Validation of simulation-based evaluation by inclusion of unlabeled data and meta-ablation

Stefan B. Ploner[1,2,✉], Linda-Sophie Schneider[1], Yunchan Hwang[2], Abdul R. Hassan[3], Nadia K. Waheed[3], James G. Fujimoto[2], and Andreas Maier[1]

[1] Pattern Recognition Lab, FAU Erlangen-Nürnberg, Germany
[2] Department of Electrical Engineering and Computer Science, and Research Laboratory of Electronics, Massachussets Institute of Technology, Cambridge, USA
[3] New England Eye Center, Tufts University, Boston, USA
stefan.ploner@fau.de

Abstract. Lack of adequate ground truth data is a common issue in various medical image processing tasks. Especially in non-uniform image registration and image reconstruction, dense manual labeling of data is both infeasible and insufficient in terms of accuracy, which prevents the use of a manually-labeled test set. In contrast, simulation can offer unlimited data with exact ground truth. However, since simulations are based on models that are limited themselves, the evaluated algorithm could be overfitted to the limited simulation model, and therefore perform worse in real data (the "realism gap"). In this paper, we present meta-ablation, a concept for label-free quantitative measurements that aid in demonstrating sufficient realism of simulated data. Our approach is intended for fusion tasks, i. e., tasks that compensate confounders in repeated measurements like motion compensation, group-wise registration, multi-frame denoising, and super resolution. Pairs of evaluations are compared: one using only simulated data and one using a combination of simulated and real data. Additional degrees of variation in real data influence the compensation of the simulated data in the combined setting and worsen the results. This worsening is an estimate of the realism gap of the simulated data. We demonstrate our approach in the evaluation of non-uniform image registration of 3D OCT data. We not only show that model advances allowed closing the realism gap, but also that inspecting residual errors from limited models in the combined setting allowed us to sequentially deduce the most important next model advance. In short: Our label-free approach validates simulated data and can provide a strategy for self-optimization.

Keywords: Evaluation · Simulation · Synthetic Data.

1 Introduction

Evaluation and validation are central aspects of automated medical data processing. Its importance increased further with the widespread use of deep learning in research and its translation to clinical care. Recently, a large-scale effort of

experts and the medical image computing community addressed the selection of metrics in tasks such as classification, object detection, and segmentation [17,11]. The impact expands to other image processing tasks, such as registration and restoration, where segmentation can be used in downstream task-based evaluations. Nevertheless, because of their abstraction, downstream task-based evaluations lose details or accuracy, which may be important if the evaluated method is, or will be, used for additional tasks. However, direct computation of suitable metrics is not possible for all tasks. In particular in image registration, the accuracy of manual labeling is inherently limited. While spatial aggregation over multiple keypoints can help in rigid registration, this is not an option when estimating non-uniform displacement fields, which are additionally infeasible for manual labeling. Similarly, manual labeling is not suited for image restoration tasks, and workarounds like averaging of repeated measurements are neither always feasible nor suited, e. g., if advanced tasks such as deblurring are desired. While phantoms can be well suited for calibration tasks, they rarely reach the complexity of real biological specimens. Reproducibility studies, specifically relevant in the compensation of motion artifacts, require the acquisition of repeated data, which is not always feasible in medical imaging due to dose constraints or patient strain. In summary, a labeled, unseen test set, which is assumed in the initially cited publications, may not always be available.

When designing an evaluation without available ground truth, metric selection is thus only the second step, while attaining a suitable ground truth is the first. In these cases, physical simulation and image synthesis offer access to arbitrary amounts of ground truth data [6]. In addition, they may have the potential to generate data for exploring the limits of a method with respect to specific effects [8]. Simulations exist for all major medical imaging modalities [1,3,5] and their extension is a topic of active research [4]. The key question in using simulation-based data for evaluation is whether the simulation captures all relevant degrees of variation to resemble the real problem's complexity. Lack of relevant effects could simplify the to-be-evaluated task, which may lead to overly optimistic results that do not generalize to real data [9]. This is particularly important when the simulated data is used to guide method development, especially in learning-based methods. Therefore, without demonstrating sufficient simulation complexity, any metric derived from a simulation-based evaluation can only provide a lower error bound, whereas an upper bound is desired.

