Evaluating registrations of serial sections with distortions of the ground truths. Supplementary material

Oleg Lobachev[a,d,*], Takuya Funatomi[e], Alexander Pfaffenroth[a], Reinhold Förster[b], Lars Knudsen[a], Christoph Wrede[c], Michael Guthe[f], David Haberthür[g], Ruslan Hlushchuk[g], Thomas Salaets[h], Jaan Toelen[h], Simone Gaffling[i], Christian Mühlfeld[a], Roman Grothausmann[a,j]

[a]*Hannover Medical School, OE 4120, Carl-Neuberg-Straße 1, 30625 Hannover, Germany*
[b]*Hannover Medical School, OE 5240, Carl-Neuberg-Straße 1, 30625 Hannover, Germany*
[c]*Hannover Medical School, OE 8840, Carl-Neuberg-Straße 1, 30625 Hannover, Germany*
[d]*Leibniz-Fachhochschule School of Business, Expo Plaza 11, 30539 Hannover, Germany*
[e]*Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan*
[f]*University of Bayreuth, 95440 Bayreuth, Germany*
[g]*University of Bern, Hochschulstrasse 6, 3012 Bern, Switzerland*
[h]*KU Leuven, Herestraat 49, 3000 Leuven, Belgium*
[i]*Friedrich-Alexander-Universität Erlangen-Nürnberg, Martensstr. 3, 91058 Erlangen, Germany*
[j]*HAWK University of Applied Sciences and Art, Hohnsen 4, 31134 Hildesheim, Germany*

Abstract

This is the supplementary material for the paper "Registration of serial sections: An evaluation method based on distortions of the ground truths". This document includes a pair-wise evaluation of the registrations.

Keywords: registration, ground truth

1. Feature matching

Figure 1 shows selected matches from our rigid feature-based registration with SURF (Lobachev et al., 2017). The input images are globally transformed and locally distorted consecutive images from LS data set (see also Krischer et al., 2021, for details on the biological acquisition).

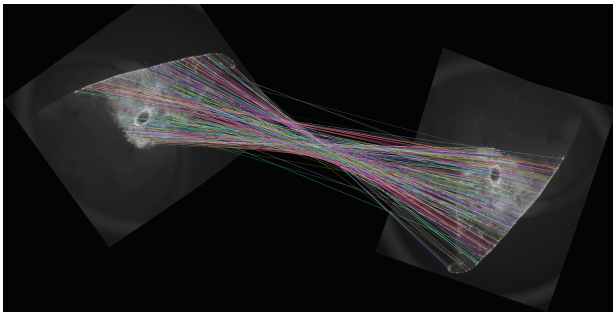


Figure 1: Feature matches for affine transform estimation.

2. Details of the EM acquisition

Our EM data set was acquired with the SBF-SEM technique (Buchacker et al., 2019). Notice that although cutting with diamond knife might induce some non-linear distortions in the block face, in a typical SBF-SEM acquisition

*Corresponding author
Email address: oleg.lobachev@leibniz-fh.de (Oleg Lobachev)
URL: https://orcid.org/0000-0002-7193-6258 (Oleg Lobachev), https://orcid.org/0000-0001-5588-5932 (Takuya Funatomi), https://orcid.org/0000-0003-3388-9187 (David Haberthür), https://orcid.org/0000-0002-6722-8996 (Ruslan Hlushchuk), http://www5.cs.fau.de/en/our-team/gaffling-simone/ (Simone Gaffling), https://orcid.org/0000-0001-5550-4239 (Roman Grothausmann)

in variable pressure mode at most the translations are to be corrected in post-processing. This alignment is often required due to movement of the whole block face in the field of view. Non-linear distortions are minimized in SBF-SEM in contrast to, e. g., serial sectioning TEM or array tomography. In SBF-SEM no distortions or foldings, but also no rotations are to be expected. During the acquisition of the data, which we use as an input in this work, *no* rotation or non-rigid alignment, but also specifically *no translation* was applied.

3. Evaluation measures

3.1. Full-size evaluation

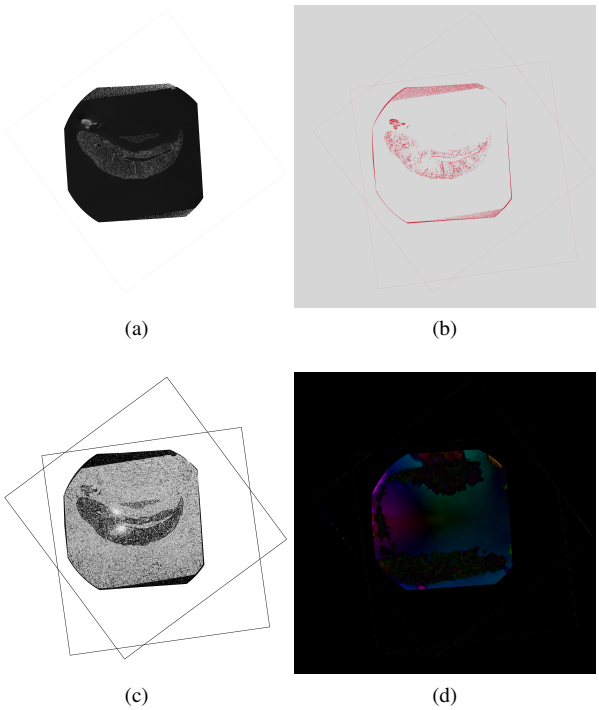
Figure 2 shows the uncropped image-based evaluation results on CT data set (Grothausmann et al., 2021, Appuhn et al. (2021)), registered with "Blending" method (Kajihara et al., 2019). Notice the border effects. Fig. 3 compares optical flow visualization, computed on the full image (similar to 2d), then cropped to middle and optical flow visualization from cropped images.

In the main text, we use SSIM (Wang et al., 2004) values from crops. We crop SSIM visualizations, as detailed below. We use Jaccard measures from crops. In contrast to the main paper, the optical flow visualization (Farnebäck, 2003) in Fig. 3a is computed from the full images.

3.2. Computation of the measures in general

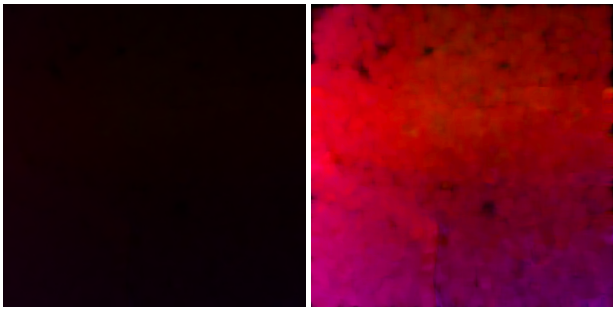
A lot of fine details impact the metrics and the visuals. In the main paper, we used 500×500 pixels crops. Due to its nature, SSIM on crops is a lesser image. We computed SSIM on the full images and cropped then. This does not change the visual distribution of the measure. SSIM values used in the numerical comparisons were computed on crops. This way, we do not let the border effects from padding impact the evaluation.

Our implementation of optical flow visualization normalizes the sizes of the displacements across the image. This leads to very boring images in the case first the flow is computed and then the visualization is cropped. Here, we computed the optical flow on crops. Arguably, different images are less comparable, when computed in this manner. However, "hot spots" and movement directions are more prominent when visualized this way.



(a) (b) (c) (d)

Figure 2: CT data set, "Blending" method. Here image-based evaluation methods of the full images are shown. (a): Registered image. (b): PSNR visualization. (c): SSIM visualization. (d): optical flow visualization.



(a) (b)

Figure 3: CT data set, an older result from the "Blending" method. (a): First optical flow then crop. (b): First crop, then optical flow. Notice the magnitude differences in the visualization.

We computed PSNR for the visuals on color images, even though all our result images were inherently grayscale.

The reason is a slightly nicer visualization. However, the PSNR values on such "fake" RGB images are misleading. The numerical PSNR values were computed on grayscale images.

The "original" ground truth images for EM are 16 bit. We evaluate them, too, but consider the 8 bit converted (and normalized) ground truth images to be a better reference value for the evaluation of registrations.

3.3. Thresholding and Dice measure visualizations

While the Dice measure can be trivially expressed in terms of the Jaccard measure, both of them use binary images as inputs. We use a simple threshold for Jaccard measure; we always state the threshold in the description of the Jaccard measure. In this case we use a global threshold. In the main paper, in the visualization of Dice measure we use Otsu's method (Otsu, 1979) on blurred images, as implemented in OpenCV (Bradski, 2000; Kaehler and Bradski, 2014).

