UNIVERSITY OF COPENHAGEN

This thesis has been submitted to the PhD School of the Faculty of Science, University of Copenhagen, in fulfillment of the requirements for the degree of

Doctor of Philosophy (Ph.D.)

Topology-Aware Image Registration

Deep Learning in Medical Imaging

Steffen Czolbe

per.sc@di.ku.dk

Supervisors: Aasa Feragen and Oswin Krause

January 2023

PREFACE

This document is a Ph.D. Thesis – a thesis submitted to obtain the degree of Ph.d. (Philosophiae Doctor) in the Danish academic system. It is form and function equivalent to a Doctor of Philosophy in English-speaking countries, or the title of Dr. rer. nat. in Germany.

I conducted the majority of the work presented here as a Ph.D. student in Denmark between 2020 and 2023. Danish Ph.D. programs are, in comparison to programs in other countries, relatively short. The danish program is limited to a duration of 3 years, which includes half a year of coursework, a half-year stay at a foreign research institution, and half a year of duty hours in the form of teaching and administrative work.

This thesis documents and discusses my research in the field of Deep Learning and Medical Imaging, with a special focus on image registration and the registration of images whose topology is not the same. The first part of the thesis offers a summary, which provides a high-level introduction to the topic. Chapter two gives a summary of this thesis' contributions to science. Chapter three gives a brief conclusion of each work and provides perspectives for further research.

As this thesis is a cumulative one as opposed to a monograph, the second part of this document contains five papers either published or currently under review. All attached papers are self-contained and can be read in any order. Some projects were published in multiple versions, for example first as a workshop paper, then a conference paper, and finally a journal paper. In these cases, only the most comprehensive and recent version is included in this thesis.

ACKNOWLEDGEMENTS

This journey would not have been the same without the help and support of amazing people. My first word has to go to my supervisors. Thank you Aasa Feragen for opening many doors for me, and standing up for your Ph.D. students at any opportunity. Thank you Oswin Krause for taking me on as a master's and Ph.D. student, and being always hands-on and unafraid of social convention. And finally, thank you Adrian Dalca for hosting me to work on an exciting project during my 6-month stay in Boston, forcing me to polish my writing skills to a level I didn't imagine possible through many rounds of feedback, and introducing me to the awesome research culture around MIT, Harvard, and MGH. I am extremely grateful for everything I learned, and leave academia with no regrets about my time spent here.

The road toward a Ph.D. is a long and arduous one, and it was sweetened by the many colleagues and friends I made along the way. A big shoutout to my fellow students Kasra Arnavaz and Paraskevas Pegios for collaborating with me on papers. Thanks to the upper floor of the IMAGE section at DIKU, who cooked together with me in 40°C offices through two corona-summers. Thanks to Aasa's Ph.D. students at DTU, who always welcomed me. And thanks to the students of the Machine Learning section at KU, who became my office mates during the last year. The sandwiches were the best section lunch I ever had!

Finally, I want to thank the people and environment outside of work who kept me sane throughout the years. My long-term partner, Jing Lin, who had incredible tolerance towards the oddities that come along with living with a Ph.D. student. And the city of Copenhagen, for being the most fun and enjoyable place on earth (at least in the summers!).

ABSTRACT

Deformable image registration, or the nonlinear alignment of images, is a fundamental preprocessing tool in medical imaging. Existing registration methods regularize this ill-posed problem by the assumption of a common topology across all images. Images are assumed to be one-to-one deformations of a common reference template. However, this assumption is frequently violated in the real world, especially among populations requiring medical intervention, where the physical anatomy can differ from a common template due to tumors or surgical resections. In such cases, the same-image assumption is often accepted as unavoidable, leading to imprecise registration and inaccuracies in the subsequent analysis. This is especially troubling as medical imaging tools are most often used to process the non-standard anatomies of patients requiring treatment, not healthy reference populations.

Over the last few years, deep learning methods have enabled fundamentally new approaches to image registration. Unsupervised learning-based registration models have achieved performance on par with classical algorithms while being many times faster. However, current deep-learning-based registration is still based on the one-to-one matching of images, thus suffering from the same shortcomings under a change of anatomy.

This thesis explores new opportunities enabled by deep learning to overcome the same-image assumption, such as implicitly inferring a solution from observed data with an unsupervised probabilistic approach and informing the optimization with semantic image representations. However, adopting deep learning to medical imaging also introduces new challenges. The thesis contributes to the open research questions of selecting a suitable model architecture for image registration, quantifying model uncertainty in the presence of annotator variability, and training few-shot models in a clinical research environment.

These strands of work combine into an unsupervised model for detecting anatomical differences alongside the registration step. The model highlights areas where the registration is inadequate due to violations of the same-image assumption, indicating where further care has to be taken in the analysis and downstream processing. This is an important milestone toward a fully topology-aware image registration and an essential building block toward a deformable registration pipeline for images with differences in anatomy.

ABSTRACT (DANISH)

Deformerbar billedregistrering, eller den ikke-lineære justering af billeder, er et grundlæggende forbehandlingsværktøj inden for medicinsk billedanalyse. Dette probelm er ikke veldefineret og reguleriseres af eksisterende registreringsmetoder ved at antage en fælles topologi på tværs af alle billeder, også kaldet "samme-billede-antagelsen". Billeder antages således at være en-til-en-deformationer af en fælles referenceskabelon. Denne antagelse bliver dog ofte overtrådt i den virkelige verden, især blandt befolkninger, der kræver medicinsk intervention, hvor den fysiske anatomi kan afvige fra en almindelig skabelon på grund af tumorer eller kirurgiske resektioner. I sådanne tilfælde anses samme-billede-antagelsen ofte som uundgåelig, hvilket fører til upræcis registrering og unøjagtigheder i den efterfølgende analyse. Dette er især bekymrende, da medicinske billedbehandlingsværktøjer oftest bruges til at behandle de ikke-standardiserede anatomier hos patienter, der kræver behandling og ikke sunde referencepopulationer.

Gennem de sidste par år har metoder baseret på deep learning muliggjort fundamentalt nye tilgange til billedregistrering. Uovervågede læringsbaserede registreringsmodeller har opnået ydeevne på niveau med klassiske algoritmer, samtidig med at de er mange gange hurtigere. Den nuværende deep-learning-baserede registrering er dog stadig baseret på en-til-en-matchning af billeder, og lider således af de samme mangler ved en ændring af anatomi.

Denne afhandling udforsker de nye muligheder som brugen af deep learning har givet til at overvinde samme-billede-antagelsen, såsom implicit at udlede en løsning fra observerede data med en uovervåget probabilistisk tilgang og informere optimeringen med semantiske billedrepræsentationer. Imidlertid introducerer brugen af deep learning til medicinsk billeddannelse også nye udfordringer. Afhandlingen bidrager til besvarelsen af de åbne forskningsspørgsmål vedrørende valg af en passende modelarkitektur til billedregistrering, kvantificering af modelusikkerhed ved tilstedeværelse af annotatorvariabilitet samt træning af few-shot-modeller i et klinisk forskningsmiljø.

