IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.1917.DOI

Evaluating registrations of serial sections with distortions of the ground truths

OLEG LOBACHEV[1,2], TAKUYA FUNATOMI[3], (Senior Member, IEEE) ALEXANDER PFAFFENROTH[1], REINHOLD FÖRSTER[7], LARS KNUDSEN[1,4], CHRISTOPH WREDE[1,4,8], MICHAEL GUTHE[14], DAVID HABERTHÜR[5], RUSLAN HLUSHCHUK[5], THOMAS SALAETS[9], JAAN TOELEN[9], SIMONE GAFFLING[16], CHRISTIAN MÜHLFELD[1,4,8], and ROMAN GROTHAUSMANN[1,18]

[1]Hannover Medical School, Institute of Functional and Applied Anatomy, OE 4120, Carl-Neuberg-Straße 1, 30625 Hannover, Germany
[7]Hannover Medical School, Institute of Immunology, OE 5240, Carl-Neuberg-Straße 1, 30625 Hannover, Germany
[4]Biomedical Research in Endstage and Obstructive Lung Disease Hannover (BREATH), Member of the German Center for Lung Research (DZL), Hannover, Germany
[8]Hannover Medical School, Research Core Unit Electron Microscopy, OE 8840, Carl-Neuberg-Straße 1, 30625 Hannover, Germany
[2]Leibniz-Fachhochschule School of Business, Expo Plaza 11, 30539 Hannover, Germany
[3]Nara Institute of Science and Technology, 8916–5 Takayama-cho, Ikoma, Nara 630–0192, Japan
[14]University of Bayreuth, 95440 Bayreuth, Germany
[5]University of Bern, Institute of Anatomy, Baltzerstrasse 2, 3012 Bern, Switzerland
[9]KU Leuven, Herestraat 49, 3000 Leuven, Belgium
[16]Chimaera GmbH, Am Weichselgarten 7, 91058 Erlangen, Germany
[18]HAWK University of Applied Sciences and Arts, Faculty of Engineering and Health, Von-Ossietzky-Str. 99, 37085 Göttingen, Germany

Corresponding author: Oleg Lobachev (e-mail: oleg.lobachev@leibniz-fh.de).

This work was supported by DFG grant MU 3118/8-1. This work was partially supported by JST, PRESTO grant number JPMJPR2025, Japan. This research was supported by a C2 grant from KU Leuven (C24/18/101) and a research grant from the Research Foundation – Flanders (FWO G0C4419N). None of the funding bodies was involved in the design or execution of the study.

ABSTRACT Registration of histological serial sections is a challenging task. Serial sections exhibit distortions and damage from sectioning. Missing information on how the tissue looked before cutting makes a realistic validation of 2D registrations extremely difficult.

This work proposes methods for ground-truth-based evaluation of registrations. Firstly, we present a methodology to generate test data for registrations. We distort an innately registered image stack in the manner similar to the cutting distortion of serial sections. Test cases are generated from existing 3D data sets, thus the ground truth is known. Secondly, our test case generation premises evaluation of the registrations with known ground truths. Our methodology for such an evaluation technique distinguishes this work from other approaches. Both under- and over-registration become evident in our evaluations. We also survey existing validation efforts.

We present a full-series evaluation across six different registration methods applied to our distorted 3D data sets of animal lungs. Our distorted and ground truth data sets are made publicly available.

INDEX TERMS registration, ground truth, histological sections, evaluation, image processing

I. INTRODUCTION

MICROSCOPY has a long tradition, and microscopic imaging is still one of the most frequently used and powerful tools in biomedical research. From light microscopic (LM) techniques, including conventional fluorescent stainings, to transmission and scanning electron microscopic (EM) methods, the last two decades have witnessed substantial methodological progress in terms of resolution, speed, and automation. Tissue clearing and super resolution LM at the one end, and serial block-face as well as focused ion beam scanning EM at the other end, have paved the way for a three-dimensional visualization of biological specimens.

Still, the use of serial sections remains an essential and cost-efficient tool to gain 3D insight into specimens for several reasons: Despite the progress in LM techniques the penetration depth of staining solutions, in particular fluorescent antibody staining, is limited, thus limiting the size of the sample that can be visualized. Genetically modified organisms, such as mice, expressing fluorescent proteins

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access

under cell-specific promoters, are available. This method, though, cannot be applied to human samples for obvious reasons. Thus, the use of serial sections in LM is often the method of choice for obtaining 3D information both from conventional and fluorescent microscopic imaging, in particular when human samples are investigated.

However, microscopic sections are inherently two-dimensional (2D) and their 3D information has to be regained from 2D images. The manual or automated cutting of thin sections for microscopy, however, induces—even in perfectly trained and experienced hands—varying distortions and deformations, such as stretching or compression. Positioning the sections on the glass slides further contributes to spatial distortion. This problem is especially evident in large sections. Further processing such as antigen retrieval and staining may further damage the section. Even digitization of the sections can be faulty and create partially corrupt representations.

The alignment or registration is an important method in medical image processing. An absence of the ground truth is a major problem during the development of new and fine-tuning of existing registration methods. While some simple synthetic data or phantoms can be generated, those would not adequately represent the problem. Firstly, some registration methods, for example, those based on feature detection, thrive from complexity of the input data. Such "sparse" methods might perform very well especially in the lung tissue, where the fraction of empty space is very high. Secondly, the distortions in phantom data might not truly represent the distortions in serial sections. Thirdly, in most real cases, no ground truths from other modalities exist in the typical acquisition resolution of the serial sections. Most "real" 3D methods either do not reach the resolution of conventional LM (e. g., micro-CT) or have typically much smaller spatial dimensions of the probe (e. g., nano-CT, EM). Aforementioned LS microscopy and tissue clearing are possible palliatives in model animals, but all those methods are still too complex, too expensive, or require a radically different biological processing pipeline that makes it impossible to apply both modalities to the same specimen. It is also much harder to apply aforementioned advanced methods in humans.

A. CONTRIBUTIONS

In this paper we present a methodology to apply typical sectioning distortions to real data sets from other modalities. We digitally "mock up" the distortions from sectioning on real biological data. Arbitrary general-purpose images can be distorted (Fig. 1, Fig. 2) and registration methods can be applied to distorted images. The results of the registrations can be immediately compared with ground truth data. Our benchmark is open to further registrations, new quality measures, and new images. As we present the method and not only the data, further data sets, even from additional modalities, can be produced by others. We focus our current presentation on animal lung images. However, our method is generic; it should be applicable to virtually any kind of innately 3D data of any organ from any species. Our goal is to enable the evaluation of registration methods for serial sections with a ground truth from real biological data. To fulfill it, we mimic sectioning distortions in an artificial, but statistically meaningful and reproducible manner. We then proceed to evaluate some existing registrations with our method. Among other approaches we present a full-series evaluation.

In this paper, we consider possible distortions during sectioning and apply those to 2D series from the innately 3D data. Our data sets originate from further modalities in bioimaging. The data sets aim to come close to LM sections of the lung in their scale—on both sides. We use both CT and LS as coarser scale and EM as a finer scale. As original, non-distorted data fit perfectly, those serve as a ground truth.

The contributions of this paper are threefold. Firstly, we provide an overview over the field with the emphasis on validations of registration. In most such validations, the problem of an absent ground truth motivates the search for further methods. Our approach is novel, we work with a present ground truth.

Secondly, we suggest a technique to generate a benchmark input from existing inherently 3D data. This way, we are, thirdly, able to compare registration methods on a common foundation by comparing the registered data with the ground truth. We perform an extensive image-based statistical evaluation of the full series.

The source code for this paper is available under https://github.com/olegl/distort, the distorted and ground truth data sets can be found under https://zenodo.org/record/4282448.

B. PAPER STRUCTURE

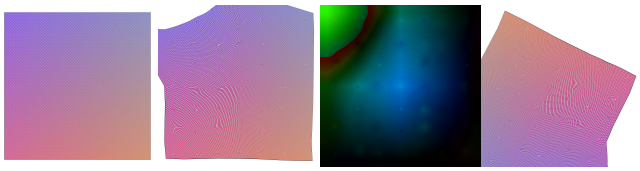
The remaining part of the paper is organized as follows: In Section II we survey existing registration methods and discuss the approaches towards validation of registration. Section III elaborates on our approach for generating distortions. The same section also presents the registration methods and the data sets we used in our benchmark. In Section IV we evaluate the results of the registration benchmark. We compare the registered images with the ground truth in this section. We present there both image-based evaluations and statistical gauges of the results. Section V discusses possible limitations and further developments of our method. Section VI concludes the manuscript.

II. RELATED WORK

A. REGISTRATION IN GENERAL

There is a lot of research on registration, esp. in the context of medical imaging. Brown [1], Zitová and Flusser [2] provide early surveys; Pluim et al. [3] and Oliveira and Tavares [4] are more focused on the medical topics. Viergever et al. [5] and Pichat et al. [6] are recent overviews of the field. Zitova [7] gives a mathematical overview of the methods. Although some manual alignment [e. g., 8] has been performed in the past, we ultimately focus on computational methods.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access



(a) Input (b) Local distortions (c) Distortion map (d) Rigid transform

Figure 1: Showcasing our method on a synthetic image of a color gradient. The distortion magnitude is increased tenfold for demonstration. The images are slightly cropped for presentation.

Sotiras et al. [9] survey non-rigid registration methods. In the context of the registration of serial sections, non-rigid registration is definitely required, as distortions during sectioning are non-linear. Even if we currently do not represent tearings or foldings of the tissue, cutting-induced local distortions are still present even in the most perfectly prepared sections. (Instead of tearings and foldings we can easily encompass missing parts of the images. Salvaging damaged or missing sections is a separate problem in our eyes.) Non-rigid registration methods include Rueckert et al. [10]; Schnabel et al. [11]; Chui et al. [12]; Hömke [13]; Zhang et al. [14]. Saalfeld et al. [15] focused on as-rigid-as-possible registration for EM.

Punithakumar et al. [16] is a recent example of a GPU-accelerated registration. Crum et al. [17] provide an overview of medical image registration, they highlight both the importance of validation and its difficulty. One of the popular software packages for registration is Elastix [18, 19] and further developments around it [20, 21]. Another popular package is ANTs [22, 23]. One of the somewhat frequent ideas is to work with images on multiple levels [see, e.g., 24, 25].

A kind of "sparse" methods involves feature detection and description. Ma et al. [26] presents a recent survey of the field. The actual detectors and descriptors include SIFT [27, 28], SURF [29, 30], AKAZE [31, 32]. A basic "sparse" registration identifies distinctive regions of both input images and then computes a correspondence between them based on the correspondence of the regions alone. In such a rigid registration RANSAC [33] is used. "Sparse" methods have been used to register medical images [e.g., 34, 35, 36, 37, 38]. Arganda-Carreras et al. [39] is the origin of ImageJ's "Register virtual stack slices" implementation. (ImageJ [40] and Fiji [41] have served as a basis for many registration and analysis approaches.) They focus strongly on various rigid approaches, although an elastic extension exists. The ImageJ plugin "TrakEM2" [42] also utilizes feature detection, but it not only performs registration, but also includes tools for 3D modeling, editing, and annotation. Ma et al. [43] and Zhang et al. [44], for example, improve the correspondence of features ("matching"). Cieslewski et al. [45] is an example of an alternative to feature descriptors. Registration of whole sections [e.g., 46] motivated the usage of feature detection in Ulrich et al. [47].

Optical flow [48, 49] is a yet another method to find "moving parts" in images [50, 51, 52]. Applications of optical flow in medical images include Dougherty et al. [53], Carata et al. [54], Lobachev et al. [55]. Feature descriptors have been used on dense, optical-flow-like data [56, 57].

A diffusion model based on thermodynamics [58] is widely used [e.g., 59, 60, 61]. Further registration approaches include graph-cut-based methods [62], smoothness assumption [63], higher-order derivatives [64], chamfer matching [65], particle swarm optimization [66], Gauss-Seidel optimization [67], Markov random fields [68, 69], over-segmentation regularization [70], elastic triangulation of a spring model [71], blending rigid transforms [72], empirical mode decomposition [73], and remote sensing [74].

Some methods register a complete stack of images at once, this approach was used, e.g., by Nikou et al. [75], Saalfeld et al. [15], Lobachev et al. [25].

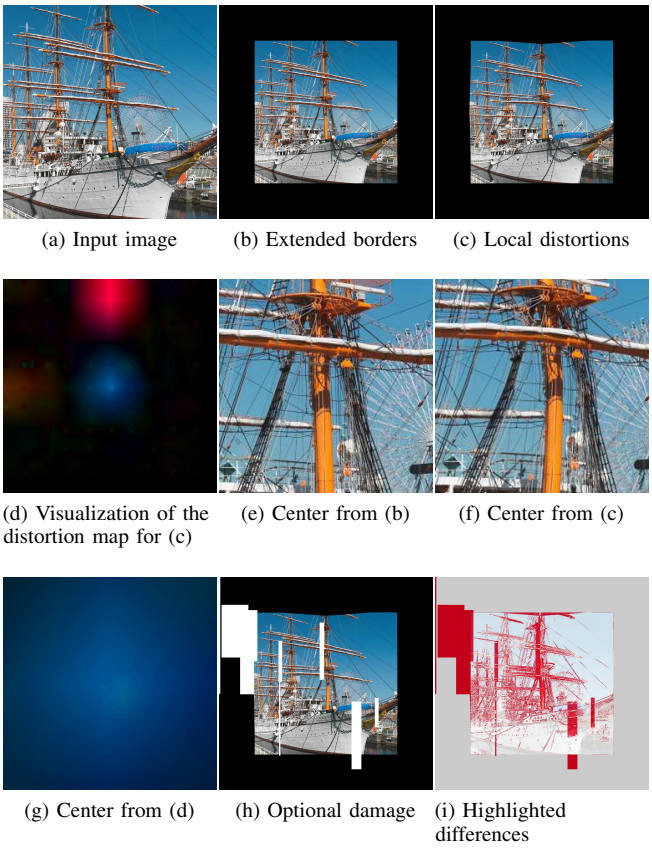
B. VALIDATION, IMAGE GENERATION, AND BENCHMARKS

In a sense, this paper is dual to Cifor et al. [63]. They thought explicitly about possible distortions during sectioning and displaced the real section images in a way that would counter this distortion—a similar idea is behind most registration methods. Our distorted ground-truth data are the input of a registration benchmark. We validate multiple existing registration methods by comparing (previously distorted and) registered images to the distorted, but not registered, and to the (not distorted, perfectly aligned) original data. Sections III-F and III-I detail on our evaluation methodology.

Pluim et al. [76] provide an overview on validations of medical registrations. Van Sint Jan et al. [77] showcase a very special kind of a registration that was validated with kinematics. The method by Delaby et al. [78] is a more typical case, where the 3D reconstructions were validated by a different modality. Schnabel et al. [11] discuss physically plausible distortions in breast MR data. Shojaii et al. [79] use block-face images and fluidical markers for validation of the registration. Kybic [80] undertakes special efforts for the evaluation of registration accuracy in absence of ground truth. In contrast to all those approaches, we suggest to use multiple image-based quality measures for the evaluation of the registrations. The essence of the present manuscript is the availability of ground truth, so no further modalities, manual interventions or external markers are required.

Generation of further images is by far not new in medical image processing, to name a few, Xue et al. [81] generate synthetic images for better T1 MRI; Duchateau et al. [82] generate pathological cardiac images; and Grova et al. [83] use computer-generated SPECT data to validate their MRI-SPECT registration. Image generation is connected to validation, because as long as we are able to generate images with given properties, these can be used to validate other image-based methods. Hamarneh et al. [84] use all kinds of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access



(a) Input image (b) Extended borders (c) Local distortions

(d) Visualization of the distortion map for (c) (e) Center from (b) (f) Center from (c)

(g) Center from (d) (h) Optional damage (i) Highlighted differences

Figure 2: Details of local distortion and damage generation. Our method works also on general-purpose images. The test image is from the Japanese ITE data set of UHD images, we took $1k \times 1k$ pixels crop from the center of the "Ship" image (U10, 2K version). A lot of straight lines allows for good identification of distortions. Global rigid transformation is omitted for its simplicity. The distortion map is in line with our usual settings. The visualization in (i) compares PSNR (peak signal-to-noise ratio) between (b) and (h); it is thresholded at 20%.

statistical and physically-based distortions, noise, artifacts but their method is focused on MRI and CT data. Even though distortions of 2D images are also possible with their framework, the method is focused on other modalities.

