

Closing the Gap between Deep and Conventional Image Registration using Probabilistic Dense Displacement Networks

Mattias P. Heinrich¹[0000–0002–7489–1972]

Institute of Medical Informatics, Universität zu Lübeck
heinrich@imi.uni-luebeck.de, github.com/multimodallelearning/pdd_net

Abstract. Nonlinear image registration continues to be a fundamentally important tool in medical image analysis. Diagnostic tasks, image-guided surgery and radiotherapy as well as motion analysis all rely heavily on accurate intra-patient alignment. Furthermore, inter-patient registration enables atlas-based segmentation or landmark localisation and shape analysis. When labelled scans are scarce and anatomical differences large, conventional registration has often remained superior to deep learning methods that have so far mainly dealt with relatively small or low-complexity deformations. We address this shortcoming by leveraging ideas from probabilistic dense displacement optimisation that has excelled in many registration tasks with large deformations. We propose to design a network with approximate min-convolutions and mean field inference for differentiable displacement regularisation within a discrete weakly-supervised registration setting. By employing these meaningful and theoretically proven constraints, our learnable registration algorithm contains very few trainable weights (primarily for feature extraction) and is easier to train with few labelled scans. It is very fast in training and inference and achieves state-of-the-art accuracies for the challenging inter-patient registration of abdominal CT outperforming previous deep learning approaches by 15% Dice overlap.

Keywords: registration · deep learning · probabilistic · abdominal

1 Introduction and Related Work

Conventional medical image registration mostly relies on iterative and multi-scale warping of a moving towards a fixed scan by minimising a dissimilarity metric together with a regularisation penalty. Deep learning based image registration (DLIR) aims to mimic this process by training a convolutional network that can predict the non-linear alignment function given two new unseen scans. Thus instead of multiple warping steps a single feed-forward transfer function has to be found using many convolution layers. The supervision for DLIR can be based on automatic or manual correspondences, semantic labels or intrinsic cost functions. It has immense potential for time-sensitive applications such as

image-guidance, fusion, tracking and shape analysis through multi-atlas registration. However, due to the large space of potential deformations that can map two corresponding anatomies onto one another, the problem is much less constrained than image segmentation and therefore remains an open challenge.

A number of approaches has been applied to brain registration [1, 17], which usually deals with localised deformations of few millimetres and for which huge labelled datasets ($\gg 100$ scans) exist. For other anatomies in the abdomen, the prostate or lungs, with shape variations of several centimetres, DLIR was mainly applied to less complex cases of intra-patient registration [6, 9]. For inhale-exhale lung registration the accuracy of DLIR is still inferior to conventional approaches: ≈ 2.5 mm in [13, 15] compared to < 1 mm in [12]. When training the state-of-the-art weakly-supervised DLIR approach Label-Reg [6] on abdominal CT [14] for inter-patient alignment, we reached an average Dice of only 42.7%, which is still substantially worse than the conventional NiftyReg algorithm [10] with a Dice of 56.1% and justifies further research.

Our hypothesis is that large and highly deformable transformations across different patients are difficult to model with a deep continuous regression network without resorting to complex multi-stage warping pipelines. Instead the use of discrete registration, which explores a large space of quantised displacements simultaneously, has been shown to capture abdominal and chest deformations more effectively [5, 12, 16] and can be realised with few or a single warping step. Unsurprisingly, discrete displacement settings have been explored in 2D vision for DLIR: namely the FlowNet-C [2]. A *correlation layer* (see Eq. 1 in [2]) is proposed that contains no trainable weights and computes a similarity metric of features from two images by shifting the moving image with a densely quantised displacement space (21×21 pixel offsets) yielding a 441-channel joint feature map. Next, a very large $441(+32) \times 256 \times 3 \times 3$ kernel is learned (followed by further convolutions) that disregards the explicit 4D geometry of the displacement space. Hence, the large number of optimisable parameters results in huge requirements of supervised training data. Extending this idea to 3D is very difficult as the dimensionality increases to 6D after dense correlation and has not been yet considered despite its benefits. Probabilistic and uncertainty modelling has been studied in DLIR, cf. [9, 17], but not in a discrete setting.