This section explains our rationale behind the above design decision visually. In our Dice measure visualizations below, the binary images were produced with multiple methods:

- global binary threshold;
- Otsu method (Otsu, 1979);
- Otsu method in blurred images;
- with Matlab's activecontour function at 300 iterations, starting from a binary thresholded image.

To give an example, consider Fig. 4 that shows a region from LS data set. The masks, such as (b), were computed using the activecontour function in Matlab with the threshold 150. We see that the Dice measure in (d) roughly midway between (c) and (e). We cannot hope to reach values similar to (e) with the method used, as no non-rigid alignment happens in this case. Thus, the result of the rigid alignment is acceptable. It manages to align the sections rigidly, despite additional non-rigid distortions induced by our generation of input data.

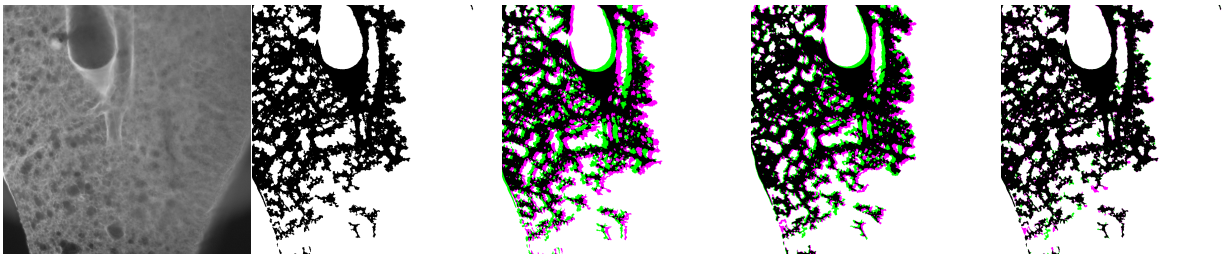
Figure 5 shows the same region as Fig. 4, the Rigid-SIFT method (d). Here, the same image pair is evaluated using different thresholding methods. We conclude that Otsu's method on blurred images produces best visualizations, although little difference is seen to Otsu without blur. For our Dice measure visualizations in the main text we use Otsu's method on blurred images. For Jaccard evaluation in the main text, we use the global thresholding for its consistency. One of the reasons for these decisions: Our input images are individually normalized. (See Section 6.) Even if they are normalized back to a "common denominator", some differences may remain. We also use the input images (with all their discrepancies) in the evaluation.

Table 1: Numerical values for the evaluation of the registrations on the CT data set, in the middle of the series (Figs. 6, 7). In Jaccard measure, we used a global threshold of 100. Low quality values in blending transforms can be explained by residual shifts. Notice the differences between the rigid-only and non-rigid methods. Notice also the lower values in the ground truth, when compared to elastix. The "Locally distorted" line shows the local distortions only, without global rigid transforms. There was some larger movement (as seen in Fig. 7), this fact affects the measures. The registrations' inputs, however, were also rigidly transformed beforehand.
The larger the values, the better—up to the ground truth value.

	CT		
Method	Jaccard	PSNR	SSIM
Rigid-SIFT	0.126	22.2	0.476
Rigid-SURF	0.135	22.7	0.508
Deform-SURF	0.375	28.5	0.772
GS	0.443	30.9	0.874
Blending	0.123	22.3	0.466
elastix	0.567	33.0	0.898
Locally distorted	0.0283	19.9	0.394
Ground truth	0.517	31.4	0.875

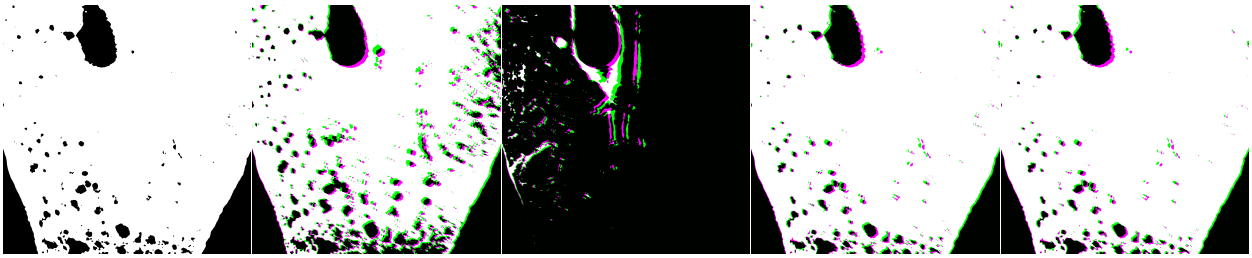
Notice, that the meaning of colors white and black in Figures 4 and 5 is different. We argue that using white for the overlap is better.

We use two consecutive not distorted images from the middle of each series as a ground truth for the evaluation of multiple registration methods. We also use two consecutive images at the same positioning from the registrations' results. (An evaluation of the full series is in the main text.) The input of the registrations is the series in its completeness after the application of distortions presented in this paper. Both local distortions and rigid transformations were applied. Two consecutive result images from the middle of the registered series are subjected to the quality measures Jaccard, PSNR and SSIM, as outlined in the main text. We also utilize two consecutive locally distorted images to produce comparison values. In the image-based



(a) Original data (b) Mask (c) Dice measure = 0.887, locally distorted data, rigid transform to follow (d) Dice measure = 0.917, after Rigid-SIFT (e) Dice measure = 0.983, ground truth

Figure 4: An example of an evaluation based on Dice measure in Matlab. (a): The region of interest, which we evaluate. It is a rat lung, obtained with LS microscopy. Using Matlab's activecontour function, the mask (b) is computed with threshold 150. (c): The Dice measure after the local distortions from our method, but before rigid transform. The rigidly transformed series serves then as an input to the registration. (d): Dice measure on the result of Rigid-SIFT method. No non-rigid steps are performed. The original, undistorted data is in (e). The higher the Dice measure is the better.



(a) Otsu+blur threshold (b) Dice on global 100 (c) Dice on global 150 (d) Dice on Otsu w/o blur (e) Dice on Otsu with blur

Figure 5: An example of an evaluation based on Dice measure in Python. The region of interest, which we evaluate, same as above. The mask (a) is computed with Otsu's method on blurred images, using OpenCV. All the measures (c)–(e) are computed for an image pair, registered with Rigid-SIFT. In contrast to Matlab, here is white the common area in both thresholded images. (b): The Dice measure visualization on global threshold of 100. (c): The Dice measure visualization on global threshold of 150. (d): The Dice measure visualization on Otsu's thresholding without a blur. (e): The Dice measure visualization on Otsu's thresholding on blurred images with blur radius 5. The higher the overlap in Dice measure the better.

evaluation we show a 500×500 pixels crop from the image center to highlight details.

The tables below show the Jaccard, PSNR and SSIM values for the ground truth (undistorted images), as well as the locally deformed images. In part, the values for the normalized ground truth images were also given, to give an idea about the impact of intensity variations on the ground truth measures. Same intensity variations as in "normalized ground truths" were also present in registrations' inputs.

Hence, the effect of the non-rigid registrations can be read as an improvement between "locally distorted" and "normalized ground truths", but there was also a global rigid transformation. Undoing it with a rigid registration might have lead to a worse starting point for the non-rigid registration than "locally distorted".

4. Results of the pair-wise evaluation

Basically, Figs. 6 and 7 show the pair-wise image-based evaluations for the CT data set; Figs. 8 and 9 show the pair-wise evaluation for the EM data set; Figs. 10 and 11 show the pair-wise evaluation for the LS data set. Table 1 shows the numerical values of the pair-wise image-based

