ANHIR: Automatic Non-rigid Histological Image Registration Challenge

Jiří Borovec, Jan Kybic *Senior Member*, Ignacio Arganda-Carreras, Dmitry V. Sorokin, Gloria Bueno, Alexander V. Khvostikov, Spyridon Bakas, Eric I-Chao Chang, Stefan Heldmann, Kimmo Kartasalo, Leena Latonen, Johannes Lotz, Michelle Noga, Sarthak Pati, Kumaradevan Punithakumar *Senior Member*, Pekka Ruusuvuori, Andrzej Skalski *Senior Member*, Nazanin Tahmasebi, Masi Valkonen, Ludovic Venet, Yizhe Wang, Nick Weiss, Marek Wodzinski, Yu Xiang, Yan Xu, Yan Yan, Paul Yushkevic, Shengyu Zhao, and Arrate Muñoz-Barrutia *Senior Member*

***Abstract*—Automatic Non-rigid Histological Image Registration (ANHIR) challenge was organized to compare the performance of image registration algorithms on several kinds of microscopy histology images in a fair and independent manner. We have assembled 8 datasets, containing 355 images with 18 different stains, resulting in 481 image pairs to be registered. Registration accuracy was evaluated using manually placed landmarks. In total, 256 teams registered for the challenge, 10 submitted the results, and 6 participated in the workshop. Here, we present the results of 7 well-performing methods from the challenge together with 6 well-known existing methods. The best methods used coarse but robust initial alignment, followed by non-rigid registration, used multiresolution, and were carefully tuned for the data at hand. They outperformed off-the-shelf methods, mostly by being more robust. The best methods could successfully register over 98 % of all landmarks and their mean landmark registration accuracy (TRE) was 0.44 % of the image diagonal. The challenge remains open to submissions and all images are available for download.**

***Index Terms*—Image registration, Microscopy**

J. Borovec and J. Kybic were with the Faculty of Electrical Engineering, Czech Technical University in Prague. J. Kybic acknowledges the support of the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 Research Center for Informatics and the Czech Science Foundation project 17-15361S. I. Arganda-Carreras is with Ikerbasque, Basque Foundation for Science, Bilbao, University of the Basque Country and Donostia International Physics Center, Donostia-San Sebastian, Spain. D. V. Sorokin and A. V. Khvostikov were with Laboratory of Mathematical Methods of Image Processing, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Russia. They acknowledge the support of the Russian Science Foundation grant 17-11-01279. Gloria Bueno is with VISILAB group and E.T.S. Ingenieros Industriales, Universidad de Castilla-La Mancha, Spain. S. Bakas, S. Pati, L. Venet and P. Yushkevich are with Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. Their work was partly supported by the National Institutes of Health under grant award numbers NIH/NCI/ITCR:U24-CA189523, NIH/NIBIB:R01EB017255, NIH/NIA:R01AG056014, and NIH/NIA:P30AG010124. E. I-C. Chang and Y. Xu are with Microsoft Research, Beijing, China. Y. Xu is also with the School of biological science and medical engineering and Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing. S. Heldmann, J. Lotz and N. Weiss are with Fraunhofer MEVIS, Lubeck, Germany. K. Kartasalo, P. Ruusuvuori and M. Valkonen are with Tampere University, Faculty of Medicine and Health Technology, Tampere, Finland. L. Latonen is with University of Eastern Finland, Institute of Biomedicine, Kuopio, Finland. P. Ruusuvuori is with Institute of Biomedicine, University of Turku, Finland and with Faculty of Medicine and Health Technology, Tampere, Finland. His work is supported by Academy of Finland projects 313921 and 314558. K. Punithakumar, M. Noga, and N. Tahmasebi are with the Department of Radiology & Diagnostic Imaging, University of Alberta, AB, Canada, and also with Servier Virtual Cardiac Centre, Mazankowski Alberta Heart Institute, AB, Canada. They acknowledge the support provided by WestGrid and Compute Canada. A. Skalski and M. Wodzinski are with AGH University of Science and Technology, Department of Measurement and Electronics, Krakow, Poland. Marek Wodzinski acknowledges the support of the Preludium project no. UMO-2018/29/N/ST6/00143 funded by the National Science Centre in Poland. Y. Xiang and Y. Wang are with Chengdu Knowledge Vision Science and Technology Co., Ltd. Chengdu, China. Y. Yan is with University of Electronic Science and Technology of China Chengdu, China. S. Zhao is with Tsinghua University, Beijing, China. A. Muñoz-Barrutia is with Bioengineering and Aerospace Engineering Department, Universidad Carlos III de Madrid and Instituto de Investigación Sanitaria Gregorio Marañón, Madrid, Spain. She acknowledges the support of the Spanish Ministry of Economy and Competitiveness (TEC2015-73064-EXP, TEC2016-78052-R and RTC-2017-6600-1) and a 2017 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation.

Jan Kybic is the corresponding author, kybic@fel.cvut.cz

I. Introduction

IMAGE registration is one of the key tasks in biomedical image analysis, aiming to find a geometrical transformation between two or more images, so that corresponding objects appear at the same position in both/all images. This research area has received a lot of attention over the last several decades and literally hundreds of different registration algorithms exist [1]–[8]. An unbiased comparison of the registration algorithms is essential to help users choose the most suitable method for a given task.

A. Goals

The goal of this study is to report the results of the Automatic Non-rigid Histological Image Registration challenge (ANHIR)[1], at the IEEE International Symposium on Biomedical Imaging (ISBI) 2019. To the best of our knowledge, ANHIR was the first open comparison of image registration algorithms on microscopy images. Moreover, we describe the associated annotated dataset of histology images, which is freely available from the challenge website, and is one of very few microscopy image datasets suitable for testing registration algorithms.

B. Previous work on image registration evaluation

One of the earliest image registration comparisons was the Retrospective Image Registration Evaluation (RIRE)

[1]https://anhir.grand-challenge.org

project [9], comparing rigid registration algorithms on computed tomography (CT), magnetic resonance (MRI), and positron emission tomography (PET) human brain data using fiducial landmarks. Later studies added more methods [10], non-linear transformations [11]–[13], and other organs, such as lung [14], or liver and prostate [15], as well as comparison on microscopy data [16]. Registration performance is usually measured using landmark distances [9], [13], [15]–[17] or segmentation overlap [11], [12], [18].

In some of the above mentioned studies, the comparison was done by the authors themselves by means of publicly available implementations [13], which is a good strategy to find the best out-of-the-box performing approach. In other cases [17], authors (re)implemented the methods themselves, eliminating the differences in the quality of the implementation.

C. Challenges

It can be argued that the authors of a method are the best equipped and motivated to implement, tune, and run their own methods. In a *challenge*, the organizers prepare the data and invite participants to apply their methods to provided data and submit the results, which are then evaluated by given criteria. It is widely believed that challenges have a very beneficial effect on the progress in a particular field by: (i) lowering the entry barrier by providing (usually large) annotated datasets, (ii) motivating participants to improve the performance of their methods, (iii) identifying the state-of-the-art approaches, and (iv) tracking the overall progress of the field.

In some of the previous image registration comparisons, only known authors of image registration methods were invited to participate. Other challenges were opened to all participants, including the following:

- The RIRE project [9] used masked fiducial points on MRI, CT and PET images. It is now fully open.
- The EMPIRE10 challenge [18] on pulmonary images.
- The recent CurIOUS challenge[2] on the registration between presurgical MRI and intra-operative ultrasound.
- The MOCO challenge asking for motion correction of myocardial perfusion MRI [19].
- The Continous Registration Challenge[3], which uses eight different datasets including lung 4D CT and MRI brain images differs in requiring participants to integrate their code to a common platform.

As far as we know, ANHIR, described here, is the only registration challenge using microscopy images.

D. Histology image registration

Microscopy images in histology need to be registered for a number of reasons [20], most frequently to create a 3D reconstruction from scanned 2D thin slices [21]–[29]. Other applications include distortion compensation [30], creating a high resolution mosaic from small 2D tiles [31], fusing information from differently stained slices [32]–[35] or even from different modalities, such as histology and MRI [36]. Registration can assist in segmenting unknown stains using known stains [37], or to combine gene expression maps from multiple specimens [38].

Histology image registration is challenging because of the following aspects: (i) large sizes of some images, e.g. $80\text{k} \times 60\text{k}$ pixels for the whole slide images (WSI), (ii) repetitive texture, making it hard to find globally unique landmarks, (iii) sometimes large non-linear elastic deformation, occlusions and missing sections due to sample preparation, (iv) differences in appearance due to different staining or acquisition methods, and (v) differences in local structure between slices. Registration methods developed for other modalities might therefore not work optimally and need to be modified for microscopy images. Changes in appearance can be handled by color normalization [39] or deconvolution [40], focusing on the parts or stains common to both images [33], segmentation [41], [42] or probabilistic segmentation [43], or using multimodal similarity features such as structural probability maps [29]. While most algorithms are intensity-based, some use keypoint features [44] or edge features [45], and some combine the two approaches [22]. The very large image size is often handled simply by working on downscaled images but better precision can be obtained by dividing the images into smaller parts (tiles), registering them separately, and combining the results [31], [46]. The availability of a large number of consecutive slices can be used to improve robustness by enforcing consistency between different pairs of slices [22].

There is a relative paucity of microscopy image datasets suitable for image registration evaluation. The exception are datasets for testing cell tracking and cell nuclei registration in fluorescence microscopy [47], [48], which differs significantly from histological images considered here. Evaluation has been performed on other datasets, which are however not publicly available, such as a comparison of histology and block-face optical images [49], and a 3D alignment of ssTEM slices [50].

This work can be regarded as a continuation of an earlier image registration evaluation [16], [51], which is significantly expanded here in terms of the number of methods, number of images and image types, image size and resolution, and the quality of annotations.

II. ANHIR Challenge

The ANHIR challenge focused on comparing the accuracy, robustness, and speed of fully automatic non-linear registration methods on microscopy histology images. The images to be registered are large and differ in appearance due to different staining.

A. Data

In total, we have assembled 8 datasets, please see Table I for an overview and Fig. 1 for examples. More details are available in Sections[4] S.I and S.IV. The appearance differences due to staining are sometimes important (see Fig. S2). In each case, high-resolution (up to $40\times$ magnification) whole slide images were acquired. The images are organized in sets such that

[2]https://curious2019.grand-challenge.org

[3]https://continuousregistration.grand-challenge.org/

[4]All references beginning with 'S' refer to the Supplementary material.