Disse forskellige spor kombineres i en uovervåget model til påvisning af anatomiske forskelle ved siden af registreringstrinnet. Modellen fremhæver områder, hvor billedregistreringen er utilstrækkelig på grund af overtrædelser af samme-billede-antagelsen, og angiver, hvor der skal udvises yderligere omhu ved efterfølgende analyse og behandling.

Dette er en vigtig milepæl i retning af billedregistrering, som er fuldt opmærksom på fuldt topologi, og er en væsentlig byggesten mod en deformerbar registreringspipeline for billeder med forskelle i anatomi.

PUBLICATIONS

During my Ph.D., I explored multiple research directions, leading to a total of 7 manuscripts either published or currently under review. This cumulative dissertation includes 5 of them, listed in the order of their appearance. The included publications are numbered. Earlier or shortened versions of attached publications are listed as sub-points.

1. **Steffen Czolbe**, Paraskevas Pegios, Oswin Krause, and Aasa Feragen (2023). "Semantic similarity metrics for image registration". *Under revision for Medical Image Analysis.*
    - **Steffen Czolbe**, Oswin Krause, and Aasa Feragen (Jul 2021). "Semantic similarity metrics for learned image registration". In: *Medical Imaging with Deep Learning (MIDL 2021).*
    - **Steffen Czolbe**, Oswin Krause, and Aasa Feragen (Dec 2020). "DeepSim: Semantic similarity metrics for learned image registration". In: *NeurIPS workshop on Medical Imaging (MedNeurIPS 2020).*
2. Paraskevas Pegios and **Steffen Czolbe** (Jul 2022). "Can Transformers capture long-range displacements better than CNNs?". In: *Medical Imaging with Deep Learning (MIDL 2022).*
3. **Steffen Czolbe**, Aasa Feragen, and Oswin Krause (Dec 2021). "Spot the Difference: Detection of Topological Changes via Geometric Alignment". In: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021).*
4. **Steffen Czolbe**, Kasra Arnavaz, Oswin Krause, and Aasa Feragen (Jun 2020). "Is segmentation uncertainty useful?". In: *International Conference on Information Processing in Medical Imaging (IPMI 2021).*
5. **Steffen Czolbe** and Adrian Dalca (2023). "Neuralizer: Neuroimage Analysis without Re-Training". *Under review for CVPR.*

During my Ph.D., I published one paper based on research I performed primarily as a Master's student. I list this paper here to show the breadth and depth of my work, but it is not included in this thesis.

6. **Steffen Czolbe**, Oswin Krause, Ingemar Cox, and Christian Igel (Dec 2020). "A Loss Function for Generative Neural Networks Based on Watson's Perceptual Model". In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020).*

CONTENTS

Part I

SUMMARY

1

INTRODUCTION

Medical imaging encompasses multiple image acquisition and processing technologies that are used to view inside a living organism in order to diagnose, monitor, or treat medical conditions. Common medical imaging methods used millions of times a day to improve patient outcomes include X-ray, Ultrasound, CT, and MRI scanning.

Computational methods are an essential component of medical imaging for both image acquisition and analysis. The raw data recorded by CT and MRI scans require extensive computational processing to be displayed as an image interpretable by humans. Many medical imaging methods produce vast amounts of visual data, either in the form of dense three-dimensional images, or video sequences, that are time-intensive, costly, and error-prone to inspect manually. The computational processing and analysis of medical images help doctors to find important signals in the data, help patients by providing a better-informed treatment, and help medical researchers to find patterns across studies involving thousands of patients.

A common pre-processing step for many processing pipelines is image registration. Image registration aligns two or more images to a common space. It can be used to rigidly overlay x-ray images recorded months apart or to track the complex deformation patterns of the lung during a breathing cycle with dozens of images recorded over a few seconds.

A long-standing problem in deformable image registration is the registration of images whose topology or anatomy is not the same. Deformable image registration methods regularize the over-parameterized problem by assuming the images to be registered show the same anatomy. However, this "same-topology assumption" is often violated in medical imaging applications. The topology of the body can change through various processes, such as tumor growth, surgical resections, and the movement of fluids, and image registration algorithms have to be able to cope with these changes.

In paper 3, this thesis provides a step towards the accurate registration of images where the same-topology assumption is violated, by proposing a model to detect local violations alongside the registration step. Papers 1 and 2 lead up to this contribution, investigating

loss functions and model architectures for image registration. Papers 4 and 5 are supporting papers that explore related questions in medical imaging, namely the handling of annotation and model uncertainty, and the use of few-shot models in a clinical research environment with limited computational resources and expertise.

Following, I give a high-level introduction to image registration, the same-topology assumption, and related challenges in medical imaging. Detailed introductions to each topic are provided in the introduction sections of the included papers. In chapter 2, I highlight the scientific contributions of each of the five papers. Finally, chapter 3 provides a summary and perspectives for further research.

1.1 IMAGE REGISTRATION

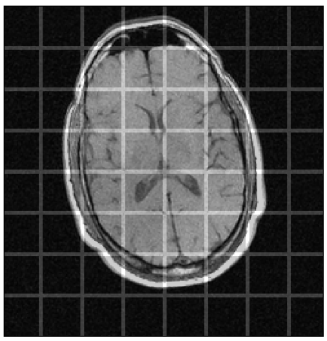
Image registration aligns two or more images through a geometric transformation on the image domain. Frequently used registration methods differ in their choice of geometric transformation, with rigid or affine registration limiting the transformation to the basic operations of translation, scale, rotation, and perspective changes (Szeliski, 2006). A wider range of motions can be modeled by non-linear transformations used in deformable image registration, which allow almost unrestricted and independent displacement of any point within the continuous image domain (Vercauteren et al., 2008; Arsigny et al., 2006). This non-linearity allows for the registration of complex movements found in the biomedical domain, such as deforming tissue or morphological shape changes. An example of a registration using affine and diffeomorphic transformations is shown in fig. 1.

Registration methods can further be divided into intensity-based (Balakrishnan et al., 2019; Avants et al., 2008) and landmark or keypoint-based (Younes, 2010; Hansen and Heinrich, 2021) methods. Intensity-based registration operates directly on the images and aims to match the pixel intensity values at all locations of the registered images. Keypoint-based methods first extract points of interest from the images and then aim to align the key points of both images.

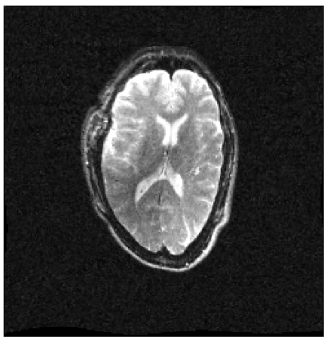
I will further focus on deformable, intensity-based image registration.