Vlachopoulos et al. [85] generate distorted CT images using landmarks and thin-plate splines. Their warping method was specially chosen to imitate aspiration. The images were used to evaluate registration methods in normal lungs and organs with interstitial lung disease. The idea of the evaluation is similar to ours, however, we focus on histological serial sections and provide an elaborate methodology to generate such distorted images. Both the nature of the deformation and the method of its implementation are different in this work. We are concerned with sectioning and processing artifacts in removed tissue. We do not use thin-plate splines and landmarks to compute distortions.

Zhang et al. [86] both generate synthetic images and use real data to compare their global registration method to others with promising results. Unlike the present work here, they distort the images using quadric 2D polynomials and focus on either homography or distortions commonly found in digital cameras. These kinds of distortions differ a lot from our approach, since they are induced by the optical pathway in the camera and not by physical sectioning of the specimen.

Related to above methods are registration benchmarks and challenges. We would like to specially mention the EMPIRE10 challenge [87]. Borovec et al. [88] benchmark registrations of differently stained serial sections [a rather distinctive kind of registration: 46, 89, 90, 91, 92]. Basically, registration of differently stained serial sections is about transferring the distortions from one kind of staining to another one. Also co-registrations across modalities are a tangent topic to our work, e.g., CT to MRI [93, 94, 95], serial sections to MRI [96, 97, 98], or MRI to ultrasound [99]. We *distort* images from other modalities (similarly to the distortions in serial sections) in order to obtain challenges for testing registration methods with a known ground truth. In this work, during each of the registration attempts we remain within a single selected modality.

A recent ANHIR challenge [100] uses multiple data sets, where the ground truth is obtained by external markers on the histological series. Their landmarks were placed by multiple human annotators. We use automatic image-based metrics in this work and do not use external markers, however our approach is open to further measures. (It would be easiest to integrate further image-based markers, though.) We eschew external markers, as we *have* a ground truth, which contrasts our work from histology-based challenges, where no direct ground truth is available. Further, the ability to fully automatically compute the "score" of a registration method from ground truths and registration output allows our approach to be used in automatic tests of registrations, such as continuous integration (Section V-G).

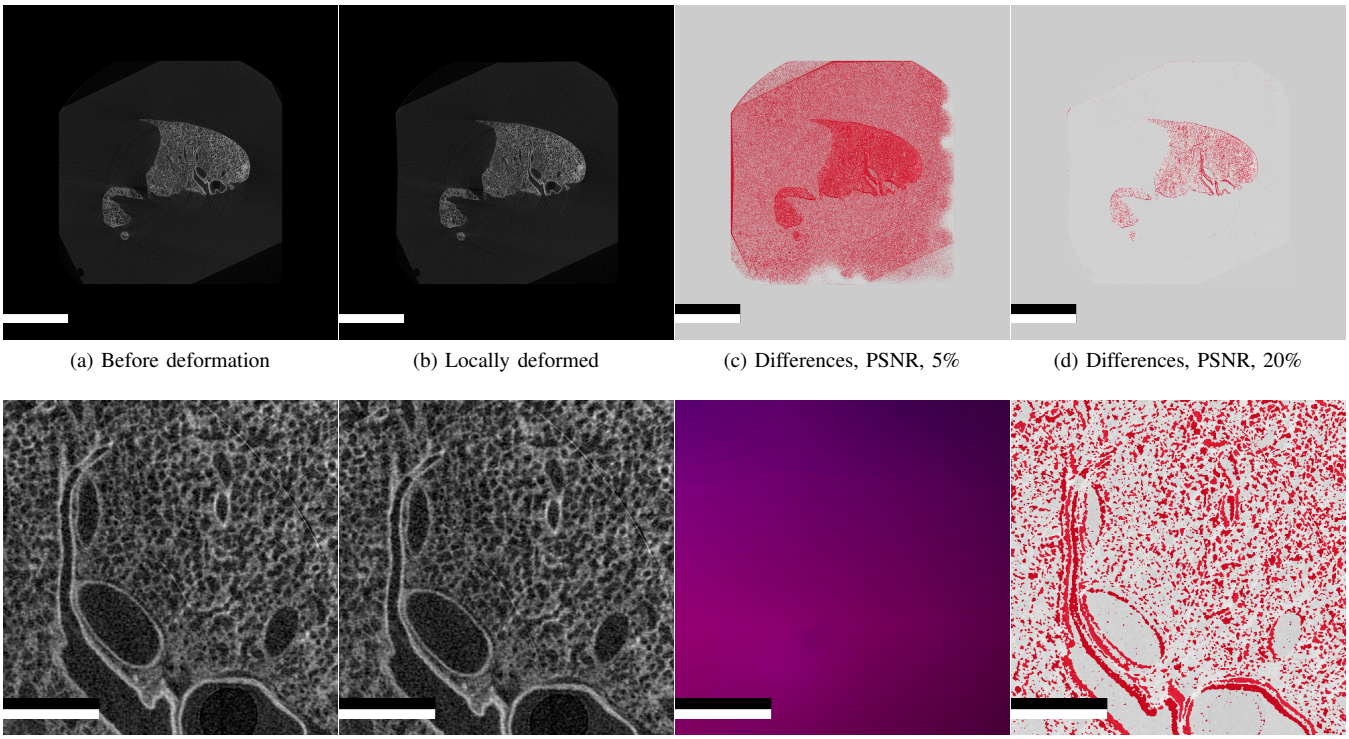
The NIREP project [101] evaluates specifically non-rigid registrations, with absent ground truth. This paper is about distorting a known ground truth for the evaluation of rigid and, mostly, non-rigid registrations, we circumvent the main problem of non-available ground truth.

To name further related papers, Pontré et al. [102] presents a cardiac perfusion MRI registration challenge focused on motion correction; Brock [103] compare accuracy of different deformable registration methods on MRI and CT data; West et al. [104] and Hellier et al. [105] are examples of the evaluations of inter-subject registrations. Klein et al. [106] and Ou et al. [107] evaluate registration methods for inter-patient brain MRI.

C. QUALITY MEASURES FOR REGISTRATION

Image registration can be seen as an optimization problem. Similarity measures are key to both good registrations and their evaluation, as a similarity measure is basically the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access



(a) Before deformation (b) Locally deformed (c) Differences, PSNR, 5% (d) Differences, PSNR, 20%

(e) Crop from (a) (f) Crop from (b) (g) Optical flow, crop (h) Crop from (d)

Figure 3: The effect of the local distortions on real data: an unfiltered micro-CT image of a rabbit lung. Fig. (a) features the extended (for later rigid transformation), but not distorted image. Fig. (b) is the result of non-linear distortions. The distortions may be hardly noticeable by a human, but they are enough to confuse computational methods. Such distortions are clearly visible in the next panels. Figs. (c) and (d) show the PSNR visualization between (a) and (b) with 5% and 20% threshold for the red color. Figs. (e)–(h) show crops, Fig. (g) shows an optical flow visualization computed from full images, but then cropped in the same manner as others. The optical flow shows the "movements" between two input images. Scale bars in (a)–(d) are 5 mm, scale bars in (e)–(h) are 1 mm.

objective of optimization. Mutual information is often used as such a measure [108, 3, 109, 110]. A related problem is the selection of the reference in a series of histological sections [111]. Nanayakkara et al. [112] introduce a metric for registration errors. Luo et al. [113] discuss the relation between registration errors and uncertainty.

In this work, we choose image-based measures as an arbiter in quality of the registration. Such approach allows not only for automatic generation of the inputs and for automatic execution of the registration, but also for an automatic evaluation of the results. In our evaluation we use the standard measures by Jaccard [114] and Wang et al. [115], as well as the dense optical flow [50]. We also use the Dice [116] measure as a visualization of the Jaccard measure—as the formulation of Dice can be converted to a formulation of Jaccard. (Details on thresholding methods are in the supplementary material.) We use a visualization based on PSNR (peak signal-to-noise ratio) as well.

Crum et al. [117] discuss generalizations of overlap-based measures, but in this work we opted for the well-known measures. Rohlfing [118] criticizes the usual image-based measures, but our method can use any measures for evaluation. Our method is not imbued with the measures we use, hence any extensions or further measures are possible. The core idea is to use (now-distorted) inherently 3D images to benchmark 2D registrations.

D. MACHINE LEARNING

With modern deep learning methods, the measure can be implicitly learned, as Krebs et al. [119] mention. Maier et al. [120] provide an introduction to deep learning in medical imaging. Li et al. [121] use style transfer to generate images from different vendors, such generated images enable better machine learning. Fu et al. [122] and Haskins et al. [123] provide an overview of machine learning in registrations. Examples of advances of deep learning in registrations include Dalca et al. [124] and Sarlin et al. [125].

We stress that our method generates distorted images without any use of machine learning. Thus, our method can be used to generate additional data sets or to augment machine learning input.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access

E. LUNG IN 3D

The methods, options, and research outcomes in 3D reconstructions of the lung using any modalities from corrosion cast and up to 3D EM methods [e.g., 126] are reviewed by Mühlfeld et al. [127]; practical applications include [128, 129]. Although, EM studies of the lung are popular and important [e.g., 130, 131, 132, 133], serial sections for LM have their place in the investigation repertoire [e.g., 134, 135, 136]. Putting stereology aside, a proper registration is paramount in any investigation of serial sections as a 3D data set.

III. METHODS

The typical sectioning-induced distortion in paraffin embedding was found by Schormann et al. [137] to be Rayleigh distributed. This basically means a normal distribution in each of the image axis. We showcase our method on a synthetic image (Fig. 1), on a standard test image (Fig. 2), and on real data, a 2D image from a micro-CT scan of an animal lung (Fig. 3). Notice that Fig. 1 overemphasizes the effect of our method: we use there 10 times larger distortions than usual.

Now, Figs. 1a, 2a show the original specimen. In real applications we extend the border (Figs. 2b, 3a). Figs. 1b, 2c, 3b demonstrate the distorted images, Figs. 1c, 2d show the color-coded distortion map. The detailed crops in Figs. 2, (e)–(g), and Figs. 3, (e)–(h), demonstrate the local movements induced by the distortion. The goal of the registration is to precisely eliminate such movements.

To simulate lost or damaged parts of the sections [which we recently learned how to repair: 91], additional arbitrarily placed "damage" can be added to the image (Fig. 2h). Finally, a global rigid transformation is applied (Fig. 1d). The rationale behind this step is that only in very rare cases the sections can be placed on the glass slide while maintaining the exact orientation. We randomly apply a rotation and translation to the images to simulate the uneven positioning. Such images serve then as inputs for the benchmark of registrations.

Fig. 4 visualizes the core approach and the complete pipeline of this paper. We present a distortion generator that is in a sense dual to a registration. We derive an evaluation of a registration method from original 3D stack and registration results.

A. LOCAL DISTORTIONS

Generation of local, non-rigid distortions is of high importance for our method. At the heart of the local distortion lies the generation of normally distributed values. Two independent normally distributed random values form the x and y coordinates of a displacement, making the displacements Rayleigh-distributed [137]. The coordinates of locations, *where* the distortions are placed, lie on a rectilinear grid.

We generate multiple distortion "levels" using a classical multiscale approach. The distortions are stored as coordinates of "new" points in a matrix holding both x and y coordinates as an element—this is a typical `remap` matrix of OpenCV. The distortions are blurred with a Gaussian kernel in each multiscale level to make the displacements smoother. This way we avoid undesirable and unrealistic foldings, as real sections folds look differently.

In the implementation, we used Mersenne Twister pseudorandom generator [138, a standard one in Python]. The distortion maps are applied to the input images with OpenCV `remap` function using bicubic interpolation. The distortion maps are saved for further analysis. We can generate those maps in a fully deterministic manner, if desired. This determinism contributes to reproducibility.

In our application, the final distortion map is visualized (Fig. 1c, 2d) using HSV colorspace. The Cartesian coordinates of the displacements are mapped to a polar angle (associated with hue); vector magnitude basically codes the intensity.

Our distortion maps are a simple, reproducible, and well-defined way to add sectioning-inspired distortions to arbitrarily registered data. We aimed to define a stable and reproducible way to model such distortions using the statistical properties of the real-world distortions.

The reproducibility is given through multiple efforts. We have the initial, "ideal" state, the ground truth. We save the exact rigid transform, the non-rigid distortion field, and the distorted result. Through the use of pre-defined, deterministic states of the pseudorandom generator, we can basically save all the transformations in form of the seed value (plus original images, of course). Above issues would be useful to ensure reproducibility, e.g., for automatic regression testing of registration methods.

B. ADDING RIGID TRANSFORM

The locally distorted images are further processed. In our application we add an image-wide rigid transform: a random rotation and a translation (Fig. 1d). The rotation angle is uniformly distributed between -180 and $+180$ degrees. The translations are also uniformly distributed, but they are chosen in a range $[-d/4, d/4]$, where d is maximal image dimension, in order to not truncate too much of the image content. The rigid transform is recorded, as it is the ground truth for the first, rigid step of the registration.

C. THE USE OF THE DISTORTED IMAGES

The distorted images serve as a starting point for the evaluation of registration methods. The registration should, basically, "undo" the distortions and transforms we applied to the original images. For the benchmark and evaluation purposes we suggest using innately registered data, i.e., original 3D images. The benefit of doing so is the available ground truth, the original undistorted 3D stack. Summarizing, we circumvent the problem of missing the real ground truth when comparing registrations of serial sections.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access

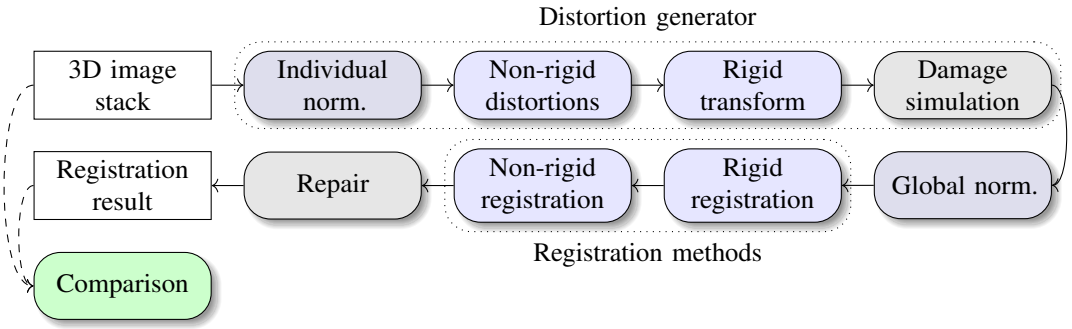


Figure 4: Illustrating the steps in this paper. Individual normalization of benchmark inputs and the application of global normalization are not mandatory. Currently, we do not use damage generation and damage repair in our benchmark images, but we could also test the repair methods in this manner. The evaluation ensues from comparison of the image series. The dotted boxes show two larger conceptual components, the distortion generator and the registration.

D. DAMAGING THE IMAGES (OPTIONAL)

An optional extension is to mimic the damage to which the sections are subjected to during the processing (Fig. 2h). Using normally-distributed pseudorandom values for dimensions and placement we iteratively generate an ImageMagick [139] script that deletes selected parts of an image. This is a quite crude approximation to the variety of possible kinds of damage to a section [91]—from an unsharp region (due to focus error) to a teared section. However, we would like to model missing parts of a section in an understandable and straightforward way. Our test subject, a registration method, would not necessarily care about the reason why some image regions are not matchable to other images. It is not our current goal to model the damaged sections realistically.

E. REGISTRATION METHODS

In order to demonstrate our methodology of the evaluation of registration, we apply the following registration methods to our distorted data sets:

1. "Rigid-SURF": Feature-based rigid-only registration based on weighted RANSAC [33, 25] and SURF feature detector [29];
2. "Rigid-SIFT": same as above, but with SIFT feature detector [27, 28];
3. "Deform-SURF": Feature-based deformable registration [25], first stage based on "Rigid-SURF", followed by multiple non-rigid stages using B-splines;
4. "Elastix": a generic Elastix [18, 19] configuration, we used rigidly registered result from "Rigid-SURF" as input. The parameter file is made available in the supplementary material. The non-rigid stage is *not* a feature-based method;
5. "GS": Registration method based on Gauss-Seidel optimization [67], we used rigidly registered result from "Rigid-SURF" as input. However, the actual non-rigid deformation is based on different principles, among others, on the gray-level co-occurrence matrices;
6. "Blending": Registration method based on blending rigid transforms in image regions [72];

The rigid methods work pair-wise on the images. We operated Elastix pair-wise on the images, hence the possible accumulation of "drift" with the progress of the series. "Deform-SURF" optimizes the whole stack at once in the non-rigid phase, "GS" does the same. The "Blending" method works backward and forward from a reference frame. In this case, the inputs were used "back and forth", for a series of $1, \ldots, n$ images, the input was $n, \ldots, 1, 1, \ldots, n$, essentially doubling the length of the series. The border interpolation was less of our concern, the images were padded before processing.