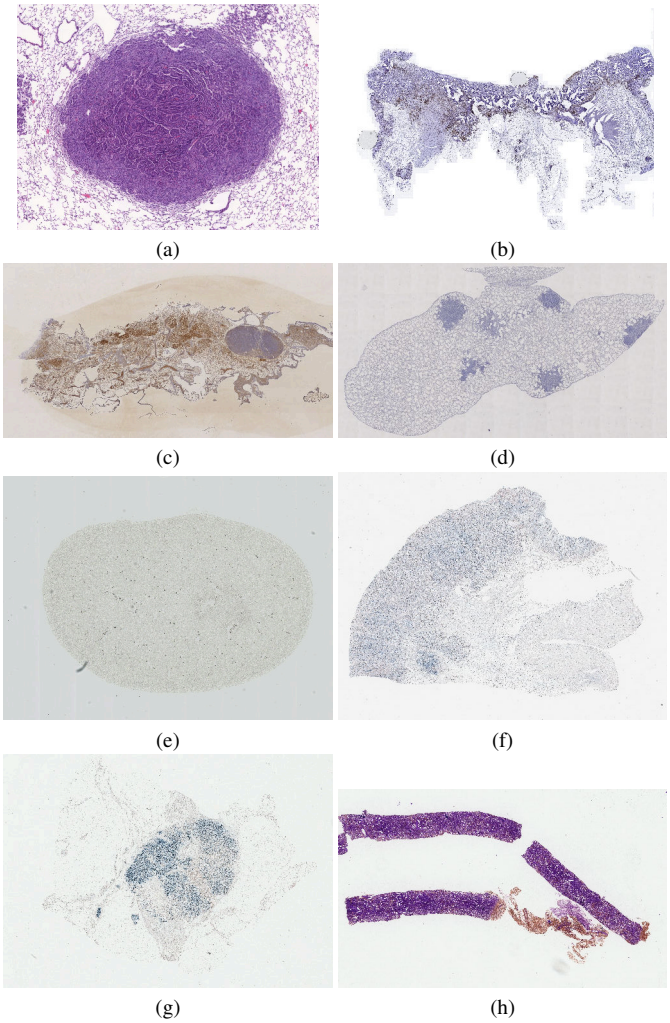


Fig. 1. Example dataset images: (a) lung lesions, (b) COAD (colon adenocarcinoma), (c) mammary glands, (d) lung lobes, (e) mouse kidney, (f) gastric mucosa and adenocarcinoma, (g) human breast, (h) human kidney. Some images were cropped for visualization. Multiple stains exist for each tissue type.

any two images within a set could be meaningfully registered, as they come from spatially close slices. Different stains are used for each image in a set and the local structure often differs, see Fig. S3. In total, we obtained 49 sets with 3 to 9 images per set, the average size was 5 slides per set. There are 355 images in total with 18 different stains[5]. We generated 481 image pairs for registration (each pair within each set but dropping symmetric pairs) and split them into 230 training and 251 testing pairs. Note that the lung lesions and lung lobes datasets were used in [16] and were therefore already completely public. We have therefore decided to treat these datasets as training only and not use them for the evaluation, except when explicitly mentioned.

Since the original images are very large, participants were suggested to register provided *medium* size images, chosen from a fixed set of scales so that their size is approximately

[5]CD1a, CD31, CD4, CD68, CD8, Cc10, EBV, ER, HE, HER2, Ki67, MAS, PAS, PASM, PR, Pro-SPC, aSMA

Name	Scanner	Mag.	μm/ pixel	Avg. size [pixels]	# sets	#train pairs	#test pairs
Lung lesions	Zeiss	40×	0.174	18k × 15k	3	30	0
Lung lobes	Zeiss	10×	1.274	11k × 6k	4	40	0
Mammary glands	Zeiss	10×	2.294	12k × 4k	2	38	0
Mouse kidney	Hamam.	20×	0.227	37k × 30k	1	15	18
COAD	3DHistech	10×	0.468	60k × 50k	20	84	153
Gastric	Leica	40×	0.2528	60k × 75k	9	13	40
Human breast	Leica	40×	0.253	65k × 60k	5	5	20
Human kidney	Leica	40×	0.253	18k × 55k	5	5	20

TABLE I
SUMMARY OF THE DATASETS AND THEIR PROPERTIES. SEE THE SUPPLEMENTARY MATERIAL FOR A MORE DETAILED DESCRIPTION. (COAD=COLON ADENOCARCINOMA).

10k × 10k pixels. Most participants have themselves reduced the image size even further as part of the preprocessing. For convenience, we also provided *small* size images at 5% of the original size, around 2k × 2k pixels.

B. Landmarks

We have annotated significant structures in the images with landmarks spread over the tissue (see Figure 2 for an example). Each landmark position is known in all images from the same set. We have placed on the average 86 landmarks per image. There were 9 annotators. Annotating a set of images took around 2 hours on average, with $20 \sim 30\%$ additional time for proofreading. In total, annotation and validation of the 355 images took about 250 hours. All images were annotated by at least two different people. The average distance between landmarks of the two annotators was $0.05\,\%$ of the image diagonal (to be compared with the relative Target Registration Error, Section II-C). This corresponds to several pixels at the original resolution. Landmarks for some of the images were withheld by the organizers. Image pairs with landmarks for both images available to participants were used for training, while image pairs with landmarks for exactly one image known to the participants formed the set of testing pairs $\mathcal{T}$. See Section S.II in the Supplementary material for details.

C. Evaluation measures

For each pair of images (i, j), we have determined the coordinates of corresponding landmarks $\mathbf{x}_l^i$, $\mathbf{x}_l^j$, where $l \in L^i$, and L^i is a set of landmarks, guaranteed to occur in both i and j. The participants were asked to submit the coordinates $\hat{\mathbf{x}}_l^j$ of landmark points in the coordinate system of image j corresponding to provided coordinates $\mathbf{x}_l^i$ in the image i. We then define the *relative Target Registration Error* (rTRE) as the Euclidean distance between the submitted coordinates $\hat{\mathbf{x}}_l^j$ and the manually determined (ground truth) coordinates $\mathbf{x}_l^j$ (withheld from participants)

$$\mathrm{rTRE}_l^{ij} = \frac{\|\hat{\mathbf{x}}_l^j - \mathbf{x}_l^j\|_2}{d_j} \tag{1}$$

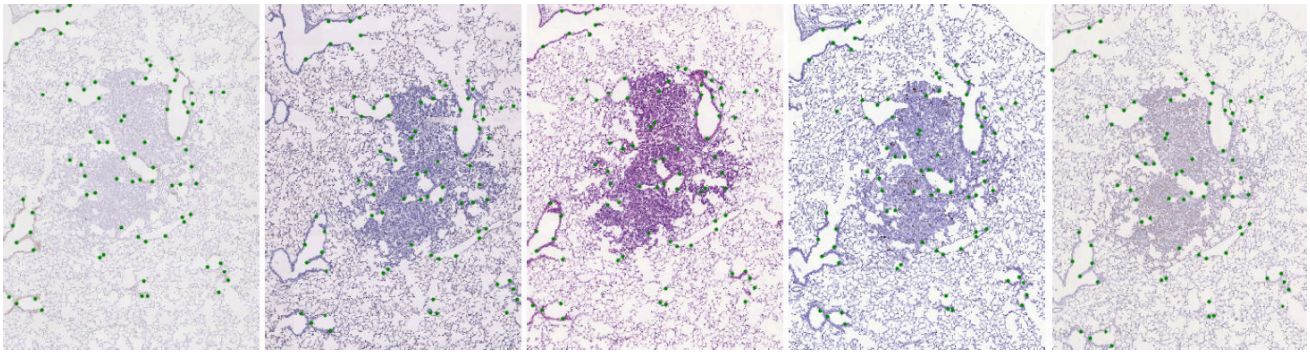
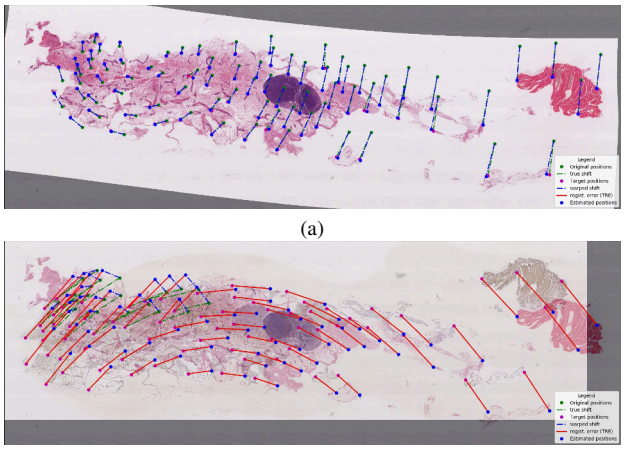


Fig. 2. An example of a set of lung lesion tissue images using Cc10, CD31, H&E, Ki67 and proSPC stains (from left to right) with landmarks shown as green dots.



(a)

(b)

Fig. 3. Examples of (a) successful and (b) unsuccessful registration. We are showing the warped image overlaid over the reference image, with the original positions and true displacements in green, estimated positions and estimated displacements in blue, target positions in magenta, and registration errors (TRE) in red.

normalized by the length of the image diagonal d_j. Figure 3 shows an example of successful and unsuccessful registrations.

Our main criterion to evaluate the registration algorithms is the *average rank of median rTRE* (ARMrTRE)

$$\text{ARMrTRE}(m) = \underset{(i,j)\in\mathcal{T}}{\text{mean}}\ \underset{m\in\mathcal{M}}{\text{rank}}\ \mu^{i,j}(m) \tag{2}$$

$$\text{where}\quad \mu^{i,j}(m) = \underset{l\in L^i}{\text{median}}\,\text{rTRE}_l^{i,j}(m), \tag{3}$$

$\mathcal{T}$ is the set of all test image pairs and $\mathcal{M}$ is the set of all methods (a method corresponds to a single submission). In other words, we evaluate the rTRE (1) for all landmarks in an image and calculate the *median* of all rTRE values in an image. Aggregation by a median was chosen to eliminate outliers. We *rank* all methods $m \in \mathcal{M}$ according to this median. Ranking was chosen to compensate for differences in difficulty between datasets and images, which may influence the scale of the rTRE. Finally, we average the ranks over all test image pairs $\mathcal{T}$.

We also calculate other criteria aggregating the rTRE, namely *average median rTRE (AMrTRE)* and *median of median rTRE (MMrTRE)*

$$\text{AMrTRE}(m) = \underset{(i,j)\in\mathcal{T}}{\text{mean}}\ \mu^{i,j}(m) \tag{4}$$

$$\text{MMrTRE}(m) = \underset{(i,j)\in\mathcal{T}}{\text{median}}\ \mu^{i,j}(m) \tag{5}$$

and similarly the *average maximum rTRE (AMxrTRE)*

$$\text{AMxrTRE}(m) = \underset{(i,j)\in\mathcal{T}}{\text{mean}}\ \underset{l\in L^i}{\max}\ \text{rTRE}_l^{i,j}(m) \tag{6}$$

and *average rank maximum rTRE (ARMxrTRE)*

$$\text{ARMxrTRE}(m) = \underset{(i,j)\in\mathcal{T}}{\text{mean}}\ \underset{m\in\mathcal{M}}{\text{rank}}\ \underset{l\in L^i}{\max}\ \text{rTRE}_l^{i,j}(m) \tag{7}$$

To evaluate robustness, we compare individual rTREs for each landmark with the relative Initial Registration Error (rIRE) before registration

$$\text{rIRE}_l^{ij} = \frac{\|\mathbf{x}_l^i - \mathbf{x}_l^j\|_2}{d_j} \tag{8}$$

This is meaningful for our images, since the images i and j share the same or similar system of coordinates, i.e. $\mathbf{x}_l^i$ is a reasonable initial approximation for $\mathbf{x}_l^j$. Let us call $K^{i,j} \subseteq L^i$ the set of successfully registered landmarks, i.e. those for which the registration error decreases, $\text{rTRE}_l < \text{rIRE}_l$. We define the robustness as the relative number of successfully registered landmarks,

$$R^{i,j}(m) = \frac{|K^{i,j}|}{|L^i|} \tag{9}$$

and then, the *average robustness* over all test image pairs is

$$\bar{R}(m) = \underset{(i,j)\in\mathcal{T}}{\text{mean}}\ R^{i,j}(m) \tag{10}$$

Finally, we asked the participants to report the registration times t^{ij} in minutes for each registration, including loading input images and writing the output files. We report the average registration time

$$\bar{t}(m) = \underset{(i,j)\in\mathcal{T}}{\text{mean}}\ t^{ij}(m) \tag{11}$$

The average registration time should only be considered as indicative, because it has not been independently validated. The times are approximately normalized with respect to participant computer performance by running an identical calibration program provided by the organizers.

