MDPI

*Article*

# Robust and realtime large deformation ultrasound registration using end-to-end differentiable displacement optimisation

**Mattias P. Heinrich** [1,†,*], **Hanna Siebert** [1,‡], **Laura Graf** [1,‡], **Sven Mischkewitz** [2] and **Lasse Hansen** [3]

[1] Institute of Medical Informatics; Universität zu Lübeck
[2] ThinkSono GmbH; Potsdam
[3] EchoScout GmbH; Lübeck
* Correspondence: heinrich@imi.uni-luebeck.de
† Current address: Ratzeburger Allee 160, 23562 Lübeck, Germany
‡ These authors contributed equally to this work.

**Abstract:** Image registration for temporal ultrasound sequences can be very beneficial for image-guided diagnostics and interventions. Cooperative human-machine systems that enable seamless assistance for both inexperienced and expert users during ultrasound examinations rely on robust, realtime motion estimation. Yet rapid and irregular motion patterns, varying image contrast and domain shifts in imaging devices pose a severe challenge to conventional realtime registration approaches. While learning based registration networks have the promise of abstracting relevant features and delivering very fast inference times, they come at the potential risk of limited generalisation and robustness for unseen data, in particular when trained with limited supervision. In this work, we demonstrate that those issues can be overcome by using an end-to-end differentiable displacement optimisation. Our method involves a trainable feature backbone, a correlation layer that evaluates a large range of displacement options simultaneously and a differentiable regularisation module that ensures smooth and plausible deformation. In extensive experiments on public and private ultrasound datasets with very sparse ground truth annotation the method shows better generalisation abilities and overall accuracy than a VoxelMorph network with the same feature backbone, while being two times faster at inference.

**Keywords:** ultrasound; image registration; deep learning; discrete optimisation

**Citation:** Heinrich, M.P.; Siebert, H.; Graf, L.; Misckewitz, S. and Hansen, L. Ultrasound registration using end-to-end differentiable optimisation. *Sensors* **2022**, *1*, 0. https://doi.org/

Received: December 31, 2022
Accepted: xx xx, 2023
Published: xx xx, 2023

## 1. Introduction

### 1.1. Motivation

Reliable realtime registration of ultrasound sequences can enable numerous practical applications in medical imaging. Tumours or organs-at-risk can be monitored and tracked in realtime to avoid unnecessary harm during radiotherapy or heat-based ablations [1], needle placement for biopsies or drug delivery can be assisted through image-guided navigation [2]. Image registration is also vital for freehand 3D reconstruction from manual ultrasound sweeps [3]. Diagnostic tasks, e.g. compression ultrasound for deep vein thrombosis (DVT) detection, that rely on dynamic human-machine interactions may be performed by frontline medical personal with help of realtime learning-based image analysis instead of transferring patients to expert centres [4].

The most prominent previous study for ultrasound motion estimation in radiotherapy applications, the CLUST 2007 MICCAI challenge [1], has focussed on tracking very few (four or less) pre-defined target anatomies over a period of time with mainly regular motion patterns. When adapting such algorithms to new ultrasound registration tasks, e.g. the evaluation of vein compression during realtime guidance for inexperienced users for DVT diagnosis, the rapid deformations and unpredictable motion can deteriorate the quality of trackers that rely on periodic motion. General purpose deep learning registration networks, such as PWC-Net [5] and VoxelMorph [6] may alleviate these problems and extend the applicability of ultrasound registration to new applications. However, they are prone to overfitting when using sparse supervision.

### *1.2. Related Work*

The comprehensive medical multi-task medical registration challenge Learn2Reg [7] has shown that nearly all state-of-the-art deep learning based registration tools, e.g. LapIRN [8] require strong supervision through densely annotated segmentation masks or keypoints. The classical optimisation-based approach of MEVIS [9] excelled at CT-based breathing motion estimation for which no manual expert supervision was available, but at the cost of very long run times of over a minute. PDD-Net [10] with metric supervision achieved the best learning-based results on multimodal ultrasound registration by employing mean-field inference, an idea borrowed from discrete optimisation, to regularise displacements.

More specialised template tracking approaches, which rely on the presence of an annotated image in the first frame (template) include COSD-CNN (Cascaded one-shot deformable CNN) [11]. During training template and instance images are processed by an unsupervised strategy to train a cross-correlation model that can roughly track the template. During inference a narrower crop is selected based on the manual annotation and a one-shot deformable convolution module is used to fit the appearance transformation of the target structures in a self-adaptive fashion. Similarly, [12] incorporate an on-line learning of a supporter model that captures the coupling of motion between image features, which are potentially found useful for predicting the target positions, and can be individually tracked. Further works, including [13], [14] and [15] aim to more explicitly incorporate temporal motion information through Conv-LSTMs, PCA motion models and a GAN-based Markov-like net that incorporates transformer modules respectively.

While basing the ultrasound registration on online learning for template appearance and/or a particular periodic motion model, helps to achieve high benchmark accuracies for datasets that fulfil those requirements, those tracking approaches may fail when unexpected motion, drift or tissue compression is present when considering a wider range of clinical applications. Furthermore, probe movement is not directly compensated leading to additional sources of errors and external optical tracking devices are cumbersome in clinical practice.

Hence, there is still a need for a general purpose dense registration tool akin to the approaches that excelled at Learn2Reg, but with the exception of being trained with very sparse annotations. To alleviate the limited availability of expert annotations a knowledge distillation framework for ultrasound registration was explored in [16], where a light-weight PDD-net was supervised by a PWC-net model trained on millions of natural image sequences. Unfortunately, the domain gap between popular computer vision datasets to clinical ultrasound appears to be too large for such approaches to yield high accuracy.

### *1.3. Contributions:*

In this work, we propose a combination of a straightforward feature backbone together with a novel differentiable convex optimisation layer that enables robust and accurate displacement predictions. This paper extends our original conference paper presented in [17] and an accepted abstract paper [18] where a preliminary version of our differentiable optimisation module was introduced, but only applied in a limited fashion to weigh and mix input channels from pre-trained or hand-crafted feature extractors or only used with a limited evaluation on dense segmentations respectively. Here, an end-to-end learning of all convolutional layers is proposed with a particular focus on learning with sparse supervision, i.e. few manually tracked points in ultrasound sequences.