Yet conversely, while models of relevant problems will always be limited, this is not necessarily an issue in the evaluation of a specific task. For example, an evaluation of image registration will have high requirements on the simulation's motion model, but certain limitations in the simulation of image noise can be tolerated, and vice-versa. A related observation is that a more accurate method in general also needs improved simulation to demonstrate its full impact, while results of an inaccurate method will not worsen from simulating effects significantly below its standard error or resolution. This raises the question: Is a simulation model complex enough? Or in learning-based synthesis, is the amount of and variation in the training data sufficient to achieve realistic evaluation results?

In this paper, we describe a general approach for demonstrating *insufficiency* of a given simulation for the evaluation of a specific image fusion-related task (image registration, reconstruction and super-resolution) and the accuracy regime defined by the evaluated method. The fundamental principle of our approach is that replacing parts of the simulated data with real data changes the derived evaluation metrics. We use this change as a surrogate for the "realism gap" between simulated and real data. Inspired by ablation studies, we vary aspects of the method *and* the simulation, coining the term meta-ablation. We demonstrate our approach in image registration, where progressive model advances allowed *closing* the realism gap, thereby demonstrating *sufficient* simulation quality of the motion model, a core part of the simulation. We furthermore show how meta-ablation guided the identification of each next most-relevant model advance. The entire approach is entirely label-free in the real data.

2 Methods

We begin with recapitulating key components of classic ablation studies in context of simulated data. We introduce a key concept that can improve and, to a certain degree, measure realism of the simulated data in data fusion tasks. We then generalize the approach to meta-ablation to evaluate the simulation itself.

In ablation studies, different setups of a method are evaluated on the same data (see Fig. 1). In our definition, a *method* only defines a function that maps a certain input to a result of the form that is expected by the evaluation. The *setup* controls all aspects of finding the solution, from tuning parameters, over switching between models of different complexity, to changing the solution method itself. Next, an *evaluation* is performed on each result independently. At the end, the derived metrics are aggregated to allow comparison between setups.

In case of simulated data, the parameters underlying a simulated image can be passed on to the evaluation as ground truth. However, the evaluation can

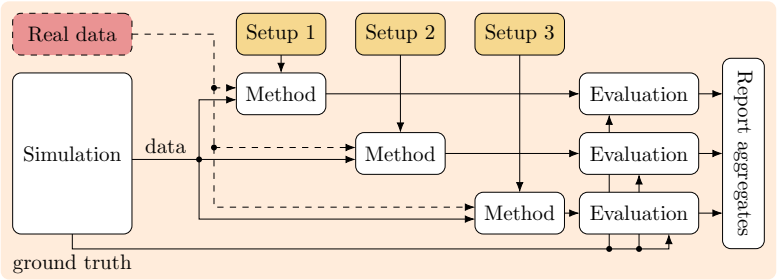

Fig. 1. Classic ablation for performance evaluation of a method. Simulation, method and evaluation blocks are consistent in each evaluation run. In data fusion tasks, simulated and real data can be combined, but this restricts evaluation metric computation to the simulated share of the data, where a ground truth is available.

then only prove an accuracy or reliability with respect to the effects and distributions that are modeled by the used simulation. If a method is overfitted to the simulation, e. g., by using an equally limited model as the simulation itself, the errors of additional or differing variation in real data will not show up. However, in fusion-based tasks, it is possible to substitute parts of the data with real data, as indicated with the dashed arrows in Fig. 1. While this limits the (ground truth-based) evaluation to the simulated share of the data, the real data will add additional degrees of variation to the problem. In fusion-based tasks, this can irritate even the compensation of the simulated data and change the distribution of the resulting evaluation metrics. An algorithm that is overfitted to a limited simulation will only perform well on the all-simulated data, but is exposed by an error increase in the combined simulated and real data. We use the resulting changes in metric distributions as a surrogate for the realism gap.