1.1.1 *Image registration framework*

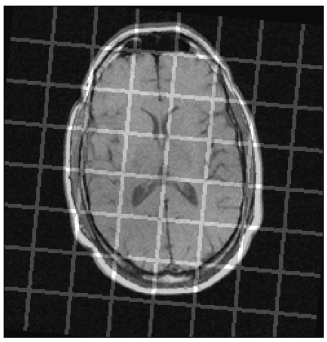
Most intensity-based deformable image registration frameworks model the problem as finding a transformation $\Phi : \Omega_T \to \Omega_S$ that aligns a moving source image $I_S : \Omega_S \to \mathbb{R}^d$ into the coordinate system of a fixed target image $I_T : \Omega_T \to \mathbb{R}^d$. The domain of the images



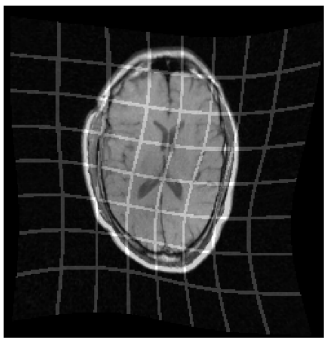
(a) Moving image I_S



(b) Target image I_T



(c) Affinely registered image $I_S \circ \Phi_{\text{affine}}$



(d) Deformably registered image $I_S \circ \Phi$

Figure 1: Example of a registration obtained with two transformation models. The moving image (a) is registered to the target image (b) using an affine transformation shown in (c) and a deformable, diffeomorphic transformation shown in (d). Figures from Guyader (2019).

is denoted by $\Omega_S, \Omega_T \subseteq \mathbb{R}^n$, with $n = 2$ for 2-dimensional images or $n = 3$ for volumetric images. The morphed moving image, obtained by application of the transformation via function composition, is denoted as $I_S \circ \Phi$.

The dissimilarity between the target and the morphed source image is expressed with a suitable distance measure $D : I_T \times I_M \to \mathbb{R}$. the mapping $\Phi \in \mathcal{G}$ is often chosen to be a field of dense displacement vectors, resulting in an over-parameterized model with multiple optimal solutions. To favor smooth transitions that avoid folds or gaps, a regularization term $R : \mathcal{G} \to \mathbb{R}$ is used. The optimal transformation Φ^* is found by minimizing the loss function

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}}\ D(I_S \circ \Phi, I_T) + R(\Phi)\,. \tag{1}$$

Most non-learned deformable registration algorithms iteratively optimize eq. (1) for each image pair (I_S, I_T) (Vercauteren et al., 2007; Faisal Beg et al., 2005; Avants et al., 2008).

Deep-learning-based image registration models instead train a parameterized model f_θ, which predicts a transformation conditioned on the moving and fixed image as

$$\Phi_\theta = f_\theta(I_S, I_T) \ . \tag{2}$$

The optimal parameters θ^* are found by minimizing eq. (1) on a dataset of image pairs $\mathcal{D}$ as

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{(I_S, I_T) \in \mathcal{D}} \left[D(I_S \circ \Phi_\theta, I_T) + R(\Phi_\theta) \right] \ , \tag{3}$$

which forms the training objective of the model. Applying a transformation to an image can be implemented efficiently and differentiably with regards to both inputs by a spatial transformer module (Jaderberg et al., 2015). The optimization of f_θ follows standard neural-network training techniques.

1.1.2 *The same-topology assumption*

To simplify the over-parameterized estimation of the transformation in eq. (1), the transformation is often constrained to be diffeomorphic, that is, bijective and continuously differentiable in both directions. In particular, diffeomorphic transformations are homeomorphic, or topology-preserving, which implies that a common topology is assumed across all images (Grenander and Miller, 1998; Faisal Beg et al., 2005). This relationship is formalized by a common template image I_{template}, from which all other images are obtained via a transformation Φ from the group of diffeomorphisms $\mathcal{G}$. Under this common topology assumption, the set of all images is given by

$$\mathcal{I} = \{ I_{\text{template}} \circ \Phi | \Phi \in \mathcal{G} \} \ .$$

In practice, this same-topology assumption is routinely violated. For example, when natural images are registered, occlusions and temporal changes introduce non-diffeomorphic changes to the image (Wang et al., 2018). In biomedical image registration, the assumption is violated when the studied anatomy differs from the "standard" anatomy, for example through pathologies or surgical intervention (Nielsen et al., 2019). As a result, most deformable registration methods can not register areas of images where the same-topology assumption is violated, potentially introducing bias in downstream analysis.

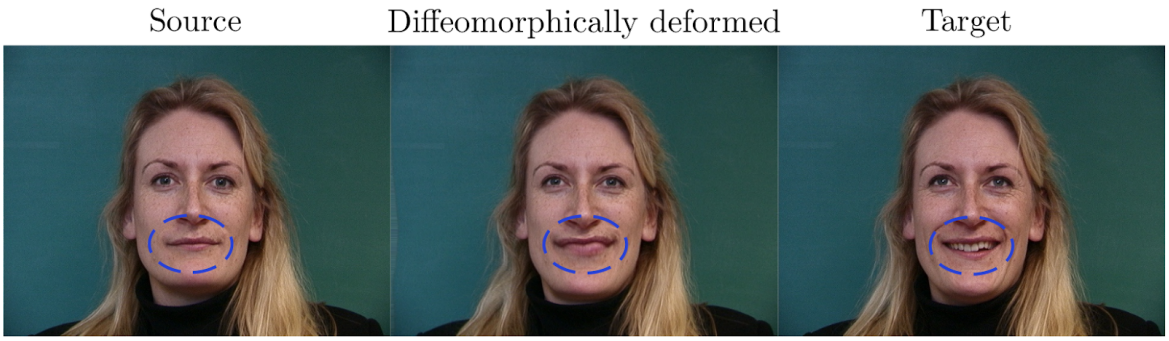


Figure 2: A source image (left) is registered to a target image (right). The same-topology assumption is violated by the teeth of the person, which alters the image topology. A diffeomorphic registration (middle) can not register this change in topology. As a consequence, the month is registered with strong local deformations, and parts of the lower lip is registered to the teeth. Example from Nielsen et al. (2019).

1.1.3 *Violations of the same-topology assumption*

An example of what happens when the same-topology assumption is violated in natural images is shown in fig. 2. The teeth of the woman in the target image are altering the image topology, and there is no fitting match for them in the source image. But a diffeomorphic transformation needs to be fully bijective and differentiable. As a result, the lips are registered to the teeth, with strong local deformations. This registration introduces both a wrong match (lips to teeth) and unnatural strong deformations (stretched lips). To obtain a correct match, the closed lips would have to be registered to the open lips, and the teeth do not have a match in the source image. However, such a transformation is not compatible with a diffeomorphism, as the transformation is no longer bijective and continuous.

1.2 DEEP LEARNING IN MEDICAL IMAGING

Many challenges to the application of deep-learning-based image registration models in medical practice are shared by the larger research field of medical image analysis. Medical image analysis is a subfield of computer vision, focusing on the computational processing and analysis of images created in a medical context. While the vision tasks in medical imaging are similar to the ones in general computer vision, additional challenges inherent to the domain distinguish medical image analysis methods and research from the larger computer vision community. These challenges are:

1. Data availability: High cost of data acquisition, patient privacy, limited data sharing.
2. Annotation availability: High cost of annotating data, high annotation-variability.

