Learning-based Affine Registration of Histological Images

Marek Wodzinski[1], Henning Müller[2]

[1]AGH University of Science and Technology
Department of Measurement and Electronics, Krakow, Poland
wodzinski@agh.edu.pl
[2]University of Applied Sciences Western Switzerland (HES-SO Valais)
Information Systems Institute, Sierre, Switzerland
henning.mueller@hevs.ch

Abstract. The use of different stains for histological sample preparation reveals distinct tissue properties and may result in a more accurate diagnosis. However, as a result of the staining process, the tissue slides are being deformed and registration is required before further processing. The importance of this problem led to organizing an open challenge named Automatic Non-rigid Histological Image Registration Challenge (ANHIR), organized jointly with the IEEE ISBI 2019 conference. The challenge organizers provided several hundred image pairs and a server-side evaluation platform. One of the most difficult sub-problems for the challenge participants was to find an initial, global transform, before attempting to calculate the final, non-rigid deformation field. This article solves the problem by proposing a deep network trained in an unsupervised way with a good generalization. We propose a method that works well for images with different resolutions, aspect ratios, without the necessity to perform image padding, while maintaining a low number of network parameters and fast forward pass time. The proposed method is orders of magnitude faster than the classical approach based on the iterative similarity metric optimization or computer vision descriptors. The success rate is above 98% for both the training set and the evaluation set. We make both the training and inference code freely available.

Keywords: Image registration · Initial alignment · Deep learning · Histology · ANHIR

1 Introduction

Automatic registration of histological images stained using several dyes is a challenging and important task that makes it possible to fuse information and potentially improve further processing and diagnosis. The problem is difficult due to: (i) complex, large deformations, (ii) difference in the appearance and partially missing data, (iii) a very high resolution of the images. The importance of the problem led to organizing an Automatic Non-rigid Histological Image

Registration Challenge (ANHIR) [1,2,3], jointly with the IEEE ISBI 2019 conference. The provided dataset [1,4,5,6,7] consists of 481 image pairs annotated by experts, reasonably divided into the training (230) and the evaluation (251) set. There are 8 distinct tissue types that were stained using 10 different stains. The image size varies from 8k to 16k pixels in one dimension. The full dataset description, including the images size and the acquisition details, is available at [3]. The challenge organizers provide an independent, server-side evaluation tool that makes it possible to perform an objective comparison between participants and their solutions.

One of the most difficult subproblems for the challenge participants was to calculate the initial, global transform. It was a key to success and all the best scoring teams put a significant effort to do this correctly, resulting in algorithms based on combined brute force and iterative alignment [8,9], or applying a fixed number of random transformations [10]. In this work, we propose a method based on deep learning which makes the process significantly faster, more robust, without the necessity to manually find a set of parameters viable for all the image pairs.

Medical image registration is an important domain in medical image analysis. Much work was done in the area, resulting in good solutions to many important and challenging medical problems. Medical image registration can be divided into classical algorithms, involving an iterative optimization for each image pair [11] or learning-based algorithms where the transformations are being learned and then the registration is performed during the inference [12]. The main advantage of the learning-based approach over the classical, iterative optimization is a fast, usually real-time registration, which makes the algorithms more useful in clinical practice. During the ANHIR challenge the best scoring teams [8,9,10] used the classical approach. However, we think that it is reasonable to solve the problem using deep networks, potentially both improving the results and decreasing the computation time.