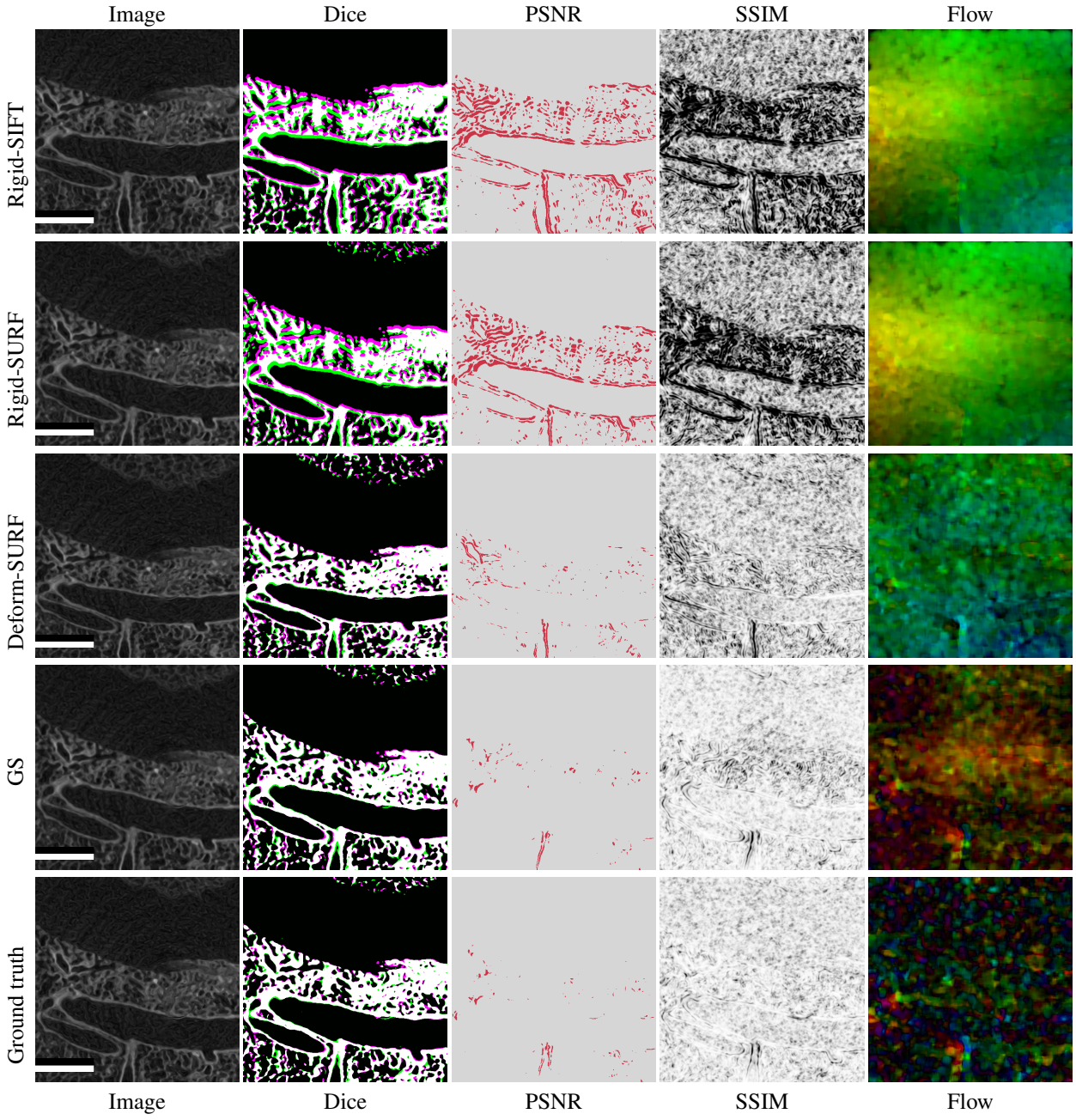


Figure 6: Evaluation of CT data set with image measures, part 1. Continued in Fig. 7. All scale bars are 1 mm.

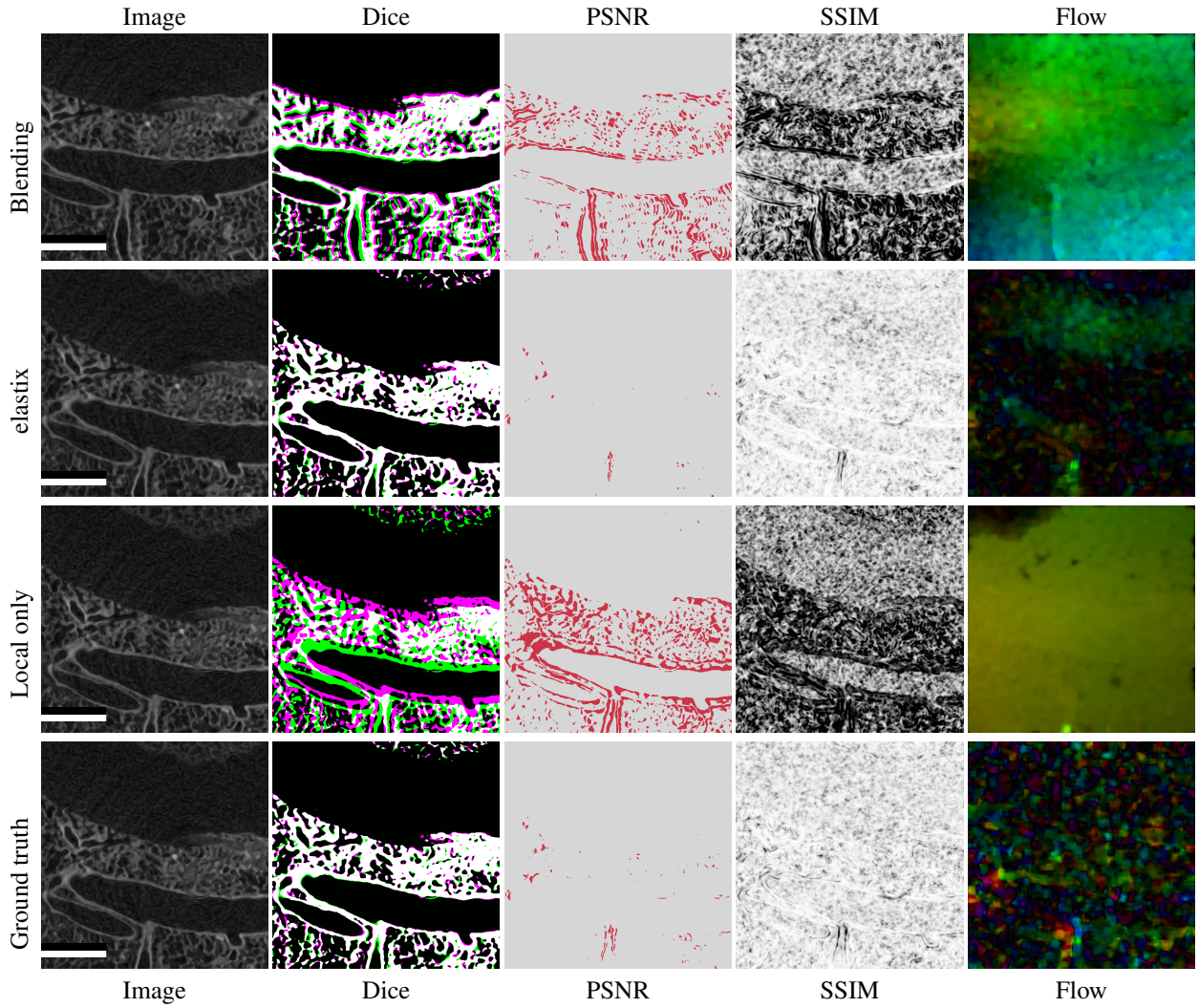


Figure 7: Evaluation of CT data set with image measures, part 2. Continued from Fig. 6. "Local only" means the distorted ground truth, but with no global transforms applied. The registrations' input had global transforms applied. All scale bars are 1 mm.

metrics for CT; Table 2 shows these for EM; Table 3 for LS. The full-series evaluation is in the main text.

4.1. CT data set

Consider Fig. 6. Both rigid-only methods leave some incongruences, these are then greatly reduced by "DeformSURF" (Lobachev et al., 2017). "GS" (Gaffling et al., 2015) also works on rigidly pre-registered input, but produces a very different image of residual movements in optical flow visualization. It reaches also very good values with respect to other measures, e. g., SSIM. The elastix-based registration (Fig. 7) reaches even better values with respect to SSIM and PSNR. We see more movement in Dice visualization, but it is not reflected in Jaccard measure (Tab. 1). The apparent reason might be the varying thresholding methods. "Blending" has some residual movement, as seen in PSNR and optical flow visualization. The goal