Why the rigid transformations? The rigid-only methods we used clearly cannot undo the non-rigid distortions. But the non-rigid distortions also make harder the search for correspondences for the rigid transformations between image features. Further, a rigid-only registration is also not necessarily perfect in what it does, in other words, a rigidly transformed (Section III-B) and then rigid-only registered series is different from the series before such transformations.

F. RESULT EVALUATION WITH QUALITY MEASURES

We propose the use of established image quality measures: structural similarity [SSIM, 115], Jaccard measure [114], often visualized here with slight implementation differences as a Dice measure [116], a visualization of dense optical flow [50]. In the latter, color stands for a direction and intensity for the magnitude of the movement. We also use PSNR visualization, as implemented in ImageMagick. There, red highlights a non-correspondence. Details of Jaccard and Dice implementations are in the supplementary material. The latter material also details on the manner in which we crop the images for evaluation in order to eliminate border effects. In our visualizations, black is "neither", magenta and green means "in one image, but not in another", white is in both.

As for image pairs, used for the evaluation, we always use two *consecutive* images from each of the series. To illustrate the exact procedure, let image pairs "ground truth 1" and "ground truth 2", as well as "registered 1" and "registered 2", be our inputs. We compare with above image measures

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access

the both "ground truth" images with *each other,* as well as both "registered" images with each other. This means, we compute, e. g., SSIM of "ground truth 1" and "ground truth 2" as well as SSIM of "registered 1" and "registered 2". While it might seem compelling to compare, e. g., "ground truth 1" with "registered 1", we would see there some larger movements that are rather irrelevant for our goals.

Namely, each, esp. non-rigid, registration has "drifts" in its results. There is a larger dissimilarity between an i^{th} registered image and i^{th} ground truth. Such drifts are, however, not quite the object of our interest. We would like to compare, how much consecutive images in the series fit *each other*. With ground truth we now know, how the real consecutive images should fit each other. When we compare each image to the ground truth, we measure the "global" accumulated registration error, but not the "local" registration error in the series itself. We deem a "local" error, such as an discontinuity in the structures, far more important as a "global" error, where a rather correct structure is merely shifted few pixels in the whole series. Of course, a throughout investigation of also the "global" errors is of interest. We would like to perform them *not* on the images themselves, but on their distortion descriptions. This is, however, an issue of the future (Section VI-A).

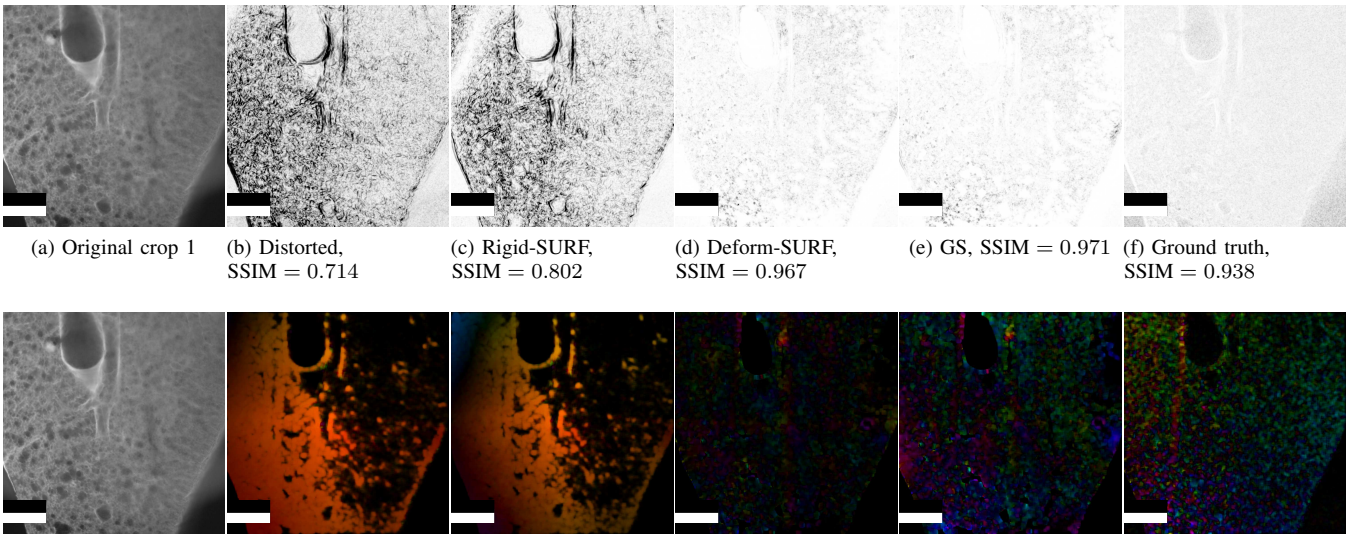
To give an example, Figure 5 shows some of the quality measures, applied to different registration of the same data. Top row (b)–(f) shows the SSIM, bottom row (h)–(l) shows the optical flow visualization. We registered a full LS series with all methods. We used the distorted images that were also globally transformed as the input. Then, a $1k \times 1k$ pixels crop from the full registration was used for evaluation to eschew the border effects. Panels (a) and (g) from Fig. 5 show regions cropped from the original images. Panels 5(b) and (h) show locally distorted images, before the global rigid transform. For all registrations, the images were locally distorted and globally transformed. Images shown in panels (c) and (i) were then registered with "Rigid-SIFT", it is the rigid transformation only. Panels (d) and (j) show the evaluation of images, registered with "Deform-SURF", both rigidly and non-rigidly. Images shown in panels (e) and (k) were rigidly registered with "Rigid-SURF", then non-rigidly registered with "GS". Panels (f) and (l) show the evaluation of original, non-distorted data, i. e., the ground truth.

G. SPECIMENS

We applied our method to micro-CT (abbreviated as "CT" in data set labels), light sheet (abbreviated LS), and EM images of animal lungs (Fig. 6). Our CT data is isotropic, our LS data set is anisotropic. The EM data set was captured anisotropically, however then resampled to be isotropic. The data was processed in the manner standard for each modality, basically, for the processing in this paper, we perceived the data as already processed and ready-to-use 3D images. The images were extended to the size specified below in order to not lose data during the global movement phase. Specifically, we used:

- A rabbit lung acquired with micro-CT (Fig. 6a). The specimen was a New Zealand White rabbit that was artificially delivered 3 days early by cesarean section and that spent 7 days in hyperoxia (95 %), the lung was perfusion fixed. The sample comes from a project studying the bronchopulmonary dysplasia in a hyperoxia preterm rabbit model [141, 142], part of a larger study of bronchopulmonary dysplasia models [143]. The experiments have been approved by the ethics committee for animal experimentation of KU Leuven, project number P081/2017.
  The sample was imaged on a Bruker SkyScan 1272 high-resolution microtomography machine (Control software version 1.1.19, Bruker microCT, Kontich, Belgium). The X-ray source was set to a tube voltage of 80 kV and a tube current of 125.0 µA, the X-ray spectrum was filtered by 1 mm of Aluminum prior to incidence onto the sample. We recorded a set of 2 stacked scans overlapping the sample height, each stack was recorded with 488 projections of 3104 × 1091 pixels (2 projections stitched laterally) at every 0.4° over a 180° sample rotation. Every single projection was exposed for 2247 ms, 5 projections were averaged to greatly reduce image noise. This resulted in a scan time of approximately 8 hours. The projection images were then subsequently reconstructed into a 3D stack of images with NRecon (Version 1.7.4.2, Bruker microCT, Kontich, Belgium) using a ring artifact correction of 7. The whole process resulted in a data set of 1135 images with an isometric voxel size of 7.0 µm (see also Fig. 3). The images were pre-processed with a 2D anisotropic diffusion denoising filter based on lattice basis reduction [140].
  We extracted 600 images from the middle of the filtered data set for our benchmark and padded them (Sec. III-H), yielding images at 3954 × 3954 pixels;
- An EM serial block-face (SBF-SEM) data set of adult mouse lung (Fig. 6b). The specimen was a 4 weeks old C57BL/6 mouse, the lung was perfusion-fixed [144]. The experiments were approved by Regierungspräsidium Karlsruhe. Overall, 5246 sections with 80 nm thickness were cut in a Zeiss Merlin VP Compact SEM (Carl Zeiss Microscopy GmbH, Jena, Germany), using a Gatan 3View2XP system (Gatan Inc., Pleasanton, CA, USA). The block-face was captured with the view port of 525 × 525 µm, yielding $15k \times 15k$ pixels with 0.5 µs dwell time, 3.0 kV acceleration voltage and variable pressure mode at 30 Pa.
  The benchmark uses a crop from the full data set with 1000 images at $1.5k \times 1.5k$ pixels. The final resolution is 0.15 µm/voxel. The data set was denoised with gradient anisotropic diffusion using ITK before usage. We did not apply any registration in post-processing, but we individually normalized the images—as detailed below;
- A lung for the light sheet (LS) data set was obtained from a male 24 week-old Fisher 344 rat with a body

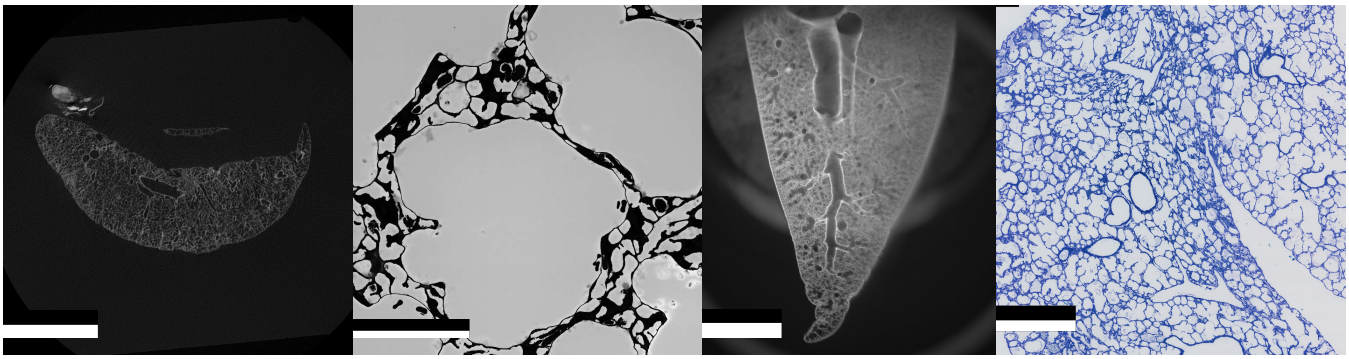
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access



(a) Original crop 1 (b) Distorted, SSIM = 0.714 (c) Rigid-SURF, SSIM = 0.802 (d) Deform-SURF, SSIM = 0.967 (e) GS, SSIM = 0.971 (f) Ground truth, SSIM = 0.938

(g) Original crop 2 (h) Distorted, flow (i) Rigid-SIFT, flow (j) Deform-SURF, flow (k) GS, flow (l) Ground truth, flow

Figure 5: Evaluating image measures on $1k \times 1k$ pixels crop from full registration of LS. The "flow" images are the visualization of optical flow. The color codes the direction of the movement. Images (f), (l) show the ground truth values: here two concecutive images of the ground truth were used to produce the quality measures.
All scale bars are 1 mm.



(a) Micro-CT, rabbit lung (b) Electron microscopy, mouse lung (c) Light sheet microscopy, rat lung (d) A fragment of a reference serial section, rabbit lung, toluidine blue

Figure 6: The data sets used in our benchmark. The micro-CT (a) is filtered with anisotropic diffusion [140]; the EM (b) and the light sheet data (c) are normalized individually; two consecutive non-registered serial sections (d) are provided as a reference. Further data sets can be created from 3D data using the methodology we present here.
Scale bars are: (a): 5 mm, (b): 50 µm, (c): 3 mm, (d): 1 mm.

weight of 320 g, which was part of a ventilation study approved by the LAVES in Oldenburg, the number of animal experiment proposal is 17/2608. The lung was fixed in an inflated state with an airway pressure corresponding to 20 cm of H_2O and perfusion fixed, compare Krischer et al. [145]. By means of a "tissue slicer" the lung was cut in slices of 2 mm thickness.
The image data was acquired with the UltraMicroscope II (LaVision BioTec GmbH, Bielefeld, Germany). The lung slices were pinned up to a mandrel in the ethyl cinnamate-filled detection chamber and illuminated unidirectionally with 6 light sheets. An sCMOS camera detected the fluorescence light with a wavelength of 490 nm, which matches the tissues autofluorescence, perpendicular to the illumination plane. Due to the large dimensions of the rat lung and the intention to depict the complete lung the only zoom factor to choose was 0.63, corresponding a 1.26-fold magnification. The series was acquired as 336 images with $5.16 \times 5.16 \times 15$ µm/voxel (Fig. 6c).
For the benchmark we use 300 images at $3{,}9k \times 3{,}9k$ pixels that were individually normalized, see below;

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access

- As a reference, we also provide two serial sections of the same rabbit lung as in micro-CT, stained with toluidine blue. Fig. 6d shows one of those sections. The images were acquired in transmitted LM with a Zeiss AxioScan.Z1 scanning microscope (Carl Zeiss Microscopy GmbH, Jena, Germany) at 0.22 µm/pixel (20× lens). The section thickness was 2 µm.

H. DATA PREPARATION, NORMALIZATION, AND AVAILABILITY

All benchmark images were extended to a larger square to reduce the loss of information during rotation and translation. Their bit depth was reduced to 8 bit, LS and EM images were normalized *individually* using ImageMagick and GNU parallel [146]. The normalization of individual images is introduced to mimic a slightly varying image intensity [111, 147, 148] due to varying thickness of slices [149] or varying penetration of the fixation and the staining solutions [e. g., 150, 151, 152].

We provide as supplementary data the original, undistorted images, the locally distorted images, the locally distorted and rigidly transformed images, the local distortions, and the values for the rigid transformations.

I. EVALUATION METHODOLOGY

Our concept of the evaluation focuses on comparing consecutive sections from the registered series to the same consecutive sections from the ground truth. We decided against comparing the images from the registered series directly to the ground truth images: Accumulated errors from the rigid transformations and non-rigid distortions impact such comparisons. We would be more interested in how the now-registered series fits to *itself* in comparison to how it *should have fit.* In a direct comparison we would have found too many rather irrelevant mismatches. To give a simple example, many methods might over-fit the registrations of their inputs for the better numerical quality values ("over-registration", "banana problem"). The "banana problem" (Fig. 7) denotes the undesired tendency of non-rigid registration methods to straighten curved structures by optimizing the individual images' similarity to its respective neighbors. With our ground truth series we would be able to find such cases.

The measures were computed on the 500 × 500 pixels crop from the middle of the images, throughout the series. Lower values in the beginning and in the end of the series can be explained with less tissue in the region. However, we advocate the use of the center crop as a simple to define and "fair" way to define a region. As detailed in the supplementary material, it is impractical to use the full image for the evaluation. Also, some methods used additional padding, making the uniform and comparable use of general cropping offsets harder. The middle of the image should arguably have meaningful tissue contents in most cases. The "drifting away" tissue, i.e., the case when different registrations accumulate the errors so differently, that we obtain fully different image regions at the same offset, is rather an exception. This way, we have meaningful content in the most of the series duration.

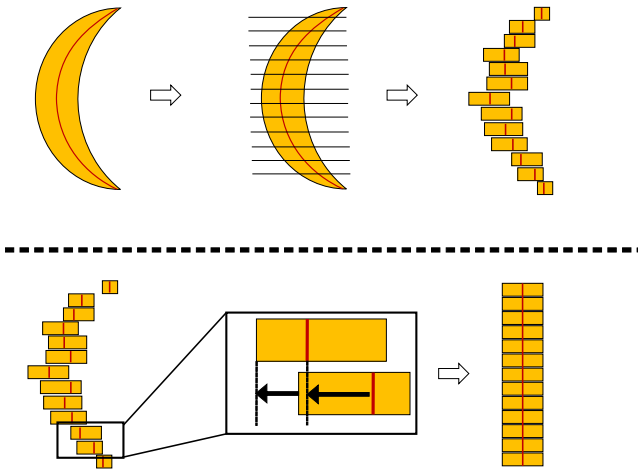


Figure 7: An illustration of the "banana problem". An attempt to maintain better similarity to the neighbors produces too large distortions, destroying the originally present curvature.

We present box plots of the appropriate values of quality measures. In this case we decided against using violin plots. Violin plots show outliers similarly, but the median and the shape of the inliers can be discerned with less clarity in most of our particular cases. (We still present some violin plots in Fig. 11, see also supplementary material.)