D. Pairwise method comparison

For each pair of methods m and m' from $\mathcal{M}$, we determine if a method m is significantly better than method m'. More specifically, we compare the median rTRE μ (3) on all test image pairs $(i, j) \in \mathcal{T}$. Our null hypothesis is that $\mu^{i,j}(m) \geq \mu^{i,j}(m')$. Clearly, the two μ values are dependent and not necessarily normally distributed, while the test image pairs are considered i.i.d. Hence, we employ the one-sided Wilcoxon signed-rank test and evaluate the p-value under the normal assumptions, since the number of samples $|\mathcal{T}| = 251$ is sufficiently large. For $p < 0.01$, we reject the null hypothesis and conclude that the method m indeed performs significantly better than method m'. The alternative would be permutation testing, as in [52], which is however much more computationally demanding.

E. Organization

The challenge was accepted to ISBI 2019 in October 2018 and officially announced in December 2018, after a website was set up and an evaluation system prepared. The lung lesions, lung lobes, and mammary glands data [16] with 9 training sets (108 image pairs with all landmark coordinates available) was immediately made accessible to enable the participants to start developing. In February 2019, we released the remaining images, training landmarks (122 image pairs) and a subset of the test landmarks (251 image pairs), and opened the web-based submission and evaluation system. The remaining test landmarks were released in the middle of March. The submission system was closed on March 31, 2019. Participants performed the registration on their own computers.

The challenge was widely advertised and totally open, anybody was free to participate. In total, 256 teams registered for the challenge and 10 submitted the results. For comparison, we also used two "internal submissions" corresponding to the bUnwarpJ and NiftyReg methods, see Section III. The teams were also asked to submit a description of their method. Seven teams complied and were invited to the workshop on April 11, during the ISBI 2019 conference, and to participate in writing this article. Winners were announced at this workshop. One of the teams (TUB) was unable to travel to the workshop, so only 6 teams presented there. The results announced at the workshop are given in Table III (column 'rank'); the results shown here may be based on submissions updated later. Note also that the teams ranked 7, 9 and 11 chose not to provide a description of their methods.

The challenge remains open to submissions and all images, as well as landmarks for the training image pairs, are available for download. The testing landmarks are kept confidential.

III. Registration methods

Let us briefly describe the evaluated registration methods. See also Table II for an overview. The participants were instructed to use the same settings for all test image pairs. Note, that the method names used here may differ from the team names found on the ANHIR webpage.

[6]N. Tahmasebi, M. Noga, K. Punithakumar

UA: (U. Alberta)[6] The images are rescaled to 1/40 of their size and converted to grayscale. A translation and rotation is estimated first, followed by a non-rigid registration using a moving mesh framework [54], which alternates gradient descent on densely represented deformation and correction steps imposing incompressibility. Both steps use a mutual-information similarity criterion.

TUNI: (U. Tampere)[7] The images are converted to grayscale and histogram equalized by Contrast Limited Adaptive Histogram Equalization (CLAHE) [55]. The registration is done by Elastic Stack Alignment (ESA) [56], which first finds a rigid transformation using Random Sample Consensus (RANSAC) from Scale-Invariant Feature Transform (SIFT) features. A non-linear registration follows using normalized correlation and virtual springs to keep the transformation close to a rigid one. Good parameters for the ESA algorithm were found by a parallel grid search on a computational cluster based on [57].

CKVST: (Chengdu)[8] The images are decomposed into channels by a stain deconvolution procedure [40] and only the hematoxylin channel is used. A rigid transformation (including scaling) is estimated first on downsampled images. The transformation is further refined by evaluating the similarity on a set of selected high resolution patches. Finally, a non-rigid registration represented by B-splines is found locally. The normalized cross correlation is used for both rigid and non-rigid registration, with gradient descent and limited-memory Broyden-Fletcher-Goldfard-Shanno (L-BFGS) optimizers, respectively.

MEVIS: (Fraunhofer)[9] Normalized gradient field (NGF) similarity [58]–[60] of grayscale versions of the images is used in a three-step approach. First, centers of mass are aligned and a number of different rotations are tried. Second, an affine transformation is found by a Gauss-Newton method. Third, a dense, non-rigid transformation is found using curvature regularization and L-BFGS optimization. The method is accelerated by optimizing memory access and by precalculation of reduced-size images.

AGH: (AGH UST)[10] The images are converted to grayscale, downsampled and histogram equalized (for some of the procedures). The key feature of the AGH method is to apply several different approaches and then automatically select the best result based on a similarity criterion. An initial similarity or rigid transformation was determined by RANSAC from feature points detected by SIFT, ORB, or SURF. If it failed, the initial transformation was found by aligning binary tissue masks calculated from the images. Second, a non-rigid transformation was found using local affine registration, various versions of the demons algorithms, or a feature-point-based thin-plate spline interpolation. For most of the cases, the most effective procedure was the MIND-based Demons algorithm.

[7]M. Valkonen. K. Kartasalo, L. Latonen, P. Ruusuvuori
[8]Y. Xiang, Y. Yan, Y. Wang
[9]J. Lotz, N. Weiss, S. Heldmann
[10]M. Wodzinski, A. Skalski, https://github.com/lNefarin/ANHIR_MW

TABLE II
EVALUATED METHODS. NGF — NORMALIZED GRADIENT FIELD, NCC — NORMALIZED CORRELATION COEFFICIENT, MI — MUTUAL INFORMATION, SSD — SUM OF SQUARED DIFFERENCES, LM — LEVENBERG-MARQUARDT, L-BFGS — LIMITED MEMORY BROYDEN–FLETCHER–GOLDFARB–SHANNO, ASGD — ADAPTIVE STOCHASTIC GRADIENT DESCENT, CNN — CONVOLUTIONAL NEURAL NETWORK. STARS MARKS WELL-KNOWN METHODS ADDED BY THE ORGANIZERS.

Method	Freely available	Grayscale	Exhaustive search	Affine/rigid prealignment	Feature points	Non-linear transformation	Criterion	Optimization/other
UA [54]		✓		✓		dense moving mesh	MI	gradient
TUNI [56]		✓		✓	✓	virtual springs	NCC	robust linear
CKVST		✓		✓		B-splines	NCC	L-BFGS
MEVIS [58]		✓	✓	✓		dense	NGF	L-BFGS
AGH	✓	✓	✓	✓	✓	various	various	various
UPENN [61]	✓	✓	✓	✓		dense	NCC	L-BFGS, gradient
TUB [53]		✓		✓		dense	NCC	CNN
bUnwarpJ⋆ [22]	✓	✓				B-splines	SSD	LM
RVSS⋆ [22]	✓	✓		✓	✓	B-splines	SSD	LM
NiftyReg⋆ [69]	✓	✓		✓		B-splines	NCC, MI	conjugated-gradients
Elastix⋆ [70]	✓	✓		✓		B-splines	NCC	ASGD
ANTs⋆ [71]	✓	✓		✓		B-splines	MI, NCC	L-BFGS
DROP⋆ [73]	✓	✓				B-splines	SSD	discrete optimization

UPENN: (U. Pennsylvania) [61][11] A stain deconvolution [40] is used to remove the DAB (diaminobenzidine) stain and make the appearance more uniform. The images are rescaled to 4% of the original size, converted to grayscale, and padded with a random noise. For robustness, 5000 random initial rigid transformations are evaluated. The registration itself is performed starting with an initial affine step using L-BFGS optimization, followed by the estimation of a detailed deformation field, using a fast multi-resolution version of the dense non-linear diffeomorphic registration with the *Greedy* tool [62] and a local NCC criterion.

TUB: (Tsinghua U.)[12] A combined standard deviation of all color channels in a Gaussian neighborhood with $\sigma = 1$ pixel is used as a scalar feature for subsequent registration. Image size is reduced to about 512×512 pixels and padded to keep the aspect ratio. The registration is performed by a convolutional neural network similar to the Volume Tweening Network [63]. First a rigid and then a dense non-linear transformation is estimated. The network is first trained in an unsupervised manner, maximizing an image correlation coefficient, then fine-tuned using provided landmark positions on the training data.

The seven methods described come from the ANHIR challenge as described in Section II-E. For completeness, we have also included several well known, freely available and often used registration methods, which we already tested in [16], to enable a better comparison with the state of the art. These methods (marked by ⋆ in Tables II and III) were not part of the ANHIR competitions but were run by the organizers on the same data and evaluation protocols.

[11]L. Venet, S. Pati, P. Yushkevich, S. Bakas, https://github.com/CBICA/HistoReg

[12]S. Zhao, Y. Xu, E. I-Chao Chang

bUnwarpJ: The images are registered using the *bUnwarpJ* ImageJ plugin [22] which calculates a non-linear B-spline-based deformation in a multiresolution way, minimizing a sum of squares difference (SSD) criterion, generalizing earlier B-spline registration methods [64], [65]. The key feature of bUnwarpJ is to simultaneously calculate the deformation in both directions, enforcing invertibility by a consistency constraint and providing strong regularization. Additional regularization is possible by penalizing the deformation field divergence and curl.

RVSS: *Register virtual stack slices* is another ImageJ registration plugin which extends the bUnwarpJ method [22] by incorporating also SIFT [66] feature points in both images to add robustness. We start by estimating a rigid transformation, which we then refine using a B-spline transformation.

NiftyReg: The NiftyReg software is used to first perform an affine registration based on block matching [67], [68] and then a B-spline non-linear registration [69].

Elastix: Elastix [70] is a registration software providing a simple to use command-line interface to the Insight Segmentation and Registration Toolkit (ITK). We use a B-spline deformation and MI similarity criteria with the adaptive stochastic gradient descent optimizer.

ANTs: Advanced Normalization Tools [71] is another registration toolkit using ITK as a backend. We first perform an affine registration by MI maximization, followed by a dense symmetric diffeomorphic registration (SyN) using cross-correlation.

DROP: Similarly to many previously mentioned approaches, DROP [72], [73] represents the deformation using B-splines, minimizing a sum of absolute differences (SAD) (ℓ_1) criterion. It differs by treating registration as a discrete optimization problem, solved efficiently in a mul-

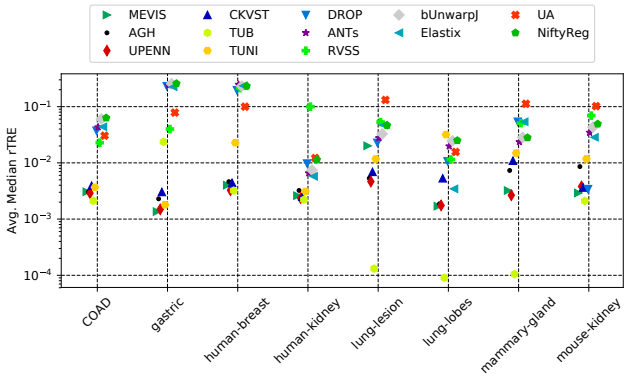


Fig. 4. Relative performance of the methods on the different datasets in terms of AMrTRE.

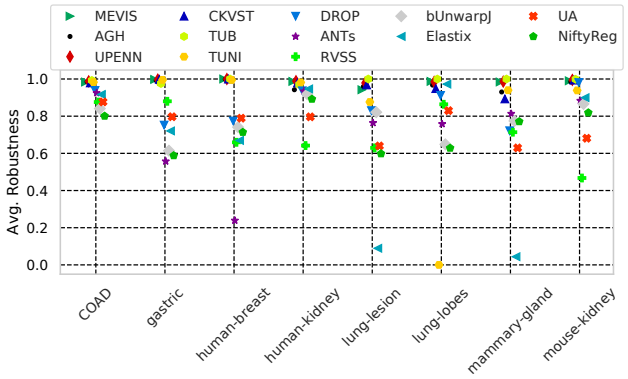


Fig. 5. Robustness (the relative number of landmarks which moved towards the ground truth location) of all methods on different tissues.

tiresolution fashion using linear programming.