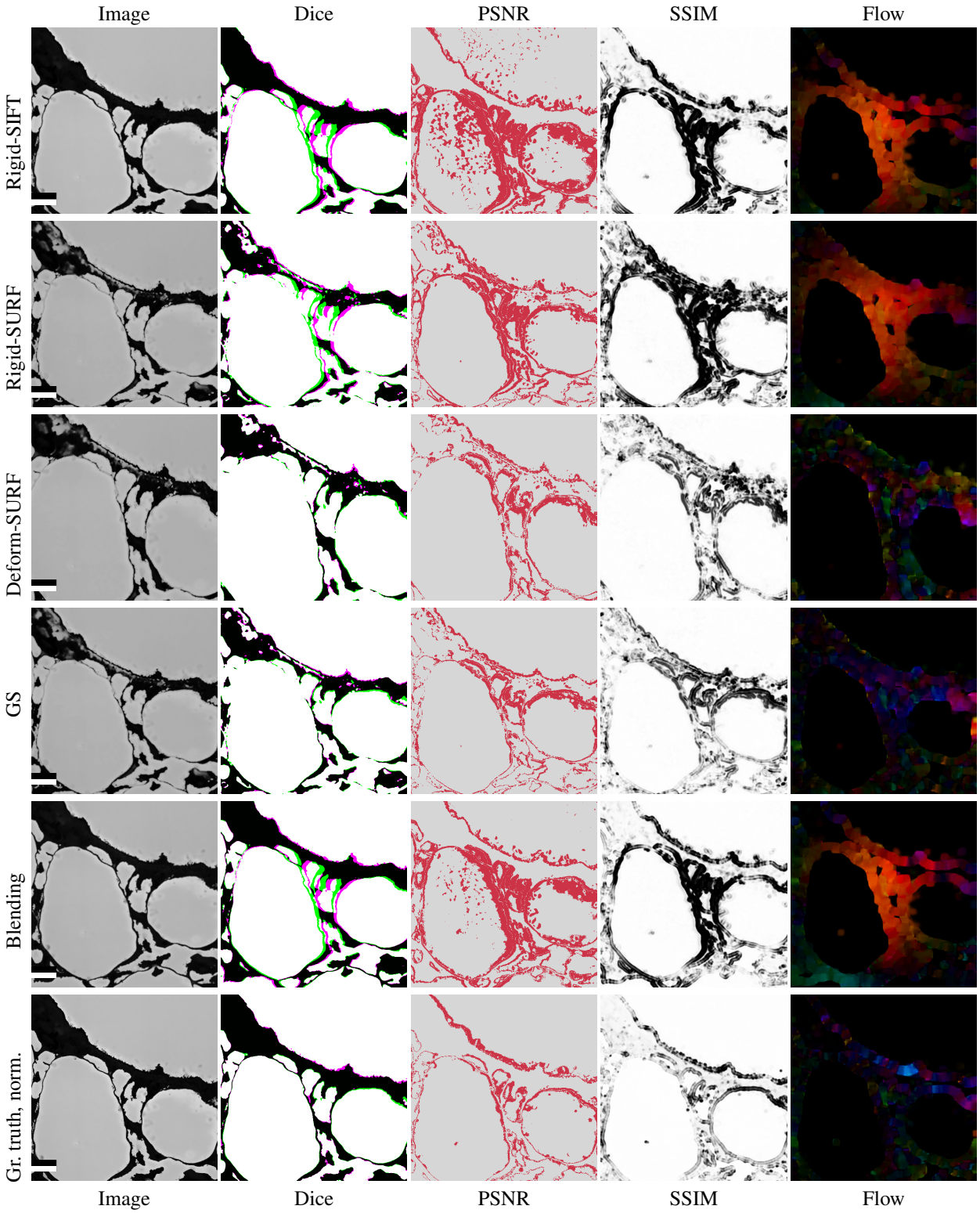


Figure 8: Evaluation of EM data set with image measures, part 1. Continued in Fig. 9. All scale bars are 10 µm.

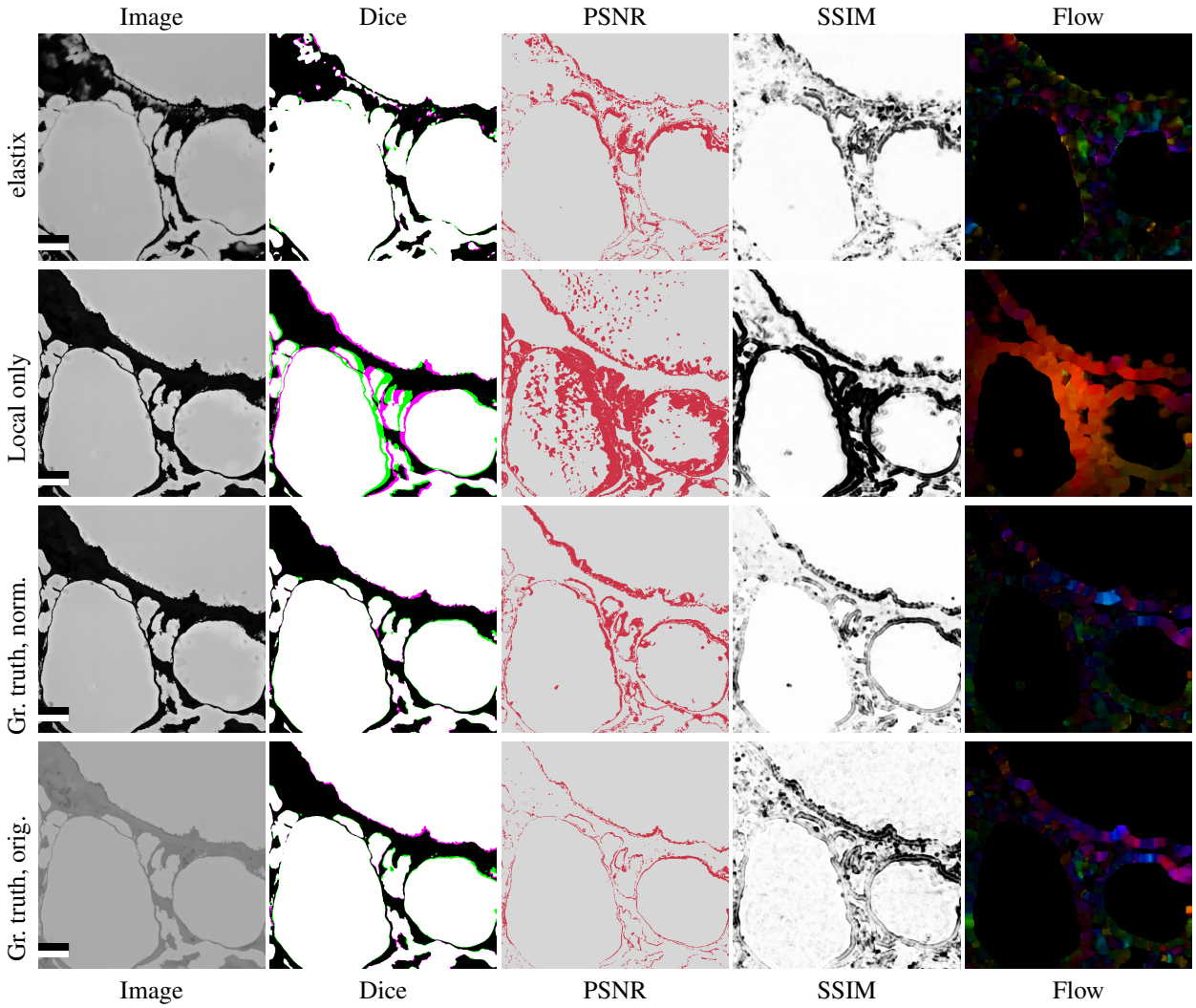


Figure 9: Evaluation of EM data set with image measures, part 2. Continued from Fig. 8. All scale bars are 10 µm.

of "Blending" was a coarse alignment, hence the lower quality measures. Notice, however, that it copes with the random global transformations on its own, without the initial "Rigid-SURF" stage. In "local only" we see some global movement with is a residue of a local distortion at a larger scale. Not surprisingly, there are still some residues of it left in rigid methods, as discussed above.

Consider the ground truth as the "ideal", fully corresponding image pair. Small signal in PSNR means little change. It is also visible in the visualizations as less red. SSIM and optical flow visualizations show that these differences are local and evenly distributed across the region. In registered images there are typically more differences; those are also more concentrated in a region. Such concentrated "change" means stronger local transformations.