Our method works very robustly **without online training, without a-priori knowledge of templates or cropping regions and without any temporal analysis** (ie. on a frame-by-frame basis). A direct comparison with tracking algorithms is therefore not purposeful, but instead we evaluate the concept in comparison to other learning-based general purpose approaches including VoxelMorph and fast optimisation with handcrafted features.

Here, a number of data augmentation and unsupervised loss strategies are explored to help cope with limited variability and sparse labels in the training data.

It is very fast with more than 370 fps (or less than 3 ms per frame on a server GPU), which will subsequently enable realtime use on mobile edge devices. The comprehensive experimental validation includes two datasets: 10 sequences of ETHZ dataset from the public CLUST challenge [1] with 4284 image frames in total that comprises breathing motion of the liver and a private vein compression dataset with several hundred sequences (each with 21 frames) that are indicative for non-invasive DVT diagnosis in the groin [4].

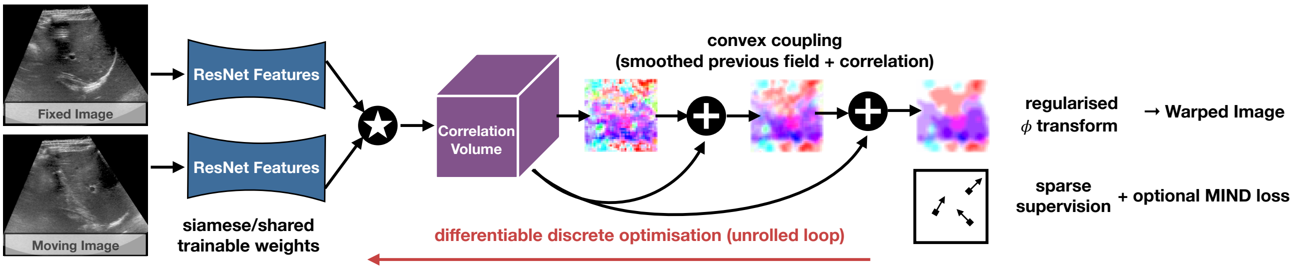
Fixed Image
Moving Image
ResNet Features
ResNet Features
siamese/shared trainable weights
Correlation Volume
convex coupling (smoothed previous field + correlation)
regularised $\phi$ transform
→ Warped Image
sparse supervision
+ optional MIND loss
differentiable discrete optimisation (unrolled loop)

**Figure 1.** Concept of differentiable convex optimisation with ResNet feature backbone. The discrete correlation layer computes the dissimilarity across feature vectors of both images over a large search window in the moving scan. The optimum of the resulting correlation volume would yield a noisy deformation estimate that is smoothed. A coupling tensor that penalises deviations from the smooth estimate is added to the correlation in the following step yielding a refined smooth estimate. Through this intrinsic regularisation a very sparse supervision with 1-4 landmarks is sufficient.

## 2. Materials and Methods:

In the following we will describe firstly all methods that are devised for the task of learning-based dense ultrasound motion estimation. Secondly, we will introduce the datasets, ablation studies, experimental and implementation details.

The main idea of our method is borrowed from our prior work ConvexAdam that ranked first in the comprehensive Learn2Reg MICCAI challenge series (2020-2021) [17]. In order to perform large-deformation image registration it employs a fast GPU based implementation of the coupled convex optimisation [19] that approximates a globally optimal solution of a discretised cost function for a densely sampled transformation. In contrast to our previous work that was not aimed at realtime performance we omit the time-consuming Adam-based instance optimisation. ConvexAdam has one important limitation, in that it does not enable backpropagation through the coupled optimisation, which is based on a non-differentiable argmin selection. This prevents end-to-end training and hence requires pre-trained or hand-crafted features that cannot be adapted to the task at hand. In the subsequent sections we will describe how a shared feature extractor for both input images can be effectively learned using a differentiable approximation of the coupled convex optimisation. As mentioned before the dense correlation enables a more robust estimation but in computer vision the PWC-Net [5] uses this complex within a highly complex architectures with millions of trainable parameters. Training the PWC-Net hence requires a much larger training dataset with dense ground truth label, which is commonly not available in medical imaging.

Here we present for the first time, a method that incorporates a fully-trainable ResNet feature backbone within an end-to-end differentiable discrete optimisation strategy and is supervised with only sparse supervision and without segmentation masks.

### *2.1. Methods:*

The conceptual overview of our approach is shown in Fig. 1. Fixed and moving input grayscale images are fed into a siamese feature extractor (that comprises two identical ResNet18 networks with shared weights). The stride of the feature maps yields a four-fold spatial reduction, while the number of channels is increased to 64 dimensions. Next a large correlation window of 9×9 voxels (equating to 33×33 pixels in original resolution) is computed using a correlation function that computes point-wise squared sum of differences of feature vectors across fixed and moving images. Finally the proposed differentiable convex optimisation is employed to yield smooth displacements and can be robustly trained in an end-to-end fashion using sparse supervision.

#### 2.1.1. Feature backbone:

We employ the ResNet-18 model as feature backbone with the following moderate modifications. We remove the fourth block and prevent too aggressive downsampling by replacing the two-fold strides in the third block with normal convolutions (see Fig. 2. The number of trainable parameters is reduced from 11 to 1 million, while the number of computations is slightly higher with 5.74 vs 2.22 billion Flops per image. We chose this model to balance network depth - to gain expressive feature abstraction to deal with challenging imaging artefacts - and maintaining high enough image resolution for the subsequent sub-pixel motion estimation.