IV. Results

A. Relative target error comparison

Our main evaluation criterion is ARMrTRE, average rank of median relative target registration error (see Section II-C, eq. 2). We show the values of ARMrTRE for all methods in Fig. 6, together with AMrTRE (4), and MMrTRE (5). We see that MEVIS is most often the best method, with AGH and UPENN within the top four, and TUB, TUNI, and CKVST more variable but close behind. On the other hand, the general methods (NiftyReg, bUnwarpJ, Elastix, DROP and ANTs), not optimized for microscopy images, are always in the second half of the ranking. It turns out that the results are almost the same when the other criteria from Section II-C are employed. A graphical comparison of those as well as other criteria based on of aggregating the relative landmark error rTRE can be found in Fig. S1 in the Supplementary material.

For quantitative values on test data, see Table III. Since some datasets contain only training and no testing data, for completeness we also provide quantitative results on all data combined in Supplementary material Table SII.

B. Pairwise method comparison

Fig. 7 shows the color coded p-values for pairwise comparison of the methods, as described in Section II-D. It turns out

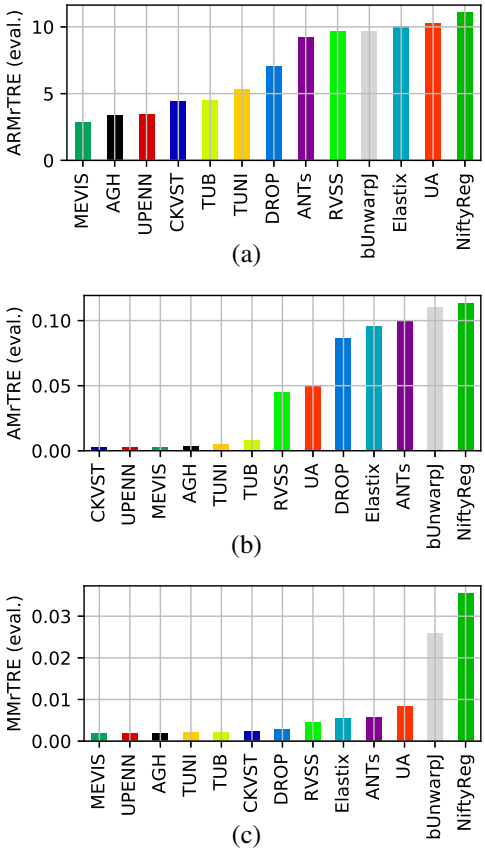


Fig. 6. Average rank of median rTRE (a), average median rTRE (b), and median of median rTRE (c) per method on the test data.

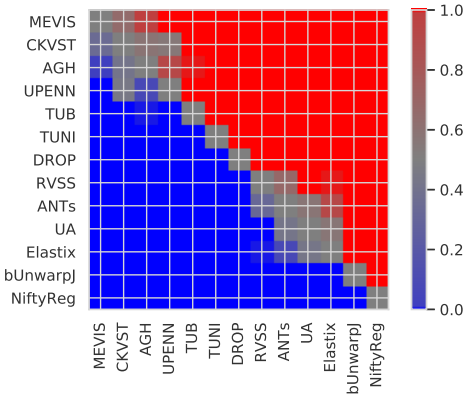


Fig. 7. Color coded p-values for pairwise comparison of methods over all test image pairs in terms of median rTRE, where the median is taken over all landmarks in an image. Red/blue indicates, that the method in the row performs statistically better/worse than the column method, respectively, according to the one-sided Wilcoxon signed-rank test at significance level $p = 0.01$. Gray and shades of gray indicate no significant differences. Methods are topologically sorted by performance from top to bottom.

that we could (topologically) sort the methods by performance from the best methods at the top to the worst performers at the bottom. All differences between methods are significant with the following exceptions: There are no significant differences between the first three methods (MEVIS, CKVST, AGH), between CKVST, AGH, and UPENN, between ANTs and RVSS, and between UA, ANTs and Elastix methods.

C. Differences between datasets

We may ask whether some datasets are more difficult than others and whether some methods are particularly suited for specific datasets. It seems that the answer is yes to both questions. For example, according to Fig. 4, the (CNN based) TUB method outperforms all other methods by an order of magnitude in terms of AMrTRE on some datasets (lung lesions, lung lobes). However, this could be also due to overfitting, since for these datasets all landmark positions were available for training since the beginning and none of them are part of the testing set. Note that evaluation was done on the testing pairs for all datasets except lung lesions and lobes, and mammary glands, which contain only training pairs.

Fig. 5 compares the robustness (8) by datasets. We see that the best six methods have a very high robustness ($> 98\%$), while others completely fail on some of the datasets (e.g. Elastix or TUNI). Fig. S4 in the Supplementary material compares the methods' performance as a function of the combination of stains being registered.

D. Training/testing

In Fig. S5, we compare the performance in terms of the distribution of the median rTRE and rank median rTRE between the training and testing data. For most methods, the distributions on training and testing data are rather similar. However, we note that TUB performs much better on training than on testing data, hinting again on possible overfitting. Interestingly, some methods, like MEVIS or CKVST, actually perform slightly better on the testing data.

V. Discussion and conclusions

Our experience with organizing the ANHIR challenge and a very positive reaction of the community, including a good attendance of the ANHIR workshop at ISBI 2019, confirmed that there is indeed a wide interest in such challenges.

However, organizing such a challenge requires a significant effort. The most time-consuming part was annotating the images with landmarks. Although the landmarks were sufficient to our purpose, there are sometimes large areas without salient corresponding structures, where very few landmarks can be identified. In the future, we would recommend that manually obtained landmarks are combined with other complementary means of evaluating the registration quality, such as measuring overlap using reference segmentations (as in NIREP [11]), or using synthetically generated (and therefore known) deformations [64]. We became aware that for the results to be meaningful, we must discourage the participants to overtrain on the provided datasets. Future challenges should also measure the speed of the algorithms more reliably, ideally by using a common hardware for all methods. Besides pairwise registration, we might consider other more advanced tasks such as 3D reconstruction by registering the histology slices, creating a large 2D mosaic from a set of smaller tile images, aligning 2D microscopy images with volumetric medical in-vivo images (e.g. MRI or ultrasound), or aligning standard histopathology with high-dimensional microscopy modalities such as MALDI [74]. For these applications, completely new evaluation approaches might be needed.

Perhaps surprisingly, most of the submitted methods (see Section III) used very similar and rather classical techniques. The best performing algorithms were originally developed for other modalities but were tuned for the data at hand. They consist of a carefully designed multi-step pipeline, which typically started with a coarse but robust prealignment, followed by a non-rigid registration step for fine tuning the results. Only one method (TUB) used the now ubiquitous CNNs; it was very fast and performed well, although not as well as the best performing methods and its generalization ability was limited. Interestingly, directly applying well-known off-the-shelf methods gave significantly worse results, although these methods contained very similar algorithmic ingredients as some of the best performing methods (e.g., MEVIS, CKVST, AGH). One explanation is that those general purpose methods were developed for other types of images, mostly 3D brain data, and possibly required more parameter tuning. Our observation is that the better performing methods often won not by being more accurate but by being more robust (see Table III), converging to the solution even from a far away initialization. The participants were clearly aware of this, since most submitted methods spend a significant effort to find a good initialization, for example by incorporating automatically detected landmarks (TUNI, AGH), trying many different starting points (MEVIS, UPENN), or even applying different methods in parallel (AGH). Another reason for the performance gap might be that the participants spent more time finding the optimal parameters for their own methods (even using a computing cluster in TUNI).

To summarize the evaluation, there was very little difference in terms of both robustness and accuracy between the first six methods (MEVIS, AGH, UPENN, CKVST, TUB, TUNI). Those six methods outperformed all others in all evaluation measures that we have tested.

On the other hand, there were large differences in the reported execution speed (Table III). With the exception of the extremely fast CNN method (TUB), the differences seem to be more due to the quality of engineering (implementation) and the choice of parameters, and not the choice of the key algorithms. While MEVIS and NiftyReg take on average a few seconds per registration, the majority of the methods need a few minutes, with the slowest method (ANTs) taking almost an hour. Note however that MEVIS uses precalculation extensively and NiftyReg might have taken advantage of the GPU.

No methods were developed explicitly for the ANHIR challenge or for this type of images. No evaluated method takes advantage of the full resolution or full color information.

TABLE III
QUANTITATIVE RESULT OF ALL METHODS ON THE EVALUATION DATA. THE FIRST AND SECOND ROWS ('AVERAGE', 'MEDIAN' ETC.) CORRESPOND TO THE AGGREGATION METHOD WITHIN EACH IMAGE PAIR AND OVER ALL IMAGE PAIRS, RESPECTIVELY. THE TABLE IS SORTED BY ARMRTRE (IN BOLD). RANK CORRESPONDS TO WORKSHOP RESULTS, MISSING NUMBERS = INSUFFICIENT INFORMATION. ⋆ = METHODS ADDED BY THE ORGANIZERS.

method	rank	Average rTRE		Median rTRE		Max rTRE		Robustness		Median rTRE	Max rTRE	*Average*
		Average	Median	Average	Median	Average	Median	Average	Median	Average Rank		*time*
		(AArTRE)		(AMrTRE)		(AMxrTRE)		R		(**ARMrTRE**)	(ARMxrTRE)	[min]
initial		*0.1340*	*0.0684*	*0.1354*	*0.0665*	*0.2338*	*0.1157*	-	-	-		-
MEVIS	1	0.0044	0.0027	0.0029	0.0018	0.0251	0.0188	0.9880	1.0000	2.84	5.04	*0.17*
AGH	3	0.0053	0.0032	0.0036	0.0019	0.0283	0.0225	0.9821	1.0000	3.42	6.00	*6.55*
UPENN	2	0.0042	0.0029	0.0029	0.0019	0.0239	0.0190	0.9898	1.0000	3.47	4.29	*1.60*
CKVST	4	0.0043	0.0032	0.0027	0.0023	0.0239	0.0189	0.9883	1.0000	4.41	5.27	*7.80*
TUB	5	0.0089	0.0029	0.0078	0.0021	0.0280	0.0178	0.9845	1.0000	4.53	3.81	*0.02*
TUNI	6	0.0064	0.0031	0.0048	0.0021	0.0287	0.0204	0.9823	1.0000	5.32	5.80	*9.73*
DROP⋆		0.0861	0.0042	0.0867	0.0028	0.1644	0.0273	0.8825	0.9892	7.06	7.43	*3.99*
ANTs⋆		0.0991	0.0072	0.0992	0.0058	0.1861	0.0351	0.7889	0.9714	9.23	7.79	*48.24*
RVSS⋆		0.0472	0.0063	0.0448	0.0046	0.1048	0.0275	0.8155	0.9928	9.65	8.42	*5.25*
bUnwarpJ⋆		0.1097	0.0290	0.1105	0.0260	0.1995	0.0727	0.7899	0.9310	9.67	9.37	*10.57*
Elastix⋆		0.0964	0.0074	0.0956	0.0054	0.1857	0.0353	0.8477	0.9722	10.04	8.88	*3.50*
UA	10	0.0536	0.0100	0.0506	0.0082	0.1124	0.0353	0.8209	0.9853	10.28	8.83	*1.70*
NiftyReg⋆		0.1120	0.0372	0.1136	0.0355	0.2010	0.0714	0.7427	0.8519	11.08	10.08	*0.14*

With a median landmark localization error measured in tens of pixels at the original resolution even for the best methods, none of the methods would be probably robust and accurate enough for routine, fully automatic use. However, these methods could certainly be used in a semi-supervised setting to substantially reduce the need for manual interaction in many tasks, such as fusing information from several differently stained images.