We also present a statistical evaluation. We performed an unpaired t-test with different means and unequal variances, a Welch two sample t-test, to be exact. We always compared a measure of registration results to the same measure of the ground truth. We deem $p < 10^{-6}$ significant; most of the time we find even smaller p values.

IV. RESULTS

Beforehand, we have established our specimens, discussed the data processing (Section III-G), data preparation (Section III-H), and evaluation methodology (Section III-I). The main result of this section is a full-series evaluation with statistical means (Section IV-A). We also include some special cases (Section IV-B). A pair-wise evaluation is included in the supplementary material.

Notice that full images were always used for the registration. We mostly look at the center crops for the consistency of the evaluation, but full images were processed beforehand.

A. FULL-SERIES EVALUATION

For an evaluation of full series, we computed the three numerical measures SSIM, Jaccard, PSNR over the whole series and evaluate these values statistically. All Jaccard values below are computed with threshold 100.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access

1) CT

Consider Fig. 8, presenting box plots of the full-series evaluations. Overall "Deform-SURF" and "GS" stand out for their good performance. Panel (a) shows SSIM values. Notice, how the median in "GS" is higher than in ground truth (0.895 17 vs. 0.877 45). In panel (a) "GS" and "Deform-SURF" are close to the ground truth, while there are a lot of outliers in "Deform-SURF" and "GS" seems to overshoot a bit, but has some lower outliers. There are few outliers in the ground truth, too, but they are rather symmetric. The latter also holds for the Jaccard measure. In panel (b) there are some outliers with *high* Jaccard values in "GS". In (c) we see, again, similar values in "GS" and "Deform-SURF" to the ground truth, but now there are many outliers with *high* and very low PSNR values in "Deform-SURF". Notice also the shape of the ground truth box content for PSNR: there are quite many values above the median. Overall, Elastix has good results, but quite tall box plots, indicating high variance. A possible reason is that Elastix operated pair-wise on the image sequence in this case. Both "Deform-SURF" and "GS" operate on a full image stack at once.

Table 1 shows a statistical evaluation. We aim to decide with a Welch t-test, if the quality measures of a registered series are similar to the ground truth. This is almost never the case. In "GS" with respect to Jaccard measure ((b)) we see the largest similarity, according to the test, but p is still quite low there, under $4.3 \cdot 10^{-4}$. Sometimes, the maximal values of the quality measures for a registration method are higher than the maximal ground truth value for the same measure. We attribute this to a wider "spread" of the variance, induced by the registration. To give an example for the SSIM measure, the variance of "Deform-SURF" for the full series is $4.29 \cdot 10^{-3}$, while the variance of the ground truth is $1.05 \cdot 10^{-4}$.

2) EM

Figure 9 presents the box plots of the quality measures for the full EM series. The normalized, 8 bit ground truth was the actual input for our distortion method. Its output was the input of the registrations. The 16 bit ground truth is the original data and is provided as a reference. Fig. 9 necessitated some adjustments. There were some very low SSIM values. We adjusted the scale of y-axis in panel (a) to show more of the relevant details around the median values and to remove some outliers. Panel (b) was plotted without any adjustments. We had to filter the PSNR data to remove infinite values in panel 9c. Also, the quality values there for the ground truth, 16 bit, were much higher than for other modalities. We adjusted the scale of y-axis to show the results from the registrations more detailed.

We see in panel (a) that "GS" almost reaches the level of the ground truth, normalized with respect to SSIM. The 16 bit ground truth has lower SSIM values because of more details. As before, "GS", "Deform-SURF", and Elastix look quite good in box plots. The Jaccard values (b) were rather high, though. The probable reason is the amount of background in

Table 1: The Welch two sample t-test for the quality measures on CT data set. Similar as above, the "locally distorted" values originate from applying only local distortions with our method. The registrations' input was also globally transformed.
Concerning the appropriate measures' values, we compare the means from each of the registration results to the mean for the ground truth. The differences are statistically significant in almost all cases, we show the p value for "not equal". The hypothesis of the equal mean was almost always refuted with a high confidence.
We additionally show the maximal value of a measure ("Max" column). The "local" row is for the local distortions only, without global movements.
The mean value marked with $*$ is the one closest to the ground per t-test. The maximum values marked with † are larger than the maximum of the ground truth. This is a "crime" many good methods commit in our evaluation.
The closer is the mean to the ground truth, the better.

(a)

CT, SSIM

Method	Mean	$p <$	Max
Rigid-SIFT	0.477	$2.2 \cdot 10^{-16}$	0.814
Rigid-SURF	0.489	$2.2 \cdot 10^{-16}$	0.827
Deform-SURF	0.852	$2.2 \cdot 10^{-16}$	0.915†
GS	0.893	$2.2 \cdot 10^{-16}$	0.920†
Blending	0.484	$2.2 \cdot 10^{-16}$	0.814
Elastix	0.755	$2.2 \cdot 10^{-16}$	0.918†
Locally distorted	0.391	$2.2 \cdot 10^{-16}$	0.699
Ground truth	0.878	–	0.905

(b)

CT, Jaccard

Method	Mean	$p <$	Max
Rigid-SIFT	0.0817	$2.2 \cdot 10^{-16}$	0.474
Rigid-SURF	0.0815	$2.2 \cdot 10^{-16}$	0.468
Deform-SURF	0.493	$5.7 \cdot 10^{-10}$	0.670
GS	0.510*	0.000 43	0.676
Blending	0.0793	$2.2 \cdot 10^{-16}$	0.463
Elastix	0.391	$2.2 \cdot 10^{-16}$	0.743
Locally distorted	0.0340	$2.2 \cdot 10^{-16}$	0.315
Ground truth	0.520	–	0.764

(c)

CT, PSNR

Method	Mean	$p <$	Max
Rigid-SIFT	23.1	$2.2 \cdot 10^{-16}$	32.1
Rigid-SURF	23.0	$2.2 \cdot 10^{-16}$	32.8
Deform-SURF	31.5*	$3.1 \cdot 10^{-7}$	35.6†
GS	31.3	$2.2 \cdot 10^{-16}$	32.6
Blending	23.4	$2.2 \cdot 10^{-16}$	34.2
Elastix	28.9	$2.2 \cdot 10^{-16}$	34.4
locally distorted	21.0	$2.2 \cdot 10^{-16}$	29.5
Ground truth	31.9	–	35.4

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access

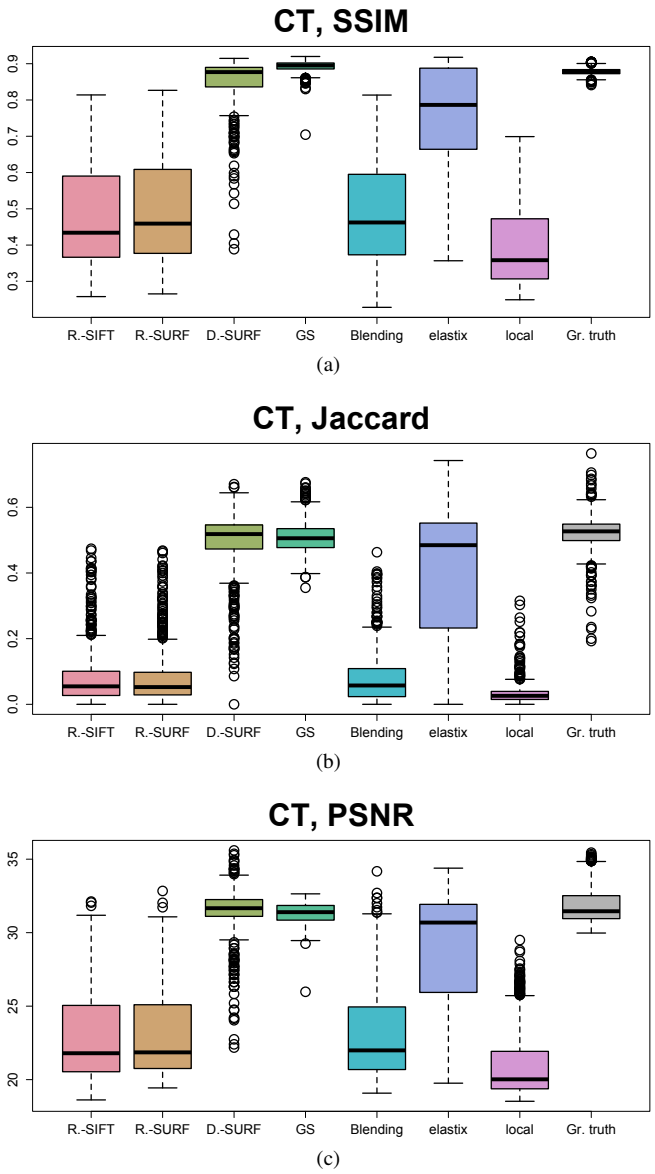


Figure 8: Box plots of quality measures for the CT data set. "R.-SIFT" stands for Rigid-SIFT, "R.-SURF" for Rigid-SURF, "D.-SURF" stands for Deform-SURF, "local" means local distortions only, without a global rigid transformation, "Gr. truth" stands for ground truth.

the EM data set. In both discussed measures there are some outliers on the lower side. PSNR (c) shows very high values for 16 bit ground truth, we disregard them as all other values are 8 bit. Concerning PSNR, we see some over-registration in "GS", less so in Elastix and "Deform-SURF": the median values and most of the box contents (the box represents 50% of the data around the median) are *higher* than in ground truth. The box plot for the normalized ground truth shows some outliers for the larger PSNR values, however.

In the statistical evaluation (Table 2), we see a slightly larger p value for "GS" with respect to SSIM, but nothing extraordinary for this measure. The high values of 1.0 for SSIM originate from the region at the end of the series with few changes because of low amount of tissue. It rather indicates a failure of the registration, as the measure is computed on a center crop. As mentioned above, the Jaccard values are rather high, again, "GS" manages to obtain a slightly higher p for its mean. Quite of interest is PNSR, where "Deform-SURF" manages a $p \leq 0.02713$, but even more spectacularly, Elastix has $p \leq 0.7992$. This value basically means that the mean of the PSNR for this data set, registered with Elastix, matches the mean of the normalized ground truth with a high probability. Such a match is an exception in our evaluations.

3) LS

Consider Fig. 10. The SSIM for LS method shows quite good values for "Deform-SURF", "GS", and also for Elastix, but in this case with some outliers. Surprisingly, the plot for the normalized ground truth is less convincing, the original, 16 bit ground truth shows even less similarity. The variance is clearly much larger in the 16 bit data. We can thus conclude, that above "good" registration method over-register. Looking into Fig. 11a, we concludes that "GS", "Deform-SURF", and less so, Elastix, produce much higher SSIM values than they should have in order to be similar with the normalized ground truth. The variance in the results of those registration methods is also lower than in the normalized ground truth. Panel (b) shows in a violin plot how the distribution of SSIM values changed between the methods.

We see quite high values for Jaccard measure in Fig. 10b, but also a lot of outliers in almost all methods. To study those further, we present a zoomed-in version in Fig. 11c. Even more interesting is panel 11d. There we have removed the values 0 and 1.0 from the evaluation. Basically, those extreme Jaccard values mean that either no correspondence at all was found or the full correspondence. The latter can be caused by too low threshold value or too little detail in the particular region. In those both panels we see a superior performance of the "Blending" method compared to all other registrations. We notice also that our local distortions do not change the Jaccard index very much. The statistical analysis (below) does not support the superiority of "Blending", however.

As also in other modalities, the PSNR value of the 16 bit ground truth is much larger, than in all other methods that utilize 8 bit images (Fig. 10c). In a zoomed-in version in Fig. 11, (e) we see that for PSNR the median of no registration method exceeds the median of the ground truth. This means that PSNR detects no over-registration in this case. The somewhat peculiar, uneven shapes of the PSNR distributions are visualized as violin plots in panel (f).

Now, consider Table 3. Statistically, no registration method matches the mean of the SSIM of normalized ground truth well. Most of the methods (namely, "Deform-SURF", "GS", Elastix) are well above and also all registrations overshoot the maximum of the ground truth SSIM. In Jaccard, we have many 1.0 values (which were not removed in this case, as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access

Table 2: The Welch two sample t-test for the quality measures on EM data set. The origin of the "locally distorted" values is as above.
We compare the means from each of the registration results to the mean for the normalized, 8 bit ground truth. The differences are statistically significant in almost all cases, we show the p value for "not equal". The hypothesis of the equal mean was almost always refuted with a high confidence.
Notice Elastix with respect to PSNR with $p < 0.7992$, it matches the mean of the normalized ground truth PSNR up to −0.2 % relative error.
The mean value marked with $*$ is the one closest to the normalized ground per t-test. We also show the maximal value of a measure (the "Max" column). The maximum values marked with † are larger than the maximum of the normalized ground truth. Notice that the quality measures are unusually high for this data set. In contrast to Fig. 9, we operate on unfiltered data for SSIM and Jaccard. We still had to remove infinite values of PSNR for a meaningful analysis. The sole method where data was individually filtered in the above manner is marked with ¶.
The closer is the mean to the ground truth, the better.

(a)

EM, SSIM

Method	Mean	$p <$	Max
Rigid-SIFT	0.901	$2.2 \cdot 10^{-16}$	1.00 †
Rigid-SURF	0.888	$2.2 \cdot 10^{-16}$	0.989
Deform-SURF	0.948	$2.2 \cdot 10^{-16}$	0.991
GS	0.957*	$1.6 \cdot 10^{-9}$	0.992
Blending	0.883	$2.2 \cdot 10^{-16}$	0.983
Elastix	0.943	$2.2 \cdot 10^{-16}$	0.983
Locally distorted	0.888	$2.2 \cdot 10^{-16}$	0.993
Ground truth, 16 bit	0.937	$2.2 \cdot 10^{-16}$	0.989
Ground truth, norm.	0.964	–	0.997

(b)

EM, Jaccard

Method	Mean	$p <$	Max
Rigid-SIFT	0.9604	$2.2 \cdot 10^{-16}$	1.000 †
Rigid-SURF	0.9610	$2.2 \cdot 10^{-16}$	0.9991
Deform-SURF	0.9864	$1.2 \cdot 10^{-11}$	1.000 †
GS	0.9883*	$3.2 \cdot 10^{-5}$	0.9995
Blending	0.9531	$2.2 \cdot 10^{-16}$	0.9963
Elastix	0.9861	$1.1 \cdot 10^{-12}$	0.9989
Locally distorted	0.9548	$2.2 \cdot 10^{-16}$	0.9996
Ground truth, 16 bit	0.9998	$2.2 \cdot 10^{-16}$	1.000 †
Ground truth, norm.	0.9902	–	0.9998

(c)

EM, PSNR

Method	Mean	$p <$	Max
Rigid-SIFT¶	18.0	$2.2 \cdot 10^{-16}$	36.55
Rigid-SURF	21.3	$2.2 \cdot 10^{-16}$	34.35
Deform-SURF	27.5	0.027 13	40.20†
GS	28.9	$2.2 \cdot 10^{-16}$	39.96
Blending	19.4	$2.2 \cdot 10^{-16}$	31.78
Elastix	26.9*	0.7992	35.42
Locally distorted	20.0	$2.2 \cdot 10^{-16}$	37.72
Ground truth, 16 bit	44.9	$2.2 \cdot 10^{-16}$	58.52
Ground truth, norm.	27.0	–	40.15

Table 3: The Welch two sample t-test for the quality measures on LS data set. The "locally distorted" values are as above. We compare the means from each of the registration results to the mean for the normalized, 8 bit ground truth. The differences are statistically significant in almost all cases, we show the p value for "not equal". The hypothesis of the equal mean was almost always refuted with a high confidence.
The maximum of the normalized ground truth for SSIM was rather low. For Jaccard, "Deform-SURF" reaches $p < 0.168$, and for PSNR the same method reaches $p < 0.7875$. For the latter, it matches the mean of the ground truth up to 0.1865 %, the 95 % confidence interval is −0.4345 to 0.5728. The mean value marked with $*$ is the one closest to the normalized ground per t-test. We also show the maximal value of a measure (the "Max" column). The maximum values marked with † are larger than the maximum of the normalized ground truth. Notice that the quality measures are unusually high for this data set. Similar to Fig. 10, we operate on unfiltered data for SSIM and Jaccard. The method marked with ¶ has no good values. The closer is the mean to the ground truth, the better.