Deep learning-based medical image registration can be divided into three main categories, depending on the training procedure: (i) a supervised training [13,14], where a known transformation is applied and being reconstructed, (ii) an unsupervised training [15,16,17], where a given similarity metric with a proper regularization or penalty terms is being optimized, (iii) an adversarial training [18,19], where both a discriminator and a generator are being trained to not only find the correct transformation but also learn a correct similarity metric. All the approaches have their strengths and weaknesses. The supervised approach does not require to define a similarity metric, however, in the case of multi-modal registration, the images must be first registered manually or by using a classical algorithm. The transformations applied during training can be both synthetic or already calculated using the state-of-the-art algorithms. However, in the case of synthetic deformations, one must ensure that they correspond to the real deformations and in case of using deformation calculated by the state-of-the-art algorithms, it is unwise to expect better results, only a lower registration time. In the case of unsupervised training, a differentiable similarity metric must be defined which for many imaging modalities is not a trivial task [20]. However, if the similarity metric can be reliably defined, unsupervised training tends to provide a better generalization [17]. The adversarial approach, just like the supervised approach, does not require defining a similarity metric but it still requires a ground-truth alignment that for many medical problems can not be determined. The adversarial training provides much better generalization than the supervised one [19]. However, the disadvantage of the adversarial approach is the fact that training this kind of network is hard and much more time-consuming than the supervised/unsupervised alternatives because finding a good balance between the generator and the discriminator is usually a difficult, trial and error procedure.

We decided to use the unsupervised approach because: (i) the state-of-the-art similarity metrics can capture the similarity of the histological images well, (ii) it does not require ground-truth to train the network, (iii) it is easy to train and has a great generalization ability. Currently, the most widely used approach for training the registration networks is to resize all the training images to the same shape using both resampling and image padding. As much as resampling the images makes sense, especially considering the initial alignment where the fine details are often not necessary, the padding is usually not a good idea, especially when the aspect ratio is high. It results in a high image resolution with much empty, unused space that then requires a deeper network to ensure large enough receptive field [21]. Therefore, we propose a network that can be trained using images with substantially different resolutions, without the necessity to perform the padding, while maintaining a relatively low number of network parameters, almost independent of the image resolution.

In this work we propose a deep network to calculate the initial affine transform between histological images acquired using different dyes. The proposed algorithm: (i) works well for images with different resolution, aspects ratios and does not require image padding, (ii) generalizes well to the evaluation set, (iii) does not require the ground-truth transform during training, (iv) is orders of magnitude faster than the iterative or descriptor-based approach, (v) successfully aligns about 98% of the evaluation pairs. We achieved this by proposing a patch-based feature extraction with a variable batch size followed by a 3-D convolution combining the patch features and 2-D convolutions to enlarge the receptive field. We make both the training and inference code freely available [22].

2 Methods

2.1 General Aspects

The proposed method adheres strictly to the ANHIR challenge requirements, namely the method is fully automatic, robust and does not require any parameter tuning during the inference time. The method can be divided into a preprocessing and the following affine registration. Both steps are crucial for the correct registration. A step by step summary of the proposed registration procedure is described in Algorithm 1.

2.2 Preprocessing

The preprocessing consists of the following steps: (i) smoothing and resampling the images to a lower resolution using the same, constant factors for each image pair, (ii) segmenting the tissue from the background, (iii) converting the images to grayscale, (iv) finding an initial rotation angle by an iterative approach.

The smoothing and resampling is in theory not strictly mandatory. However, since the fine details are not necessary to find a correct initial alignment, it is unwise to use the full resolution due to high computational time and memory consumption. Both the resampling and the smoothing coefficients were determined empirically, without an exhaustive parameter tuning. The resampling preserves the aspect ratio. After the resampling, the size across the larger dimension varies from ∼600 to ∼2000 pixels, depending on the tissue type.

The next step is to remove the background. This procedure significantly improves the results for mammary glands or mice kidneys because there are staining artifacts in the background that have a strong influence on the similarity metric. In this work, we remove the background using smoothed Laplacian thresholding with a few morphological operations. It works for all the cases and more advanced background removal algorithms are not necessary. Nonetheless, this is data specific step. For other digital pathology data sets, this step may be unnecessary or can look differently (e.g. a stain deconvolution or deep learning-based segmentation).

Finally, after converting both images to grayscale, an initial rotation angle is being optimized. We decided to use a simple procedure similar to [8,9] because optimization of a single parameter can be done extremely fast and does not require any advanced optimization techniques. As a result, the network architecture can be much simpler and requires fewer parameters to capture the possible transformations. The initial rotation angle is being optimized by the iterative rotation of the source image around the translated center of mass, with a given, pre-defined angle step. Then, the angle with the largest global normalized cross-correlation (NCC) is used as the best one. In practice, this step calculation time depends on the predefined angle step and can be optimized by performing it using a GPU. However, even considering an unoptimized, single-core CPU implementation, the computational time of this step is negligible compared to the data loading, initial resampling, and background removal. The affine registration network was trained using the preprocessed data and therefore this step is required during inference.