The visualized differences in the ground truth appear to be less than in most of the registration results (all in Fig. 6)

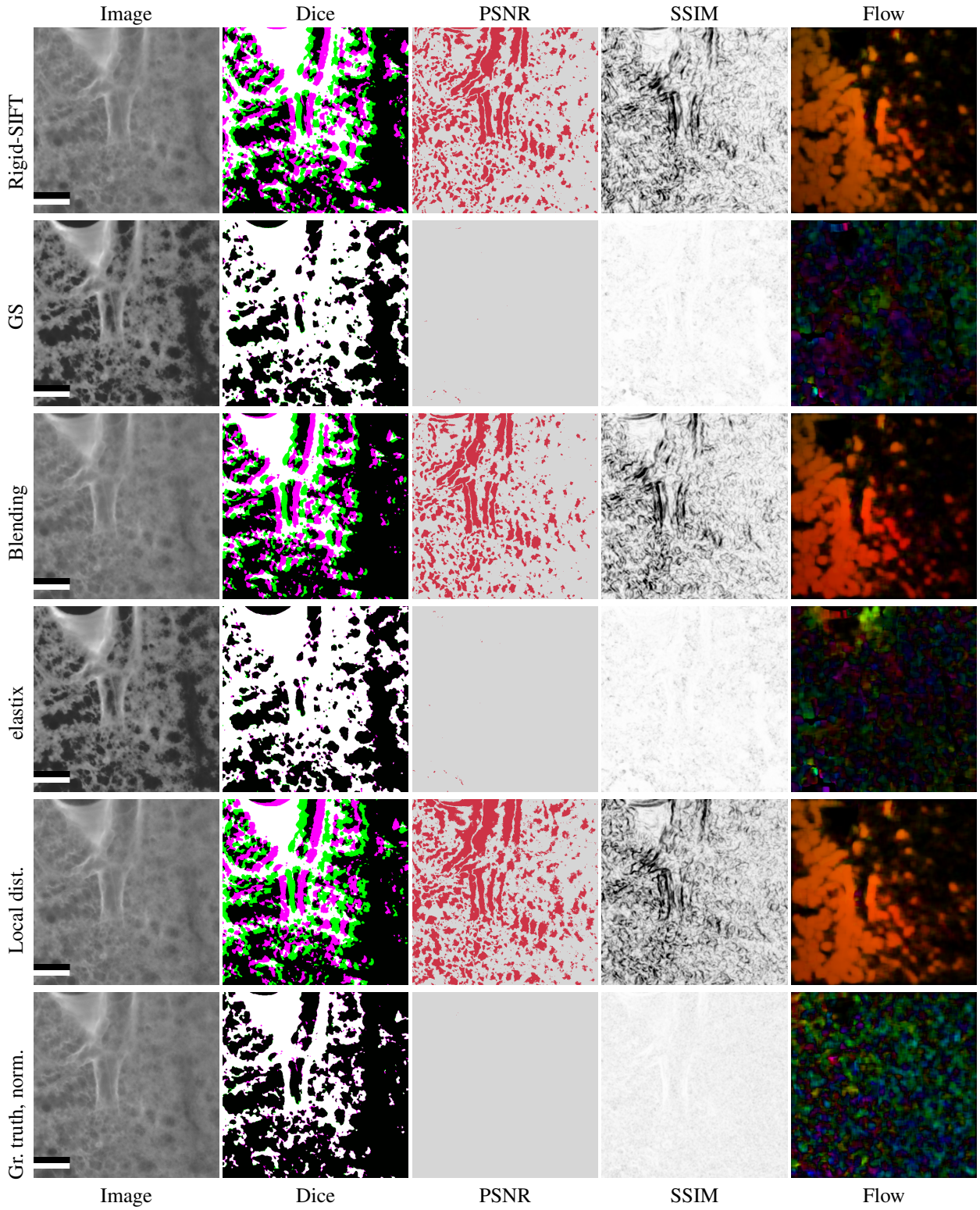


Figure 10: Evaluation of LS data set with image measures, part 1. The ground truth is normalized. Continued in Fig. 11. All scale bars are 500 µm.

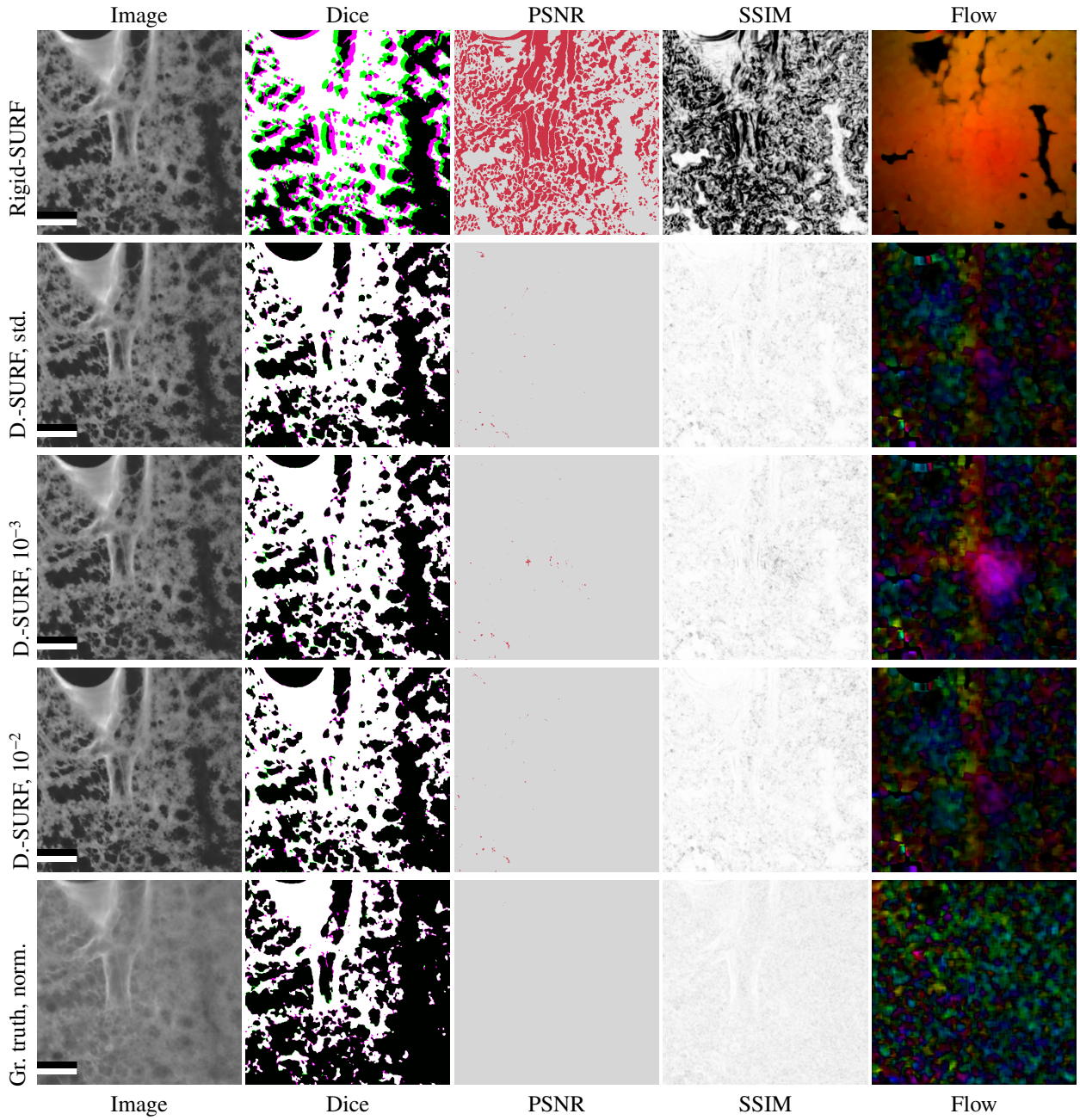


Figure 11: Evaluation of LS data set with image measures, part 2, mostly variations over "Deform-SURF", abbreviated here as "D.-SURF". We differentiate the "stretch" parameter of the non-rigid phase. The default value is the factor $5 \cdot 10^{-3}$ of the image size. We also test "Deform-SURF" with the values 10^{-3} and 10^{-2}. The ground truth is normalized. Continued from Fig. 10. All scale bars are 500 µm.

Table 2: Numerical values for the evaluation of the registrations on the EM data set, in the middle of the series (Figs. 8, 9). In Jaccard measure, we use global thresholds of 100 and 150, as indicated with "J 100" and "J 150". "Locally distorted" means images resulting from our method applied to normalized ground truth for *local* distortions, but no global rigid transformations were used in this case. The actual registrations operate on images that were also rigidly transformed. Notice the differences between the rigid-only and non-rigid methods. Notice also the differences between both rigid versions as well as the impact of normalization and 8 bit discretization on the ground truth values.
The larger the values, the better—up to the ground truth value.

EM

Method	J 100	J 150	PSNR	SSIM
Rigid-SIFT	0.928	0.923	15.6	0.813
Rigid-SURF	0.925	0.925	16.5	0.771
Deform-SURF	0.975	0.976	22.4	0.885
GS	0.964	0.967	21.8	0.885
Blending	0.925	0.919	16.3	0.816
elastix	0.979	0.981	23.4	0.900
Locally distorted	0.910	0.906	13.9	0.772
Ground truth, norm'ed	0.974	0.974	21.8	0.925
Ground truth, original	1.000	0.982	37.1	0.869

Table 3: Numerical values for the evaluation of the registrations on the LS data set, in the middle of the series (Figs. 10, 11.). In Jaccard measure, we used a global threshold of 100. Low values in "Blending" and in rigid-only methods can be explained by residual shifts. Notice the improvements in non-rigid methods. The shorthand s stands for "stretch". The larger the values, the better—up to the ground truth value.