Contributions We propose a new learning model for DLIR that better leverages the advantages of probabilistic dense displacement sampling by introducing strong regularisation with differentiable constraints that explicitly considers the 6D nature of the problem. We hence decouple convolutional feature learning from the fitting of a spatial transformation using mean-field inference for regularisation [8, 18] and approximate min-convolutions [3] for computing inter-label compatibilities. Our feature extractor uses 3D deformable convolutions [4] and is very lightweight. To our knowledge this is the first approach that combines discrete DLIR with the differentiable use of mean-field regularisation. In contrast to previous work, our model requires fewer trainable weights, captures larger deformations and can be trained from few labelled scans to high accuracy. We

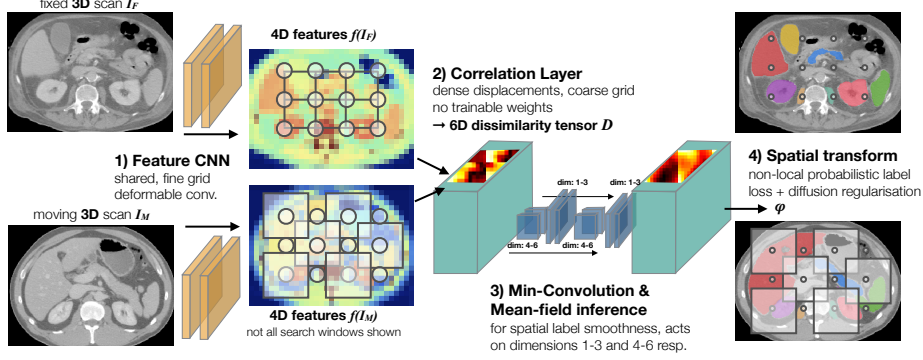


Fig. 1. Concept of probabilistic dense displacement network: 1) deformable convolution layers extract features for both fixed and moving image. 2) the correlation layer evaluates for each 3D grid point a dense displacement space yielding a 6D dissimilarity map. 3) spatial filters that promote smoothness act on dimensions 4-6 (min-convolutions) and dim. 1-3 (mean-field inference) in alternation. 4) the probabilistic transform distribution obtained using a softmax (over dim. 4-6) is used in a non-local label loss and converted to 3D displacements for a diffusion regularisation and to warp images.

also introduce a new non-local label loss for improved guidance instead of the more widely used spatial transformer based loss.

2 Methods

We aim to align a fixed I_F and moving I_M 3D scan by finding a spatial transformation ϕ based on a learned feature mapping f of I_F and I_M subject to constraints on the regularity of ϕ . In order to learn a suitable feature extraction that is invariant to noise and uninformative contrast-variations, we provide a supervisory label during training for both volumes ℓ_F and ℓ_M , for which $\ell_F \approx \phi \circ \ell_M$ should hold after registration. We define spatial coordinates as continuous variables $\mathbf{x} \in (-1, +1)^3$ and use trilinear interpolation to sample from discrete grids. ϕ is parameterised with a set of $\mathbf{k} \in |K| \in \mathbb{R}^3$ (a few thousands) control points on a coarser grid. The range of displacements \mathbf{d} is constrained to a discrete displacement space, with linear spacing e.g. $\mathcal{L} = q \cdot \{-1, -\frac{6}{7}, -\frac{5}{7}, \dots, +\frac{5}{7}, +\frac{6}{7}, +1\}^3$, where q is a scalar that defines the capture range and in our case $|\mathcal{L}|$ is 3375. The network model should predict a 6D tensor of displacement probabilities $K \in \mathbb{R}^3 \times \mathcal{L} \in \mathbb{R}^3$, where the sum over the dimensions 4-6 of \mathcal{L} for each control point is 1. The (inner) product of the probabilities with the displacements \mathcal{L} yields the weighted average of these probabilistic estimates to obtain 3D displacements for ϕ during inference.