The above mentioned weaknesses will likely be addressed in future research, e.g. by creating algorithms capable of working on full resolution images or avoiding the overfitting in CNNs. We therefore believe that the need for objective performance evaluation will remain and that challenges such as ANHIR should be continued to stimulate and evaluate research in image registration.

ACKNOWLEDGEMENTS

The authors are grateful to Ben Glocker for his help with tuning the DROP parameters.

REFERENCES

[1] B. Zitová and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, no. 21, pp. 977–1000, 2003.
[2] L. Brown, "A survey of image registration techniques," *ACM Computing Surveys*, vol. 24, no. 4, pp. 326–376, Dec. 1992.
[3] J. Maintz and M. A. Viergever, "A survey of medical image registration," *Medical Image Analysis*, vol. 2, no. 1, pp. 1–36, 1998.
[4] H. Lester and S. R. Arridge, "A survey of hierarchical non-linear medical image registration," *Pattern Recognit.*, vol. 32, no. 1, pp. 129–149, 1999.
[5] J. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: A survey," *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.
[6] F. P. M. Oliveira and J. M. R. S. Tavares, "Medical image registration: a review," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 17, no. 2, pp. 1–21, 2012.
[7] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable Medical Image Registration: A Survey," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1153–1190, July 2013.
[8] B. Zitova, "Mathematical Approaches for Medical Image Registration," in *Reference Module in Biomedical Sciences*, 2018.
[9] J. West, J. Fitzpatrick, and M. Wang, "Comparison and evaluation of retrospective intermodality brain image registration techniques," *Journal of Computer Assisted Tomography*, vol. 21, no. 4, pp. 554–568, 1997.
[10] P. Hellier, *et al.*, "Retrospective evaluation of intersubject brain registration," *IEEE Trans. Med. Imag.*, vol. 22, no. 9, pp. 1120 – 1130, 2003.
[11] G. Christensen, *et al.*, "Introduction to the non-rigid image registration evaluation project (NIREP)," in *Biomedical Image Registration*. Springer, 2006, vol. 4057, pp. 128–135.
[12] A. Klein, J. Andersson, B. Ardekani, *et al.*, "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," *Neuroimage*, vol. 46, no. 3, pp. 786–802, 2009.
[13] Y. Ou, H. Akbari, M. Bilello, X. Da, and C. Davatzikos, "Comparative evaluation of registration algorithms in different brain databases with varying difficulty: Results and insights," *IEEE Trans. Med. Imag.*, vol. 33, no. 10, pp. 2039–2065, 2014.
[14] D. Sarrut, B. Delhay, P.-F. Villard, V. Boldea, M. Beuve, and P. Clarysse, "A Comparison Framework for Breathing Motion Estimation Methods From 4-D Imaging," *IEEE Trans. Med. Imag.*, vol. 26, pp. 1636–1648, 2007.
[15] K. K. Brock, "Results of a multi-institution deformable registration accuracy study (MIDRAS)," *International Journal of Radiation Oncology · Biology · Physics*, vol. 76, no. 2, pp. 583 – 596, 2010.
[16] J. Borovec, A. Munoz-Barrutia, and J. Kybic, "Benchmarking of Image Registration Methods for Differently Stained Histological Slides," in *Proc. Int. Conf. Image Process.*, Athens, 2018, pp. 3368–3372.
[17] R. Castillo, E. Castillo, R. Guerra, *et al.*, "A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets." *Phys. Med. Biol.*, vol. 54, no. 7, pp. 1849–1870, 2009.
[18] K. Murphy, B. Van Ginneken, J. Reinhardt, *et al.*, "Evaluation of registration methods on thoracic CT: The EMPIRE10 challenge," *IEEE Trans. Med. Imag.*, vol. 30, no. 11, pp. 1901–1920, 2011.
[19] B. Pontre, *et al.*, "An open benchmark challenge for motion correction of myocardial perfusion MRI." *IEEE J. Biomedical and Health Informatics*, vol. 21, no. 5, pp. 1315–1326, 2017.
[20] M. T. McCann, J. A. Ozolek, C. A. Castro, B. Parvin, and J. Kovacevic, "Automated histology analysis: Opportunities for signal processing," *IEEE Signal Processing Magazine*, vol. 32, no. 1, pp. 78–87, Jan 2015.
[21] J. Pichat, J. E. Iglesias, T. Yousry, S. Ourselin, and M. Modat, "A survey of methods for 3D histology reconstruction." *Medical Image Analysis*, vol. 46, pp. 73–105, 2018.
[22] I. Arganda-Carreras *et al.*, "Consistent and elastic registration of histological sections using vector-spline regularization," in *Computer Vision Approaches to Medical Image Analysis*, vol. 4241, 2006, pp. 85—-95.
[23] M. Feuerstein, T. H. Heibel, J. Gardiazabal, N. Navab, and M. Groher, "Reconstruction of 3-D histology images by simultaneous deformable registration." in *MICCAI*, vol. 6892, 2011, pp. 582–589.

[24] P. A. Yushkevich, *et al.*, "3D mouse brain reconstruction from histology using a coarse-to-fine approach." in *WBIR*, vol. 4057, 2006, pp. 230–237.

[25] Y. Song, D. Treanor, A. Bulpitt, *et al.*, "3D reconstruction of multiple stained histology images," *J. Path. Inf.*, vol. 4, no. 2, p. 7, 2013.

[26] M. Tang, "Automatic registration and fast volume reconstruction from serial histology sections." *Computer Vision and Image Understanding*, vol. 115, no. 8, pp. 1112–1120, 2011.

[27] J. Pichat, M. Modat, T. Yousry, and S. Ourselin, "A multi-path approach to histology volume reconstruction," in *Int. Symp. Biomed. Imag.*, April 2015, pp. 1280–1283.

[28] M. Gibb, *et al.*, "Resolving the three-dimensional histology of the heart." in *Comp. Meth. Syst. Biol.*, vol. 7605. Springer, 2012, pp. 2–16.

[29] M. Mueller, M. Yigitsoy, H. Heibel, and N. Navab, "Deformable reconstruction of histology sections using structural probability maps," in *MICCAI*, 2014.

[30] A. du Bois D'Aische and M. Craene, "Efficient multi-modal dense field non-rigid registration: alignment of histological and section images," *Medical Image Analysis*, vol. 9, no. 6, pp. 538–546, 2005.

[31] S. Saalfeld, A. Cardona, V. Hartenstein, and P. Tomančák, "As-rigid-as-possible mosaicking and serial section registration of large sstem datasets," *Bioinformatics*, vol. 26, no. 12, pp. –57, 2010.

[32] G. J. Metzger, S. C. Dankbar, J. Henriksen, *et al.*, "Development of multigene expression signature maps at the protein level from digitized immunohistochemistry slides." *PloS One*, vol. 7, no. 3, p. e33520, 2012.

[33] D. Obando, A. Frafjord, I. Oynebraten, *et al.*, "Multi-staining registration of large histology images," in *Int. Symp. Biomed. Imag.*, 2017, pp. 345–348.

[34] O. Déniz, D. Toomey, *et al.*, "Multi-stained whole slide image alignment in digital pathology," in *SPIE Med. Imag.*, vol. 9420, 2015, p. 94200Z.

[35] N. Trahearn, D. Epstein, I. Cree, D. Snead, and N. Rajpoot, "Hyper-stain inspector: A framework for robust registration and localised co-expression analysis of multiple whole-slide images of serial histology sections," *Scientific Reports*, vol. 7, 2017.

[36] C. Ceritoglu and L. Wang, "Large deformation diffeomorphic metric mapping registration of reconstructed 3D histological section images and in vivo MR images," *Front. Hum. Neurosci.*, vol. 4, p. 43, 2010.

[37] L. Gupta, B. M. Klinkhammer, P. Boor, D. Merhof, and M. Gadermayr, "Stain independent segmentation of whole slide images: A case study in renal histology," in *Int. Symp. Biomed. Imag.*, 2018, pp. 1360–1364.

[38] J. Borovec and J. Kybic, "Binary pattern dictionary learning for gene expression representation in drosophila imaginal discs," in *Workshop Math. Comp. Meth. Biomed. Imag. Image Anal.*, 2017, pp. 555–569.

[39] C. Wang, S. Ka, and A. Chen, "Robust image registration of biological microscopic images," *Scientific Reports*, vol. 4, no. 1, p. 6050, 2015.

[40] A. C. Ruifrok and D. A. Johnston, "Quantification of histochemical staining by color deconvolution," *Anal. Quant. Cytol. Histol.*, vol. 23, no. 4, pp. 291–299, 2001.

[41] J. Kybic and J. Borovec, "Automatic simultaneous segmentation and fast registration of histological images," in *Int. Symp. Biomed. Imag.*, 2014, pp. 774–777.

[42] J. Borovec, J. Kybic, M. Bušta, C. Ortiz-de Solorzano, and A. Muñoz-Barrutia, "Registration of multiple stained histological sections," in *Int. Symp. Biomed. Imag.*, San Francisco, 2013, pp. 1034–1037.

[43] Y. Song, *et al.*, "Unsupervised content classification based nonrigid registration of differently stained histology images." *IEEE Trans. Biomed. Engineering*, vol. 61, no. 1, pp. 96–108, 2014.

[44] O. Lobachev, C. Ulrich, B. Steiniger, V. Wilhelmi, V. Stachniss, and M. Guthe, "Feature-based multi-resolution registration of immunostained serial sections." *Medical Image Analysis*, vol. 35, pp. 288–302, 2017.

[45] J. Kybic, M. Dolejsi, and J. Borovec, "Fast registration of segmented images by normal sampling," in *Bio Image Computing (BIC) workshop at CVPR*, 2015, pp. 11–19.

[46] L. Solorzano, G. Almeida, B. Mesquita, D. Martins, C. Oliveira, and C. Wählby, "Whole slide image registration for the study of tumor heterogeneity," in *COMPAY: Comp. Pathol. workshop*, 2018, pp. 95–102.

[47] D. V. Sorokin, I. Peterlik, M. Tektonidis, K. Rohr, and P. Matula, "Non-rigid contour-based registration of cell nuclei in 2-D live cell microscopy images using a dynamic elasticity model," *IEEE Trans. Med. Imag.*, vol. 37, no. 1, pp. 173–184, 2017.

[48] D. V. Sorokin, J. Suchánková, E. Bártová, and P. Matula, "Visualizing stable features in live cell nucleus for evaluation of the cell global motion compensation," *Folia biologica*, vol. 60, no. Suppl. 1, pp. 45–49, 2014.

[49] R. Shojaii, T. Karavardanyan, M. J. Yaffe, and A. L. Martel, "Validation of histology image registration." in *SPIE Med. Imag.*, vol. 7962. SPIE, 2011, p. 79621E.

[50] C. W. Wang, E. Budiman Gosno, and Y. S. Li, "Fully automatic and robust 3D registration of serial-section microscopic images," *Scientific Reports*, vol. 5, 2015.

[51] J. Borovec, "BIRL: Benchmark on image registration methods with landmark validation," Czech Tech. Univ. in Prague, Tech. Rep., 2019.

[52] S. Bakas, *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 11 2018, arXiv:1811.02629.

[53] T. F. Lau, J. Luo, S. Zhao, E. I.-C. Chang, and Y. Xu, "Unsupervised 3D end-to-end medical image registration with volume tweening network." *CoRR*, vol. abs/1902.05020, 2019.

[54] K. Punithakumar, P. Boulanger, and M. Noga, "A GPU-accelerated deformable image registration algorithm with applications to right ventricular segmentation," *IEEE Access*, vol. 5, pp. 20 374–20 382, 2017.

[55] S. M. Pizer, *et al.*, "Adaptive histogram equalization and its variations," *Comput. Vision Graph. Image Process.*, vol. 39, no. 3, pp. 355–368, 1987.