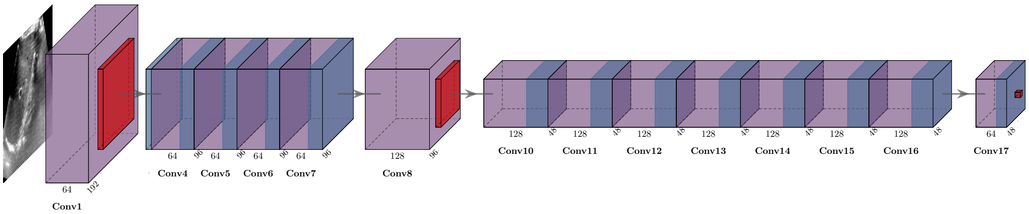

**Figure 2.** Visualisation of the modified ResNet-18 backbone for an example input image of 160×190, with reduced number of blocks, and importantly fewer downsampling steps to preserve spatial resolution of features. Residual connections are present in the model but not drawn.

#### 2.1.2. Differentiable convex optimisation:

Our proposed approach is straightforward and well designed for realtime estimation of large motion for ultrasound sequences with temporal dynamic and easier to implement because it does not require multiple warping steps, attention mechanisms or cascaded architectures.

The differentiable convex optimisation proposed here is an extension of a non-differentiable convex-discrete method [19], which aims to find a deformation field $\mathbf{u}$ by solving a cost function that simultaneously optimises smoothness (weighted by $\alpha$) and feature similarity

$$E(\mathbf{v},\mathbf{u}) = CorrVol(\mathbf{v}) + \theta(\mathbf{v}-\mathbf{u})^2 + \alpha|\nabla\mathbf{u}|^2 \tag{1}$$

Here the displacement correlation volume *CorrVol* is computed from the above mentioned ResNet features from fixed and moving image. Note that the correlation assigns a cost to each pixel and each potential displacement (here 81) in a search window that slides over the moving image. That means an intermediate deformation field $\mathbf{v}$ can be obtained through the softmin operation followed by an integral regression over all potential displacements. See also the concept of integral heatmap regression [20] for further details and motivation of this last step. To encourage smoothness of the resulting transformation an additional term is added to *CorrVol* by iterating a few times over Eq. 1. The hyperparameter $\theta$ is increased after each step to adapt the coupling of similarity and regularisation penalty, which converts the non-convex optimisation problem into two coupled convex ones. This enables fast convergence to a global optimum in the space of potential displacements (which are predefined for the given 9×9 window). The coupling term models parabolas rooted at the current displacement estimation, updated by choosing the minimal cost solution in each step and providing a robust regularisation. To solve the second part of Eq. 1 a spatial Gaussian smoothing to the previous displacement field is applied, implicitly solving Green's function and hence $\alpha$ is indirectly defined through the Gaussian kernel. Crucially we also incorporate a spatial B-spline smoothing filter at the end to estimate a plausible displacement field [17]. This complete module has no trainable parameters and requires less run time and memory than the feature extractor.

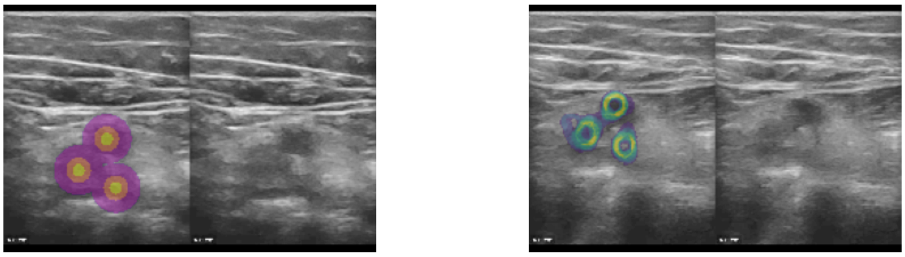

**Figure 3.** Concept to employ sparse landmarks as spatial transformer (warping) loss by generating pseudo segmentation labels that can be directly used in affine augmentation and Dice loss functions. Left: heatmap labels along with original CoCoAI frame. Right: 10th temporal frame in sequence with overlay of all warped heatmaps (middle label) averaged

### 2.1.3. Sparse Supervision:

As mentioned before, our proposed approach excels when limited and sparse ground truth annotations are available for training. We rely on only few (1-5) 2D landmarks in each frame of our training database. To enable the use of landmark supervision as drop-in replacement for spatial transformer loss functions, which rely on applying the full estimated deformation to a (multi-channel) label image, we create three concentric rings with increasing label number around each annotated 2D position (see. example image in Fig. **??**) yielding three-class heatmaps around the manual landmarks. Subsequently, a standard soft Dice loss can be employed and the use of affine augmentation during training becomes easier. To improve the coverage of the estimated transform and avoid overfitting of the features towards specific target anatomies that were annotated in the training dataset, we add the unsupervised MIND loss [21] with the following extension. Instead of using only 8 immediate neighbours to compute local contrast-invariant self-similarities, we sample 64 2D offsets for patch comparisons that provide a better context representation. Finally, a loss on the standard deviation of the Jacobian determinant of the estimated displacements is used, motivated by the fact that this metric was used in Learn2Reg [7] as quantitative smoothness measure.

### 2.1.4. Extended ResNet-VoxelMorph baseline:

The second method, which we compare as baseline and state-of-the-art benchmark is an extension of the simplistic U-Net proposed in the VoxelMorph framework [6]. Different to the original version we do not feed the images directly into the VoxelMorph module, but employ the same ResNet-18 backbone as feature extractor. This is hence more similar to the successful approaches that proposed siamese tracking networks for ultrasound applications, e.g. [22]. The final convolutional layer regresses a two channel displacement field, which is upsampled to the original image dimension using a B-spline function for improved smoothness, efficiently implemented as consecutive average pooling layers without stride (following the theory of recursive cardinal splines as in [23]).

## *2.2. Datasets:*

### 2.2.1. ETHZ:

The public CLUST dataset [1] comprises several 2D ultrasound sequences of healthy volunteers freely breathing with multiple different scanners and transducers [1]. Here, we focus on ten sequences of several hundred frames each from the ETHZ dataset (see

[1] https://clust.ethz.ch/data.html

Fig. 4). The temporal resolution is 15 Hz but higher frequencies are easily achievable with newer ultrasound probes. The image dimensions are approx. 400x400 pixels with a square resolution of 0.4mm. Either one or two anatomical landmarks are manually annotated by experts along each sequence to serve as both training objective and for cross-validation of the test accuracy. About a tenth of frames are labeled by three annotators to compute the inter-rater variance. The expected motion can reach several centimetres in each direction, i.e. motion of 16 pixels or more has to be estimated. So far nearly all state-of-the-art approaches used either temporal consistency (often coupled with online learning or fine-tuning) or restricted the search to a predefined template region. In our experiments the whole image is used without cropping and bilinearly downsampled to 160×192 pixels.