1) Convolutional feature learning network: To learn a meaningful nonlinear mapping from input intensities to a dense feature volume (with $|c| = 16$ channels and a stride of 3), we employ the Obelisk approach [4], which comprises

a 3D deformable convolution with trainable offsets followed by a simple 1×1 MLP and captures spatial context very effectively. We extend the authors’ implementation by adding a normal $5 \times 5 \times 5$ convolution kernel with 4 channels prior to the Obelisk layer to also learn edge-like features. The network has 64 spatial filter offsets and in total 120k trainable parameters, shared for fixed and moving scan to yield $f(I_F)$ and $f(I_M)$.

2) Correlation layer for dense displacement dissimilarity: Given the feature representation of the first part, we aim to find a regularised displacement field that assigns a vector \mathbf{d} to every control point for a nonlinear transform $\varphi(\mathbf{k}) \leftarrow \mathbf{d}$ that maximises the (label) similarity between fixed and warped moving scan. As done in conventional discrete registration [5] and the correlation layer of [2], we perform a dense evaluation of a similarity metric over the displacement search space $\mathbf{d} \in \mathcal{L}$. The negated mean squared error (MSE) across the feature dimension c of learned descriptors is used to obtain the 6D tensor of dissimilarities $\mathcal{D}(\mathbf{k}, \mathbf{d}) = -\frac{1}{|c|} \sum_c (f_c(I_F)_{\mathbf{k}} - f_c(I_M)_{\mathbf{k}+\mathbf{d}})^2$. Different metrics such as the correlation coefficient could be employed. Due to the sparsity of the control points the displacement similarity evaluation requires less than 2 GFlops in our experiments. The capture range of displacements q is set to 0.4.

3) Regularisation using min-convolutions and mean-field inference: Since nonlinear registration is usually ill-posed, additional priors are used to keep deformations spatially smooth. In contrast to other work on DLIR, which in principle learn an unconstrained deformation and only enforce spatial smoothness as loss term, we propose to model regularisation constraints as part of the network architecture. A diffusion-like regularisation penalty for displacements based on their squared difference $\mathcal{R}(\mathbf{d}_i, \mathbf{d}_j) = \|\mathbf{d}_i - \mathbf{d}_j\|^2$ is often used in Markov random field (MRF) registration [3] and e.g. optimised with loopy belief propagation (LBP). [7] and [18] integrated smoothness constraints of graphical models into end-to-end learned segmentation networks. Since, LBP requires many iterations to yield an optimum and is hence not well suited as unrolled network layers, we use the fast mean-field inference (two iterations) used for discrete optimisation in [8]. It consists of two alternating steps: a label-compatibility transform that acts on spatial control points independently and a filter-based message passing implemented using average pooling layers with a stride of 1.

As noted in [3] the diffusion regularisation for a dense displacement space can be compute using min-convolutions with a lower envelope of parabolas rooted at the (3D) displacement offsets with heights equalling to the sum of dissimilarity term and the previous iteration of the mean-field inference. This lower envelope is not directly differentiable, but we can obtain a very accurate approximation using first, a min-pooling (with stride=1) that finds local minima in the cost tensor followed by two average pooling operations (with stride=1) that provide a quadratic smoothing. As shown with blue blocks in Fig. 1, the novel regularisation part of our approach comprises min- and average-pooling layers that act on the 3 displacement dimensions (min-convolution) followed by average filtering on the 3 spatial dimensions (mean-field inference). Before each operation,

scaling and bias factors $\alpha_1 - \alpha_6$ are introduced and optimised together with the feature layers during end-to-end training.