Turning to meta-ablation, which is depicted in Fig. 2, we now add variations to the simulation. Here, the complexity of the inherent *model* can be varied. Motion models could vary in the degrees of freedom of their transform (rigid, affine, homography), whereas image degradation could model additional noise sources or structuredness of the noise. In simulations that utilize measurements of real data, the setting of these *measurements* can be varied. For example, motion simulation can integrate motion measurements from an additional modality with higher tracking accuracy, or the simulation of image degradation could be based on a device with higher resolution. In general, overly complex simulation models worsen evaluation results in comparison to real data, whereas limited models improve them. However, detailed domain knowledge is needed to predict its magnitude, or even just the direction of bias.

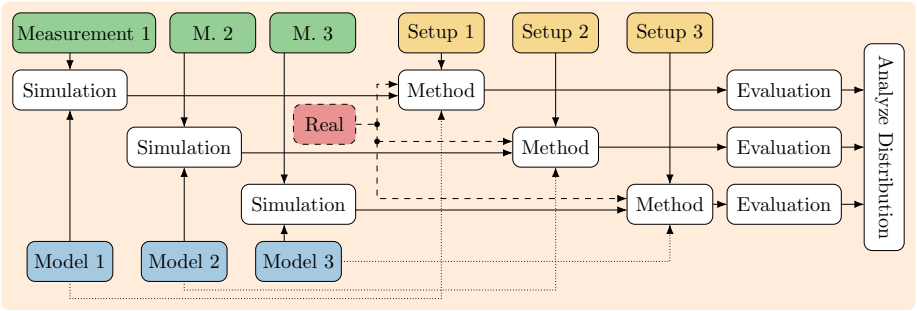

Fig. 2. In contrast to regular ablation, meta-ablation either varies model assumptions, or measurements that are utilized for simulation. The key point is that partial inclusion of real data reduces the potential for overfitting, even when identical limited models are used in both simulation and the method. This creates a difference compared to only using simulated data: the realism gap. The ability to "resolve" the gap depends on the accuracy of the method. Therefore, the gap is also setup-dependent: More accurate methods require better simulation to close the realism gap.

Combining meta-ablation with the inclusion of real data, which shifts results toward increased realism, enables inferring the direction of the bias. However, since only parts of the data can be replaced to maintain simulated ground truth for evaluation, it is not possible to measure its magnitude. For this, it is necessary to refine the simulation's realism until the discrepancies between only simulated and combined data vanish.

At this point, another aspect of our approach becomes relevant: While we previously only discussed looking at metrics after aggregation, errors can also be investigated locally for a more granular investigation. This could, for example, reveal presence of low amplitude illumination bias fields or the existence of additional subtle global distortion, which may be hard to spot in qualitative analysis. The challenge is that these effects cannot be observed directly in terms of parameter change because, by definition, these residual errors arise from unmodeled degrees of variation. However, they are mapped into the existing (non-fully orthogonal) degrees of freedom that are compensated by the method. This makes the estimation differ from the ground truth, even if global change in illumination or displacement is compensated for the existing parameters. Additionally, the magnitudes of these residuals can be aggregated over multiple datasets to reduce the estimation noise and, in a relative sense, to improve visibility of biases.

Since the goal is typically to develop a method rather than the simulation, the best-performing model found should be used in the method, if feasible. Seen in one way, this limits the simulation to use (at best) the same model to generate the test data. From a different perspective, an enhancement of the simulation model also allows us to improve the model of the method, which allows us to measure residual errors and the realism gap more accurately. While only improving the simulation model would close the realism gap up to the accuracy level of the original method, joint improvement of both the simulation and method models reduces bias and therefore allows demonstration of realistic simulation at the best possible accuracy level. This is indicated by the dotted lines in Fig. 2.