LS

Method	Jaccard	PSNR	SSIM
Rigid-SIFT	0.929	24.9	0.759
Rigid-SURF	0.599	19.0	0.479
Deform-SURF, $s = 10^{-3}$	0.930	36.0	0.962
Deform-SURF, $s = 5 \cdot 10^{-3}$	0.933	36.4	0.965
Deform-SURF, $s = 10^{-2}$	0.933	36.4	0.964
GS	0.940	37.1	0.973
Blending	0.928	25.3	0.777
elastix	0.941	37.2	0.968
Locally distorted	0.909	23.3	0.710
Ground truth, norm'ed	0.981	40.3	0.946

and also more scattered across the image: compare, e. g., the optical flow visualization between ground truth, "GS", and "Deform-SURF". As for elastix (Klein et al., 2010; Shamonin et al., 2014), there are *less* differences between two consecutive images than in the ground truth. The SSIM and optical flow visualization of elastix results show the differing parts to be more concentrated in the certain areas (e. g., the bottom part and the first third part of the crop) when compared to the same areas of the ground truth. Such a behavior is hinting at overfitting. In the ground truth the "change" is more evenly distributed across the image. (We can easily make statements about the localization of the movement or local differences with the visualization of optical flow. Optical flow highlights the movement across consecutive images. However, in its present form, it is rather hard to make statements about the *magnitude* of the movement across multiple visualizations. In this work we need to rely on other measures for a comparable assessment.)

If we look at the numerical values, we observe that all three measures are *lower* for the ground truth than for elastix. The verdict appears to be that elastix (with our parameter file) over-registers the CT data set. (The over-registration or over-fitting of the registration is basically the "banana problem", too much correspondence is created.) However, there is still some noise in the data that is irrelevant for real-world tasks. The noise, however, still contributes to the measures and is also the subject of registration—we did not use a mask. Such "unnecessary" alignments of the noise might explain the too large values for elastix. Still, the GS method is remarkable in how close it gets to the values of the ground truth quality measures.

4.2. EM data set

We have applied a prior individual normalization in EM data set (Buchacker et al., 2019), hence it might be less fair to compare the registered images to original data, that has additionally 16 bit depth. Hence, we also show the normalized ground truth images.

The differences between rigidly registered images with SIFT and SURF in Fig. 8 appear to be the effect of different normalizations—we were attempting to undo the individual normalization of the image in inputs of the "Rigid-SURF" method. Again, we see residual incongruences with rigid-only methods. Both "Deform-SURF" (Fig. 8) and elastix (Fig. 9) have multiple small "movements" in optical flow visualization, as opposed to a more or less uniform color of a global shift. "Blending" (Fig. 8) shows somewhat uncompensated distortion in the middle of the

image, in Dice, SSIM, and optical flow visualizations. It appears that in this method the effect of our local distortion (Fig. 9) has not been corrected completely, but the random global transformation was undone well.

From the visualizations, "GS" (Fig. 8) and normalized ground truth have definitely less movement than, e. g., "Deform-SURF". The movement in "GS" seems more evenly distributed, normalized ground truth has a "hot spot" in the optical flow visualization. Curiously, the original ground truth has both more details (because it is not clamped-down 16 bit data) and more movement can be detected therein (compare flows and SSIM of the normalized and original ground truths). At the same time, PSNR is higher with the original image pair.

Numerically, elastix has best Jaccard measures (both at 100 and 150 thresholds, indicated with postfix in the following), PSNR, and SSIM among all registrations. Especially with Jaccard 100, "Deform-SURF" is a very close follow-up, at 99.6 % of the performance of elastix. Now, if we look at the measures' values of the objective ground truths—the luxury we did not have before in an evaluation of registrations of serial sections—an issue is apparent. Both Jaccard 100 and Jaccard 150 values for elastix and (less so) for "Deform-SURF" are *higher* than for the normalized ground truth! This issue might indicate an over-optimization by those registrations.

To give some values, Jaccard 150 measure for elastix is 0.748 % *higher* than normalized ground truth, same method has also 7.34 % *higher* PSNR than normalized ground truth; "Deform-SURF" reaches 95.7 % of the normalized ground truth SSIM. Notably, from the visualizations we would deem "GS" as a very good registration, comparable, if not beating elastix. But numerically, elastix has much higher values.

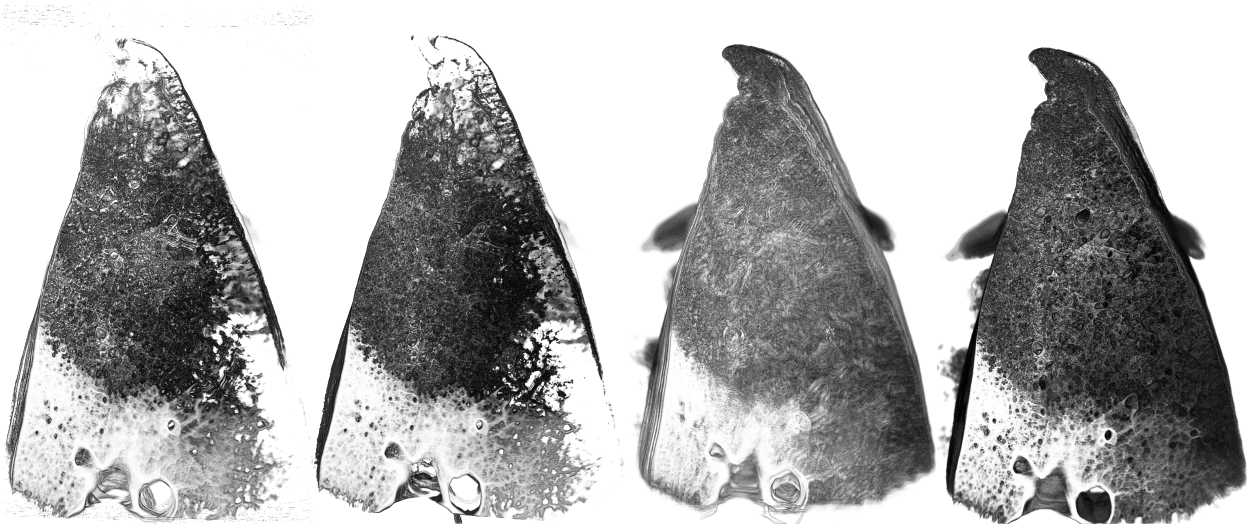
4.3. LS data set

In LS data set the individual normalization was also used for challenge data, in methods with SURF we also aimed to fix those discrepancies that we created in the challenge data to ensure better rigid registration. This explains the visual differences between the results. As we compare images from each of the series to each other and not across the series, this problem is less an issue. Still, it highlights the importance of the input normalization for good registration results. We also used this data set to study the effect of different distortion magnitudes in "Deform-SURF".

In Figure 10 we see a larger movement in the rigid method. It originates from our local distortions in the region of interest ("Local dist."), but cannot be undone with rigid methods only. "Blending" still suffers from these distortions, but restores the coarse correspondence well. "GS" recovers very well from those local movements, however it might over-register, based on visual comparison with normalized ground truth. Both in optical flow visualizations for "GS" and for the ground truth we see multiple small "movements", originating from the fact that we compare two consecutive non-equal images. However, "GS" shows *less* movement than ground truth, hence our above suspicion. Numerical values would provide more clarity, we will look at them next. As for elastix, Dice, PSNR, and SSIM visualizations look very good, similar as in "GS". In elastix, the visualization of optical flow appears in general to be lower than the ground truth, but it also shows a "hot spot" at the top part of the image. However, as already mentioned, we should be careful with such comparisons between optical flow visualizations.

Figure 11 shows quite confident results from "Deform-SURF", but the "Rigid-SURF" is even worse than "Rigid-SIFT" (Fig. 10). We use three different values for the non-linear "stretch" in the non-rigid feature-based registration method "Deform-SURF". The "stretch" values we state are a factor of image size in pixels, limiting the magnitude of the "movement" in the non-linear registration. When the deformation magnitude is apparently too small (10^{-3}), SSIM (and less so, PSNR) visualizations are slightly worse, but the optical flow shows the problem "hot spot" near the center of the image. The default magnitude of $5 \cdot 10^{-3}$ is already much better and does not have the aforementioned problem. An even larger magnitude 10^{-2} also looks similar to the standard setting. Whether those two parameters in "Deform-SURF" over-register can be determined from the numerical values below. For now, we can say that normalized ground truth is slightly better in PSNR, Dice, and SSIM, but appears more "noisy", but also more "even" than "Deform-SURF". We attribute the variations in intensity of the registered images in different methods to the normalization.