3. Image size: Large size of volumetric images limit parameter count and network complexity.

4. Accountability: Understanding the accuracy, reliability, fairness, reasoning, and uncertainty of methods is essential to inform medical treatment.

Many concepts in medical image analysis are an adaptation of general computer vision concepts but give special attention to these challenges unique to the medical domain.

1.2.1 *Annotation and model uncertainty*

In many medical imaging applications, the solution is often ambiguous, and sometimes multiple solutions to a problem can be correct. When asked individually, a group of doctors frequently disagrees in their assessment and labeling of a condition (Codella et al., 2019). This introduces uncertainty into the medical imaging model, and is a challenge for model evaluation.

Next to annotator variability, uncertainty can be introduced at multiple stages of the processing chain: Image acquisition methods can introduce noise (Quay et al., 2018), and the approximation models we build to perform a prediction can introduce uncertainty, often as a result of limited training data (Nguyen et al., 2019). Sources of uncertainty are often split into the uncertainty present in the data and annotations, called aleatoric uncertainty, and the uncertainty introduced by the model, called epistemic uncertainty. Understanding and quantifying uncertainty and variability are important steps toward analyzing deep learning models and their predictions.

1.2.2 *Multi-task and few-shot models*

A drawback of most deep-learning-based approaches is that each model is limited to solving the task it has been trained on, on the data it has been trained on. Yet, medical imaging contains a lot of different tasks and biomedical images from many domains, while the availability of computational resources, human expertise, and annotated data for training dedicated models for each setting is limited. In practice, this is a barrier to the adoption of deep learning methods, and classical algorithmic approaches to image segmentation and registration remain common in medical practice (Li et al., 2020).

Multi-Task Learning (MTL) models attempt to alleviate this problem by solving multiple prediction tasks with a single model. It exploits similarities between related tasks, thus achieving synergistic effects (Caruana et al., 1997). MTL can improve performance, lower data requirements, and reduce computational cost compared to designing task-specific solutions (Evgeniou and Pontil, 2004; Sener and Koltun, 2018). In medical imaging, MTL networks are frequently used to solve multiple prediction tasks on the same input image, for example simultaneous segmentation and classification of an image (Gupta et al., 2021; Díaz-Pernas et al., 2021).

Few-shot models attempt to lower the data requirement of adopting a deep learning model to a new task. They infer predictions from just a few labeled examples provided at test-time (Ravi and Larochelle, 2017; Wang et al., 2020; Liu et al., 2019b; Schonfeld et al., 2019). Several methods pass a query image, along with an additional set of support images and labels as input to the model to inform a prediction (Bian et al., 2022; Feyjie et al., 2020; Feng et al., 2021). For example, few-shot image segmentation methods (Liu et al., 2020; Zhang et al., 2019b) use single image-label pairs (Zhang et al., 2019a; Li et al., 2021) as support, thus requiring as little as one annotated image for segmenting an image of a pre-trained domain.

2

SCIENTIFIC CONTRIBUTIONS

After giving a general introduction to the topics of this thesis, this section summarizes the scientific contributions of each paper. The core contributions are grouped into three areas, shown in fig. 3. Papers 1 and 2 are on challenges in general image registration, leading up to paper 3 on the detection of topological changes alongside the registration. Paper 4 investigates uncertainty quantification with applications in image segmentation, which inspired the approach in paper 3. Paper 5, written during my stay at the Martinos Center, proposes a novel method for few-shot, multi-task generalization.

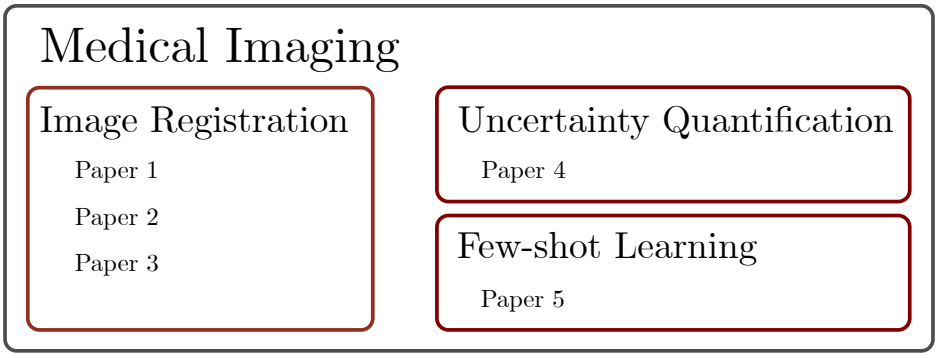


Figure 3: The presented papers are categorized into the areas of image registration, uncertainty quantification, and few-shot learning.

2.1 PAPER 1: A SEMANTIC SIMILARITY METRIC FOR IMAGE REGISTRATION

Image registration models find correspondences between images. Most algorithmic and deep learning-based methods solve the registration problem by the minimization of a loss function consisting of a similarity metric and a regularization term encuraging the smoothness of the transformation. The similarity metric is essential to the optimization, as it judges the quality of the match between registered images and has a strong influence on the result.

Pixel-based similarity metrics like euclidean distance and patch-wise cross-correlation are widely used within algorithmic and deep-learning-based image registration. These metrics assume that if the image intensities are aligned, or strongly correlated, the images are well aligned. This assumption can be incorrect if noise or contrast differences are present in the data. Further shortcomings of pixel-based similarity metrics have been studied in the image generation community, including my published master's thesis (Czolbe et al., 2020). In the wider computer vision community, pixel-based metrics have been superseded by deep similarity metrics (Hou et al., 2017; Zhang et al., 2018).

Contribution

I propose a data-driven similarity metric for image registration based on the alignment of learned, task-specific semantic features. The experimental results illustrate that the method is robust toward image noise and achieves consistently favorable tradeoffs between registration accuracy and transformation smoothness. I evaluate the method using deep-learning-based image registration with U-Net (Ronneberger et al., 2015; Balakrishnan et al., 2019) and Transformer models (Chen et al., 2022), and classical registration using the SyN algorithm (Avants et al., 2008).

To learn filters of semantic importance to the dataset, I present both an unsupervised approach using auto-encoders, and a semi-supervised approach using a segmentation model. I use the learned features to construct a similarity metric used for training a registration model, and validate my approach on three biomedical datasets of different image modalities and applications.

Finally, I perform an extensive ablation study to evaluate the influence of individual feature layers, model architectures, order of operations of the proposed loss, and the possibility of using transfer learning in the absence of a dataset-specific semantic model. I re-used this semantic similarity metric to great effect in paper 3.

2.2 PAPER 2: MODEL ARCHITECTURES FOR IMAGE REGISTRATION

One consequence of introducing deep learning into image registration is the need to choose an architecture for the model. While early works on applying deep learning to the image registration problem stayed clear of exploring architectures (Yang et al., 2017; Balakrishnan et al., 2019), instead mostly focusing on setting up the training objective and finding the right representation of the transformation, the model architecture soon became another factor to consider for optimizing learned registration methods.