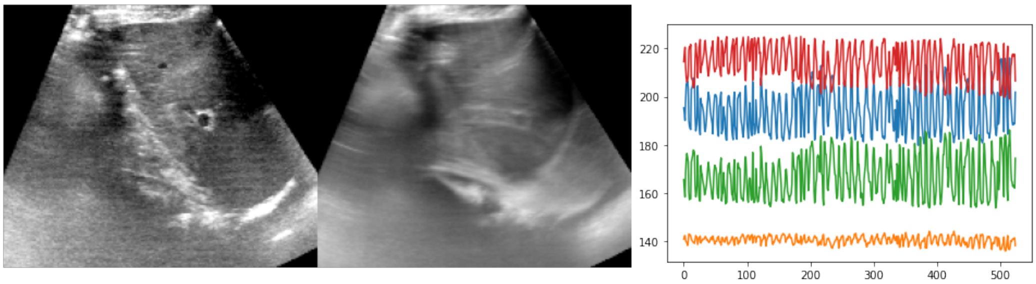

**Figure 4.** Example of reference frame and temporal average across one sequence from the ETHZ dataset along with a plot of the observed motion based on expert annotations (xy-coordinates for two landmarks)

For many clinical applications, not only the location but the precise propagation of the contour or segmentation of an anatomical structure is required in each frame. In image-guided radiotherapy organs-at-risk or a targeted tumour could deform during patient motion. In our second use case, the realtime guidance for inexperienced clinicians to perform accurate compression ultrasound analysis for deep vein thrombosis is considered. Here, the area of the investigated vein has to be monitored over the course of a manual compression to exclude a possible blockage that would point to an increased risk of a thrombosis clot and is potentially life threatening. During the examination the vein has to kept with the field of view despite very strong deformations, making the image interpretation particularly challenging. Providing a realtime guidance through overlaying an automatically propagated segmentation of the contour can be of great practical help and avoid user-dependent diagnostic errors [4]. Previous work relied on instance segmentation in each individual frame of the ultrasound video, which can lead to temporal instabilities, e.g. rapid switching between vein and artery label, sudden drop-outs (missing to segment a structure) and inaccuracies during compression. This will hence reduce user friendliness and high prevent clinical adoption. Therefore robust and fast image registration methods can help to improve the temporal continuity in the provided segmentations.

### 2.2.2. CoCoAI

The second dataset, which is part of the *Cooperative and Communicating AI methods for medical image-guided diagnostics* (CoCoAI) project [2], comprises 338 image sequences from handheld ultrasound showing one to five leg veins and arteries of healthy volunteers. All sequences show a temporal view inside the leg at a diagnostically relevant anatomical landmark, where a vein is compressed by pressure applied on the probe, either to full disappearance or in case of a thrombus to the size of the occlusion. We cropped all sequences to a common length of 21 frames and image dimensions of 160x160 pixels. While expert pixel-level segmentation have been curated to measure registration accuracy, we only use a simpler centre point annotation of each vessel to train our algorithms. This is much more

[2] https://www.imi.uni-luebeck.de/en/research/p46-cocoai.html

scalable for the generation of larger datasets that e.g. cover further anatomical positions in future work. The compression also moves the background structures in a non-uniform pattern and causes changes in contrast and ambiguous vessel boundaries, making this a particularly hard registration problem.

## 3. Experiments:

We perform a large number of experiments to evaluate the performance gains that result from replacing the popular VoxelMorph network for displacement prediction with our novel differentiable convex optimisation module.

Different to the participants of CLUST [1] we do not address the more restricted task of tracking a known template sequentially over time, but treat the alignment of the ETHZ dataset as a dense registration problem where each pair within a sequence has to be registered without any temporal information - **ETHZ dense unordered**. This adds challenges but also yields practical advantages: First, since we register any pair of frames from a long temporal sequence regardless of any specific ordering, our approach cannot learn to overfit on a particular breathing motion pattern and is therefore more robust against sudden changes and motion drift. Second, because our method predicts dense deformations, the template does not have to be known a priori and due to the unsupervised MIND loss all structures (including e.g. organs-at-risk) are registered accurately. For completeness we also evaluate our methods at inference time for a sequential ordering of frames **ETHZ dense sequential**, which is slightly easier since the average misalignment is smaller and should result in higher accuracy, despite the fact that we still do not use temporal guidance. In each experiment, we perform a 5-fold cross validation with 8 training and 2 test sequences and ensure that the same patient is not at the same time in both sets (i.e. ETHZ-01-1 and ETHZ-01-2 have to be in the same set).

For the **CoCoAI** dataset for vein compression ultrasound [], we perform the same setting in which all images from one sequence are randomly drawn in pairs during training to avoid overfitting on simple incremental motion and deformation. The estimation of a dense displacement field is required to accurately not only detect translational but also shape transformations. As outlined before we extract a three-class heatmap around the manual landmarks to define our sparse loss. Different to ETHZ we employ the Dice overlap based on 6'888 pixel-wise annotated frames with two classes (vein and artery) for evaluating the registration quality at test time. Note that this is more challenging since no supervision (in form of pixelwise segmentation labels) is provided for training and the models need to generalise well enough to align the object boundaries nevertheless. The sequences show substantial deformation due to compression with average initial Dice values of only 66%.