(a)

LS, SSIM

Method	Mean	$p <$	Max
Rigid-SIFT	0.796	$2.2 \cdot 10^{-16}$	0.968†
Rigid-SURF	0.678	$2.2 \cdot 10^{-16}$	0.992†
Deform-SURF	0.964	$2.2 \cdot 10^{-16}$	0.993†
GS	0.971	$2.2 \cdot 10^{-16}$	0.994†
Blending	0.819	$2.2 \cdot 10^{-16}$	0.966†
Elastix	0.945*	$1.0 \cdot 10^{-10}$	0.985†
Locally distorted	0.721	$2.2 \cdot 10^{-16}$	0.914
Ground truth, 16 bit	0.572	$2.2 \cdot 10^{-16}$	0.851
Ground truth, norm.	0.915	–	0.950

(b)

LS, Jaccard

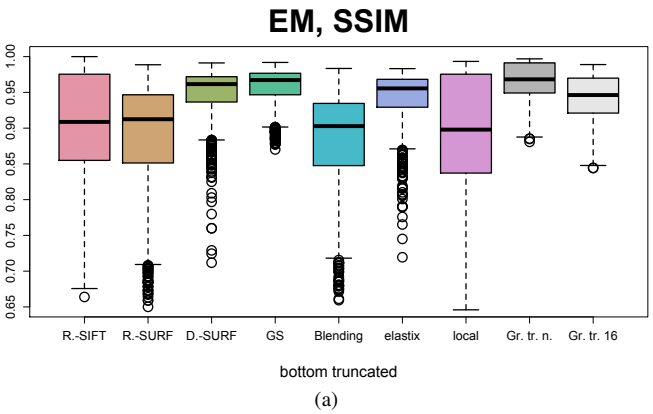
Method	Mean	$p <$	Max
Rigid-SIFT	0.8399	0.097 31	1.000
Rigid-SURF	0.6347	$2.2 \cdot 10^{-16}$	1.000
Deform-SURF	0.8487*	0.1680	0.9890
GS	0.8444	0.1245	0.9882
Blending	0.8353	0.068 07	1.000
Elastix	0.8337	0.045 40	0.9893
Locally distorted	0.9700	$1.3 \cdot 10^{-7}$	1.000
Ground truth, 16 bit¶	1.000	$2.2 \cdot 10^{-16}$	1.000
Ground truth, norm.	0.8795	–	1.000

(c)

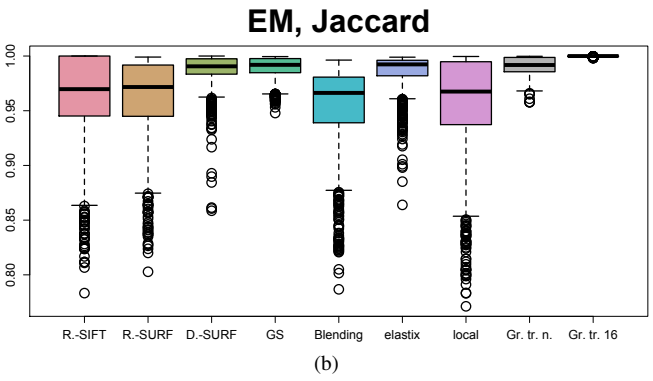
LS, PSNR

Method	Mean	$p <$	Max
Rigid-SIFT	26.9	$2.2 \cdot 10^{-16}$	42.29†
Rigid-SURF	24.6	$2.2 \cdot 10^{-16}$	48.96†
Deform-SURF	37.1	0.7875	50.04†
GS	37.3	0.3101	50.45†
Blending	27.8	$2.2 \cdot 10^{-16}$	42.35†
Elastix	35.7	$5.8 \cdot 10^{-5}$	47.30†
Locally distorted	27.0	$2.2 \cdot 10^{-16}$	39.69
Ground truth, 16 bit	56.3	$2.2 \cdot 10^{-16}$	69.20†
Ground truth, norm.	27.0	–	41.36

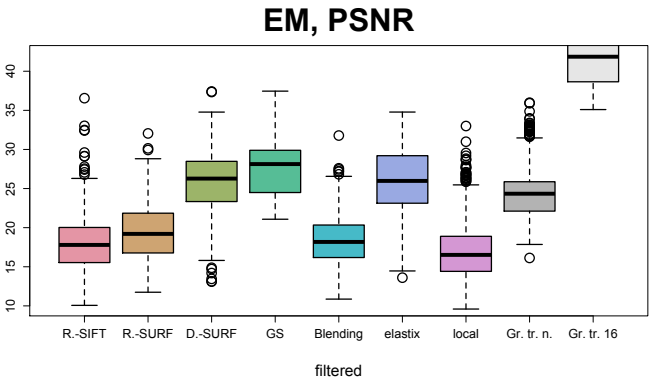
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access



(a)

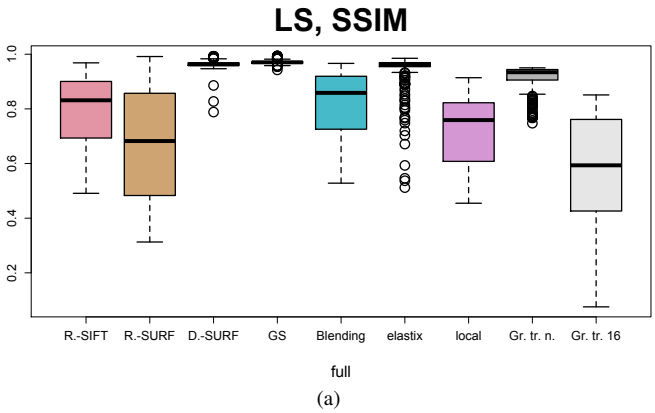


(b)

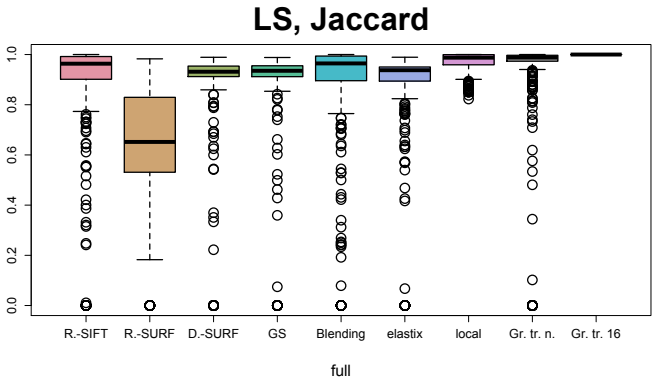


(c)

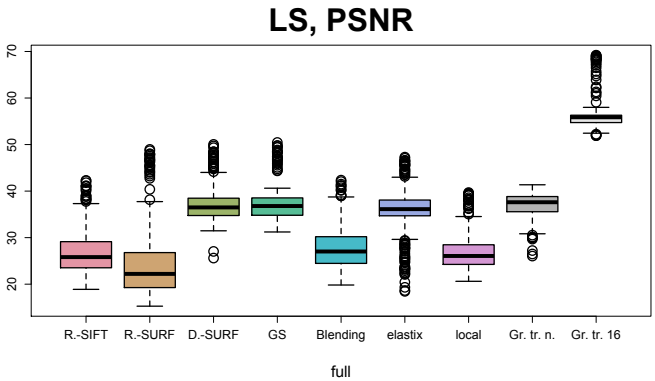
Figure 9: Box plots of quality measures for the EM data set. “R.-SIFT” stands for Rigid-SIFT, “R.-SURF” for Rigid-SURF, “D.-SURF” for Deform-SURF, “local” for only local distortions, “Gr. tr. n.” for “ground truth, normalized”, “Gr. tr. 16” for “ground truth, 16 bit”.



(a)



(b)



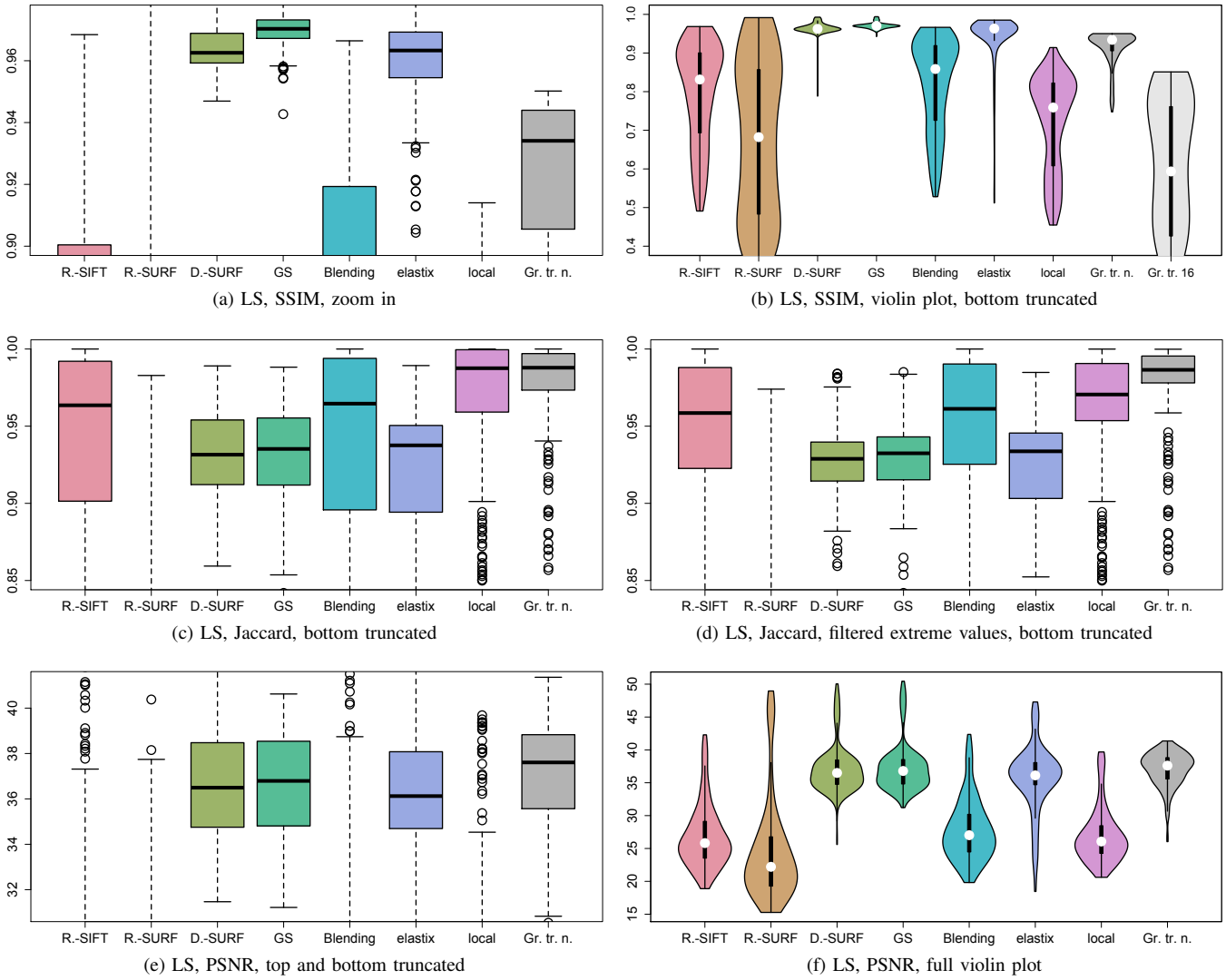
(c)

Figure 10: Box plots of quality measures for the LS data set. “R.-SIFT” stands for Rigid-SIFT, “R.-SURF” for Rigid-SURF, “D.-SURF” for Deform-SURF, “local” for only local distortions, “Gr. tr. n.” for “ground truth, normalized”, “Gr. tr. 16” for “ground truth, 16 bit”. Fig. 11 shows some further details.

we want to contrast those statistics to the box plots). We see a match in the means with $p < 0.168$ in “Deform-SURF”, however we would be quite cautions in this case because of some “invalid”, too low or too high Jaccard values. With PSNR, “Deform-SURF” manages to match the mean with $p < 0.7875$.

Some methods (e. g., “GS”, Elastix) have even larger mean values of PSNR. We would deem those methods as better than “Deform-SURF” in this case, if we did not have the ground truth. We have to note, however, that *all* registrations overshoot the maximum of the PSNR in the ground truth, as evident from Table 3c. A repeated look on Fig. 11, panels (e), (f) shows that the distributions of the PSNR values for “GS” and “Deform-SURF” are much more similar to each other than to that of the normalized ground truth.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access



(a) LS, SSIM, zoom in

(b) LS, SSIM, violin plot, bottom truncated

(c) LS, Jaccard, bottom truncated

(d) LS, Jaccard, filtered extreme values, bottom truncated

(e) LS, PSNR, top and bottom truncated

(f) LS, PSNR, full violin plot

Figure 11: More detailed plots of LS quality measures. Abbreviations are same as above. "Truncated" means that not the full range of the measure was plotted. In Jaccard, "filtered extreme values" means that numerous zero and 1.0 values were removed. They can be attributed to missing tissue or completely full region, were no meaningful comparison can be made at current threshold. We also show some violin plots where those might bring additional insight through their shape.
Concerning the "Rigid-SURF" method, its quality was much lower. As this figure aims to provide a more detailed view, we scale the box plots to better visualize the differences between other methods. Fig. 10 as well as the violin plot (f) show the full picture.
As before, "R.-SIFT" stands for Rigid-SIFT, "R.-SURF" for Rigid-SURF, "D.-SURF" for Deform-SURF, "local" for only local distortions, "Gr. tr. n." for "ground truth, normalized", "Gr. tr. 16" for "ground truth, 16 bit".

B. SPECIAL CASES

For the pair-wise evaluation in supplementary material we intentionally picked the center of a series and a clearly defined region. Here we would like to separately highlight some especially good of bad consecutive image pairs in Fig. 12. This is a subjective selection of image pairs, as contrasted with the previous section.

In panel (a) a larger global shift in non-linear distortion phase of the registrations' input is shown. It is CT data set, sections 95–96, Dice visualization. Fig. (b) shows uncorrected global movement in Elastix-based registration, same data set, sections 99–100, Dice visualization. There is a "floppy end", a movement, in Elastix result (c), from same data set, sections 305–306, Dice visualization. Mostly the registration is good, but in the depicted region the offsets are much larger.

In panels (d), (e) the measures of an interesting image pair from "Deform-SURF" are depicted. We show Dice (d) and optical flow (e) visualizations from the same region. There *is* some movement, but is it all explainable with the "natural"

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access

differences in consecutive images?

The areas with little tissue (as found in our EM series near its end) are a greater challenge for the feature-based methods, as Figs. (f), (g) show. We see there a failure of the feature-based SIFT method to find a correct rigid alignment. Full images from EM data set, sections 702–703, are shown. The probable reason is the low number of viable key points found by the feature detection.

Subfigures (h)–(j) show a small evaluation of a particularly good GS-registered image pair, LS data set, images 161–162. We show there Dice visualization, PSNR, and SSIM, respectively, for the same region. Notice the low values and very few differences.

V. DISCUSSION

A. THE RESULTS AND THE "BANANA PROBLEM"

We have mostly discussed the results in the previous section, still there is an issue we would like to specially highlight. We have quite often seen that methods which produce better image-based metrics also seem to over-register the series. It would be very hard to find such an over-registration (a "banana problem", Fig. 7) without a ground truth. CT scans of the specimens before sectioning might help, but, as mentioned above, they lack on the resolution. Basically, our ground truth bounds from above the amount of correspondence between consecutive images. Such challenges as ours would help to develop better registrations that try to reach such a bound, but not to overstep it.

Occasionally, we have found the values of our quality measures for a particular registration method *higher* than for the ground truth. How is this possible? Our reasoning is that the ground truth does not constitute a *perfect* correspondence of the consecutive input images. It is merely their real correspondence drawn from inherently 3D data. This means that those registrations might create *too much* correspondence, they over-register their inputs. We have also seen an interesting effect, where the correspondences after a registration where more *heterogeneous* than in ground truth. It appears to us that some areas were over-registered and some areas were under-registered. Again, without a ground truth such observations would be impossible.

In other words, to obtain good values overall or on average (while high variance cannot be reduced), some registrations seem to attempt to shift all of the correspondences between the images towards their maximums. Naturally, this behavior forces the maximal values to exceed the maximum of the ground truth while the mean is still under the mean of the ground truth. This leads to the "banana problem".

B. DISTORTION MECHANISM

Our distortions "stretch" and "shrink" the images, but their positions are organized in a rectilinear grid, even if the distortions themselves are random and Rayleigh-distributed. It would also make sense to choose the distortion positions randomly, too. However, we opted for a grid for a better reproducibility of the appearance of the test images: a human would know where to look. Still, the actual distorted images do not show that much "bad" regularity, the above grid is not visible, so we argue that no problems arise from such a rectilinear grid placement. If our method is used to benchmark registrations using neural networks with a direct assignment of neurons to either pixels or distortions, the randomization of the distortion positions might be required.

We use a rigid transformation to model the inaccuracy in section placement. If a fully affine or an even more generic transform is needed, our code can be easily extended to incorporate it. Indeed, some registration methods [e. g., 42, 153] use affine global transformations to model wedge-shaped sections.