[56] S. Saalfeld, R. Fetter, A. Cardona, and P. Tomancak, "Elastic volume reconstruction from series of ultra-thin microscopy sections," *Nat. Meth.*, vol. 9, no. 7, pp. 717–720, July 2012.

[57] K. Kartasalo, L. Latonen, J. Vihinen, T. Visakorpi, M. Nykter, and P. Ruusuvuori, "Comparative analysis of tissue reconstruction algorithms for 3D histology." *Bioinformatics*, vol. 34, no. 17, pp. 3013–3021, 2018.

[58] J. Lotz, N. Weiss, and S. Heldmann, "Robust, fast and accurate: a 3-step method for automatic histological image registration," 2019. [Online]. Available: http://arxiv.org/abs/1903.12063

[59] O. Schmitt, J. Modersitzki, S. Heldmann, S. Wirtz, and B. Fischer, "Image registration of sectioned brains," *Int. J. Comp. Vis.*, vol. 73, no. 1, pp. 5–39, Sept. 2006.

[60] J. Modersitzki, *FAIR: Flexible Algorithms for Image Registration*. SIAM, 2009.

[61] L. Venet, S. Pati, P. Yushkevich, and S. Bakas, "Accurate and robust alignment of variable-stained histologic images using a general-purpose greedy diffeomorphic registration tool," 2019, arXiv:1904.11929.

[62] P. A. Yushkevich, J. Pluta, H. Wang, L. E. Wisse, S. Das, and D. Wolk, "Fast automatic segmentation of hippocampal subfields and medial temporal lobe subregions in 3 Tesla and 7 Tesla T2-weighted MRI," *Alzheimer's & Dementia*, vol. 12, no. 7, Supp., pp. P126 – P127, 2016.

[63] S. Zhao, T. Lau, J. Luo, E. I. Chang, and Y. Xu, "Unsupervised 3D end-to-end medical image registration with volume tweening network," *IEEE J. Biomed. Health Informa.*, 2019.

[64] J. Kybic and M. Unser, "Fast parametric elastic image registration," *IEEE Trans. Imag. Proc.*, vol. 12, no. 11, pp. 1427–1442, 2003.

[65] C. Sánchez Sorzano, P. Thévenaz, and M. Unser, "Elastic registration of biological images using vector-spline regularization," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 4, pp. 652–663, Apr. 2005.

[66] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comp. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[67] S. Ourselin, A. Roche, G. Subsol, X. Pennec, and N. Ayache, "Reconstructing a 3D structure from serial histological sections," *Image Vision Comput.*, vol. 19, no. 1-2, pp. 25–31, 2001.

[68] M. Modat, D. Cash, P. Daga, G. Winston, J. S. Duncan, and S. Ourselin, "A symmetric block-matching framework for global registration," in *SPIE Med. Imag.*, vol. 9034, 03 2014, p. 90341D.

[69] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[70] S. Klein, M. Staring, and K. Murphy, "Elastix: a toolbox for intensity-based medical image registration," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, 2010.

[71] B. Avants, C. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[72] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios, "Dense image registration through MRFs and efficient linear programming," *Medical Image Analysis*, vol. 12, no. 6, pp. 731–741, 2008.

[73] B. Glocker *et al.*, "Deformable medical image registration: Setting the state of the art with discrete methods," *Ann. Rev. Biomed. Eng.*, vol. 13, no. 1, pp. 219–244, 2011.

[74] M. Machálková, *et al.*, "Drug penetration analysis in 3D cell cultures using fiducial-based semiautomatic coregistration of MALDI MSI and immunofluorescence images," *Anal. Chem.*, vol. 91, no. 21, pp. 13 475–13 484, 2019.

Supplementary material for "ANHIR: Automatic Non-rigid Histological Image Registration Challenge"

Jiří Borovec, Jan Kybic* *Senior Member*, Ignacio Arganda-Carreras, Dmitry V. Sorokin, Maria Gloria Bueno Garcia, Alexander V. Khvostikov, Spyridon Bakas, Eric I-Chao Chang, Stefan Heldmann, Kimmo Kartasalo, Leena Latonen, Johannes Lotz, Michelle Noga, Sarthak Pati, Kumaradevan Punithakumar *Senior Member*, Pekka Ruusuvuori, Andrzej Skalski *Senior Member*, Nazanin Tahmasebi, Masi Valkonen, Ludovic Venet, Yizhe Wang, Nick Weiss, Marek Wodzinski, Yu Xiang, Yan Xu, Yan Yan, Paul Yushkevic, Shengyu Zhao, and Arrate Muñoz-Barrutia *Senior Member*

S.I. Dataset descriptions

Lung lesions: Unstained adjacent $3\,\mu$m formalin-fixed paraffin-embedded sections were cut from the blocks and stained with Hematoxylin and Eosin (H&E) or by immunohistochemistry with a specific antibody for platelet endothelial cell adhesion molecule (PECAM-1, also known as CD31), prosurfactant protein C (proSPC), clara cell 10 protein (CC10) or antigen ki-67 (Ki67). Images of three mice lung lesions (adenoma or adenocarcinoma) were acquired with a Zeiss Axio Imager M1 microscope (Carl Zeiss, Jena, Germany) equipped with a dry Plan Apochromat objective (numerical aperture NA=0.95, magnification $40\times$, pixel size $0.174\,\mu$m/pixel). See also Fig. 2.

Lung lobes: Images of four whole mouse lung lobes, corresponding to the same set of histological samples as in the lung lesions dataset. They were also acquired with a Zeiss Axio Imager M1 microscope (Carl Zeiss, Jena, Germany) equipped with a dry EC Plan-Neofluar objective (NA=0.30, magnification $10\times$, pixel size $1.274\,\mu m/pixel$).

Mammary glands: The sections are cuts from two mouse mammary glands blocks stained with H&E in even sections and in odd sections alternatively with an antibody against the estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). The images were acquired with the same microscope and set of acquisition parameters as the mouse lung lobes, the pixel size was $2.294\,\mu$m/pixel).

COAD: The COlon ADenocarcinoma (COAD) set assembles series of histological sections from colon cancer samples, scanned with a 3DHistech Pannoramic MIDI II scanner at $10\times$ magnification, for a resolution of $0.468\,\mu$m/pixel with a white-balance set to auto. Each series consists of one H&E histopathology section (first cut) followed by a variable number (4–7) of immunohistopathology sections stained with hematoxylin and DAB, with antibodies binding to proteins expressed by various immune cells (T-cells and macrophages). The information about antibodies used was not disclosed on a per-image basis by the owner of the data.

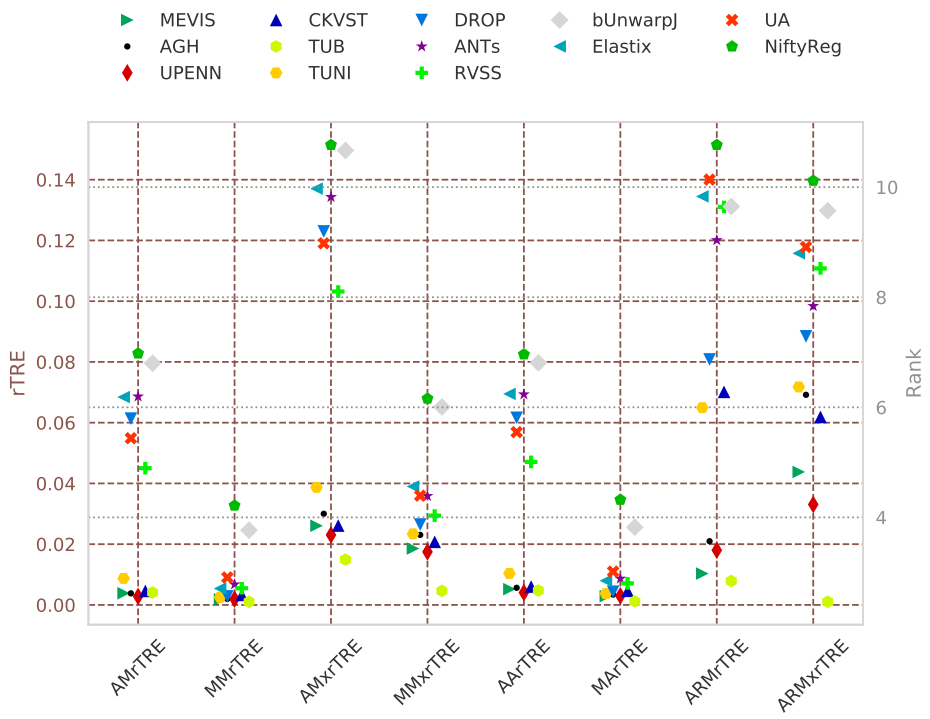
Mouse kidney: The set consists of resected healthy mouse kidneys which show high similarity to human kidneys. We used nine consecutive whole slide images having similar tissue structures. Whole slides were digitized with a NanoZoomer 2.0HT scanner (Hamamatsu) and a $20\times$ objective lens. The images were each roughly of $37\text{k}\times30\text{k}$ pixel size. Each image was dyed with one of the three stains — periodic acid-Schiff (PAS), smooth muscle actin (SMA) or CD31, such that every alternate slide is a PAS image.

Gastric: Surgical material from patients with a histologically verified diagnosis (gastric adenocarcinoma) were used for routine staining with Hematoxylin and Eosin (H&E) or for immunophenotyping. IHC-staining for latent membrane protein 1 (LMP-1) was used for Epstein-Barr virus (EBV) identification. The study of the cellular composition of the tumour tissue infiltrate was performed by immunohistochemical staining on the markers CD4, CD8, CD68 and CD1a. Deparaffinization and antigen recovery was performed by using Thermo Dewax and HIER Bufer L, a pH 6 buffer. The preparations were acquired with a Leica DM LB2 microscope.

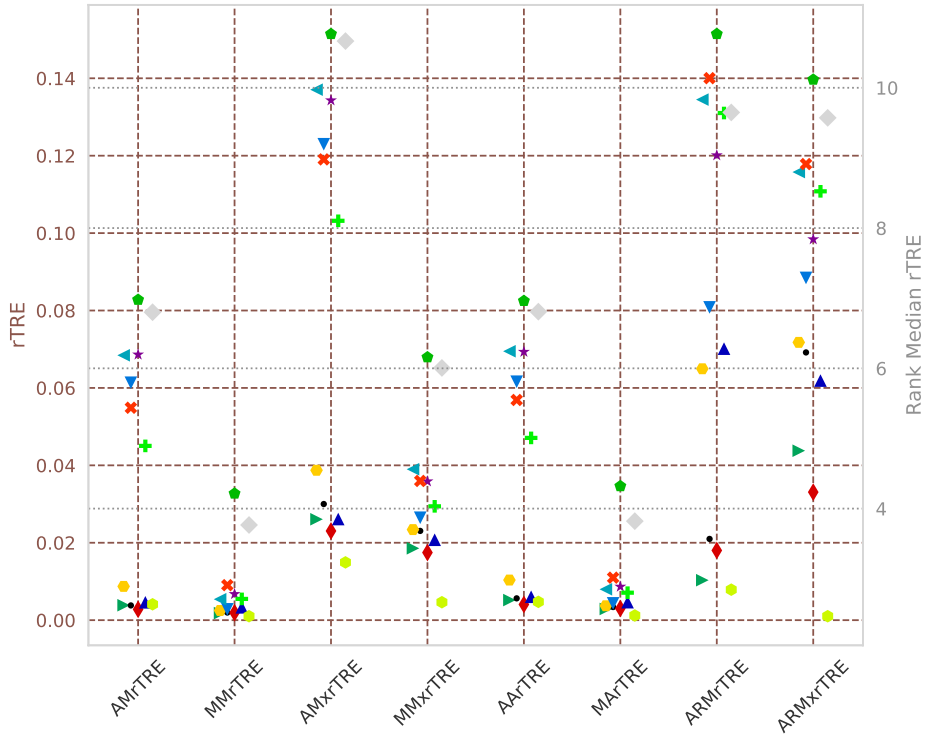
Human breast: Unstained adjacent $3\,\mu$m formalin-fixed paraffin-embedded sections were cut from the blocks, stained with Hematoxylin and Eosin (H&E), and with immunohistochemistry (IHC) with an antibody against ER, PR, and HER2, and imaged with Leica Biosystems Aperio AT2.

