Sabuncu, M.R.; Guttag, J.; Dalca, A.V. Voxelmorph: A learning framework for deformable medical image registration. *IEEE Trans. Med. Imaging* **2019**, *38*, 1788–1800. [CrossRef] [PubMed]