Finally, note that accuracy is not only limited by model bias, but also by parameter estimation accuracy (i. e., their "variance"). Since this affects the resolvability of the realism gap, the *setup* nodes are still included in this figure.

We demonstrate our approach on motion correction of repeated retinal OCT volume images as described in [14,15,13]. Since this acts only as an example, we only restate algorithm properties that are relevant to the demonstration of our approach. The method estimates a non-uniform displacement field by reducing an image similarity metric via numerical optimization. The object is assumed to be rigid, but moves over time during the scanned OCT image acquisition. Therefore, the transform is given by time-dependent parameters as follows:

$$\begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cos(\alpha) + a_{x,x} & -\sin(\alpha) & 0 & t_x \\ 0 & 0 & \sin(\alpha) & \cos(\alpha) + a_{y,y} & 0 & t_y \\ m_{xx} & m_{yy} & m_x & m_y & 1 & t_z \end{pmatrix} \begin{pmatrix} 1.5x^2 - 0.5 \\ 1.5y^2 - 0.5 \\ x \\ y \\ z \\ 1 \end{pmatrix} \tag{1}$$

Besides translation (t_x, t_y, t_z) and rotation (α), additional axial sheering (m_x, m_y), axial curvature (m_{xx}, m_{yy}) and transverse sheering $(a_{x,x}, a_{y,y})$ distortion arises from the optical setup of the imaging process [7,13]. Aliasing in the similarity metric is prevented by sampling the objective function at pseudo-random locations to improve accuracy. Robustness and accuracy is improved by jointly compensating multiple images.

Simulation is performed by resampling corrected volumes with a known displacement field. Following our previously described approach, the displacement field is computed based on the same transform model as the method. Since OCT is the only modality that allows estimation of the full 3D transform, realistic eye motion traces are extracted from previous runs of the same method, except transverse translation. The latter contains frequency components (ocular tremor [12]) beyond what can be estimated by the method. Therefore, in this case, we use openly accessible data from Bowers et al. instead, which is measured by high frame-rate (2D) adaptive optics scanning laser ophthalmoscopy [2]. Before comparing the estimated displacement field with the simulated ground truth, global registration (i. e., no time-dependence along the scan) to the reference is performed to compensate for non-zero mean average displacements.

3 Results

15 high-resolution spectral-domain OCT datasets were acquired from 15 eyes and subjects (average age 72.7, most eyes affected by age-related macular degeneration) with a prototype device at the New England Eye Center in Boston, USA [10] (A-scan rate 128 kHz, 3 and 13 µm axial and transverse resolutions, beam diameter 1.4 mm at $\frac{1}{e^2}$). 6 volumes of 500×500 A-scans with orthogonally alternating fast scan directions were acquired over 6 mm × 6 mm. All imaging was performed in agreement with the declaration of Helsinki and in accordance with the Tufts University Health Sciences Institutional Review Board.

Figure 3 shows the results of the meta-ablation. The simulation and method transform models are jointly varied to compensate for 3D translation and axial sheering (*Kraus14* [7]), added transverse rotation (*Ploner22* [14]), additionally added quadratic curvature (*Ploner25* [13]), and using the full *Eq. 1.* Blue bars are computed based on motion correction of all single simulated volumes in combination with five additional simulated volumes, whereas red bars are computed based on the same simulated volumes and five real volumes (resulting in $15 \cdot 6 = 90$ values each). The constituent measurements are the L2-norm position errors, aggregated per dataset via the median, representing the typical 3D registration error. We also evaluated the percentage of outliers for multiple thresholds (1 µm, 2 µm, .5 px, 1 px). We also evaluated each coordinate axis, as well as the fast and slow scan axis, individually. Lastly, we evaluated different combinations of simulated and real volumes. These additional evaluations were consistent and therefore omitted for brevity. Significant difference is determined by the paired sign test. Three stars represent $p < 0.001$, whereas no stars represent $p > 0.5$. The overfitted simulation performance (blue bars) is defined by parameter estimation