2.3 Affine Registration

We propose a network architecture that is able to calculate the correct affine transformation in a single pass, independently of the image size and the aspect ratio. The idea behind the network is as follows. First, the images are passed to the network independently. They are unfolded to a grid of patches with a given, predefined size (224x224 in our case) and stride equal to the patch size, the patches do not overlap. Then, the patches are combined to a single tensor where

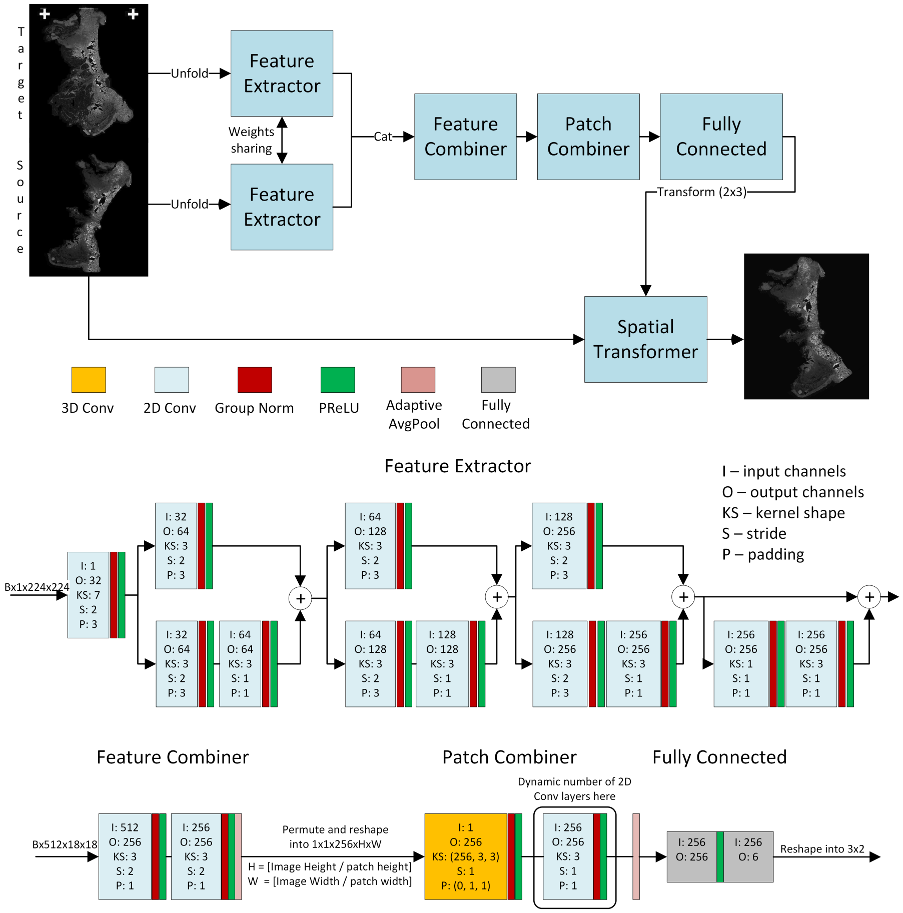
Target
Source
Unfold
Unfold
Feature Extractor
Weights sharing
Feature Extractor
Cat
Feature Combiner
Patch Combiner
Fully Connected
Transform (2x3)
Spatial Transformer
3D Conv
2D Conv
Group Norm
PReLU
Adaptive AvgPool
Fully Connected
Feature Extractor
I – input channels
O – output channels
KS – kernel shape
S – stride
P – padding
Bx1x224x224
I: 1 O: 32 KS: 7 S: 2 P: 3
I: 32 O: 64 KS: 3 S: 2 P: 3
I: 32 O: 64 KS: 3 S: 2 P: 3
I: 64 O: 64 KS: 3 S: 1 P: 1
I: 64 O: 128 KS: 3 S: 2 P: 3
I: 64 O: 128 KS: 3 S: 2 P: 3
I: 128 O: 128 KS: 3 S: 1 P: 1
I: 128 O: 256 KS: 3 S: 2 P: 3
I: 128 O: 256 KS: 3 S: 2 P: 3
I: 256 O: 256 KS: 3 S: 1 P: 1
I: 256 O: 256 KS: 1 S: 1 P: 1
I: 256 O: 256 KS: 3 S: 1 P: 1
Feature Combiner
Patch Combiner
Fully Connected
Bx512x18x18
I: 512 O: 256 KS: 3 S: 2 P: 1
I: 256 O: 256 KS: 3 S: 2 P: 1
Permute and reshape into 1x1x256xHxW
H = [Image Height / patch height]
W = [Image Width / patch width]
I: 1 O: 256 KS: (256, 3, 3) S: 1 P: (0, 1, 1)
Dynamic number of 2D Conv layers here
I: 256 O: 256 KS: 3 S: 1 P: 1
I: 256 O: 256
I: 256 O: 6
Reshape into 3x2