To contrast phantoms to our approach: the argument on not fully representing the distortions might also hold for our distortion generation, but we still work with real data. Hence, the reasoning on lacking data complexity does not hold. This issue is especially prominent in methods based on feature detection.

C. FURTHER IDEAS FOR THE DISTORTION MODELING

Our method currently does not directly account for tissue folding and tearing. Such damaged areas can be represented with "holes" in the images, but they currently would not correlate with larger distortions in the connected areas. A realistic modeling of section damage was not our current goal. Nowadays, methods to bridge section damage exist [91]. Basically, any kinds of damage can be assessed with masks for the repair, similar to the masks we use here to simulate the damage. Some further recent works assess cracks and discontinuities [154, 155].

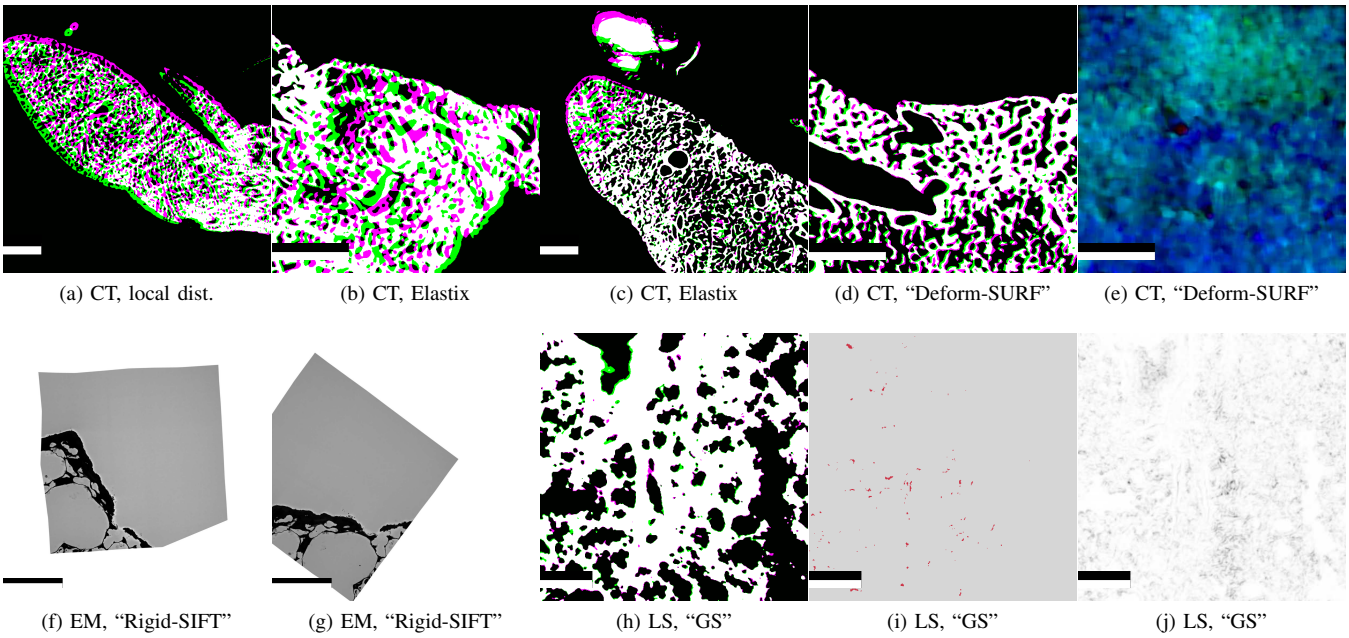
A convolutional neural network, transforming "clean" images into damaged ones with some kind of a style transfer [156, 157, 158, 159] is an interesting idea. We sought for a functional and well-defined image deformation that allowed for using transformed images as a benchmark input. Neural-network-generated images might have some unnoticeable for humans drawbacks that would obscure and throw off-track some other (probably, also deep-learning-based) registrations—detection of adversarial examples is a separate problem. Our test images are produced through simple, robust, reproducible, and well-understood image transformations.

D. IMPACT OF A 3D SERIES

We use full-blown, inherently 3D images as a series of 2D images. Those 2D inputs are used for our benchmark for a reason. Some registrations do not operate on image pairs, but optimize the spatial positions of the full image stack at one.

Next, the presence of already three-dimensional images as the starting point enables us to state how the final image stack should look like. We digitally simulate the distortions by sectioning and further section handling. Then we apply a registration method (we evaluate multiple of them in this work). The discrepancy between the distorted series is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3124341, IEEE Access



(a) CT, local dist. (b) CT, Elastix (c) CT, Elastix (d) CT, "Deform-SURF" (e) CT, "Deform-SURF"

(f) EM, "Rigid-SIFT" (g) EM, "Rigid-SIFT" (h) LS, "GS" (i) LS, "GS" (j) LS, "GS"

Figure 12: Special cases, especially good or bad image pairs.
Scale bars are: (a)=1 mm, (b)=500 µm, (c)=1 mm, (d)=500 µm, (e)=500 µm, (f)=50 mm, (g)=50 mm, (h)–(j)=500 µm.

larger, than in the registered series; this is the whole idea of registration. But, contrary to the usual 2D registrations of serial sections, we still have the initial starting point, the 3D images. They are in the same resolution as the registered series. We call those initial images the ground truth. We can compare the registered series to the ground truth and find out, what was wrong with the registration method in question.

E. CHALLENGES IN PROJECT EXECUTION

It was quite hard for some methods to cope with large angles in rigid transformations, in those cases we used SURF-based rigid registrations as an initial phase. Many registrations are inherently trimmed for the most used input data kinds and modalities. Adaptation to further images is possible, but requires more or less tuning. In the best case, the tuning can be commenced with parameter files, such as with Elastix.

The present project was quite large. A decent automation of the workflow (we used Python, bash, and GNU Make) was key for fast and error-free processing. This issue was of especial importance in case of the evaluations.

F. EVALUATIONS

The different normalization issues, esp. in EM data set, might also explain the observed variations in the measurements. This is a typical trade-off: a better normalization allows for better registration, but a normalization also changes the data set, so a direct comparison with non-normalized data might be harder.

The visualizations of optical flow we used as one of the measures is, on the one hand, a valuable tool. Those visualizations show issues less visible otherwise, the "hot spots". On the other hand, a direct comparison of such visualizations with each other in their present form might be misleading because of individual magnitudes.

One of the ideas for further improvement of our work is to compare not images, but deformation fields from various methods. However, multiple implementation questions would arise. One of the issue is the registration "drift" that would be different in various methods: Currently, our challenge is complete open: the participants need to obtain the input images and produce the result images, the registration method itself does not need to be adapted or changed, it can remain closed-source or even a commercial secret. If we would like to compare the deformation fields, we would need to provide a consistent way to output comparable distortions across all the implementations, libraries, programming languages the participants use. The code for the actual method needs to be changed, which means it should be available and human resources for the change need to be allocated.

G. CONTINUOUS INTEGRATION

Our visualizations and measures are computed automatically. No human interaction what so ever is needed: the distortions of the ground truth, registrations, and evaluations can happen fully automatically. Thus, our evaluation method is a gateway to wide-scale registration challenges and to regression testing of further developing registration methods. Our method can be applied as a part of a continuous integration workflow [160]. With our approach, better and more thorough regression testing of registration becomes possible.

Basically, this paper shows a further path towards automat-

ically testing different regularizations in image registrations. The goal would be to reduce the magnitude of the "banana problem", while still maintaining good registration results. Such testing can be done with image-based metrics, as we do here, but any other metric would work too.

VI. CONCLUSIONS

We introduce an approach to generate individual 2D distortions applied to existing 3D medical data. Those distortions have the basic statistic properties of the cutting-induced variations in serial sections. The distortions are computed in a straightforward, understandable, and reproducible manner. We also apply a global rigid transform to mimic the inexact placement of a section on glass slide. Modeling damaged sections is also possible. Combined, we can simulate the transition from a tissue block to a set of serial sections. Thus, we are able to test registration methods on such simulated data originating from real tissues.

We provide an overview of the existing registration and evaluation efforts. To our knowledge, the approach, we pursue in this work, has not been undertaken previously.

The key contribution of this work is the utilization of the original, undistorted data for the evaluation. Previously, it was impossible to evaluate registrations of serial sections with ground truth, as the tissue block is destroyed by sectioning. Micro-CT scans currently do not have sufficient resolution and can serve only as a coarse guide in a co-registration of micro-CT to real serial sections. Phantoms and synthetic images might not have the needed complexity. We use real, micrometer-scale lung images from other modalities in this work. By using real images of animal lungs we affirm that the kind of the images used is comparable to real serial sections of the same tissue. Our method is, however, directly applicable to other organs or generic images (Figs. 1, 2).

In this work we evaluate six registration methods on three distorted data sets. In each of them, a ground truth is present. With the ground truth, we can not only compare the registrations with each other, but also with the inherent 3D data, in other words: with original data, with how the 2D "slices" should have been aligned if no cutting took place. We address the quality of registrations with four visualizations of image metrics and three image-based quality measures. In this work, we both look at a specific image pair (in the supplementary material) and provide an evaluation over the full range of the series. Our method can be applied in a continuous integration workflow.

We make the source code and the data sets publicly available. Further contributions, both in form of additional registration results and further data sets, are welcome.

A. FUTURE WORK

Utilization of our method, statistical analysis of real sections [similar to 137], and an introduction of better quality measures may lead to better registrations, both utilizing deep learning and not. We definitely look forward to more comparisons of registration methods. If the sectioning distortions in resin embeddings are found to be similar to the model we use here, or if a different model is derived, our method can be also adapted to simulate distortions in such embeddings. Further tests on CT or MRT data sets with artificially increased anisotropy can be of interest.

It would be very interesting to find a way to compare the registration result to ground truth directly. (In this work we compare consecutive images from each of the results and evaluate the resulting measures.) Presently, the accumulation of registration "drift" and some global offsets make a well-founded assessment more complicated than our present evaluation.

ACKNOWLEDGMENTS

We thank Adrian V. Dalca (MIT, MA, USA), Markus Wedekind (Technische Universität Braunschweig, Germany), and Birte S. Steiniger (Philipps-University of Marburg, Germany) for helpful discussions. Susanne Kuhlmann and Susanne Faßbender (MHH, Hannover, Germany) provided excellent technical support. Anja Bubke (MHH, Hannover, Germany) was involved in LS acquisition. A GPU (Quadro P5000) used for this research was donated by the NVIDIA Corporation. We used the SSIM implementation by Zhou Wang. We modified the code by Jean Francois Pambrun for Dice measure visualizations.

References

[1] L. G. Brown, "A survey of image registration techniques," *ACM Comput. Surv.*, vol. 24, no. 4, pp. 325–376, 1992.

[2] B. Zitová and J. Flusser, "Image registration methods: a survey," *Image Vis. Comput.*, vol. 21, no. 11, pp. 977–1000, 2003.

[3] J. Pluim, J. Maintz, and M. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE T. Med. Imaging*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.

[4] F. P. Oliveira and J. M. R. Tavares, "Medical image registration: a review," *Comput. Method. Biomec.*, vol. 17, no. 2, pp. 73–93, Jan. 2014.

[5] M. A. Viergever, J. B. A. Maintz, S. Klein, K. Murphy, M. Staring, and J. P. W. Pluim, "A survey of medical image registration – under review," *Med. Image Anal.*, vol. 33, pp. 140–144, Oct. 2016.

[6] J. Pichat, J. E. Iglesias, T. Yousry, S. Ourselin, and M. Modat, "A survey of methods for 3D histology reconstruction," *Med. Image Anal.*, vol. 46, pp. 73–105, May 2018.

[7] B. Zitova, "Mathematical approaches for medical image registration," in *Encyclopedia of Biomedical Engineering*, R. Narayan, Ed. Oxford: Elsevier, Jan. 2019, pp. 21–32.

[8] J. H. van Krieken, J. Te Velde, J. Hermans, and K. Welvaart, "The splenic red pulp; a histomorphometrical study in splenectomy specimens embedded in

methylmethacrylate," *Histopathol.*, vol. 9, no. 4, pp. 401–416, 1985.

[9] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE T. Med. Imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.

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

[11] J. Schnabel, C. Tanner, A. Castellano-Smith, A. Degenhard, M. Leach, D. Hose, D. Hill, and D. Hawkes, "Validation of nonrigid image registration using finite-element methods: application to breast MR images," *IEEE T. Med. Imaging*, vol. 22, no. 2, pp. 238–247, Feb. 2003.

[12] H. Chui, L. Win, R. Schultz, J. S. Duncan, and A. Rangarajan, "A unified non-rigid feature registration method for brain mapping," *Med. Image Anal.*, vol. 7, no. 2, pp. 113–130, Jun. 2003.

[13] L. Hömke, "A multigrid method for anisotropic PDEs in elastic image registration," *Numer. Linear Algebr.*, vol. 13, no. 2-3, pp. 215–229, Mar. 2006.

[14] Y.-L. Zhang, S.-J. Chang, X.-Y. Zhai, J. S. Thomsen, E. I. Christensen, and A. Andreasen, "Non-rigid landmark-based large-scale image registration in 3-D reconstruction of mouse and rat kidney nephrons," *Micron*, vol. 68, pp. 122–129, Jan. 2015.

[15] S. Saalfeld, A. Cardona, V. Hartenstein, and P. Tomancak, "As-rigid-as-possible mosaicking and serial section registration of large ssTEM datasets," *Bioinformatics*, vol. 26, no. 12, pp. i57–i63, Jun. 2010.

[16] 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.

[17] W. Crum, L. Griffin, D. Hill, and D. Hawkes, "Zen and the art of medical image registration: correspondence, homology, and quality," *NeuroImage*, vol. 20, no. 3, pp. 1425–1437, Nov. 2003.

[18] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "elastix: A toolbox for intensity-based medical image registration," *IEEE T. Med. Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

[19] D. P. Shamonin, E. E. Bron, B. P. F. Lelieveldt, M. Smits, S. Klein, and M. Staring, "Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease," *Front. Neuroinform.*, vol. 7, 2014.

[20] F. F. Berendsen, K. Marstal, S. Klein, and M. Staring, "The design of SuperElastix — A unifying framework for a wide range of image registration methodologies," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 498–506.

[21] K. Marstal, F. Berendsen, M. Staring, and S. Klein, "SimpleElastix: A user-friendly, multi-lingual library for medical image registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jun. 2016, pp. 134–142.

[22] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ANTs similarity metric performance in brain image registration," vol. 54, no. 3, pp. 2033–2044, 2011.

[23] B. B. Avants, N. J. Tustison, M. Stauffer, G. Song, B. Wu, and J. C. Gee, "The Insight ToolKit image registration framework," vol. 8, p. 44, 2014.

[24] U. Bağci, X. Chen, and J. Udupa, "Hierarchical scale-based multiobject recognition of 3-D anatomical structures," *IEEE T. Med. Imaging*, vol. 31, no. 3, pp. 777–789, Mar. 2012.

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

[26] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *Int. J. Comput. Vis.*, Aug. 2020.

[27] D. G. Lowe, "Object recognition from local scale-invariant features," in *Processings of IEEE International Conference on Computer Vision*, ser. ICCV '99, vol. 2. IEEE, 1999, pp. 1150–1157.

[28] ——, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[29] H. Bay, T. Tuytelaars, and L. Gool, "SURF: Speeded up robust features," in *Computer Vision – ECCV 2006*, ser. LNCS. Springer, 2006, vol. 3951, pp. 404–417.

[30] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Und.*, vol. 110, no. 3, pp. 346–359, 2008.

[31] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE features," in *Computer Vision – ECCV 2012*, ser. LNCS, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Springer, 2012, vol. 7577, pp. 214–227.

[32] P. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," in *Procedings of the British Machine Vision Conference 2013*. Bristol: British Machine Vision Association, 2013, pp. 13.1–13.11.

[33] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[34] A. Can, C. Stewart, B. Roysam, and H. Tanenbaum, "A feature-based, robust, hierarchical algorithm for registering pairs of images of the curved human retina," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 24, no. 3,

pp. 347–364, 2002.

[35] M. Urschler, J. Bauer, H. Ditt, and H. Bischof, "SIFT and shape context for feature-based nonlinear registration of thoracic CT images," in *Computer Vision Approaches to Medical Image Analysis*, ser. LNCS, R. R. Beichel and M. Sonka, Eds. Berlin, Heidelberg: Springer, 2006, vol. 4241, pp. 73–84.

[36] A. Ruiz, M. Ujaldon, L. Cooper, and K. Huang, "Non-rigid registration for large sets of microscopic images on graphics processors," *J. Signal Process. Sys.*, vol. 55, no. 1–3, pp. 229–250, 2009.