accuracy and remains approximately constant. Slight worsening is possible for more complex models, where errors from more parameters can add up. Large realism gaps are revealed for the limited models, but they decrease by adding relevant parameters until they approximately vanish with our full model. In the right part, the same models are evaluated without compensating for aliasing in the method's objective function [14], which reduces the accuracy of the method. As a consequence, the realism gap is resolved considerably worse and already vanishes for the *Ploner22* model, when measured based on the L2-norm error, which combines the errors along all axes. In this example of 3D registration, it is possible to evaluate the error on each axis separately, which still exposes a realism gap in the axial dimension. This realism gap is then also closed with the *Ploner25* model.

Omission of noise simulation shows further improvement, especially in the all-simulated case. Combined data exposes this non-realistic simulation, but since the simulated parts of the data are still noise-free, the error increase does not reach the level of the realistic noise used in the results labeled *Eq. 1*.

Realistic transverse eye motion contains frequency components beyond those resolvable with our motion estimation. Their error shows up in the metric because we simulate and compute the displacement error at the high sampling rate of the A-scans. Consequently, simulation with motion that is band-limited up to the motion estimation frequency ("OCT motion") shows a lower error. However, higher frequency variations in real data are not mapped to the estimate of our method because it is band-limited. The higher frequency components of the displacements are orthogonal in the sense of a Fourier decomposition and are discarded without confusing the lower frequency part. Therefore, in this case, a realism gap does not appear.

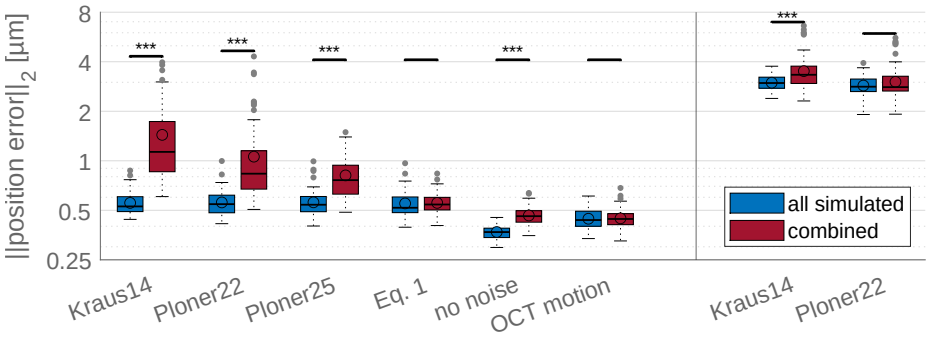

Fig. 3. Boxplots show log-scale median A-scan position errors based on simulated data only (blue bars) and combined simulated and real data (red bars). Simplified models that lead to overfitted results are exposed by a realism gap between both distributions. A vanished realism gap can indicate realistic simulation in the accuracy regime of the evaluated method. Combined data only shows a tendency, not the full real error (no noise). Not all relevant limitations create a realism gap (OCT motion). The right part uses a method with reduced accuracy, which reduces resolvability of the realism gap.

Figure 4 shows residual errors in combined data compensated with a method that uses the same limited transform model as the simulation. Unmodeled variation of the real data is revealed, which aids in the identification of the parameter with the next-largest residual error. Errors vanish when modeled (black arrows).