Table 3 shows numerical values. The Jaccard measure is computed with threshold 100. "Rigid-SIFT" was very successful with respect to Jaccard measure, but PSNR and SSIM are lower than in a typical non-rigid registration. In a contrast, "Rigid-SURF" is quite bad with respect to all



(a) elastix (b) Deform-SURF (c) local distortions (d) ground truth, normalized

Figure 12: Selected volume renderings for LS data set.

three measures. Still, it manages to be a good input for the non-rigid methods "Deform-SURF", GS, and elastix. Comparing different "stretch" values with "Deform-SURF" numerically, we see that the default value of $5 \cdot 10^{-3}$ is slightly better than the larger value. Too small "stretch" decreases the measures slightly, but it is still better than all rigid methods, for PSNR and SSIM with a larger margin. The "Blending" method has some problems, also evident in visualizations. We discussed this issue above.

The methods "GS" and elastix are very similar with respect to Jaccard measure. Interestingly, PSNR is slightly higher with elastix, but SSIM is higher with "GS". Both "GS" and elastix have better numerical values than all "Deform-SURF" methods tested here—but let us consider the ground truth!

When compared to the (normalized) ground truth, Jaccard and PSNR measures in all registrations are lower. Still, the Jaccard measures for "GS" and elastix are closer to the ground truth than in other methods. However, the SSIM value for "Deform-SURF" is less than 2 % *larger* than the actual SSIM for the normalized ground truth image pair. Still, "GS" and elastix have *even larger* SSIM values than the ground truth: 2.31 % and 2.80 % correspondingly. Such larger SSIM values might signal over-registration.

5. Visual comparisons for LS

This section presents selected volume renderings of the full stacks. While 3D representations typically convey more information, a 3D overview can show only the most coarse incongruences. We demonstrate here the volume renderings, but base out actual evaluation on a sequence of objective measures, as presented above and also in the main paper.

Our volume renderings were produced with ImageVis3D (Fogal and Krüger, 2010). Fig. 12 shows the frontal views of the LS data set. We used the same, standard settings for all the images.

6. Effect of individual normalizations

The *individual* normalizations have an effect of varying intensity of the images through the series. Such variations should mimic the effect of varying section thickness that is normally countered by a *series-wide* normalization prior to the registrations. Fig. 13 showcases the individual normalizations in form of a z-stack from EM series.

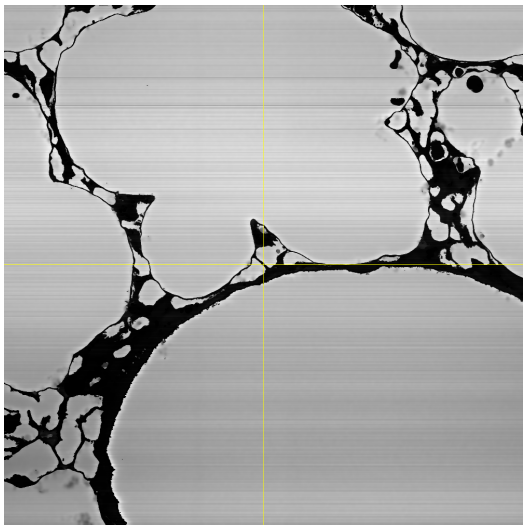


Figure 13: Individual normalizations of EM series, z stacks. We show the original data, the ground truth, after individual normalizations.

7. Effect of distortions on LS

Figure 14 shows full optical flow and SSIM visualization of a consecutive image pair from the LS data set. Concerning the locally distorted images (a), (c), notice that there is some global movement that creates problems for the rigid-only methods. Some traces of these problems are still evident in the results of, e. g., "Blending", see main text. Panels (b), (d) show the normalized ground truth. We use the same image pair (150–151) that is used everywhere else for the pair-wise evaluation.

8. A violin plot

For the statistical evaluation, instead of box plots, violin plots can be used. We advocated against them in the main paper, as we would be more interested in the inliers. For the completeness, we show here an example violin plot in Fig 15. It is the Jaccard evaluation of the whole CT series. We see there some interesting consequences on the distribution of the outliers, but in this case little cannot be inferred from the corresponding box plot in the main paper. It is less clearer to see in this vioplot plot without training which method has a higher median, though.

9. Gallery of distortion visualizations

Figure 16 shows the visualizations of our generated distortions for the LS data set. The images are normalized; HSV colorspace is used to visualize directions.

References

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.

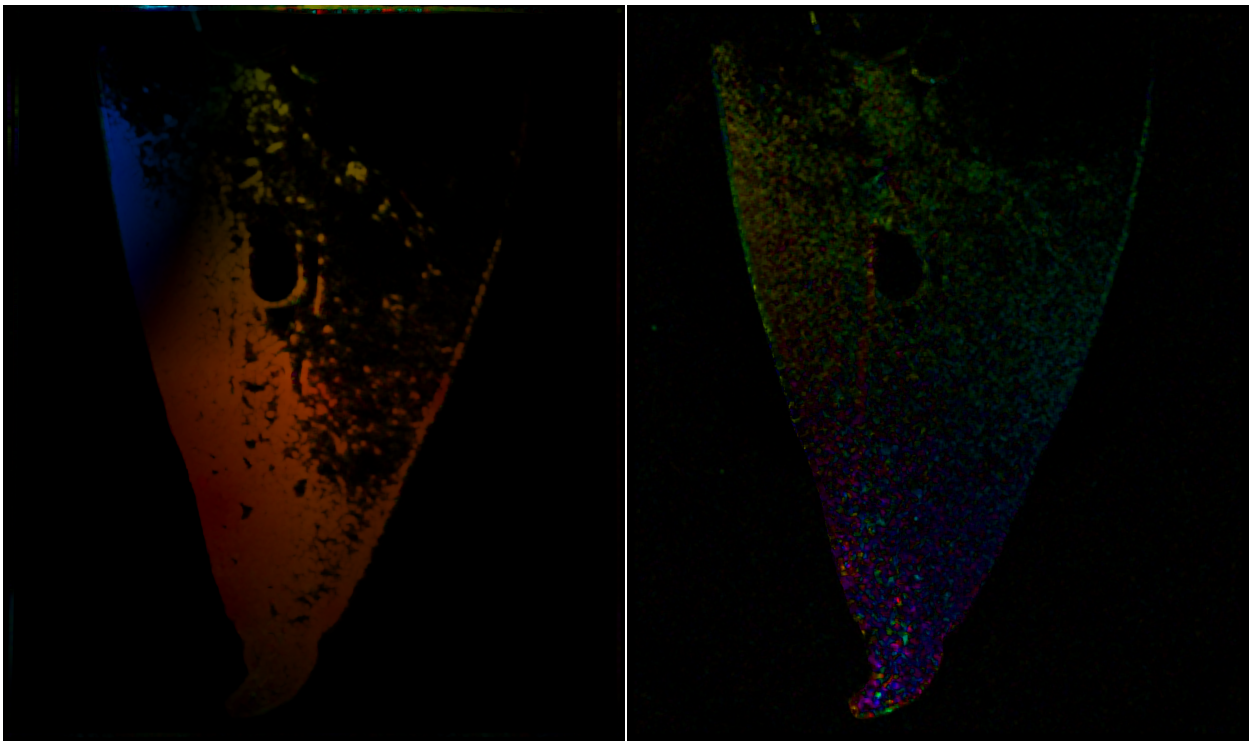
J.-M. Krischer, K. Albert, A. Pfaffenroth, E. Lopez-Rodriguez, C. Ruppert, B. J. Smith, and L. Knudsen, "Mechanical ventilation-induced alterations of intracellular surfactant pool and blood–gas barrier in healthy and pre-injured lungs," *Histochem. Cell. Biol.*, vol. 155, no. 2, pp. 183–202, 2021.

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.