During my work on image registration, vision transformer models utilizing the attention mechanism came into prominence. When trained on large amounts of data, these transformer models outperformed prior CNN-based networks (Chen et al., 2021a, 2022; Wang and Delingette, 2021; Zhang et al., 2021). These papers argue that this performance difference results from the transformers' improved ability to accurately predict long-range displacement vectors, something that CNN-based methods struggle with unless a hierarchical approach is taken (Hering et al., 2019; Liu et al., 2019a; Hu et al., 2019).

Contribution

In paper 2, I and my co-author provide the first experimental evaluation of this claim. We find no evidence to support the claim.

We compare the U-Net-based Voxelmorph model (Balakrishnan et al., 2019) to two adaptations of transformer models to image registration: the ViT-net (Dosovitskiy et al., 2021) based approach titled "ViT-V-Net" (Chen et al., 2021a), and the shifted-window (Liu et al., 2021) based approach titled "Transmorph" (Chen et al., 2022).

To allow for the evaluation of the registration accuracy as a function of displacement length, we select a dataset with annotated keypoint coordinates for this study. We use the "Learn2Reg: CT Lung Registration" dataset of 30 image pairs of inhaling and exhaling CT thorax images (Hering et al., 2022). The key point annotations are only used for evaluation, and not for training the models.

In the experimental evaluation, we find no evidence of the claim that transformers are better than CNNs at registering long displacements. While the transformers outperformed the CNN slightly on average registration error, we find this difference stemmed from a more accurate registration of smaller displacements. On large displacements of $> 20mm$, no statistically significant difference in registration performance was observed.

These results refute a common claim made in the registration transformer literature, by providing the first experimental investigation into registration error by displacement length. Since the paper was published, other works have further questioned the validity of many of the claims made about transformer architectures over CNNs in computer vision applications (Park and Kim, 2021; Liu et al., 2022). The findings of this paper informed and validate my choice of network architecture in papers 1, 3, and 4.

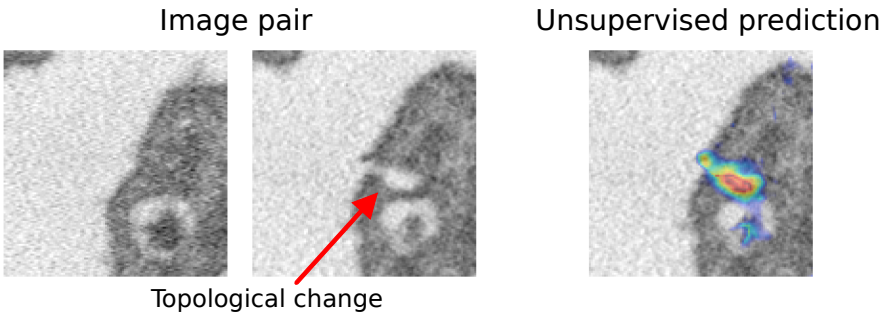


Figure 4: In paper 3, I propose a model for the unsupervised detection of topological changes alongside the registration. Left: Highlighted change of topology between two adjacent slices of a 3d image. Right: Heatmap of the likelihood of topological changes predicted by the unsupervised model.

2.3 PAPER 3: DETECTING TOPOLOGICAL CHANGES IN IMAGE REGISTRATION

A challenge in image registration is the alignment of domains whose topology is not the same. This violates the same-image assumption laid out in the introduction. In biomedical image registration, this problem can be caused by a variety of processes. For example, when the studied anatomy differs from "standard" anatomy (Nielsen et al., 2019), image slices obtained from a volume do not all contain the same elements, or after the surgical removal of tissue. Despite being extremely common, this problem is routinely ignored or accepted as inevitable, potentially introducing bias in downstream analysis.

Contribution

In paper 3, I propose an unsupervised algorithm for the detection of changes in image topology. To achieve this, I train a conditional variational autoencoder for predicting image-to-image alignment, obtaining a per-target-pixel probability of being obtained from the moving image via diffeomorphic transformation. I combine the semantic loss function from paper 1 with a learnable prior on the space of transformations (Dalca et al., 2018), allowing me to incorporate both the reconstruction error, as well as knowledge about the expected transformation strength. Following the learning of paper 2, the model is entirely convolution based. An example is given in fig. 4.

I test the validity of my approach on a dataset of cell slices with annotated topological changes and on the proxy task of unsupervised brain tumor detection. I also validate my approach by investigating a spatial "topological inconsistency likelihood", and showing that this likelihood is higher in regions where topological inconsistencies are known to be common. My model is able to detect topological inconsistencies with a purely registration-

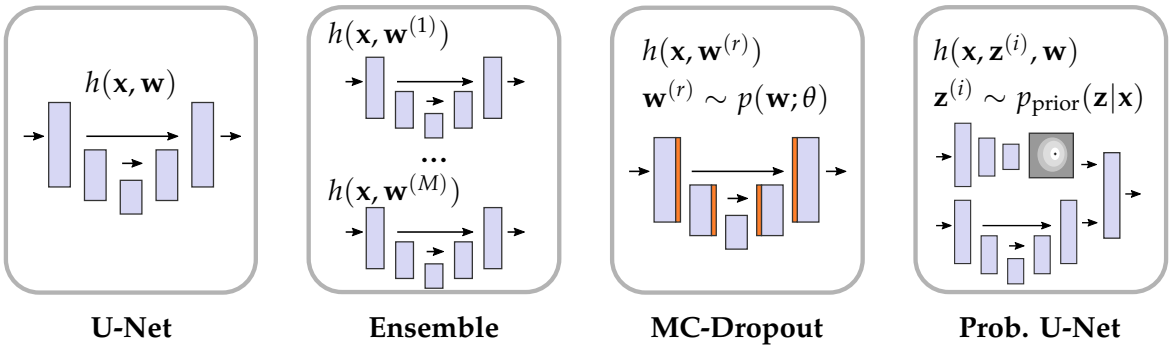


Figure 5: Models for estimating segmentation uncertainty alongside the segmentation. Figure adapted from Kohl et al. (2018). Blue: residual blocks (He et al., 2016). Orange: Dropout layers (Srivastava et al., 2014) essential to the networks' functionality.

driven framework, and thus allows me to know where the assumption of a common topology is violated.

2.4 PAPER 4: UNCERTAINTY IN IMAGE SEGMENTATION

The deep-learning methods used in papers 1-3 rely on manual annotations of medical images to train or evaluate the networks. Manual annotations are costly and time intensive to obtain, and annotations obtained from different annotators often show large variations (Armato III et al., 2011). In paper 4, I investigate how deep learning models can learn from uncertain annotations and quantify uncertainty in their own predictions. My co-author uses the uncertainty estimates as a sample selection strategy to reduce total annotation needs with an active learning framework.