Fig. 1: An overview of the proposed network architecture. The source and target are unfolded and passed independently to the feature extractor where the batch size is equal to the number of patches after unfolding. Then, the extracted features are concatenated and passed to the feature combiner, patch combiner, and fully connected layers respectively. The whole network has slightly above 30 million parameters, independently of the input image size.

the number of patches defines the batch size. This step is followed by a feature extraction by a relatively lightweight, modified ResNet-like architecture [23]. The feature extractor weights are shared between the source and the target. Then, the features are concatenated and passed through additional 2-D convolutions to combine the source and target into a single representation. Finally, the global correspondence is extracted by a 3-D convolution followed by a variable number of 2-D convolutions using the PyTorch dynamic graphs. The final step allows getting global information from the unfolded patches. The number of final 2-D convolutions depends on the image resolution and can be extended dynamically to enlarge the receptive field. In practice, on the resampled ANHIR dataset (the larger dimension contains from ∼600 to ∼2000 pixels) a single convolutional layer is sufficient. Eventually, the features are passed to adaptive average pooling and fully connected layers, which output the transformation matrix. The network architecture and the forward pass procedure is presented in Figure 1. The number of parameters is slightly above 30 million, the forward pass memory consumption depends on the image resolution.

The proposed network is trained in a relatively unusual way. The batch is not strictly the number of pairs during a single pass through the network. The image pairs are given one by one and the loss is being backwarded after each of them. However, the optimizer is being updated only after a gradient of a given number of images (the real batch size) was already backpropagated. This approach makes its possible to use any real batch size during training but it requires an architectural change. Since all the image pairs have a different resolution, they are divided into a different number of patches during unfolding. As a result, it is incorrect to use the batch normalization layers because during inference they are unable to automatically choose the correct normalization parameters and strong overfitting is observed. Therefore, we replaced all the batch normalization layers by a group normalization [24], which solved the problem. One can argue that this approach significantly increases the training time. This is not the case because the batch size dimension after unfolding is sufficiently large to utilize the GPU correctly.

Algorithm 1: Algorithm Summary.

```
Input  : $\mathbf{M_p}$ (moving image path), $\mathbf{F_p}$ (fixed image path)
Output: $\mathbf{T}$ (affine transformation (2x3 matrix)
$\mathbf{M}$, $\mathbf{F}$ = load both the images from $\mathbf{M_p}$ and $\mathbf{F_p}$
$\mathbf{M}$, $\mathbf{F}$ = smooth and resample the images to a lower resolution using the same,
  constant factors for each image pair
$\mathbf{M}$, $\mathbf{F}$ = segment the tissues from the background
$\mathbf{M}$, $\mathbf{F}$ = convert the $\mathbf{M}$, $\mathbf{F}$ images to the grayscale and invert the intensities
$\mathbf{T_{rot}}$ = find the initial rotation angle by an iterative approach which
  maximizes the NCC similarity metric between $\mathbf{M}$ and $\mathbf{F}$
$\mathbf{M_{rot}}$ = warp $\mathbf{M}$ using $\mathbf{T_{rot}}$
$\mathbf{T_{aff}}$ = pass $\mathbf{M_{rot}}$ and $\mathbf{F}$ through the proposed network to find the affine
  matrix
$\mathbf{T}$ = compose $\mathbf{T_{rot}}$ and $\mathbf{T_{aff}}$
return $\mathbf{T}$
```