Human kidney: Unstained adjacent $3\,\mu$m formalin-fixed paraffin-embedded sections were cut from the glomerulopathies blocks, stained with Hematoxylin and Eosin (H&E) and PAS, Masson and Methenamine, and imaged with Leica Biosystems Aperio AT2.

See Fig. S4 for examples of the appearance differences due to staining between pairs of images from the same sets to be registered. Fig. S3 shows examples of the differences of the local structure.



(a)



(b)

Fig. S1. Relative performance on all data (a) and on test data (b) of all methods measured by AMrTRE (average median rTRE), MMrTRE (median of the median rTRE), AMxrTRE (average maximum rTRE), MMxTRE (median of the maxium rTRE), AArTRE (average of the average rTRE), MArTRE (median of the average rTRE), ARMrTRE (average rank of the median rTRE), and ARMxrTRE (average rank of the maximum rTRE), all based on aggregating rTRE.

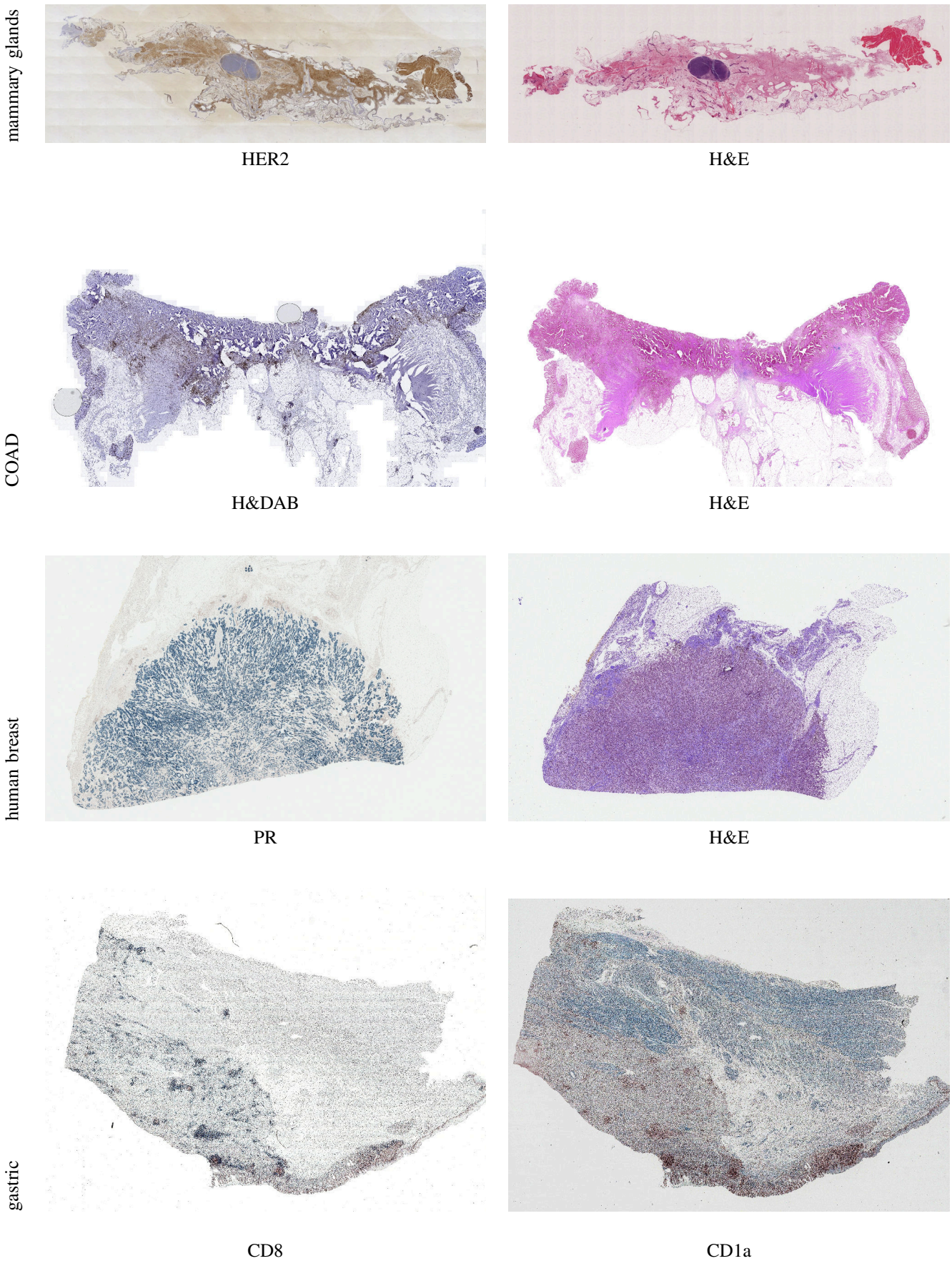
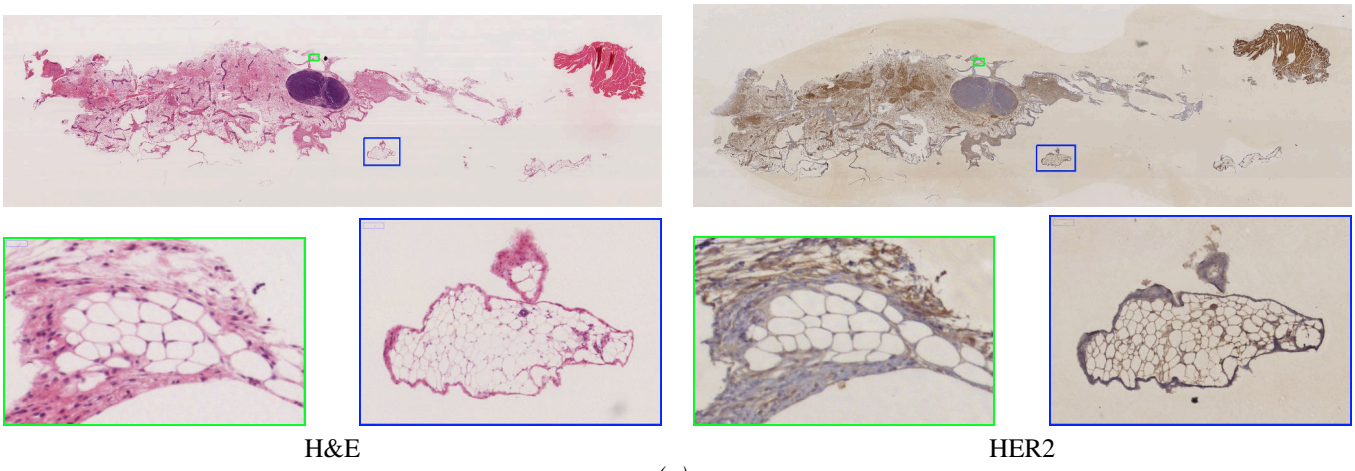
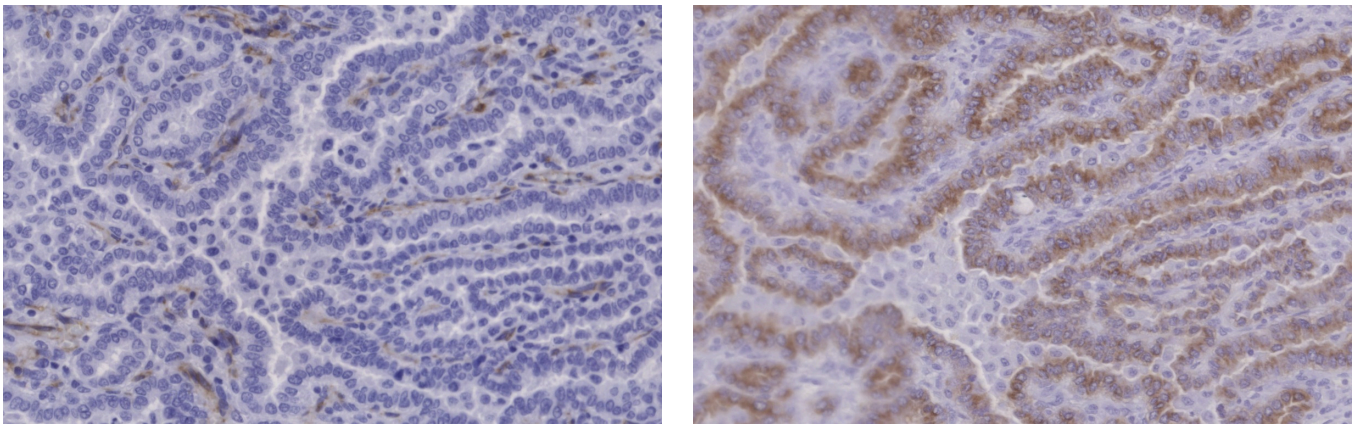


Fig. S2. Examples showing the differences between images to be registered. Each row shows two images from the same dataset with different stains. Some images were clipped and color-enhanced for visualization purposes.



(a)



(b)

Fig. S3. *(a)* Two differently stained images of the mammary gland. The contents of the blue and green rectangles in the first row are shown magnified in the second row. *(b)* Differences in the local structure of two differently stained images of the lung tissue at the same locations from the same set of images.

TABLE SI
GIVEN A SET OF FIVE IMAGES, THE LANDMARKS FOR EVERY THIRD IMAGE (IN BOLD) ARE WITHHELD AND THE REST IS PROVIDED TO PARTICIPANTS. THIS LEADS TO 3 IMAGE PAIRS FOR TRAINING (MARKED BY △) AND 6 IMAGE PAIRS FOR TESTING (MARKED BY ⋆). INVERTED PAIRS ARE NOT COUNTED.

		target image				
		1	2	3	**4**	5
source image	**1**		⋆	⋆		⋆
	2			△		△
	3					△
	4		⋆	⋆		⋆
	5					

S.II. LANDMARK ANNOTATION

Annotation was performed in ImageJ [S1][1] with the help of simple custom macros and scripts, which we provide[2]. The correctness of the landmarks was checked visually in high magnification.

Within each set, landmarks for every third image were withheld by the organizers and the remaining ones were made available to the participants to be used for training. For example, for 5 images in the set, we would withhold landmarks for images 1 and 4, resulting in three image pairs for training $((2,3),(2,5),(3,5))$ and 6 for testing $((1,2),(1,3),(1,5),(4,2),(4,3),(4,5))$, as shown in Table SI.