Contribution

To estimate the uncertainty along with the segmentation, multiple modifications to segmentation networks have been proposed, shown in fig. 5. I evaluated a simple U-Net (Ronneberger et al., 2015), an ensemble of U-Nets, MC-Dropout (Gal, Yarin and Ghahramani, 2016) and the probabilistic U-Net (Kohl et al., 2018). While all these models are able to estimate per-pixel uncertainty, the last three models are also able to also propose alternative segmentation masks.

I investigate the degree to which the predicted uncertainty correlates with the error of the prediction. I find that the uncertainty estimates of all models correlate strongly with both segmentation errors and the uncertainty among a set of expert annotators. Surprisingly, the model architecture used does not have a strong influence on the quality of estimates, with even a simple U-Net giving good pixel-level uncertainty estimates.

Mu co-author investigated the potential for uncertainty estimates to be used for selecting samples for annotation in an active learning framework. We find that there are many pitfalls to an uncertainty-based data selection strategy. For example, in an experiment with multiple annotators, the images with the highest model uncertainty were precisely those images where the annotators were also uncertain. Labeling these ambiguous images by a group of expert annotators yielded conflicting ground truth annotations, providing little certain evidence for the model to learn from, which meant that adding more samples of those uncertain images could not reduce model uncertainty. A more differentiated view of aleatoric and epistemic uncertainty (uncertainty stemming from the data and annotations, instead of the model) might be necessary.

2.5 PAPER 5: FEW-SHOT MULTI-TASK GENERALIZATION FOR CLINICAL RESEARCH

While deep learning methods in medical imaging are more accurate and faster than classical approaches, the adoption of deep learning within clinical research is hindered by the large up-front investment and knowledge required to develop and train deep learning models. This problem is amplified by the many processing tasks present in clinical research, and image characteristics and quality vary depending on the image acquisition site and method. The generalization of deep learning models to new tasks and domains, such as different acquisition protocols or new segmentation targets, remains a barrier to adoption (Li et al., 2020).

Each deep learning model is limited to solving the task it has been trained on, on the data it has been trained on. Performing tasks like segmentation, registration, or reconstruction require different models for each processing step, despite operating on the same input data and methods exhibiting strong similarities in network architecture (Ronneberger et al., 2015; Hoffmann et al., 2020; Billot et al., 2021). Yet, designing and training models to solve these tasks in each application domain is expensive, and the resources required to do so – clinical expertise, deep learning knowledge, large annotated datasets, and specialized graphics processing hardware – are often not present.

Contribution

To remove the need for training or fine-tuning task or dataset-specific models in clinical research, I introduce a general-purpose few-shot multi-task model that, given a set of examples at inference, can solve a broad range of image processing tasks without the need for task-specific training or fine-tuning.

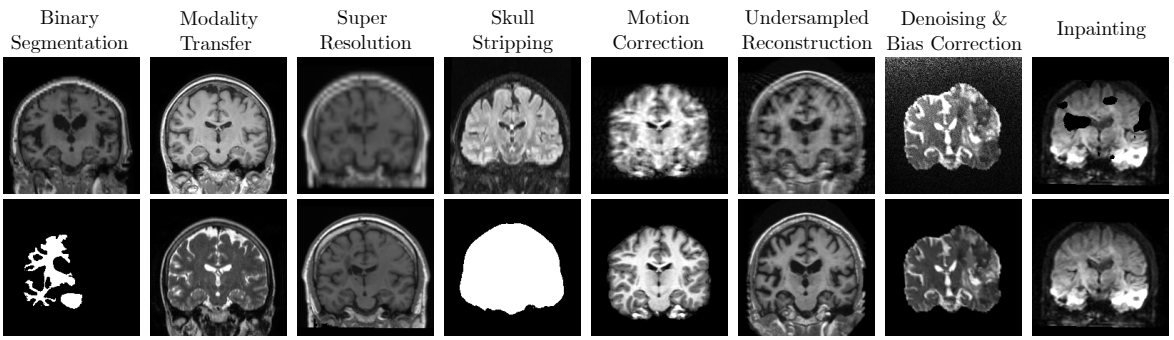


Figure 6: In clinical research, researchers have to solve a large range of image-processing tasks, often on multiple input domains. As an example, the image pictures 8 tasks from a Brain-MRI processing pipeline, covering images from multiple acquisition sites, modalities, and protocols. Training dedicated machine learning models for each task and input modality is infeasible, necessitating new solutions.

The model uses a novel model architecture, that takes as input a context set of examples to inform the processing task, and thus does not require the prior definition of the tasks. The method enables single-pass generalization during inference and can process any number of reference images in a single pass to inform the prediction.

I apply the method to neuroimaging, where I solve 8 brain-MRI processing tasks across images of 7 modalities with a single model and generalize with minimal loss in performance to non-trained tasks. A representative set of tasks is given in fig. 6.

I evaluate the model by comparing its single-pass, multi-task generalization performance to task-specific baselines conditioned on an equivalent amount of data. I find that the method outperforms task-specific baselines on tasks where ≤ 32 labeled examples are available, despite never training on the task. When generalizing across segmentation protocols, the method matches the performance of baselines trained directly on the dataset.

3

CONCLUSIONS AND PERSPECTIVES FOR FURTHER RESEARCH

In this chapter, I briefly summarize the main contribution of each paper in the context of the larger research field and offer perspectives for further research directions.

3.1 IMAGE REGISTRATION

In paper 1, I compare multiple similarity metrics for image registration, and propose a new, semantic metric, utilizing dataset-specific learned features. My experiments differ from most of the published work by evaluating the method on multiple datasets, and showing improved performance on noisy data. The proposed method lays the groundwork for paper 3, where the use of this semantic similarity metric boosted performance by a large margin.

Similarity metrics for image registration are a well-established research field. Even before deep learning was used for the registration, multiple works explored learning image descriptors to drive algorithmic registration, often using a supervised approach (Haskins et al., 2019; Cheng et al., 2018; Simonovsky et al., 2016). The work by Pielawski et al. (2020), published parallel to mine, proposed a related approach using an unsupervised semantic similarity metric. They extended their method to multimodal image registration, however, their registration is performed by classical algorithmic method.

While each of these papers shows that deep similarity metrics can improve registration using standard registration methods and limited datasets, the use of these metrics in image registration competitions remains low (Hering et al., 2022). Combining the advanced and often dataset-specific models winning these challenges with semantic similarity metrics is a promising future avenue for increasing performance.

In paper 2, I and my co-author compare CNN and transformer-type architectures for image registration networks. The paper was well received by the community, as it offers a novel comparison of these architectures for image registration by evaluating error as a

function of displacement length. Overall skepticism towards transformer models in the vision community remains high. Park and Kim (2021) supports the finding that long-range data dependency does not provide an observable advantage in practice. Liu et al. (2022) show that well-designed CNNs can match, or even outperform, vision transformer models. Finding and evaluating new model architectures for specific tasks remains a staple research direction with plenty of opportunity to publish, but well-founded insights remain elusive in this area.