Probabilistic transform losses and label supervision: We can make further use of the probabilistic nature of our displacement sampling and specify our supervised label loss term based on a non-local means weighting [11]. I.e., we first convert the negated output of the regularisation part (scaled by α_6) into pseudo-probabilities using a softmax computed over the displacements. Next, one-hot representations of the moving segmentation are sampled at the same spatially displaced locations and these vectors are multiplied by the estimated probabilities to compute the label loss as MSE w.r.t. the ground truth (one-hot) segmentation. The continuous valued 3D displacement field φ is obtained by a weighted average of the probabilistic estimates multiplied with the displacement labels and followed by trilinear interpolation to the image resolution. A diffusion regularisation penalty over all 3 spatial gradients $\lambda \cdot (|\nabla \varphi_1|^2 + |\nabla \varphi_2|^2 + |\nabla \varphi_3|^3)$ of the displacement field is employed to encourage a smooth transform.











3 Experimental Validation

To demonstrate the ability of our method to capture very large deformations across different patients within the human abdomen, we performed experiments with a 3-fold cross validation on 10 contrast-enhanced 3D CT scans of the VIS-CERAL3 training data [14] with each nine anatomical structures manually segmented: ■ liver, ■ spleen, ■ pancreas, ■ gallbladder, ■ urinary bladder, ■ right kidney, ■ left kidney, ■ right psoas major muscle (psoas) and ■ left psoas (see Fig. 2). The images were resampled to isotropic voxel sizes of 1.5 mm^3 with dimensions of $233 \times 168 \times 286$ voxels and without any manual pre-alignment.

We compare our probabilistic dense displacement network (**pdd-net**)¹ with the two conventional algorithms **NiftyReg** [10] and **deeds** [5] that performed best in the inter-patient abdominal CT registration study of [16], a task not yet tackled by DLIR. NiftyReg was used with mutual information and a 5-level multi-resolution scheme to capture large deformations and has a run-time of 40-50 sec. Deeds was considered with a single scale dense displacement space (which takes about 4-6 sec) and then extended to three-levels of discrete optimisation (25-35 sec run-time). Next, we trained the weakly-supervised DLIR method **Label-Reg** [6] on our data (in >24 hours per fold). To reduce memory requirements below 32 GBytes, the resolution was reduced to 2.2 mm and the base channel number halved to 16. Further small adjustments were made to optimise for inter-patient training. We implemented a 3D extension of **FlowNet-C** [2] in pytorch with Obelisk feature extraction, a dense correlation layer and a regularisation network that has $|\mathcal{L}| = 3375$ input channels, comprises five 3D conv. layers with batch-norm and PReLU. It has 2 million weights and outputs a (non-probabilistic) 3D displacement field. In order to obtain meaningful results it was necessary to add a semantic segmentation loss to the intermediate output of the Obelisk layers. Our

¹ our code with all implementation details will be made publicly available.

Table 1. Quantitative comparative evaluation of cross-validation for 10 scans of the VISCERAL anatomy 3 dataset, based on 24 combinations of test scans not seen in training (numbers are Dice scores). Our method **pdd-net** outperforms the considered DLIR methods, Label-Reg and FlowNet-C, by a margin of about 15% and closes the gap to conventional methods, NiftyReg and deeds, for this task. Our ablation study shows the benefits of 1) the use of learned Obelisk features vs handcrafted SSC descriptors, 2) employing mean-field inference and 3) the use of our new non-local label loss. °Additionally a fast instance level optimisation was implemented for **pdd-net+inst..** *FlowNet-C is our 3D extension of [2] with Obelisk features and a trainable regularisation network. To compare: the Dice before registration was 30.0% on average.