[37] R. Shojaii and A. L. Martel, "Optimized SIFTFlow for registration of whole-mount histology to reference optical images," *J. Med. Imag.*, vol. 3, no. 4, pp. 047 501–1–10, Oct. 2016.

[38] C. Wang, M. Oda, Y. Hayashi, B. Villard, T. Kitasaka, H. Takabatake, M. Mori, H. Honma, H. Natori, and K. Mori, "A visual SLAM-based bronchoscope tracking scheme for bronchoscopic navigation," *Int. J. Comput. Ass. Rad.*, Aug. 2020.

[39] I. Arganda-Carreras, R. Fernández-González, A. Muñoz-Barrutia, and C. Ortiz-De-Solorzano, "3D reconstruction of histological sections: Application to mammary gland tissue," *Microsc. Res. Techniq.*, vol. 73, no. 11, pp. 1019–1029, 2010.

[40] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "NIH image to ImageJ: 25 years of image analysis," *Nat. Methods*, vol. 9, no. 7, pp. 671–675.

[41] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona, "Fiji: an open-source platform for biological-image analysis," *Nat. Methods*, vol. 9, no. 7, pp. 676–682, 2012.

[42] A. Cardona, S. Saalfeld, J. Schindelin, I. Arganda-Carreras, S. Preibisch, M. Longair, P. Tomancak, V. Hartenstein, and R. J. Douglas, "TrakEM2 software for neural circuit reconstruction," *PLOS ONE*, vol. 7, no. 6, p. e38011, Jun. 2012.

[43] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, "Locality preserving matching," *Int. J. Comput. Vis.*, vol. 127, no. 5, pp. 512–531, May 2019.

[44] F. Zhang, Y. Gao, and L. Xu, "An adaptive image feature matching method using mixed Vocabulary-KD tree," *Multimed. Tools Appl.*, vol. 79, no. 23, pp. 16 421–16 439, Jun. 2020.

[45] T. Cieslewski, M. Bloesch, and D. Scaramuzza, "Matching features without descriptors: Implicitly matched interest points," in *2019 30th British Machine Vision Conference*, ser. BMVC '2019. BMVA Press, Aug. 2019, p. 32. [Online]. Available: http://arxiv.org/abs/1811.10681

[46] D. Mueller, D. Vossen, and B. Hulsken, "Real-time deformable registration of multi-modal whole slides for digital pathology," *Comput. Med. Imag. Grap.*, vol. 35, no. 7, pp. 542–556, Oct. 2011.

[47] C. Ulrich, O. Lobachev, B. Steiniger, and M. Guthe, "Imaging the vascular network of the human spleen from immunostained serial sections," in *Proceedings of the 4th Eurographics Workshop on Visual Computing for Biology and Medicine*, ser. VCBM '14. Goslar, Germany: Eurographics, 2014, p. 69–78.

[48] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence – Volume 2*, ser. IJCAI '81. Vancouver, BC, Canada: Morgan Kaufmann, Aug. 1981, pp. 674–679.

[49] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, no. 1-3, pp. 185–203, Aug. 1981. [Online]. Available: https://dspace.mit.edu/handle/1721.1/6337

[50] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, ser. LNCS, J. Bigun and T. Gustavsson, Eds. Berlin Heidelberg: Springer, 2003, vol. 2749, pp. 363–370.

[51] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Computer Vision — ECCV 2004*, ser. LNCS, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, T. Pajdla, and J. Matas, Eds. Berlin, Heidelberg: Springer, 2004, vol. 3024, pp. 25–36.

[52] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "EpicFlow: Edge-preserving interpolation of correspondences for optical flow," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, Jun. 2015, pp. 1164–1172.

[53] L. Dougherty, J. C. Asmuth, and W. B. Gefter, "Alignment of CT lung volumes with an optical flow method," *Acad. Radiol.*, vol. 10, no. 3, pp. 249–254, Mar. 2003.

[54] L. Carata, D. Shao, M. Hadwiger, and E. Groeller, "Improving the visualization of electron-microscopy data through optical flow interpolation," in *Proceedings of the 27th Spring Conference on Computer Graphics*, ser. SCCG '11. New York, NY, USA: ACM, 2013, pp. 103–110.

[55] O. Lobachev, B. S. Steiniger, and M. Guthe, "Compensating anisotropy in histological serial sections with optical flow-based interpolation," in *Proceedings of the 33rd Spring Conference on Computer Graphics*, ser. SCCG '17. New York, NY, USA: ACM, 2017, p. 11.

[56] C. Liu, J. Yuen, and A. Torralba, "SIFT Flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[57] G. Eilertsen, P.-E. Forssén, and J. Unger, "BriefMatch: Dense binary feature matching for real-time optical flow estimation," in *Image Analysis*, ser. LNCS, P. Sharma and F. M. Bianchi, Eds. Cham: Springer, 2017, vol. 10269, pp. 221–233.

[58] J.-P. Thirion, "Image matching as a diffusion process: an analogy with Maxwell's demons," *Med. Image Anal.*, vol. 2, no. 3, pp. 243–260, 1998.

[59] A. Cifor, L. Risser, D. Chung, E. M. Anderson, and J. A. Schnabel, "Hybrid feature-based diffeomorphic registration for tumor tracking in 2-D liver ultrasound images," *IEEE T. Med. Imaging*, vol. 32, no. 9, pp. 1647–1656, 2013.

[60] S. Lan, Z. Guo, and J. You, "Non-rigid medical image registration using image field in Demons algorithm," *Pattern Recogn. Lett.*, vol. 125, pp. 98–104, Jul. 2019.

[61] Y. Zhang, L. Zhang, L. E. Court, P. Balter, L. Dong, and J. Yang, "Tissue-specific deformable image registration using a spatial-contextual filter," *Comput. Med. Imag. Grap.*, vol. 88, p. 101849, 2021.

[62] H. Lombaert, Y. Sun, and F. Cheriet, "Landmark-based non-rigid registration via graph cuts," in *ICIAR 2007: Image Analysis and Recognition*, ser. LNCS. Berlin Heidelberg: Springer, 2007, vol. 4633, pp. 166–175.

[63] A. Cifor, L. Bai, and A. Pitiot, "Smoothness-guided 3-D reconstruction of 2-D histological images," *NeuroImage*, vol. 56, no. 1, pp. 197–211, 2011.

[64] S. Wirtz, N. Papenberg, B. Fischer, and O. Schmitt, "Robust and staining-invariant elastic registration of a series of images from histologic slices," in *Medical Imaging 2005: Image Processing*, ser. Proc. SPIE 5747, 2005, pp. 1256–1262.

[65] K. Becker, M. Stauber, F. Schwarz, and T. Beißbarth, "Automated 3D–2D registration of X-ray microcomputed tomography with histological sections for dental implants in bone using chamfer matching and simulated annealing," *Comput. Med. Imag. Grap.*, vol. 44, pp. 62–68, Sep. 2015.

[66] M. Tang, "Automatic registration and fast volume reconstruction from serial histology sections," *Comput. Vis. Image Und.*, vol. 115, no. 8, pp. 1112–1120, Aug. 2011.

[67] S. Gaffling, V. Daum, S. Steidl, A. Maier, H. Kostler, and J. Hornegger, "A Gauss-Seidel iteration scheme for reference-free 3-D histological image reconstruction," *IEEE T. Med. Imaging*, vol. 34, no. 2, pp. 514–530, Feb. 2015.

[68] B. Glocker, A. Sotiras, N. Komodakis, and N. Paragios, "Deformable medical image registration: Setting the state of the art with discrete methods," *Annu. Rev. Biomed. Eng.*, vol. 13, no. 1, pp. 219–244, Jul. 2011.

[69] M. Feuerstein, H. Heibel, J. Gardiazabal, N. Navab, and M. Groher, "Reconstruction of 3-D histology images by simultaneous deformable registration," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011*, ser. LNCS, G. Fichtinger, A. Martel, and T. Peters, Eds. Berlin, Heidelberg: Springer, 2011, pp. 582–589.

[70] B. W. Papież, J. Franklin, M. P. Heinrich, F. V. Gleeson, and J. A. Schnabel, "Liver motion estimation via locally adaptive over-segmentation regularization," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. LNCS, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer, 2015, vol. 9351, pp. 427–434.

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

[72] T. Kajihara, T. Funatomi, H. Makishima, T. Aoto, H. Kubo, S. Yamada, and Y. Mukaigawa, "Non-rigid registration of serial section images by blending transforms for 3D reconstruction," *Pattern Recogn.*, vol. 96, p. 106956, Dec. 2019.

[73] F. Guryanov and A. Krylov, "Fast medical image registration using bidirectional empirical mode decomposition," *Signal Process.-Image*, vol. 59, pp. 12–17, Nov. 2017.

[74] H.-H. Chang, G.-L. Wu, and M.-H. Chiang, "Remote sensing image registration based on modified SIFT and feature slope grouping," *IEEE Geosci. Remote S.*, vol. 16, no. 9, pp. 1363–1367, Sep. 2019.

[75] C. Nikou, F. Heitz, A. Nehlig, I. J. Namer, and J.-P. Armspach, "A robust statistics-based global energy function for the alignment of serially acquired autoradiographic sections," *Journal of Neuroscience Methods*, vol. 124, no. 1, pp. 93 – 102, 2003.

[76] J. P. Pluim, S. E. Muenzing, K. A. Eppenhof, and K. Murphy, "The truth is hard to make: Validation of medical image registration," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. Cancun: IEEE, Dec. 2016, pp. 2294–2300.

[77] S. Van Sint Jan, P. Salvia, I. Hilal, V. Sholukha, M. Rooze, and G. Clapworthy, "Registration of 6-DOFs electrogoniometry and CT medical imaging for 3D joint modeling," *J. Biomech.*, vol. 35, no. 11, pp. 1475–1484, Nov. 2002.

[78] A. Delaby, L. Espinosa, C. Lépolard, C. Capo, and J.-L. Mège, "3D reconstruction of granulomas from transmitted light images implemented for long-time microscope applications," *J. Immunol. Methods*, vol. 360, no. 1-2, pp. 10–19, Aug. 2010.

[79] R. Shojaii, T. Karavardanyan, M. Yaffe, and A. L. Martel, "Validation of histology image registration," in *Medical Imaging 2011: Image Processing*, ser. Proc. SPIE 7962, B. M. Dawant and D. R. Haynor, Eds. Lake Buena Vista, Florida: SPIE, Mar. 2011, p. 79621E.

[80] J. Kybic, "Fast no ground truth image registration accuracy evaluation: Comparison of bootstrap and Hessian approaches," in *2008 5th IEEE International*

Symposium on Biomedical Imaging: From Nano to Macro, May 2008, pp. 792–795.

[81] H. Xue, S. Shah, A. Greiser, C. Guetter, A. Littmann, M.-P. Jolly, A. E. Arai, S. Zuehlsdorff, J. Guehring, and P. Kellman, "Motion correction for myocardial T1 mapping using image registration with synthetic image estimation," *Magn. Reson. Med.*, vol. 67, no. 6, pp. 1644–1655, 2012.

[82] N. Duchateau, M. Sermesant, H. Delingette, and N. Ayache, "Model-based generation of large databases of cardiac images: Synthesis of pathological cine MR sequences from real healthy cases," *IEEE T. Med. Imaging*, vol. 37, no. 3, pp. 755–766, Mar. 2018.

[83] C. Grova, A. Biraben, J.-M. Scarabin, P. Jannin, I. Buvat, H. Benali, and B. Gibaud, "A methodology to validate MRI/SPECT registration methods using realistic simulated SPECT data," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2001*, ser. LNCS, W. J. Niessen and M. A. Viergever, Eds., vol. 2208. Berlin, Heidelberg: Springer, 2001, pp. 275–282.

[84] G. Hamarneh, P. Jassi, and L. Tang, "Simulation of ground-truth validation data via physically- and statistically-based warps," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008*, ser. LNCS, D. Metaxas, L. Axel, G. Fichtinger, and G. Székely, Eds., vol. 5241. Springer, 2008, pp. 459–467.

[85] G. Vlachopoulos, P. Korfiatis, S. Skiadopoulos, A. Kazantzi, C. Kalogeropoulou, I. Pratikakis, and L. Costaridou, "Selecting registration schemes in case of interstitial lung disease follow-up in CT: Registration schemes in ILD CT follow-up analysis," *Med. Phys.*, vol. 42, no. 8, pp. 4511–4525, Jul. 2015.

[86] X. Zhang, C. Gilliam, and T. Blu, "All-pass parametric image registration," *IEEE T. Image Process.*, vol. 29, pp. 5625–5640, 2020.

[87] K. Murphy, B. van Ginneken, J. Reinhardt, S. Kabus, K. Ding, X. Deng, K. Cao, K. Du, G. Christensen, V. Garcia, T. Vercauteren, N. Ayache, O. Commowick, G. Malandain, B. Glocker, N. Paragios, N. Navab, V. Gorbunova, J. Sporring, M. de Bruijne, X. Han, M. Heinrich, J. Schnabel, M. Jenkinson, C. Lorenz, M. Modat, J. McClelland, S. Ourselin, S. Muenzing, M. Viergever, D. De Nigris, D. Collins, T. Arbel, M. Peroni, R. Li, G. Sharp, A. Schmidt-Richberg, J. Ehrhardt, R. Werner, D. Smeets, D. Loeckx, G. Song, N. Tustison, B. Avants, J. Gee, M. Staring, S. Klein, B. Stoel, M. Urschler, M. Werlberger, J. Vandemeulebroucke, S. Rit, D. Sarrut, and J. Pluim, "Evaluation of registration methods on thoracic CT: The EMPIRE10 challenge," *IEEE T. Med. Imaging*, vol. 30, no. 11, pp. 1901–1920, 2011.

[88] J. Borovec, A. Munoz-Barrutia, and J. Kybic, "Benchmarking of image registration methods for differently stained histological slides," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct. 2018, pp. 3368–3372.

[89] Y. Song, D. Treanor, A. J. Bulpitt, N. Wijayathunga, N. Roberts, R. Wilcox, and D. R. Magee, "Unsupervised content classification based nonrigid registration of differently stained histology images," *IEEE T. Bio-Med. Eng.*, vol. 61, no. 1, pp. 96–108, 2014.

[90] 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," *Sci. Rep.*, vol. 7, no. 1, Dec. 2017.

[91] O. Lobachev, "The tempest in a cubic millimeter: Image-based refinements necessitate the reconstruction of 3D microvasculature from a large series of damaged alternately-stained histological sections," *IEEE Access*, vol. 8, pp. 13 489–13 506, 2020.

[92] M. Wodzinski and H. Müller, "DeepHistReg: Unsupervised deep learning registration framework for differently stained histology samples," *Comput. Meth. Prog. Bio.*, vol. 198, p. 105799, Jan. 2021.

[93] E. Bardinet, S. Ourselin, D. Dormont, G. Malandain, D. Tandé, K. Parain, N. Ayache, and J. Yelnik, "Co-registration of histological, optical and MR data of the human brain," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2002*, ser. LNCS, T. Dohi and R. Kikinis, Eds., vol. 2488. Berlin, Heidelberg: Springer, 2002, pp. 548–555.

[94] L. Tang, A. Hero, and G. Hamarneh, "Locally-adaptive similarity metric for deformable medical image registration," in *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, May 2012, pp. 728–731.

[95] M. Gehrung, M. Tomaszewski, D. McIntyre, J. Disselhorst, and S. Bohndiek, "Co-registration of optoacoustic tomography and magnetic resonance imaging data from murine tumour models," *Photoacoustics*, vol. 18, p. 100147, Jun. 2020. [Online]. Available: https://www.biorxiv.org/content/10.1101/636035v1

[96] M. A. Jacobs, J. P. Windham, H. Soltanian-Zadeh, D. J. Peck, and R. A. Knight, "Registration and warping of magnetic resonance images to histological sections," *Med. Phys.*, vol. 26, no. 8, pp. 1568–1578, 1999.

[97] A. du Bois d'Aische, M. D. Craene, X. Geets, V. Gregoire, B. Macq, and S. K. Warfield, "Efficient multi-modal dense field non-rigid registration: alignment of histological and section images," *Med. Image Anal.*, vol. 9, no. 6, pp. 538–546, Dec. 2005.

[98] B. Foster, R. Boutin, S. Henrichon, D. Noblett, C. Bayne, R. Szabo, A. Borowsky, and A. Chaudhari, "MRI – histopathology registration for osteoarthritis biomarker evaluation," *Osteoarthr. Cartilage*, vol. 25, pp. S229–S230, Apr. 2017.