In paper 3, I develop a novel way for the detection of local violations to the same-image assumption, which is inherent to all fully diffeomorphic image registration models. Currently, to treat violations to this assumption, piecewise diffeomorphic models use explicit annotations (Nielsen et al., 2019; Li et al., 2012) or domain-specific, often supervised modeling (Risser et al., 2013; Delmon et al., 2013; Pace et al., 2013; Chen et al., 2021b; Schmidt-Richberg et al., 2012) of areas of topological change to exclude these areas from the registration. My method improves on unsupervised and domain-independent detection methods (Li and Wyatt, 2010) by a large margin, providing a building block towards the first fully end-to-end trained piecewise-diffeomorphic registration model. Building such a model remains an interesting research direction, however, compounding errors from combining multiple error-prone methods, and the many degrees of freedom of such a model remain a challenge.

3.2 DEEP LEARNING IN MEDICAL IMAGING

Papers 4 and 5 focus on more general problems in medical imaging.

In paper 4, I investigate uncertainty quantification methods for image segmentation and my co-author attempts to use these uncertainty estimates as a sample selection strategy for active learning. The experiments show that even a simple U-net is competitive for assessing segmentation uncertainty as a proxy for likely segmentation error. Other comparison studies come to a similar result (Jungo and Reyes, 2019). Using the uncertainty estimates for active learning was unsuccessful, as the model would repeatedly select samples of high data uncertainty, not high model uncertainty. Nguyen et al. (2019) argue that a future research direction would be to separate aleatoric and epistemic uncertainty, and only use the epistemic, model-induced uncertainty as a selection strategy.

In paper 5, I proposed a novel model for rapid few-shot, single-pass, multi-task generalization to solve a wide range of medical imaging tasks in neuroimaging. The work improves on prior multi-task models by solving more tasks than any prior publication in the neuroimaging domain and requiring no prior definition of or training on the set of tasks. The research was performed alongside Dalca and Others (2023), who continue to work on the method to solve image segmentation for segmentation tasks across all biomedical domains with a single model. As the first paper demonstrating the potential of this new method, I made simplifying assumptions, such as affinely pre-aligning images, excluding lesion segmentation from the scope of tasks, and performing the experiments on 2d slices. More engineering, compute, and research is required to develop it into a product useable by clinical researchers.

4

REFERENCES

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, and Eric A Hoffman. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics*, 38(2):915–931, 2011.

Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 924–931. Springer Verlag, 2006.

B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, feb 2008.

Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, feb 2019.

Cheng Bian, Chenglang Yuan, Kai Ma, Shuang Yu, Dong Wei, and Yefeng Zheng. Domain Adaptation Meets Zero-Shot Learning: An Annotation-Efficient Approach to Multi-Modality Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 41(5): 1043–1056, may 2022.

Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, and Juan Eugenio Iglesias. SynthSeg: Domain Randomisation for Segmentation of Brain Scans of any Contrast and Resolution. jul 2021.

Rich Caruana, Lorien Pratt, and Sebastian Thrun. Multitask Learning. *Machine Learning 1997 28:1*, 28(1):41–75, 1997.

Junyu Chen, Yufan He, Eric C. Frey, Ye Li, and Yong Du. ViT-V-Net: Vision Transformer for Unsupervised Volumetric Medical Image Registration. *Medical Imaging with Deep Learning*, 2021a.

Junyu Chen, Yong Du, Yufan He, William P. Segars, Ye Li, and Eric C. Frey. TransMorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82, nov 2022.

Xiang Chen, Nishant Ravikumar, Yan Xia, and Alejandro F Frangi. A Deep Discontinuity-Preserving Image Registration Network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, jul 2021b.

Xi Cheng, Li Zhang, and Yefeng Zheng. Deep similarity learning for multimodal medical images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):248–252, may 2018.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). feb 2019.

Steffen Czolbe, Oswin Krause, Ingemar Cox, and Christian Igel. A Loss Function for Generative Neural Networks Based on Watson's Perceptual Model. *Advances in Neural Information Processing Systems*, jun 2020.

Adrian V. Dalca and Others. Universeg: Universal Medical Image Segmentation. *Draft Document*, 2023.

Adrian V. Dalca, Guha Balakrishnan, John Guttag, and Mert R. Sabuncu. Unsupervised Learning for Fast Probabilistic Diffeomorphic Registration. *Medical Image Computing and Computer Assisted Intervention*, pages 729–738, may 2018.

V Delmon, S Rit, R Pinho, and D Sarrut. Registration of sliding objects using direction dependent B-splines decomposition Registration of sliding objects using direction dependent B-splines decomposition *. *Phys. Med. Bio*, 58(5):1303–1314, 2013.

Francisco Javier Díaz-Pernas, Mario Martínez-Zarzuela, David González-Ortega, and Míriam Antón-Rodríguez. A Deep Learning Approach for Brain Tumor Classification and Segmentation Using a Multiscale Convolutional Neural Network. *Healthcare*, 9 (2):153, feb 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *Proceedings of the 38th International Conference on Machine Learning*, oct 2021.

Theodoras Evgeniou and Massimiliano Pontil. Regularized multi-task learning. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117, 2004.

Mirza Faisal Beg, Michael I Miller, Alain Trouvétrouv, and Laurent Younes. Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms. *International Journal of Computer Vision*, 61(2):139–157, 2005.

Ruiwei Feng, Xiangshang Zheng, Tianxiang Gao, Jintai Chen, Wenzhe Wang, Danny Z. Chen, and Jian Wu. Interactive Few-Shot Learning: Limited Supervision, Better Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 40(10):2575–2588, oct 2021.

Abdur R Feyjie, Reza Azad, Marco Pedersoli, Claude Kauffman, Ismail Ben Ayed, and Jose Dolz. Semi-supervised few-shot learning for medical image segmentation. mar 2020.

Zoubin Gal, Yarin and Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning*, pages 1050—-1059, 2016.

Ulf Grenander and Michael I Miller. Computational anatomy: An emerging discipline. *Quarterly of applied mathematics*, 56:617—-694, 1998.

Sachin Gupta, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. MAG-Net: Multi-task Attention Guided Network for Brain Tumor Segmentation and Classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13147 LNCS:3–15, 2021.

Jean-marie Guyader. *Advanced Medical Image Registration Methods for Quantitative Imaging and Multi-Channel Images*. PhD thesis, University Medical Centre Rotterdam, the Netherlands, 2019.

Lasse Hansen and Mattias P. Heinrich. GraphRegNet: Deep Graph Regularisation Networks on Sparse Keypoints for Dense Registration of 3D Lung CTs. *IEEE Transactions on Medical Imaging*, 40(9):2246–2257, sep 2021.

Grant Haskins, Jochen Kruecker, Uwe Kruger, Sheng Xu, Peter A. Pinto, Brad J. Wood, and Pingkun Yan. Learning deep similarity metric for 3D MR–TRUS image registration. *International Journal of Computer Assisted Radiology and Surgery*, 14(3):417–425, mar 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Alessa Hering, Bram van Ginneken, and Stefan Heldmann. mlVIRNET: Multilevel Variational Image Registration Network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11769 LNCS: 257–265, 2019.