S.III. WELL KNOWN METHODS

Here we provide pointers to the implementation of existing registration methods which we used for comparison (see Section III of the main article):

bUnwarpJ: https://imagej.net/BUnwarpJ
RVSS: https://imagej.net/Register_Virtual_Stack_Slices
NiftyReg: https://github.com/jonclayden/RNiftyReg
Elastix: http://elastix.isi.uu.nl, based on ITK (https://itk.org/)
ANTs: http://stnava.github.io/ANTs/
DROP: https://www.mrf-registration.net/, https://github.com/biomedia-mira/drop2

[1] https://imagej.net/ImageJ
[2] https://borda.github.io/dataset-histology-landmarks/

TABLE SII

QUANTITATIVE RESULT OF ALL METHODS ON ALL DATA (TRAINING AND TESTING COMBINED), TO BE COMPARED WITH TABLE III OF THE MAIN ARTICLE. THE FIRST AND SECOND ROWS ('AVERAGE', 'MEDIAN' ETC.) CORRESPOND TO THE AGGREGATION METHOD WITHIN EACH IMAGE PAIR AND OVER ALL IMAGE PAIRS, RESPECTIVELY. THE TABLE IS SORTED BY ARMRTRE (IN BOLD). ⋆ = METHODS ADDED BY THE ORGANIZERS.

method	Average rTRE		Median rTRE		Max rTRE		Robustness		Median rTRE	Max rTRE	*Average*
	Average	Median	Average	Median	Average	Median	Average	Median	Average Rank		*time*
	(AArTRE)		(AMrTRE)		(AMxrTRE)		R		(**ARMrTRE**)	(ARMxrTRE)	[min]
initial	*0.1340*	*0.0684*	*0.1354*	*0.0665*	*0.2338*	*0.1157*	-	-	-		-
TUB	0.0047	0.0012	0.0041	0.001	0.0149	0.0046	0.9919	1.0000	2.84	2.47	0.02
MEVIS	0.0052	0.0029	0.0039	0.0018	0.0261	0.0186	0.9845	1.0000	2.98	4.83	0.15
UPENN	0.0041	0.0030	0.0028	0.0019	0.0230	0.0175	0.9888	1.0000	3.40	4.23	1.45
AGH	0.0056	0.0034	0.0038	0.0020	0.0300	0.0231	0.9770	1.0000	3.57	6.23	6.86
TUNI	0.0104	0.0037	0.0087	0.0025	0.0387	0.0234	0.8899	1.0000	5.99	6.37	10.32
CKVST	0.0060	0.0047	0.0046	0.0033	0.0261	0.0208	0.9730	1.0000	6.28	5.83	7.13
DROP⋆	0.0616	0.0043	0.0613	0.0028	0.1230	0.0265	0.8861	0.9907	6.87	7.29	3.41
ANTs⋆	0.0693	0.0087	0.0686	0.0067	0.1343	0.0359	0.8137	0.9718	9.04	7.84	43.09
RVSS⋆	0.0471	0.0071	0.0450	0.0055	0.1032	0.0294	0.7958	0.9875	9.64	8.52	4.72
bUnwarpJ⋆	0.0797	0.0256	0.0796	0.0246	0.1496	0.0652	0.7940	0.9310	9.65	9.57	9.15
Elastix⋆	0.0695	0.0080	0.0684	0.0054	0.1371	0.0390	0.7668	0.9706	9.83	8.80	2.96
UA	0.0569	0.0110	0.0549	0.0090	0.1190	0.0360	0.8076	0.9737	10.14	8.91	1.47
NiftyReg⋆	0.0825	0.0346	0.0828	0.0327	0.1514	0.0679	0.7495	0.8519	10.77	10.12	0.15

TABLE SIII

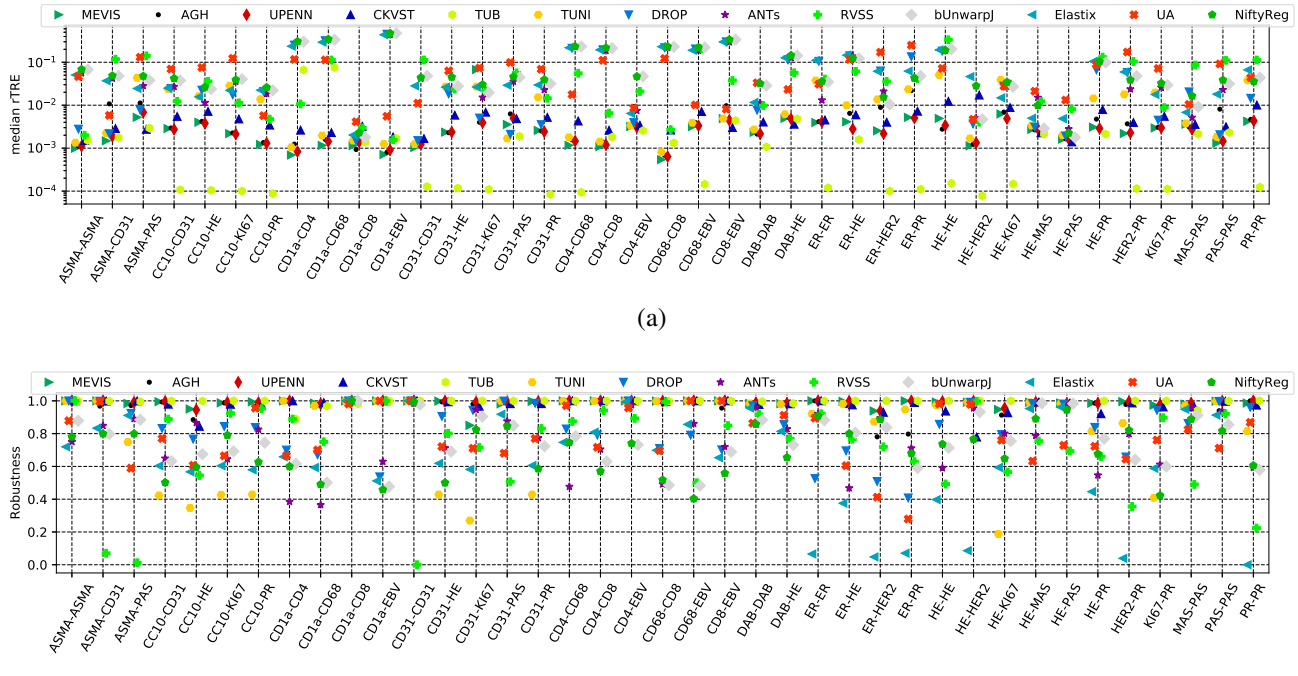
NUMBER OF IMAGE PAIRS BY STAINING FOR THE EVALUATION DATASETS.

reference/moving	ASMA	CC10	CD1a	CD31	CD4	CD68	CD8	DAB	EBV	ER	HE	HER2	KI67	MAS	PAS	PR	*sum*
ASMA	1	-	-	4	-	-	-	-	-	-	-	-	-	-	6	-	11
CC10	-	-	-	1	-	-	-	-	-	-	3	-	6	-	-	3	13
CD1a	-	-	-	-	8	7	1	-	2	-	-	-	-	-	-	-	18
CD31	-	6	-	1	-	-	-	-	-	-	3	-	6	-	2	6	24
CD4	-	-	1	-	-	7	2	-	2	-	-	-	-	-	-	-	12
CD68	-	-	1	-	1	-	1	-	2	-	-	-	-	-	-	-	5
CD8	-	-	-	-	7	7	-	-	2	-	-	-	-	-	-	-	16
DAB	-	-	-	-	-	-	-	171	-	-	5	-	-	-	-	-	176
EBV	-	-	-	-	1	-	1	-	-	-	-	-	-	-	-	-	2
ER	-	-	-	-	-	-	-	-	-	1	12	1	-	-	-	4	18
HE	-	4	-	4	-	-	-	61	-	6	9	3	6	5	5	14	117
HER2	-	-	-	-	-	-	-	-	-	2	2	-	-	-	-	2	6
KI67	-	1	-	1	-	-	-	-	-	-	1	-	-	-	-	3	6
MAS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5	-	5
PAS	2	-	-	8	-	-	-	-	-	-	-	-	-	5	14	-	29
PR	-	4	-	1	-	-	-	-	-	1	11	1	4	-	-	1	23
sum	3	15	2	20	17	21	5	232	8	10	46	5	22	10	32	33	481

S.IV. DATASET ACKNOWLEDGEMENTS

The lesions, lung-lobes and mammary-gland images were provided by Prof. Carlos Ortiz de Solórzano and Dr. Arrate Munoz Barrutia, Center for Applied Medical Research (CIMA), University of Navarra, Pamplona Spain [S2, S3]. The mice kidney images were provided by Prof. Peter Boor and Dr. Barbara M. Klinkhammer, Institute of Pathology, University Hospital Aachen, RWTH Aachen University [S4]. The colorectal cancer images were provided by Dr. Rudolf Nenutil (Masaryk Memorial Cancer Institute Brno), and Dr. Eva Budinska and Dr. Vlad Popovici (Masaryk University Brno) and were collected under grant nr.16-31966A by Ministry of Health of the Czech Republic. Gastric mucosa and gastric adenocarcinoma tissue images were provided by Prof. Pavel G. Malkov, Dr. Natalya V. Danilova, Dr. Nina A. Oleynikova and Ilya A. Mikhailov, Department of Pathology, Lomonosov Moscow State University [S5]. The kidney and breast cancer whole slide images were provided by Dr. Gloria Bueno and Dr. Oscar Deniz from Grupo VISILAB, Universidad de Castilla-La Mancha (UCLM). The images were obtained and prepared thanks to the AIDPATH European project[3] coordinated by UCLM.

[3]http://aidpath.eu



(a)

(b)

Fig. S4. Quantitative comparison of methods performance — *(a):* median rTRE and *(b):* robustness R) — as a function of the staining combination.

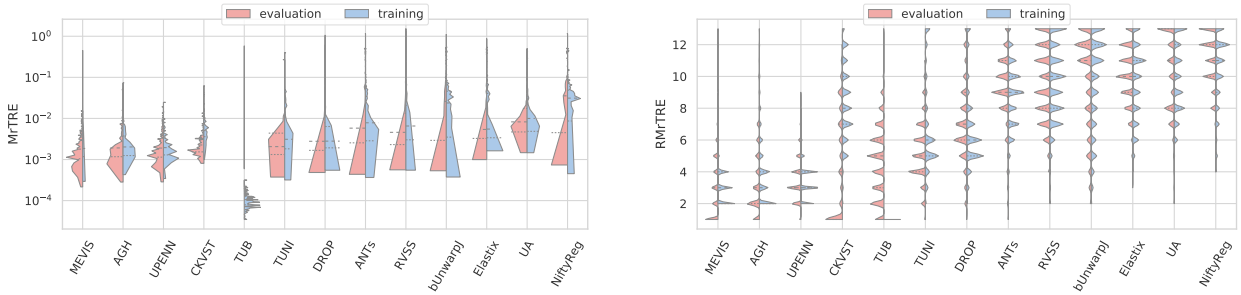


Fig. S5. Comparison of method performances, shown as histogram of the median rTRE and rank median rTRE, between training and testing (evaluation) datasets, to detect overfitting.

References

[S1] C. Schneider, W. Rasband, and K. W. Eliceiri, "NIH Image to ImageJ: 25 years of image analysis," *Nature Methods*, vol. 9, pp. 671–675, 2012.

[S2] J. Borovec, A. Munoz-Barrutia, and J. Kybic, "Benchmarking of Image Registration Methods for Differently Stained Histological Slides," in *Proc. Int. Conf. Image Process.*, Athens, 2018, pp. 3368–3372.

[S3] R. Fernandez-Gonzalez, A. Jones, E. Garcia-Rodriguez, P. Chen, A. Idica, S. Lockett, M. Barcellos-Hoff, and C. Ortiz de Solórzano, "System for combined three-dimensional morphological and molecular analysis of thick tissue specimens," *Microscopy Research & Techniques*, no. 59, pp. 522–530, 2002.

[S4] L. Gupta, B. M. Klinkhammer, P. Boor, D. Merhof, and M. Gadermayr, "Stain independent segmentation of whole slide images: A case study in renal histology," in *Int. Symp. Biomed. Imag.*, 2018, pp. 1360–1364.

[S5] I. Mikhailov, N. Danilova, and P. Malkov, "The immune microenvironment of various histological types of EBV-associated gastric cancer," presented at European Congress on Pathology, 2018.