4 Discussion

We described a label-free approach for quantitative "meta"-evaluation of the realism of simulated data. It directly allows statistical demonstration of insufficiency of a simulation. Such detection can be improved by using multiple metrics and aggregation functions, and by individual evaluation of composed errors, e. g., by applying displacement metrics on x, y and z components separately. In case of insufficient simulation, we demonstrated how residual errors generated with our approach can aid model improvement, which allowed closing the realism gap. In case of a closed realism gap, for example, reducing a parameter density until it shows up again allows label-free estimation of real data properties (at parameter estimation accuracy which can be determined from all-simulated data).

Although a contrary conclusion of sufficiency is not permissible, a closed realism gap limits the need for theoretical analysis of potential remaining errors to modes of variation that do not propagate into the method's estimation. While a general conclusion is hard, in our example, this is the case for frequency ranges of motion that are not estimated. If the spectral densities of these frequency ranges are known, which is the case for eye motion [2], an error bound can be derived. In our case, there are also measurements of a higher frame-rate motion estimation available, which even allow direct integration of such motion in the simulation. In general, the more complex a model is, especially when estimating non-uniform displacements (image registration) or individual pixel intensities (image reconstruction), the better missing components can show up as a realism gap. While the opposite is true if a model is significantly underparameterized, this case can be exposed by qualitative evaluation. In our example, the model advances derived from this approach can be confirmed by the alignment of individual photoreceptors, as we demonstrate in a concurrent submission [16, Fig. 3].

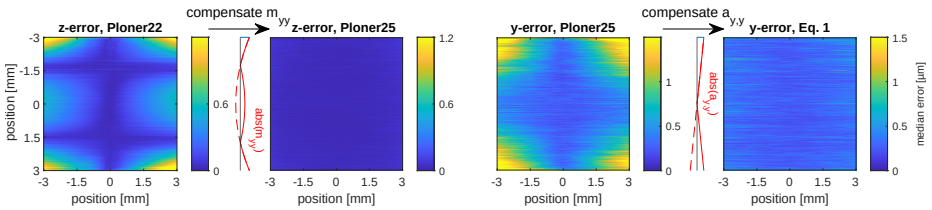

Fig. 4. Combined-evaluation residual errors reveal unmodeled variation of real data. Shown are median-of-median errors over all 45 horizontally fast-scanned volumes. The real targets' axial curvature (left) and transverse sheer variations (right) appear along the vertical axis. When added to the transform and compensated, the errors vanish.

Limitations of our approach include its inability to exactly quantify the realism gap and its exclusive applicability to fusion-based tasks. However, tasks like motion correction are classic examples for those without a feasible ground truth. A related issue is that the simulated data must be consistent with the real data, e. g., there should be no discontinuity between simulated and real data. Lastly, cancellation of effects must be prevented. Artifacts in the simulation could worsen the simulated results and suggest a closed realism gap. Therefore, our approach does not eliminate the need for domain knowledge, should be used with multiple and fine-grained metrics, and only in combination with qualitative evaluation. Besides these limitations, the approach is a significant aid in understanding model limitations of both simulation and the method, and can provide direct insight in how to overcome them.

Author contributions Conceptualization, methodology, implementation, visualization, and original manuscript: SP. Manuscript review: SP, LS. Data curation: YH. Consenting, Imaging: AH. Funding: SP, AM, JF, NW.

Acknowledgments. The authors acknowledge funding from the German Research Foundation (project 508075009) and the National Institutes of Health (5-R01-EY011289, 5-R01-EY034080).

Disclosure of Interests. Author JF has IP on orthogonal scan motion correction.