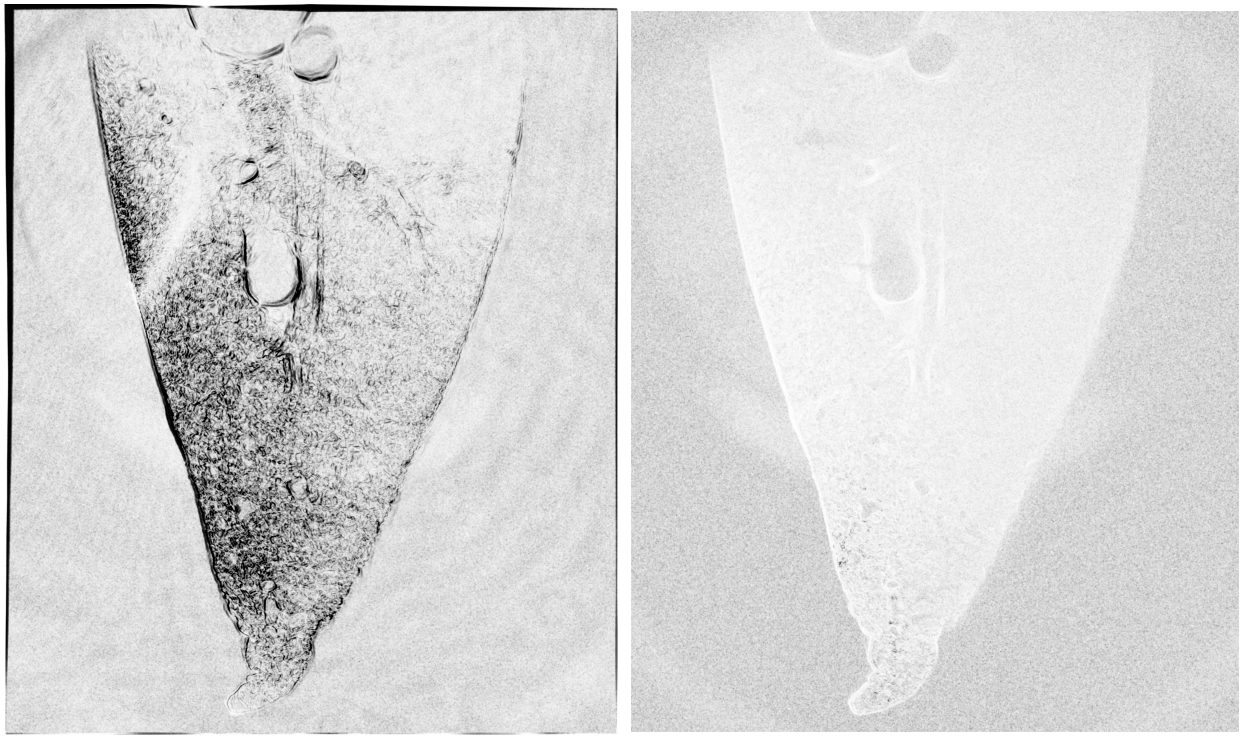
R. Grothausmann, J. Labode, P. Hernandez-Cerdan, D. Haberthür, R. Hlushchuk, O. Lobachev, C. Brandenberger, A. G. Gie, T. Salaets, J. Toelen, W. L. Wagner, and C. Mühlfeld, "Combination of μCT and light microscopy for generation-specific stereological analysis of pulmonary arterial branches: a proof-of-concept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

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.

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.



(a) (b)



(c) (d)

Figure 14: Consecutive images from LS data set after local distortions with our method (a), (c), and before (b), (d), i. e., on ground truth. Figures (a), (b) show optical flow visualizations, (c), (d) show SSIM.

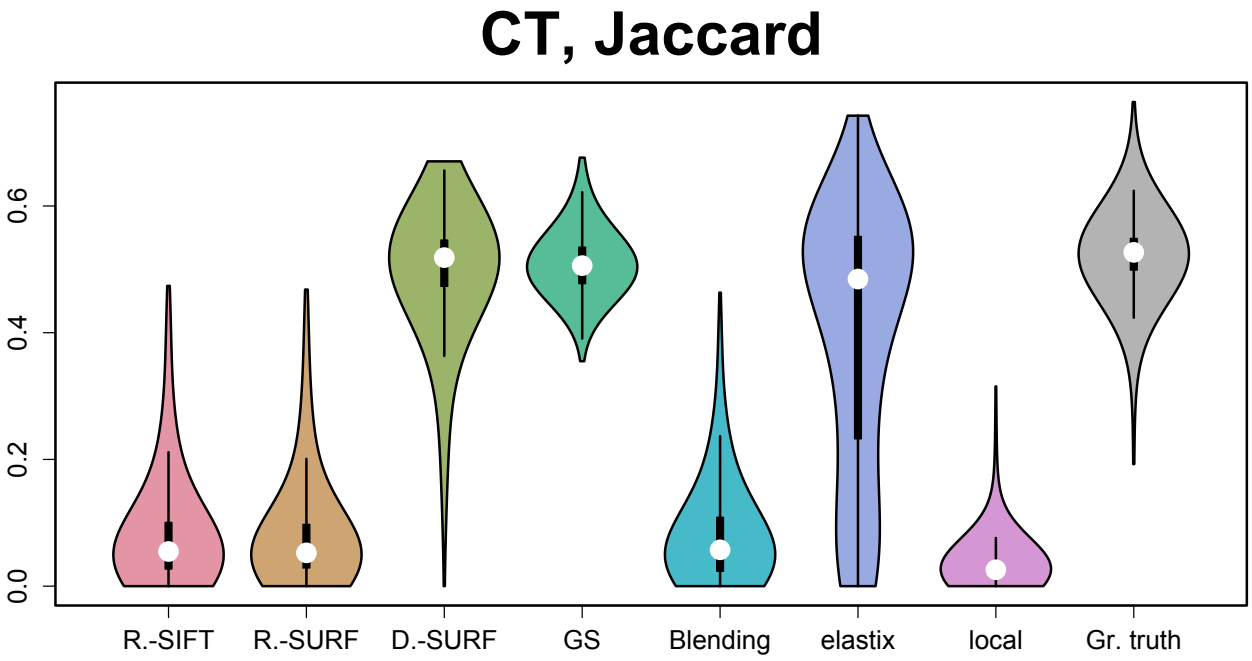


Figure 15: Violin plot of the Jaccard measures over the whole CT data set, registered with various methods.

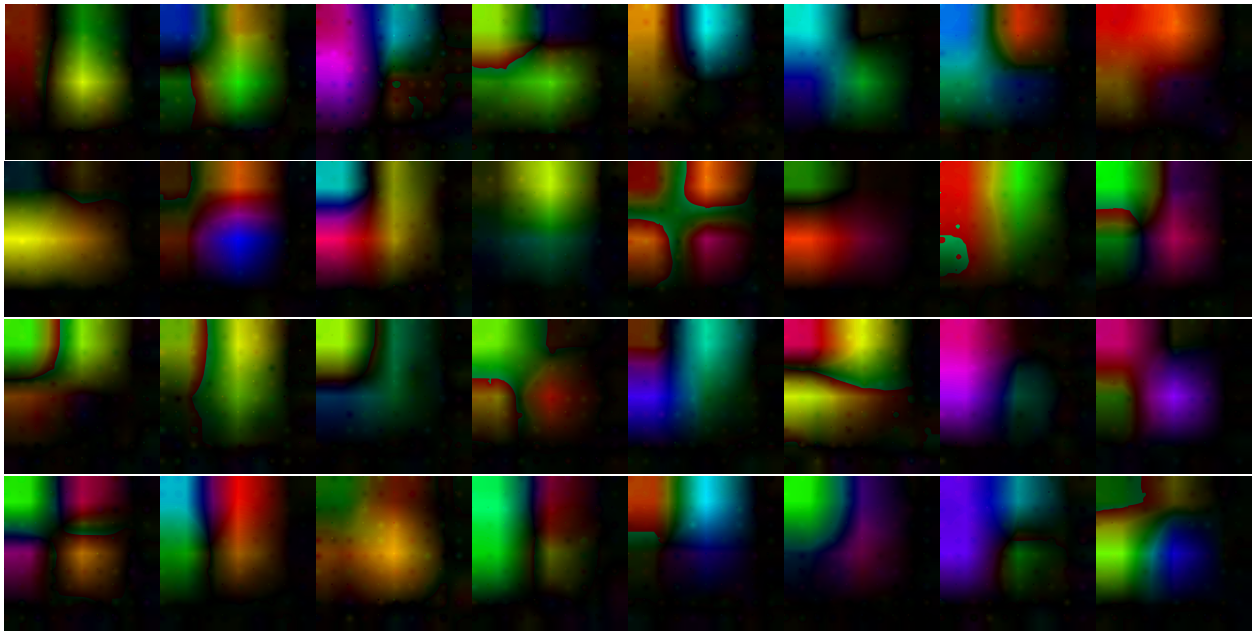


Figure 16: Visualizations of our local distortions on LS data set.

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.

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.

N. Otsu, "A threshold selection method from gray-level histograms," *IEEE T. Syst. Man. Cyb.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

G. Bradski, "The OpenCV library," *Dr. Dobbs J.*, vol. 25, pp. 120–125, 2000.

A. Kaehler and G. Bradski, *Learning OpenCV, 2Nd Edition*. O'Reilly Media, Inc., 2014.

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.

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.

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.

T. Fogal and J. Krüger, "Tuvok, an architecture for large scale volume rendering," in *Proceedings of the 15th International Workshop on Vision, Modeling, and Visualization*, ser. VMV '10. Eurographics, 2010. [Online]. Available: http://www.sci.utah.edu/tfogal/academic/tuvok/Fogal-Tuvok.pdf