Method	1) 2) 3)	         	∅ 9 lab.	std(Jac)	runtime
Label-Reg		71 51 7 5 38 53 59 45 55	42.7±5.5	0.58	4 sec.
FlowNet-C*	✓ ✓ ✓	73 45 9 7 40 48 53 49 52	41.8±7.2	0.34	0.13 sec.
pdd+SSC	✗ ✓ ✓	69 47 8 6 49 56 56 59 57	45.2±5.4	0.54	1.38 sec.
pdd w/o MF	✓ ✗ ✓	74 53 7 8 49 65 63 56 60	48.2±4.8	0.38	0.45 sec.
pdd w/o NL	✓ ✓ ✗	83 62 11 8 47 69 68 60 60	51.9±7.1	0.39	0.57 sec.
pdd-net	✓ ✓ ✓	84 62 13 12 57 76 70 69 68	56.7±6.0	0.40	0.57 sec.
pdd+inst.°	✓ ✓ ✓	83 66 18 14 58 77 74 68 68	58.4±5.9	0.26	0.71 sec.
deeds+SSC	1 level	72 50 14 13 51 54 58 62 60	48.0±6.8	0.67	4 sec.
deeds+SSC	3 level	78 62 18 19 60 71 67 70 69	57.0±8.2	0.27	25 sec.
NiftyReg NMI	5 level	77 58 19 27 56 70 65 67 66	56.1±18	1.30	42 sec.

proposed method employs the same feature learning part (with 200k parameters) but now uses min-convolutions, mean-field inference (no semantic guidance) and the non-local label loss, which adds only 6 trainable weights (and not 2 million). The influence of these three choices is analysed with an ablation study, where a replacement of Obelisk feature learning with handcraft self-similarity context features [5] is also considered. We use a diffusion regularisation weight of $\lambda = 1.5$ for control grids of size 32^3 throughout and trained our networks with Adam (learning rate of 0.01) for 1500 iterations in about 90 minutes and ≈ 16 GByte of GPU memory with checkpointing. We implemented an instance-wise gradient descent optimiser that refines the feed-forward predictions. [1] also used this idea, but in our case it is a hundred times faster (0.24 sec. vs 24 sec.), since we can directly operate on the pre-computed displacement probabilities and require no iterative back-propagation through the network.

4 Results and Discussions

The inference time of **pdd-net** is only 0.57 sec, yielding plausible displacement fields with a standard deviation of the Jacobian determinants of 0.40 and <1% folding voxels (negative Jacobians). Table 1 shows the average Dice scores across 24 registrations of the cross-validation, where no labelled training scan was used for any evaluated test registration. Our method outperforms the two compared DL approaches, **Label-Reg** and **FlowNet-C**, by a margin of about 15% points

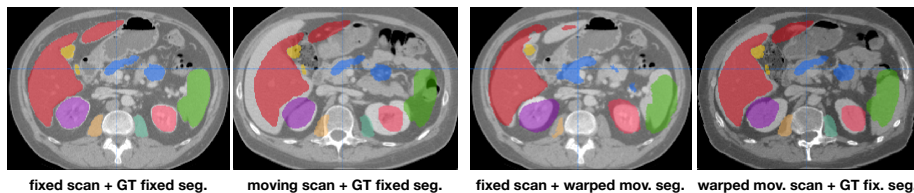


Fig. 2. Visual outcome of proposed **pdd-net** method to register two patients and transfer a segmentation (moderate example). Most organs have been very well aligned and also anatomies that are not labelled in training (stomach, vertebras) can be registered.

and achieves 56.7% Dice for this challenging inter-patient task with an initial alignment of only 30%. It is 10% better than a comparable setting of the conventional discrete registration **deeds** with one grid-level. In particular the labels ■, ■, ■, ■, ■, ■ and ■ are very well aligned. Our instance-wise (per scan-pair) optimisation requires 0.24 sec, reduces foldings (to less than 0.6%) and further increases the accuracy to 58.4%, which is above the level of the conventional multi-level registrations **deeds** and **NiftyReg**.

Comparing **deeds**+SSC with one grid-level to our variant **pdd**+SSC, which uses the same self-similarity features and only adapts the α parameters of the regularisation part, we get a similar accuracy and deformation complexity. This suggests that the proposed regularisation layers with min-convolutions and two mean-field inference steps can nearly match the capabilities of the full sequential MRF optimisation in [5]. Using weak supervision to learn features results in more than 20% increased Dice. The non-local loss term and our instance-wise fine-tuning, contribute further gains of 5% and 2% Dice overlap, respectively. The importance of the mean-field inference is clear, given the inferior quality of an unconstrained FlowNet-C with more trainable weights or our variant that only uses min-convolutions but no filtering in spatial domain. We achieve a more robust alignment quality (lower std.dev. of Dice) than conventional registration. Visual registration examples are shown in Fig. 2 and as surface-rendered video files in the supplementary material.