### *3.1. Implementation details:*

All methods and experiments are implemented using pytorch v1.10 (Cuda Version 11.5) and models were trained and evaluated with a single Nvidia RTX A4000. Each training was run for 3000 iterations with a batch-size of 8, that means 4 epochs for CoCoAI and 8 epochs for ETHZ. We used the Adam optimiser with initial learning rate of 0.001 and cosine and annealing scheduling with warm restarts every 375 iterations. When employing augmentation a random contrast variation field using a B-spline with 20 ×24 control points and a geometric affine transformation with random standard deviation of 0.15 was used. MIND features were multiplied by 10 and used with an MSE loss, a soft-Dice loss for landmark heatmaps with $\epsilon = 10^{-8}$ and a weighting of 2 as well as a Jacobian determinant standard deviation loss with weighting 0.5 was used as regularisation. The random augmentation and Jacobian loss were phased in with a sigmoid ramp-up to avoid early underfitting in particular for the VoxelMorph models. The regression output of VoxelMorph was obtained using a final hyperbolic tangent activation with multiplication of 0.2. For the differentiable convex optimisation we employed a grid-spacing of 4 and $\alpha = 20$ for the softmin operator. The coupling term $\theta$ was set to $(0, 0.3, 1, 3, 10)$ and we found that employing only one iteration during training and five at inference gave the best performance.

To replicate our results for the public data and build upon the method we released open-source code at https://github.com/multimodallearning/differentiableConvex.

### *3.2. Ablation Studies:*

We perform numerous ablation studies on the **ETHZ** dataset to evaluate the influence of each of our contributions.

**MIND+convex:** First, we reimplement a version of [14] which reached competitive results in the CLUST challenge using handcrafted MIND features together with a PCA motion model and block-matching optimisation. Since we are here interested in a general purpose registration, we do not consider model-based or temporal regularisation but use only the proposed coupled-convex discrete optimisation with MIND features. This approach requires no training and is therefore directly employable for inference on unseen data.

**VoxelMorph:** Second, we use a standard VoxelMorph model [6] that uses a direct concatenation of the two images as input and estimates the displacements using a U-Net. We use four levels and 64 channels as hyperparameters and an input size of half the image resolution. To improve smoothness of the transformations and ease learning with sparse supervision we append a B-spline deformation model (see also Sec. 2.1.4). This constitutes the fasted model.

**ResNet+VoxelMorph:** Next, we extend the VoxelMorph model by employing the ResNet-18 feature backbone described in Sec. 2.1.1. This enables a separate processing of both inputs in two streams and hence enables a higher abstraction of features before concatenating them as input to the U-Net for displacement estimation.

**ResNet+convex:** Finally, we evaluate the proposed approach, which balances complexity between the former two, because the differentiable correlation module is leaner than the U-Net and extends upon the first approach by introducing end-to-end trainable features. It also trains fastest among all learning based methods.

**Augmentation:** In addition to the different choices in network design, we also evaluate the influence of input data augmentation. Due to the large inter-sequence variation in the ETHZ dataset and the small number of scans (despite a great number of frames per sequence) overfitting can become a serious problem. By adding both geometric affine transformations that are applied to both input images and label maps the same way and random field intensity augmentations the models can be made more generalisable.

**MIND loss:** While the sparse landmark annotations can already guide the models to register the most relevant features the remaining parts of the images could remain misaligned. Hence, the influence of the unsupervised MIND loss that can be evaluated in the absence of spatially close landmarks is also of interest.

**Capture range of correlation layer:** We explore whether the robustness of our approach can be further improved by using a larger window of potential displacements within the correlation layer and also set this to 13×13 (roughly twice as many displacements). We do not retrain the models that were optimised for 9×9.

## 4. Results

The quantitative results show large difference across the aforementioned methods. Considering the more challenging unordered pairwise evaluation of frames we see a reduction in target registration error (TRE) of 2.56 mm or 48% of our best approach (without MIND loss and capture range 13×13) compared to the baseline VoxelMorph. Even against the extended baseline with the same ResNet backbone our differentiable convex optimisation an improvement of 1.97 mm or 42% can be seen. When considering the 90% percentile of errors our approach reaches 1.83 mm, which is 49% lower than the VoxelMorph baseline. The classic non-learning approach that directly employs MIND features with convex optimisation - and not only indirectly through the loss function - comes relatively close to our approach (+0.22 mm) based on the lower 90% percentile and is 24% worse for the full set of landmarks. This indicates that the sparsely supervised training can indeed improve the robustness of the motion estimation. A sequential evaluation, where a single central frame is defined as template and all others are aligned towards thus one all results improve as expected and the trend is similar in that our approach is at least 47% superior to VoxelMorph.

Considering the ablation studies, the augmentation is found to be of clear importance - with more than 35% increase of error when omitting it and removing the MIND loss leads to a slight decrease in TRE for our approach, but a very substantial increase of 31% for the ResNet+VoxelMorph. This shows that VoxelMorph struggles particularly strongly for sparse supervision, while the intrinsic regularisation of our approach is more stable. The increased capture range of 13×13 lead to a moderate improvement of 5%, which are most likely found for very challenging large deformations (hence the lower 90th percentile error does not benefit).

**Table 1.** Quantitative evaluation of five-fold cross-validation of all approaches on the **ETHZ dataset**. Target registration error (TRE) is shown as average in mm as well as the 90th percentile (in brackets) based on initial misalignment. Best results per category are set in bold. Our approach is superior to all compared methods in terms of accuracy. It is much faster than the second and third best approach and requires the least amount of training time (*except the untrained MIND variant). Note the first three columns all show the **unordered frame pairs** setup.

| **TRE in mm** | **w/o augment** | **w/o MIND** | **augment + MIND** | **best sequential** | **train time** | **inference speed** |
|---|---|---|---|---|---|---|
| initial | | | 6.84<br>(5.21) | 5.04<br>(3.92) | | |
| MIND+convex(13x13) | | | 3.59<br>(2.05) | 1.99<br>(1.40) | **0 min** | 92 fps |
| VoxelMorph | | | 5.28<br>(3.59) | 3.34<br>(2.27) | 12 min | **470 fps** |
| ResNet+VoxelMorph | | 6.14<br>(4.53) | 4.69<br>(3.04) | 2.80<br>(1.91) | 18 min | 280 fps |
| ResNet+convex(9x9) | 3.35<br>(2.11) | 2.86<br>**(1.62)** | 3.09<br>(1.72) | 1.54<br>**(1.21)** | 9 min | 305 fps |
| **ResNet+convex(13x13)** | 3.72<br>(2.64) | **2.72**<br>(1.83) | 2.87<br>(1.87) | **1.48**<br>(1.28) | 9 min | 210 fps |

**Computational efficiency:** The employed server GPU (RTX A4000) has a theoretical FP32 peak throughput of 38.35 TFlops when using mainly Tensor Core operations (counting multiply and add as two operations) and our best model reaches 210 fps. This indicates a moderate utilisation of ≈17% of the peak performance. When translating this to a mobile device, e.g. a current iPhone with 3rd generation Neural Engine and 15.8TFlops we expect 90 fps at FP16 precision, which would enable seamless integration into a user-friendly realtime guidance system.