References

1. Abadi, E., Harrawood, B., Sharma, S., Kapadia, A., Segars, W.P., Samei, E.: DukeSim: A Realistic, Rapid, and Scanner-Specific Simulation Framework in Computed Tomography. IEEE Transactions on Medical Imaging **38**(6), 1457–1465 (2019). https://doi.org/10.1109/TMI.2018.2886530
2. Bowers, N.R., Boehm, A.E., et al: The effects of fixational tremor on the retinal image. J. Vis. **19**(11), 8–8 (09 2019). https://doi.org/10.1167/19.11.8
3. Duncan-Gelder, P., O'Keeffe, D., Bones, P., Marsh, S.: PhoenixMR: A GPU-based MRI simulation framework with runtime-dynamic code execution. Medical Physics **51**(9), 6120–6133 (2024). https://doi.org/10.1002/mp.17273
4. Fernandez, V., Wiesner, D., Zuo, L., Casamitjana, A., Remedios, S.W.: Simulation and Synthesis in Medical Imaging, vol. 16085. Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-05573-6
5. Garcia, D.: SIMUS: An open-source simulator for medical ultrasound imaging. Part I: Theory & examples. Computer Methods and Programs in Biomedicine **218**, 106726 (2022). https://doi.org/10.1016/j.cmpb.2022.106726
6. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. pp. 2672—2680. NIPS'14, MIT Press, Cambridge, MA, USA (2014)
7. Kraus, M.F., Liu, J.J., et al: Quantitative 3D-OCT motion correction with tilt and illumination correction, robust similarity measure and regularization. Biomed. Opt. Express **5**(8), 2591–2613 (Aug 2014). https://doi.org/10.1364/BOE.5.002591

8. Kwan, R.S., Evans, A., Pike, G.: MRI simulation-based evaluation of image-processing and classification methods. IEEE Transactions on Medical Imaging **18**(11), 1085–1097 (1999). https://doi.org/10.1109/42.816072
9. Köhler, T., Bätz, M., Naderi, F., Kaup, A., Maier, A., Riess, C.: Toward bridging the simulated-to-real gap: Benchmarking super-resolution on real data. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(11), 2944–2959 (2020). https://doi.org/10.1109/TPAMI.2019.2917037
10. Lee, B., Chen, S., Moult, E.M., Yu, Y., Alibhai, A.Y., Mehta, N., Baumal, C.R., Waheed, N.K., Fujimoto, J.G.: High-speed, ultrahigh-resolution spectral-domain oct with extended imaging range using reference arm length matching. Transl. Vis. Sci. Technol. **9**(7), 12–12 (06 2020). https://doi.org/10.1167/tvst.9.7.12, https://doi.org/10.1167/tvst.9.7.12
11. Maier-Hein, L., Reinke, A., Godau, e.a.: Metrics reloaded: recommendations for image analysis validation. Nature Methods **21**(2), 195–212 (2 2024). https://doi.org/10.1038/s41592-023-02151-z
12. Martinez-Conde, S., Macknik, S.L., Hubel, D.H.: The role of fixational eye movements in visual perception. Nat. Rev. Neurosci. **5**(3), 229–240 (2004). https://doi.org/10.1038/nrn1348
13. Ploner, S., Babiker, F., Hwang, Y., et al: Quantification and compensation of 2nd order image distortions in high resolution OCT volume fusion and longitudinal registration. Invest. Ophthalmol. Vis. Sci. **66**, PB0088 (7 2025)
14. Ploner, S., Chen, S., Won, J., et al: A Spatiotemporal Model for Precise and Efficient Fully-Automatic 3D Motion Correction in OCT. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 517–527. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_50
15. Ploner, S., Won, J., Schottenhamml, J., et al: A Spatiotemporal Illumination Model for 3D Image Fusion in Optical Coherence Tomography. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2023). https://doi.org/10.1109/ISBI53787.2023.10230526
16. Ploner, S.B., Fan, W., Zhang, H.F., et al.: Volume-Fused Super-Resolution OCT Enables Non-Mydriatic Cellular Imaging in the Outer Segment across Expanded Fields. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2026. Springer Nature Switzerland, Cham (2026), under review.
17. Reinke, A., Tizabi, M.D., Baumgartner, M., et al.: Understanding metric-related pitfalls in image analysis validation. Nature Methods **21**(2), 182–194 (2024). https://doi.org/10.1038/s41592-023-02150-0