The network was trained using an Adam optimizer, with a learning rate equal to 10^{-4} and a decaying scheduler after each epoch. The global negative NCC was used as the cost function. No difference was observed between the global NCC and the patch-based NCC. Moreover, the results provided by NCC were better than MIND or NGF since the latter two are not scale-resistant and would require additional constraints. The dataset was augmented by random affine transformations applied both to the source and the target, including translating, scaling, rotating and shearing the images. The network was trained using only the training dataset consisting of 230 image pairs. The evaluation dataset consisting of 251 image pairs was used as a validation set. However, no decision was made based on the validation results. The network state after the last epoch was used for testing. Thanks to the augmentation, no overfitting was observed. Moreover, the loss on the validation set was lower than on the training set. No information about the landmarks from both the training and the validation set was used during the training. The source code, for both the inference and training, is available at [22].

3 Results

The proposed algorithm was evaluated using all the image pairs provided for the ANHIR challenge [1,2,3]. The data set is open and can be freely downloaded, so results are fully reproducible. For a more detailed data set description, including the tissue types, the procedure of the tissue staining and other important information, we refer to [3].

We evaluated the proposed algorithm using the target registration error between landmarks provided by the challenge organizers, normalized by the image diagonal, defined as:

$$rTRE = \frac{TRE}{\sqrt{w^2 + h^2}}, \tag{1}$$

where TRE denotes the target registration error, w is the image width and h is the image height. We compare the proposed method to the most popular computer vision descriptors (SURF [25] and SIFT [26]) as well as the intensity-based, iterative affine registration [27]. All the methods were applied to the dataset after the preprocessing and the parameters were tuned to optimize the results. Unfortunately, we could not compare to initial alignment methods used by other challenge participants because the submission system reports only the final results after nonrigid registration. The cumulative histogram of the target registration error for the available landmarks is shown in Figure 2. In Table 1 we summarize the rTRE for the evaluation set using the evaluation platform provided by the challenge organizers. We also show the success ratio and the affine registration time, excluding data loading and preprocessing time, which is the same for all the methods. As the success ratio, we define cases that are registered in a manner that can we followed by a converging, generic, nonrigid registration algorithm like B-Splines free form deformations or Demons. In Figure 3 we show an exemplary case for which the proposed method is successful and the remaining methods failed or were unable to converge correctly.

4 Discussion and Conclusion

The proposed method works well for more than 98% of the ANHIR image pairs. It calculates a transformation that can be a good starting point for the following nonrigid registration. The registration time is significantly lower than using the iterative or feature-based approach. However, it should be noted that currently

Table 1: Quantitative results of the rTRE calculated using the ANHIR submission website [3] as well as the average processing time for the affine registration step. The success rate for the initial state shows the ratio of pairs not requiring the initial alignment.

	rTRE			Time [ms]	Success Rate
	Median	Average	Max (Avg)	Average	[%]
Initial	0.056	0.105	0.183	-	31.15
Preprocessed	0.023	0.035	0.069	-	67.36
Proposed	0.010	0.025	0.060	4.51	98.34
SIFT [26]	0.005	0.085	0.174	422.65	79.21
SURF [25]	0.005	0.100	0.201	169.59	78.38
Iterative [27]	0.004	0.019	0.050	3241.15	97.30