Due to this efficiency, also the training is very fast with 9 minutes per fold in our 5-fold cross validation. The UNet of VoxelMorph was trained twice as long since it requires longer to converge.

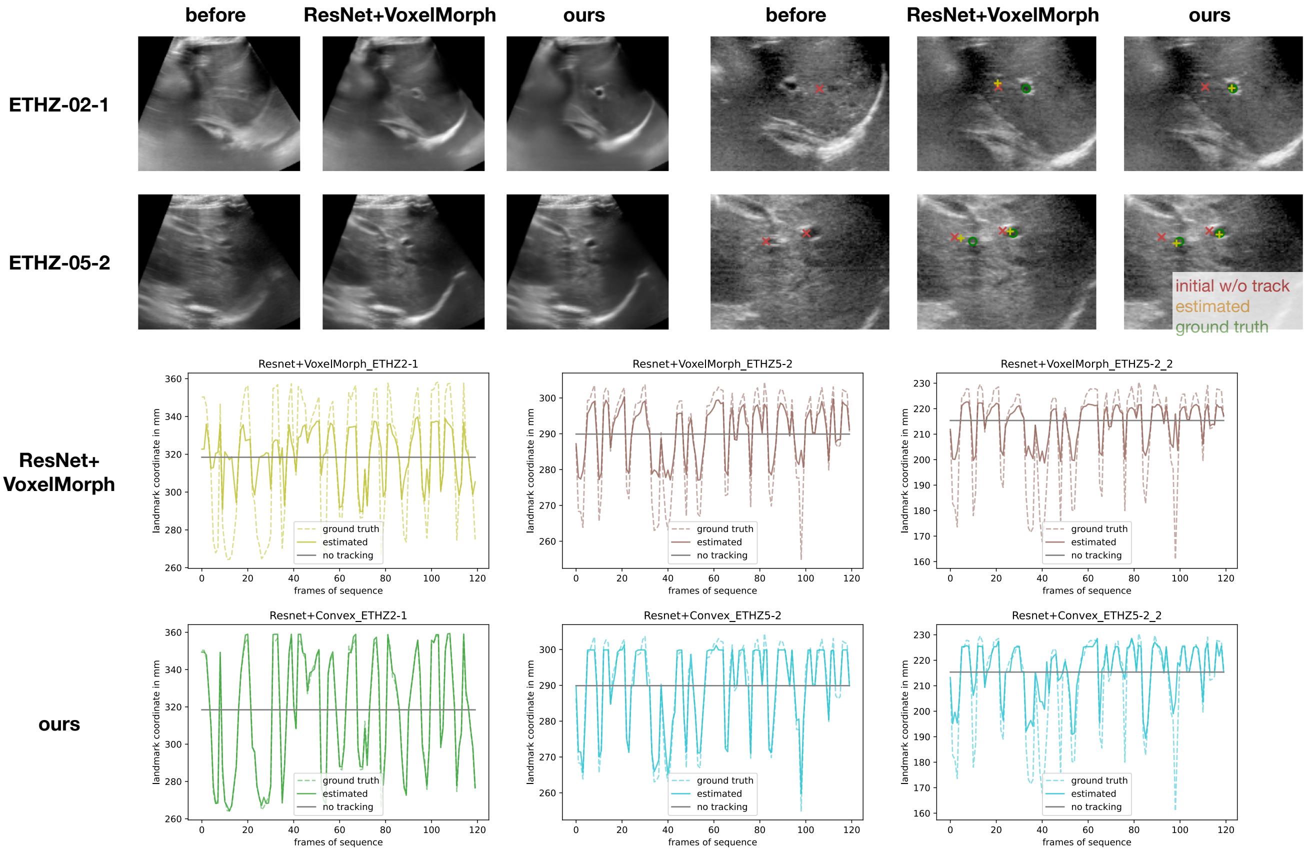

**Figure 5.** Comprehensive overview of selected qualitative results for the ETHZ dataset comparing the best VoxelMorph method with our proposed method. The averaged warped images (over 120 frames) clearly demonstrate a very good motion compensation and sharp mean of our approach, whereas VoxelMorph's result shows residual motion blur. The plots of tracking landmarks without temporal information show that VoxelMorph fails to capture larger displacement, while the convex optimisation excels at sequences ETHZ-02-1 and ETHZ-05-2 for the first landmark and only slightly underestimates the sudden shifts in the second landmark of ETHZ-05-2.

Finally, we also evaluate the two top performing models (from our proposed method and the VoxelMorph baselines) with one another on the **CoCoAI dataset**. As described above the challenges of this dataset arise from strong deformation during vein compression, limited contrast of vessels and the potential of confusing artery and vein. Both models are trained with sparse supervision of only the centre points of each vessel across the 21 considered frames. For evaluation only we used segmentation overlap by propagating the manual multi-class annotation between a fixed and moving frame using the estimated displacments. Tab. 2 shows the quantitative results for training on 300 sequences and testing on 38 held-out sequences. It becomes clear that the advantages seen in the previous experiments are also be supported by a second dataset and our approach outperforms VoxelMorph by a large margin of 7.3%points and reaches a lower standard deviation. In addition, we also train an nnUNet [24] for individual frame segmentation. This requires much more cumbersome annotation but could potentially help for difficult image quality at boundaries. Instead of aligning pairs of images the validation model is run in feed-forward mode and a direct Dice overlap can be obtained. Despite a slower computational speed (approx. 140 fps vs ours with 210 fps) the segmentation approach is substantially less robust with 8%points lower Dice scores. That means the registration-based guidance is likely more user-friendly and can improve the adoption in clinical practice.