5 Conclusion

Our novel **pdd-net** combines probabilistic dense displacements with differentiable mean-field regularisation to achieve one-to-one accuracies of over 70% Dice for 7 larger anatomies for inter-patient abdominal CT registration. It outperforms the previous deep learning-based image registration (DLIR) methods, Label-Reg and FlowNet-C, by a margin of 15% points and can be robustly trained with few labelled scans. It closes the quality gap of DLIR to state-of-the-art conventional methods, exemplified by NiftyReg and **deeds**, while being extremely fast (0.5 sec). Our concept offers a clear new potential to enable the use of DLIR in image-guided interventions, diagnostics and atlas-based shape

analysis beyond the currently used pixel segmentation networks that lack geometric interpretability. Future work could yield further gains by using multiple alignment stages and a more adaptive sampling of control points.

References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. *IEEE Trans medical imaging* (2019)
2. 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: *Proc. ICCV*. pp. 2758–2766 (2015)
3. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient belief propagation for early vision. *Int J Computer Vision* **70**(1), 41–54 (2006)
4. Heinrich, M.P., Oktay, O., Bouteldja, N.: Obelisk-net: Fewer layers to solve 3D multi-organ segmentation with sparse deformable convolutions. *Medical image analysis* **54**, 1–9 (2019)
5. Heinrich, M.P., Jenkinson, M., Papież, B.W., Brady, M., Schnabel, J.A.: Towards realtime multimodal fusion for image-guided interventions using self-similarities. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 187–194. Springer (2013)
6. Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., et al.: Weakly-supervised convolutional neural networks for multimodal image registration. *Medical image analysis* **49**, 1–13 (2018)
7. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis* **36**, 61–78 (2017)
8. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with gaussian edge potentials. In: *NeurIPS*. pp. 109–117 (2011)
9. Krebs, J., Mansi, T., Mailhé, B., Ayache, N., Delingette, H.: Unsupervised probabilistic deformation modeling for robust diffeomorphic registration. In: *MICCAI DLMIA*, pp. 101–109. Springer (2018)
10. Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S.: Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine* **98**(3), 278–284 (2010)
11. Rousseau, F., Habas, P.A., Studholme, C.: A supervised patch-based approach for human brain labeling. *IEEE Trans Medical Imaging* **30**(10), 1852–1862 (2011)
12. Rühaak, J., Polzin, T., Heldmann, S., Simpson, I.J., Handels, H., Modersitzki, J., Heinrich, M.P.: Estimation of large motion in lung ct by integrating regularized keypoint correspondences into dense deformable registration. *IEEE Trans Medical Imaging* **36**(8), 1746–1757 (2017)
13. Sentker, T., Madesta, F., Werner, R.: Gdl-fire 4D: Deep learning-based fast 4D CT image registration. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 765–773. Springer (2018)
14. Jimenez-del Toro, O., Müller, H., Krenn, M., et al.: Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks. *IEEE Trans Medical Imaging* **35**(11), 2459–2475 (2016)

15. 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. *Medical image analysis* **52**, 128–143 (2019)
16. Xu, Z., Lee, C., Heinrich, M.P., Modat, M., Rueckert, D., Ourselin, S., Abramson, R.G., Landman, B.: Evaluation of 6 registration methods for the human abdomen on clinically acquired CT. *IEEE Trans Biomed Eng* **63**(8), 1563–1572 (2016)
17. Yang, X., Kwitt, R., Styner, M., Niethammer, M.: Quicksilver: Fast predictive image registration—a deep learning approach. *NeuroImage* **158**, 378–396 (2017)
18. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: *Proc. ICCV*. pp. 1529–1537 (2015)