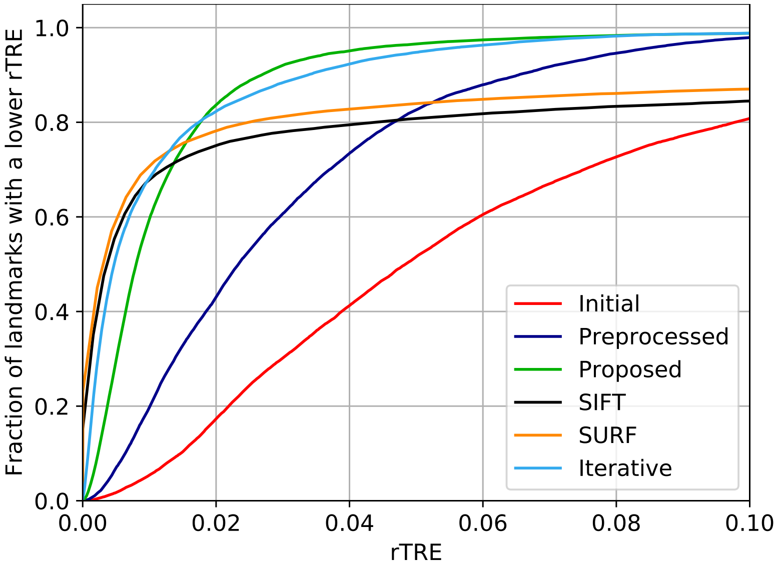

Fig. 2: The cumulative histogram of the target registration error for the proposed and compared methods. Please note that all the compared methods use the same preprocessing pipeline to make them comparable. We experimentally verified that the preprocessing does not deteriorate the results for the feature-based approach and significantly improves the results for the iterative registration.

more than 99% of the computation time is spent on the data loading, initial smoothing, and resampling. This step could be significantly lowered by proposing a different data format, which already includes the resampled version of the images.

It can be noticed that both the iterative affine registration and the feature-based alignment provide slightly better results when they can converge correctly. However, the registration accuracy achieved by the proposed method is sufficient

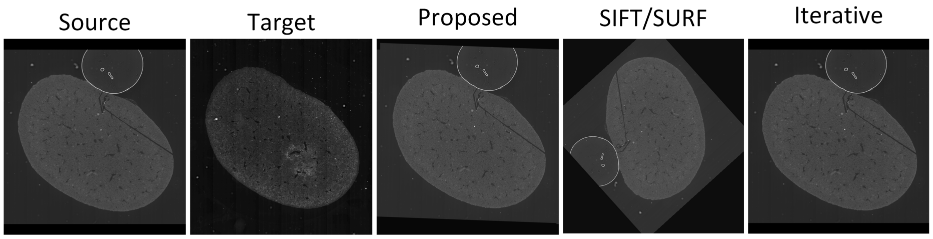

Fig. 3: An exemplary failure visualization of the evaluated methods. Please note that the calculated transformations were applied to the images before the preprocessing. It is visible that the feature-based approach failed and the iterative affine registration was unable to converge correctly.

for the following nonrigid registration for which the gap between the proposed method and the iterative alignment is not that important. The proposed method is significantly faster and more robust, resulting in a higher success ratio, which in practice is more important than the slightly lower target registration error. The feature-based methods often fail and without a proper detection of the failures they cannot be used in a fully automatic algorithm. On the other hand, the proposed method does not suffer from this problem.

To conclude, we propose a method for an automatic, robust and fast initial affine registration of histology images based on a deep learning approach. The method works well for images with different aspect ratios, resolutions, generalizes well for the evaluation set and requires a relatively low number of the network parameters. We make the source code freely available [22]. The next step involves a deep network to perform the non-rigid registration, using the highest resolution provided by the challenge organizers. We think it is possible to solve this problem efficiently, even though a single image can take up to 1 GB of the GPU memory.

Acknowledgments

This work was funded by NCN Preludium project no. UMO-2018/29/N/ST6/00143 and NCN Etiuda project no. UMO-2019/32/T/ST6/00065.