**Table 2.** Quantitative evaluation for the **CoCoAI dataset**. We compare the quality of segmentation propagation using motion estimation our proposed registration approach against the best ResNet-VoxelMorph baseline. Both are trained with only centre point annotations of vessels. In addition we also evaluated a segmentation model (nnUNet) that was trained on the full pixelwise annotation, but performs substantially worse, highlighting the advantages of registration for user-guidance.

| Dice overlap in % | vein | artery | overall |
|---|---|---|---|
| initial | 66.0±12.5 | 66.8±15.4 | 66.4±13.4 |
| nnUNet | 72.06 | 78.96 | 75.51 |
| ResNet+VoxelMorph | 74.9±11.6 | 77.6±13.4 | 76.3±12.0 |
| **ResNet+convex(13x13)** | **81.5±7.4** | **85.7±7.1** | **83.6±6.2** |

The visual results are shown in Fig. 6 along with a plot of the cumulative distribution of Dice accuracies across all 798 validation samples. The temporal averaging of all warped frames in Fig. 6 b) should ideally be a sharp mean, as expected this is not fulfilled without registration - but also the VoxelMorph result appears blurry while our approach finds consistent mappings of images and labels. When considering the plausibility of the deformed segmentation it appears that in contrast to our approach is VoxelMorph not able to extrapolate from the sparse (centre) annotations to dense displacement that align all objects with their boundaries as well as the proposed differentiable convex optimisation.

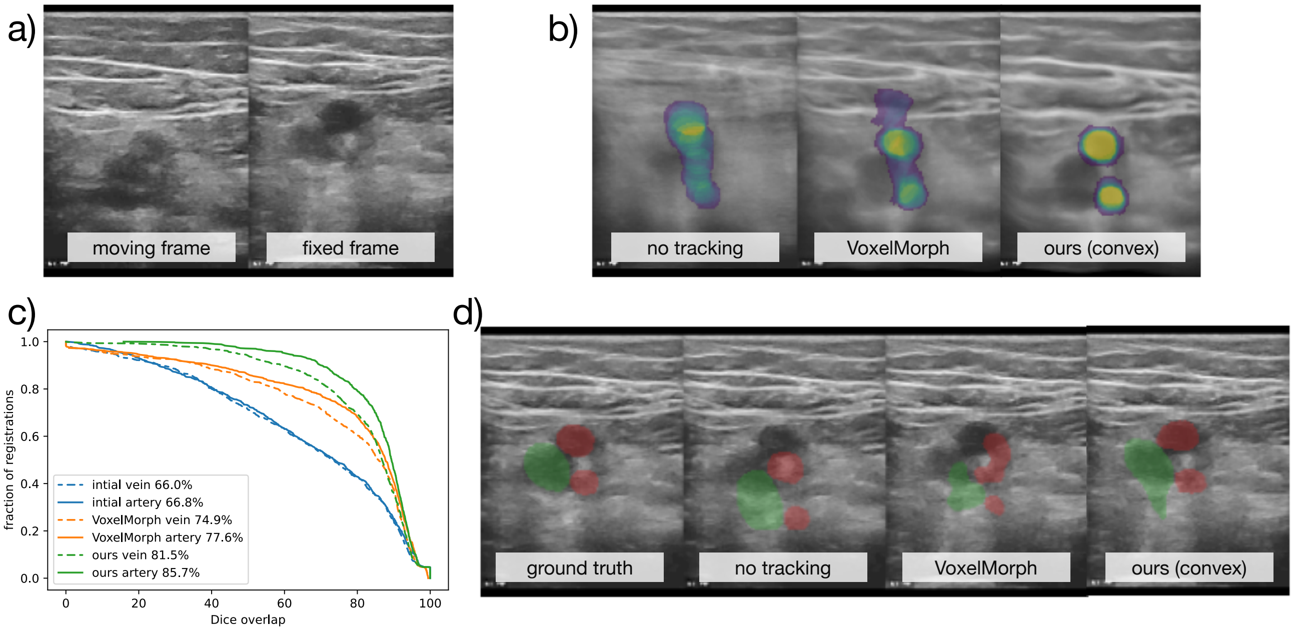

**Figure 6.** Qualitative results for the **CoCoAI** dataset. a) shows two original frames from a compression sequence, b) across all registrations with same fixed frame the label probabilities for arteries after registration are averaged - ideally two non-blurry circles should be seen, c) the cumulative accuracy plot for all compared methods shows the advantage of our approach, d) an example segmentation propagation (vein in green, arteries in red), shows that VoxelMorph's deformation is not as plausbile as ours.

## 5. Discussion

In this work we have presented a powerful methodological alternative for current state-of-the-art dense medical image registration techniques by combining our new differentiable convex optimisation module with an adapted ResNet-18 backbone for feature extraction. Contrary to our prior work [17] no dense segmentations are required to pre-train a feature extractor and hence our new approach can be learned in an end-to-end fashion from scratch on a wider range of problems. For inter-frame ultrasound registration, where the number of structures can vary across sequences and often only sparse annotations of few centre points are available, the intrinsic regularising effect of the differentiable discrete optimisation that incrementally adds a stronger penalty for non-smooth displacement fields to the correlation volume. Thus more robust and accurate registration results are obtained. We do not assume any prior knowledge of specific targets or templates or periodic motion and therefore the algorithm can be seen as general purpose registration tool.

We employ two challenging datasets with thousands of ultrasound frames and perform extensive comparison experiments to baseline networks, classical methods and ablations of the proposed method. Our approach excels in both applications and provides very low computational complexity with inference speeds of over 200 frames per second. A direct application on mobile devices without delay for user-friendly automatic guidance through our model can be obtained when assuming a current flagship phone or tablet.

In future, we would like to explore an extension of our method for a coarse-to-fine (multi-stage) motion estimation that could further improve precision. Pre-training with unsupervised data (through our image-based MIND loss) could further improve robustness across scanner types and motion patterns. Due to the flexibility of the models, we could also extend the evaluation to further modalities.

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

**Funding:** This research was funded by the German Ministry of Education and Research (BMBF) under grant number 01GP1908.