[99] H. Guo, M. Kruger, S. Xu, B. J. Wood, and P. Yan, "Deep adaptive registration of multi-modal prostate images," *Comput. Med. Imag. Grap.*, vol. 84, p. 101769, 2020.

[100] J. Borovec, J. Kybic, I. Arganda-Carreras, D. V. Sorokin, G. Bueno, A. V. Khvostikov, S. Bakas, E. I.-C. Chang, S. Heldmann, K. Kartasalo, L. Latonen, J. Lotz, M. Noga, S. Pati, K. Punithakumar, P. Ruusuvuori, A. Skalski, N. Tahmasebi, M. Valkonen, L. Venet, Y. Wang, N. Weiss, M. Wodzinski, Y. Xiang, Y. Xu, Y. Yan, P. Yushkevic, S. Zhao, and A. Muñoz-Barrutia, "ANHIR: Automatic non-rigid histological image registration challenge," *IEEE T. Med. Imaging*, 2020.

[101] G. E. Christensen, X. Geng, J. G. Kuhl, J. Bruss, T. J. Grabowski, I. A. Pirwani, M. W. Vannier, J. S. Allen, and H. Damasio, "Introduction to the non-rigid image registration evaluation project (NIREP)," in *Biomedical Image Registration*, ser. LNCS, J. P. W. Pluim, B. Likar, and F. A. Gerritsen, Eds., vol. 4057. Berlin, Heidelberg: Springer, 2006, pp. 128–135.

[102] B. Pontré, B. R. Cowan, E. DiBella, S. Kulaseharan, D. Likhite, N. Noorman, L. Tautz, N. Tustison, G. Wollny, A. A. Young, and A. Suinesiaputra, "An open benchmark challenge for motion correction of myocardial perfusion MRI," *IEEE J. Biomed. Health.*, vol. 21, no. 5, pp. 1315–1326, Sep. 2017.

[103] K. K. Brock, "Results of a multi-institution deformable registration accuracy study (MIDRAS)," *Int. J. Radiat. Oncol.*, vol. 76, no. 2, pp. 583–596, Feb. 2010.

[104] J. West, J. M. Fitzpatrick, M. Y. Wang, B. M. Dawant, C. R. Maurer, R. M. Kessler, R. J. Maciunas, C. Barillot, D. Lemoine, A. Collignon, F. Maes, P. Suetens, D. Vandermeulen, P. A. van den Elsen, S. Napel, T. S. Sumanaweera, B. Harkness, P. F. Hemler, D. L. G. Hill, D. J. Hawkes, C. Studholme, J. B. A. Maintz, M. A. Viergever, G. Malandain, X. Pennec, M. E. Noz, G. Q. Maguire, M. Pollack, C. A. Pelizzari, R. A. Robb, D. Hanson, and R. P. Woods, "Comparison and evaluation of retrospective intermodality brain image registration techniques," *J. Comput. Assist. Tomo.*, vol. 21, no. 4, pp. 554–568, Jul. 1997.

[105] P. Hellier, C. Barillot, I. Corouge, B. Gibaud, G. Le Goualher, D. Collins, A. Evans, G. Malandain, N. Ayache, G. Christensen, and H. Johnson, "Retrospective evaluation of intersubject brain registration," *IEEE Trans. Med. Imaging*, vol. 22, no. 9, pp. 1120–1130, Sep. 2003.

[106] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, J. H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R. P. Woods, J. J. Mann, and R. V. Parsey, "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," *NeuroImage*, vol. 46, no. 3, pp. 786–802, Jul. 2009.

[107] 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 T. Med. Imaging*, vol. 33, no. 10, pp. 2039–2065, Oct. 2014.

[108] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *Int. J. Comput Vision*, vol. 24, no. 2, pp. 137–154, Sep. 1997.

[109] O. Déniz, D. Toomey, C. Conway, and G. Bueno, "Multi-stained whole slide image alignment in digital pathology," in *Medical Imaging 2015: Digital Pathology*, vol. 9420. SPIE, Mar. 2015, p. 94200Z.

[110] M. Polfliet, S. Klein, W. Huizinga, M. M. Paulides, W. J. Niessen, and J. Vandemeulebroucke, "Intra-subject multimodal groupwise registration with the conditional template entropy," *Med. Image Anal.*, vol. 46, pp. 15–25, May 2018.

[111] U. Bağci and L. Bai, "Automatic best reference slice selection for smooth volume reconstruction of a mouse brain from histological images," *IEEE T. Med. Imaging*, vol. 29, no. 9, pp. 1688–1696, 2010.

[112] N. D. Nanayakkara, B. Chiu, and A. Fenster, "A surface-based metric for registration error quantification," in *2009 International Conference on Industrial and Information Systems (ICIIS)*. Peradeniya, Sri Lanka: IEEE, Dec. 2009, pp. 349–353.

[113] J. Luo, S. Frisken, D. Wang, A. Golby, M. Sugiyama, and W. Wells III, "Are registration uncertainty and error monotonically associated?" in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, ser. LNCS, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds., vol. 12263. Springer, Sep. 2020, pp. 264–274. [Online]. Available: http://arxiv.org/abs/1908.07709

[114] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytol.*, vol. 11, no. 2, pp. 37–50, 1912.

[115] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE T. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[116] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[117] W. Crum, O. Camara, and D. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE T. Med. Imaging*, vol. 25, no. 11, pp. 1451–1461, Nov. 2006.

[118] T. Rohlfing, "Image similarity and tissue overlaps as surrogates for image registration accuracy: Widely used but unreliable," *IEEE T. Med. Imaging*, vol. 31, no. 2, pp. 153–163, Feb. 2012.

[119] J. Krebs, T. Mansi, H. Delingette, L. Zhang, F. C. Ghesu, S. Miao, A. K. Maier, N. Ayache, R. Liao,

and A. Kamen, "Robust non-rigid registration through agent-based action learning," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, ser. LNCS, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham: Springer, 2017, vol. 10433, pp. 344–352.

[120] A. Maier, C. Syben, T. Lasser, and C. Riess, "A gentle introduction to deep learning in medical image processing," *Z. Med. Phys.*, vol. 29, no. 2, pp. 86–101, May 2019.

[121] L. Li, V. A. Zimmer, W. Ding, F. Wu, L. Huang, J. A. Schnabel, and X. Zhuang, "Random style transfer based domain generalization networks integrating shape and spatial information," in *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, ser. LNCS, E. Puyol Anton, M. Pop, M. Sermesant, V. Campello, A. Lalande, K. Lekadir, A. Suinesiaputra, O. Camara, and A. Young, Eds., vol. 12592. Springer, Jan. 2021, pp. 208–218. [Online]. Available: http://arxiv.org/abs/2008.12205

[122] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "Deep learning in medical image registration: a review," *Phys. Med. Biol.*, 2020. [Online]. Available: https://arxiv.org/abs/1912.12318

[123] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Mach. Vision. Appl.*, vol. 31, no. 1, p. 8, Jan. 2020. [Online]. Available: http://arxiv.org/abs/1903.02026

[124] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, ser. LNCS, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds., vol. 11070. Cham: Springer, 2018, pp. 729–738. [Online]. Available: https://arxiv.org/abs/1805.04605

[125] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp. 4938–4947. [Online]. Available: https://arxiv.org/abs/1911.11763

[126] J. P. Schneider, J. Hegermann, and C. Wrede, "Volume electron microscopy: analyzing the lung," vol. 155, no. 2, pp. 241–260, 2021.

[127] C. Mühlfeld, C. Wrede, L. Knudsen, T. Buchacker, M. Ochs, and R. Grothausmann, "Recent developments in 3-D reconstruction and stereology to study the pulmonary vasculature," *Am. J. Physiol. Lung Cell Mol. Physiol.*, vol. 315, no. 2, pp. L173–L183, 2018.

[128] T. M. Mayhew, C. Mühlfeld, D. Vanhecke, and M. Ochs, "A review of recent methods for efficiently quantifying immunogold and other nanoparticles using TEM sections through cells, tissues and organs," *Ann. Anat.*, vol. 191, no. 2, pp. 153–170, Apr. 2009.

[129] R. Grothausmann, L. Knudsen, M. Ochs, and C. Mühlfeld, "Digital 3D reconstructions using histological serial sections of lung tissue including the alveolar capillary network," *Am. J. Resp. Cell Mol.*, vol. 312, no. 2, pp. L243–L257, Dec. 2016.

[130] M. Ochs, "The closer we look the more we see? Quantitative microscopic analysis of the pulmonary surfactant system," *Cell. Physiol. Biochem.*, vol. 25, no. 001, pp. 027–040, 2010.

[131] M. Ochs, L. Knudsen, J. Hegermann, C. Wrede, R. Grothausmann, and C. Mühlfeld, "Using electron microscopes to look into the lung," *Histochem. Cell. Biol.*, vol. 146, no. 6, pp. 695–707, Dec. 2016.

[132] J. P. Schneider, C. Wrede, J. Hegermann, E. R. Weibel, C. Mühlfeld, and M. Ochs, "On the topological complexity of human alveolar epithelial type 1 cells," *Am. J. Respir. Crit. Care Med.*, vol. 199, no. 9, pp. 1153–1156, May 2019.

[133] C. Mühlfeld, C. Wrede, V. Molnár, A. Rajces, and C. Brandenberger, "The plate body: 3d ultrastructure of a facultative organelle of alveolar epithelial type II cells involved in SP-a trafficking," *Histochem. Cell Biol.*, vol. 155, no. 2, pp. 261–269, 2021.

[134] J. D. Woodward and J. N. Maina, "Study of the structure of the air and blood capillaries of the gas exchange tissue of the avian lung by serial section three-dimensional reconstruction," *J. Microsc.-Oxford*, vol. 230, no. 1, pp. 84–93, 2008.

[135] C. Mühlfeld, R. Grothausmann, and M. Ochs, "Visualization and quantitative analysis of the alveolar capillary network – implications for lung developmental biology," *Eur. Respir. J.*, vol. 50, no. suppl 61, p. PA4189, Sep. 2017.

[136] S. A. K. Pentinga, K. Kwan, S. A. Mattonen, C. Johnson, A. Louie, M. Landis, R. Inculet, R. Malthaner, D. Fortin, G. Rodrigues, B. Yaremko, D. A. Palma, and A. D. Ward, "3D human lung histology reconstruction and registration to in vivo imaging," in *Medical Imaging 2018: Digital Pathology*, vol. 10581. SPIE, Mar. 2018, p. 105810V.

[137] T. Schormann, A. Dabringhaus, and K. Zilles, "Statistics of deformations in histology and application to improved alignment with MRI," *IEEE T. Med. Imaging*, vol. 14, no. 1, pp. 25–35, Mar. 1995.

[138] M. Matsumoto and T. Nishimura, "Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator," *ACM Trans. Model. Comput. Simul.*, vol. 8, no. 1, pp. 3–30, Jan. 1998.

[139] The ImageMagick Development Team, "Imagemagick," Jan. 2021. [Online]. Available: https://imagemagick.org

[140] J. Fehrenbach and J.-M. Mirebeau, "Sparse non-negative stencils for anisotropic diffusion," *J. Math.*

Imaging Vis., vol. 49, no. 1, pp. 123–147, May 2014.

[141] R. Grothausmann, J. Labode, P. Hernandez-Cerdan, D. Haberthür, R. Hlushchuk, O. Lobachev, C. Brandenberger, A. G. Gie, T. Salaets, J. Toelen, W. L. Wagner, and C. Mühlfeld, "Combination of µCT and light microscopy for generation-specific stereological analysis of pulmonary arterial branches: a proof-of-concept study," *Histochemistry and Cell Biology*, vol. 155, no. 2, pp. 227–239, Feb. 2021.

[142] C. Mühlfeld, H. Schulte, J. C. Jansing, C. Casiraghi, F. Ricci, C. Catozzi, M. Ochs, F. Salomone, and C. Brandenberger, "Design-based stereology of the lung in the hyperoxic preterm rabbit model of bronchopulmonary dysplasia," *Oxid. Med. Cell. Longev.*, vol. 2021, p. e4293279, Oct. 2021.

[143] S. V. Appuhn, S. Siebert, D. Myti, C. Wrede, D. E. Surate Solaligue, D. Pérez-Bravo, C. Brandenberger, J. Schipke, R. E. Morty, R. Grothausmann, and C. Mühlfeld, "Capillary changes precede disordered alveolarization in a mouse model of bronchopulmonary dysplasia," *Am. J. Respir. Cell. Mol. Biol.*, Mar. 2021.

[144] T. Buchacker, C. Mühlfeld, C. Wrede, W. L. Wagner, R. Beare, M. McCormick, and R. Grothausmann, "Assessment of the alveolar capillary network in the postnatal mouse lung in 3D using serial block-face scanning electron microscopy," *Front. Physiol.*, vol. 10, p. 1357, 2019.

[145] J.-M. Krischer, K. Albert, A. Pfaffenroth, E. Lopez-Rodriguez, C. Ruppert, B. J. Smith, and L. Knudsen, "Mechanical ventilation-induced alterations of intracellular surfactant pool and blood–gas barrier in healthy and pre-injured lungs," *Histochem. Cell. Biol.*, vol. 155, no. 2, pp. 183–202, 2021.

[146] O. Tange, "GNU parallel," Sep. 2021. [Online]. Available: https://zenodo.org/record/5523272

[147] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE T. Biomed. Eng.*, vol. 61, no. 6, pp. 1729–1738, 2014.

[148] S. Nadeem, T. Hollmann, and A. Tannenbaum, "Multimarginal Wasserstein barycenter for stain normalization and augmentation," *arXiv:2006.14566 [cs, eess]*, Jun. 2020, arXiv: 2006.14566. [Online]. Available: http://arxiv.org/abs/2006.14566

[149] P. Hanslovsky, J. A. Bogovic, and S. Saalfeld, "Post-acquisition image based compensation for thickness variation in microscopy section series," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. IEEE, Apr. 2015, pp. 507–511.

[150] D. Bulmer, "Observations on histological methods involving the use of phosphotungstic and phosphomolybdic acids, with particular reference to staining with phosphotungstic acid/haematoxylin," *J. Cell. Sci.*, vol. s3-103, no. 63, pp. 311–323, Sep. 1962.

[151] Y. Lanir, J. Walsh, and R. W. Soutas-Little, "Histological staining as a measure of stress in collagen fibers," *J. Biomech. Eng.*, vol. 106, no. 2, pp. 174–176, May 1984.

[152] C. Dullin, R. Ufartes, E. Larsson, S. Martin, M. Lazzarini, G. Tromba, J. Missbach-Guentner, D. Pinkert-Leetsch, D. M. Katschinski, and F. Alves, "μCT of ex-vivo stained mouse hearts and embryos enables a precise match between 3D virtual histology, classical histology and immunochemistry," *PLOS ONE*, vol. 12, no. 2, p. e0170597, Feb. 2017.

[153] Y. Xu, J. G. Pickering, Z. Nong, E. Gibson, J.-M. Arpino, H. Yin, and A. D. Ward, "A method for 3D histopathology reconstruction supporting mouse microvasculature analysis," *PLoS ONE*, vol. 10, no. 5, p. e0126817, 2015.

[154] H. O. Aggrawal, M. S. Andersen, and J. Modersitzki, "An image registration framework for discontinuous mappings along cracks," in *Biomedical Image Registration*, ser. LNCS, Ž. Špiclin, J. McClelland, J. Kybic, and O. Goksel, Eds., vol. 12120. Cham: Springer, 2020, pp. 163–173.

[155] E. Ng and M. Ebrahimi, "An unsupervised learning approach to discontinuity-preserving image registration," in *Biomedical Image Registration*, ser. LNCS, Ž. Špiclin, J. McClelland, J. Kybic, and O. Goksel, Eds., vol. 12120. Cham: Springer, 2020, pp. 153–162.

[156] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2414–2423. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2016/html/Gatys_Image_Style_Transfer_CVPR_2016_paper.html

[157] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, "Staingan: Stain style transfer for digital histological images," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Apr. 2019, pp. 953–956.

[158] H. Liang, K. N. Plataniotis, and X. Li, "Stain style transfer of histopathology images via structure-preserved generative learning," *arXiv:2007.12578 [cs, eess]*, Jul. 2020. [Online]. Available: http://arxiv.org/abs/2007.12578

[159] A. Khan, M. Atzori, S. Otálora, V. Andrearczyk, and H. Müller, "Generalizing convolution neural networks on stain color heterogeneous data for computational pathology," in *Medical Imaging 2020: Digital Pathology*, vol. 11320. SPIE, Mar. 2020, p. 113200R.

[160] M. Meyer, "Continuous integration and its tools," *IEEE Software*, vol. 31, no. 3, pp. 14–16, Mar. 2014.