References

1. Borovec, J., Munoz-Barrutia, A., Kybic, J.: Benchmarking of Image Registration Methods for Differently Stained Histological Slides. IEEE International Conference on Image Processing (2018) 3368–3372
2. Borovec, J., et al.: ANHIR: Automatic Non-rigid Histological Image Registration Challenge. IEEE Transactions on Medical Imaging (2020)
3. Borovec, J., et al.: ANHIR Website. `https://anhir.grand-challenge.org`
4. Fernandez-Gonzalez, R., et al.: System for combined three-dimensional morphological and molecular analysis of thick tissue specimens. Microscopy Research and Technique **59**(6) (2002) 522–530
5. Gupta, L., Klinkhammer, B., Boor, P., Merhof, D., Gadermayr, M.: Stain independent segmentation of whole slide images: A case study in renal histology. IEEE ISBI (2018) 1360–1364

6. Mikhailov, I., Danilova, N., Malkov, P.: The immune microenvironment of various histological types of ebv-associated gastric cancer. Virchows Archiv (2018)
7. Bueno, G., Deniz, O.: AIDPATH: Academia and Industry Collaboration for Digital Pathology. `http://aidpath.eu`
8. Lotz, J., Weiss, N., Heldmann, S.: Robust, fast and accurate: a 3-step method for automatic histological image registration. arXiv:1903.12063 (2019)
9. Wodzinski, M., Skalski, A.: Automatic Nonrigid Histological Image Registration with Adaptive Multistep Algorithm. arXiv:1904.00982 (2019)
10. Venet, L., Pati, S., Yushkevich, P., Bakas, S.: Accurate and Robust Alignment of Variable-stained Histologic Images Using a General-purpose Greedy Diffeomorphic Registration Tool. arXiv:1904.11929 (2019)
11. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: A survey. IEEE Transactions on Medical Imaging **32**(7) (2013) 1153–1190
12. Haskins, G., Kruger, U., Yan, P.: Deep Learning in Medical Image Registration: A Survey. arXiv:1903.02026 (2019)
13. DeTone, D., Malisiewicz, T., Rabinovich, A.: Deep Image Homography Estimation. arXiv:1606.03798 (2016)
14. Chee, E., Wu, Z.: AIRNet: Self-Supervised Affine Registration for 3D Medical Images using Neural Networks. arXiv:1810.02583 (2018)
15. de Vos, B., Berendsen, F., Viergever, M., Sokooti, H., Staring, M., Isgum, I.: A deep learning framework for unsupervised affine and deformable image registration. Medical Image Analysis **52** (2019) 128–143
16. Balakrishnan, G., Zhao, A., Sabuncu, M., Guttag, J., Dalca, A.: VoxelMorph: A Learning Framework for Deformable Medical Image Registration. IEEE Transactions on Medical Imaging **38**(8) (2019) 1788–1800
17. Dalca, A., Balakrishnan, G., Guttag, J., Sabuncu, M.: Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Medical Image Analysis **57** (2019) 226–236
18. Fan, J., Cao, X., Wang, Q., Yap, P., Shen, D.: Adversarial learning for mono- or multi-modal registration. Medical Image Analysis **58** (2019)
19. Mahapatra, D., Antony, B., Sedai, S., Garnavi, R.: Deformable medical image registration using generative adversarial networks. IEEE ISBI (2018) 1449–1453
20. Xiao, Y., et al.: Evaluation of MRI to ultrasound registration methods for brain shift correction: The CuRIOUS2018 Challenge. IEEE Transactions on Medical Imaging (2019)
21. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. Advances in Neural Information Processing Systems (2016) 4905–4913
22. Wodzinski, M.: The Source Code. `https://github.com/lNefarin/DeepHistReg`
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. IEEE CVPR (2016) 770–778
24. Wu, Y., He, K.: Group Normalization. arXiv:1803.084943 (2018)
25. Bay, H., TuytelaarsT., and Van Gool L.: SURF: Speeded up robust features. European Conference on Computer Vision (2006) 404–417
26. Lowem D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
27. Klein, S., Staring, M., Murphy, K., Viergever, M., Pluim, J.: elastix: A Toolbox for Intensity-Based Medical Image Registration. IEEE Transactions on Medical Imaging **29**(1) (2010) 196–205