**Institutional Review Board Statement:** Ethical review and approval were waived for this study due to the fact that only healthy volunteers participated in the image acquisition and the study was non-invasive. I.e. no ionising radiation, nor contrast medium was used.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Please refer to suggested Data Availability Statements in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics. If the study did not report any data, you might add "Not applicable" here.

**Acknowledgments:** We thank all the organisers and data providers of the CLUST challenge in particular Valeria De Luca and Christine Tanner for making available the open source dataset.

## References

1. De Luca, V.; Banerjee, J.; Hallack, A.; Kondo, S.; Makhinya, M.; Nouri, D.; Royer, L.; Cifor, A.; Dardenne, G.; Goksel, O.; et al. Evaluation of 2D and 3D ultrasound tracking algorithms and impact on ultrasound-guided liver radiotherapy margins. *Medical physics* **2018**, *45*, 4986–5003.
2. De Silva, T.; Fenster, A.; Cool, D.W.; Gardi, L.; Romagnoli, C.; Samarabandu, J.; Ward, A.D. 2D-3D rigid registration to compensate for prostate motion during 3D TRUS-guided biopsy. *Medical physics* **2013**, *40*, 022904.
3. Prevost, R.; Salehi, M.; Jagoda, S.; Kumar, N.; Sprung, J.; Ladikos, A.; Bauer, R.; Zettinig, O.; Wein, W. 3D freehand ultrasound without external tracking using deep learning. *Medical image analysis* **2018**, *48*, 187–202.
4. Kainz, B.; Heinrich, M.P.; Makropoulos, A.; Oppenheimer, J.; Mandegaran, R.; Sankar, S.; Deane, C.; Mischkewitz, S.; Al-Noor, F.; Rawdin, A.C.; et al. Non-invasive diagnosis of deep vein thrombosis from ultrasound imaging with machine learning. *npj Digital Medicine* **2021**, *4*, 1–18.
5. Sun, D.; Yang, X.; Liu, M.Y.; Kautz, J. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8934–8943.
6. Balakrishnan, G.; Zhao, A.; Sabuncu, M.R.; Guttag, J.; Dalca, A.V. VoxelMorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging* **2019**, *38*, 1788–1800.
7. Hering, A.; Hansen, L.; Mok, T.C.; Chung, A.C.; 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. *IEEE Transactions on Medical Imaging* **2022**.

8. Mok, T.C.; Chung, A. Large deformation diffeomorphic image registration with laplacian pyramid networks. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020, pp. 211–221.
9. Häger, S.; Heldmann, S.; Hering, A.; Kuckertz, S.; Lange, A. Variable fraunhofer MEVIS RegLib comprehensively applied to Learn2Reg challenge. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2021, pp. 74–79.
10. 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. Springer, 2019, pp. 50–58.
11. Liu, F.; Liu, D.; Tian, J.; Xie, X.; Yang, X.; Wang, K. Cascaded one-shot deformable convolutional neural networks: Developing a deep learning model for respiratory motion estimation in ultrasound sequences. *Medical image analysis* **2020**, *65*, 101793.
12. Ozkan, E.; Tanner, C.; Kastelic, M.; Mattausch, O.; Makhinya, M.; Goksel, O. Robust motion tracking in liver from 2D ultrasound images using supporters. *International journal of computer assisted radiology and surgery* **2017**, *12*, 941–950.
13. Huang, P.; Yu, G.; Lu, H.; Liu, D.; Xing, L.; Yin, Y.; Kovalchuk, N.; Xing, L.; Li, D. Attention-aware fully convolutional neural network with convolutional long short-term memory network for ultrasound-based motion tracking. *Medical physics* **2019**, *46*, 2275–2285.
14. Ha, I.Y.; Wilms, M.; Handels, H.; Heinrich, M.P. Model-based sparse-to-dense image registration for realtime respiratory motion estimation in image-guided interventions. *IEEE Transactions on Biomedical Engineering* **2018**, *66*, 302–310.
15. Dai, X.; Lei, Y.; Roper, J.; Chen, Y.; Bradley, J.D.; Curran, W.J.; Liu, T.; Yang, X. Deep learning-based motion tracking using ultrasound images. *Medical Physics* **2021**, *48*, 7747–7756.
16. Nicke, T.; Graf, L.; Lauri, M.; Mischkewitz, S.; Frintrop, S.; Heinrich, M.P. Realtime Optical Flow Estimation on Vein and Artery Ultrasound Sequences Based on Knowledge-Distillation. In Proceedings of the Biomedical Image Registration; Springer International Publishing: Cham, 2022; pp. 134–143.
17. Siebert, H.; Heinrich, M.P. Learn to Fuse Input Features for Large-Deformation Registration with Differentiable Convex-Discrete Optimisation. In Proceedings of the Biomedical Image Registration; Springer International Publishing: Cham, 2022; pp. 119–123.
18. Graf, L.F.; Siebert, H.; Mischkewitz, S.; Heinrich, M.P. Highly accurate deep registration networks for large deformation estimation in compression ultrasound. In Proceedings of the SPIE Medical Imaging, 2023, pp. 1–4. accepted in print.
19. Heinrich, M.P.; Papież, B.W.; Schnabel, J.A.; Handels, H. Non-parametric discrete registration with convex optimisation. In Proceedings of the International Workshop on Biomedical Image Registration. Springer, 2014, pp. 51–61.
20. Sun, X.; Xiao, B.; Wei, F.; Liang, S.; Wei, Y. Integral human pose regression. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 529–545.
21. Heinrich, M.P.; Jenkinson, M.; Bhushan, M.; Matin, T.; Gleeson, F.V.; Brady, M.; Schnabel, J.A. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical image analysis* **2012**, *16*, 1423–1435.
22. Bharadwaj, S.; Prasad, S.; Almekkawy, M. An upgraded siamese neural network for motion tracking in ultrasound image sequences. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control* **2021**, *68*, 3515–3527.
23. Ha, I.Y.; Wilms, M.; Heinrich, M. Semantically guided large deformation estimation with deep networks. *Sensors* **2020**, *20*, 1392.
24. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **2021**, *18*, 203–211.