Alessa Hering, Lasse Hansen, Tony C.W. Mok, Albert C.S. Chung, Hanna Siebert, Stephanie Hager, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, Sulaiman Vesal, Mirabela Rusu, Geoffrey Sonn, Theo Estienne, Maria Vakalopoulou, Luyi Han, Yunzhi Huang, Pew Thian Yap, Mikael Brudfors, Yael Balbastre, Samuel Joutard, Marc Modat, Gal Lifshitz, Dan Raviv, Jinxin Lv, Qiang Li, Vincent Jaouen, Dimitris Visvikis, Constance Fourcade, Mathieu Rubeaux, Wentao Pan, Zhe Xu, Bailiang Jian, Francesca De Benetti, Marek Wodzinski, Niklas Gunnarsson, Jens Sjolund, Daniel Grzech, Huaqi Qiu, Zeju Li, Alexander Thorley, Jinming Duan, Christoph Grossbrohmer, Andrew Hoopes, Ingerid Reinertsen, Yiming Xiao, Bennett Landman, Yuankai Huo, Keelin Murphy, Nikolas Lessmann, Bram Van Ginneken, Adrian V. Dalca, and Mattias P. Heinrich. Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 2022.

Malte Hoffmann, Benjamin Billot, Douglas N. Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V. Dalca. SynthMorph: learning contrast-invariant registration without acquired images. *IEEE Transactions on Medical Imaging*, apr 2020.

Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *Winter Conference on Applications of Computer Vision*, pages 1133–1141. IEEE, 2017.

Xiaojun Hu, Miao Kang, Weilin Huang, Matthew R. Scott, Roland Wiest, and Mauricio Reyes. Dual-Stream Pyramid Registration Network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 382–390. 2019.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial Transformer Networks. *Advances in Neural Information Processing Systems*, 2015.

Alain Jungo and Mauricio Reyes. Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11765 LNCS:48–56, 2019.

Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A Probabilistic U-Net for Segmentation of Ambiguous Images. *Advances in Neural Information Processing Systems*, 31:6965–6975, jun 2018.

Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive Prototype Learning and Allocation for Few-Shot Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8334–8343, 2021.

Xiaoxiao Li, Yufeng Gu, Nicha Dvornek, Lawrence H. Staib, Pamela Ventola, and James S. Duncan. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Medical Image Analysis*, 65:101765, oct 2020.

Xiaoxing Li and Chritopher Wyatt. Modeling topological changes in deformable registration. In *2010 7th IEEE International Symposium on Biomedical Imaging*, pages 360–363, 2010.

Xiaoxing Li, Xiaojing Long, Christopher Wyatt, and Paul Laurienti. Registration of Images with Varying Topology Using Embedded Maps. *IEEE Transactions on Medical Imaging*, 31 (3):749–765, mar 2012.

Lihao Liu, Xiaowei Hu, Lei Zhu, and Pheng-Ann Heng. Probabilistic Multilayer Regularization Network for Unsupervised 3D Brain Image Registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 346–354. 2019a.

Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-Shot Unsupervised Image-to-Image Translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10551–10560, 2019b.

Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. CRNet: Cross-Reference Networks for Few-Shot Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4165–4173, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.

Proceedings of the IEEE/CVF international conference on computer vision, pages 10012–10022, mar 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, Saining Xie, and Facebook AI Research. A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

Vu Linh Nguyen, Sébastien Destercke, and Eyke Hüllermeier. Epistemic Uncertainty Sampling. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11828 LNAI:72–86, 2019.

Rune Kok Nielsen, Sune Darkner, and Aasa Feragen. TopAwaRe: Topology-Aware Registration. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 364–372, 2019.

Danielle F. Pace, Stephen R. Aylward, and Marc Niethammer. A locally adaptive regularization based on anisotropic diffusion for deformable image registration of sliding organs. *IEEE Transactions on Medical Imaging*, 32(11):2114–2126, 2013.

Namuk Park and Songkuk Kim. How Do Vision Transformers Work? *International Conference on Learning Representations (ICLR 2021)*, 2021.

Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, Jiahao Lu, Carolina Wählby, Joakim Lindblad, and Nataša Sladoje. CoMIR: Contrastive Multimodal Image Representation for Registration. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, jun 2020.

Matthew Quay, Zeyad Emam, Adam Anderson, and Richard Leapman. Designing deep neural networks to automate segmentation for serial block-face electron microscopy. In *International Symposium on Biomedical Imaging*, volume 2018-April, pages 405–408. IEEE Computer Society, may 2018.

Sachin Ravi and Hugo Larochelle. Optimization as a Model for Few-Shot Learning. In *International Conference on Learning Representations (ICLR)*, jul 2017.

Laurent Risser, François Xavier Vialard, Habib Y. Baluwala, and Julia A. Schnabel. Piecewise-diffeomorphic image registration: Application to the motion estimation between 3D CT lung images with sliding conditions. *Medical Image Analysis*, 17(2):182–193, feb 2013.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 9351, pages 234–241. Springer Verlag, 2015.

Alexander Schmidt-Richberg, Jan Ehrhardt, René Werner, and Heinz Handels. Fast explicit diffusion for registration with direction-dependent regularization. *Biomedical Image Registration*, 7359:220–228, 2012.

Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8247–8255, 2019.

Ozan Sener and Vladlen Koltun. Multi-Task Learning as Multi-Objective Optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Martin Simonovsky, Benjamín Gutiérrez-Becker, Diana Mateus, Nassir Navab, and Nikos Komodakis. A deep metric for multimodal registration. In *Lecture Notes in Computer Science*, volume 9902 LNCS, pages 10–18. Springer Verlag, oct 2016.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

Richard Szeliski. Image Alignment and Stitching: A Tutorial. *Foundations and Trends in Computer Graphics and Vision*, 2(1):1–104, 2006.

Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Non-parametric diffeomorphic image registration with the demons algorithm. In *Lecture Notes in Computer Science*, volume 4792 LNCS, pages 319–326, 2007.

Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Symmetric Log-Domain Diffeomorphic Registration: A Demons-based Approach. *Medical Image Computing and Computer Assisted Intervention*, pages 754–761, 2008.

Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion Aware Unsupervised Learning of Optical Flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4884–4893, 2018.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a Few Examples. *ACM Computing Surveys*, 53(3), jun 2020.

Zihao Wang and Hervé Delingette. Attention for Image Registration (AiR): an unsupervised Transformer approach. may 2021.

Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration - A deep learning approach. *NeuroImage*, 158:378–396, sep 2017.

Laurent Younes. *Shapes and Diffeomorphisms*. 2010.

Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid Graph Networks With Connection Attentions for Region-Based One-Shot Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9587–9595, 2019a.

Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. CANet: Class-Agnostic Segmentation Networks With Iterative Refinement and Attentive Few-Shot Learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5217–5226, 2019b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *Conference on Computer Vision and Pattern Recognition*, pages 586–595, jan 2018.

Yungeng Zhang, Yuru Pei, and Hongbin Zha. Learning Dual Transformer Network for Diffeomorphic Registration. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 129–138, sep 2021.

Part II

PUBLICATIONS

5

SEMANTIC SIMILARITY METRICS FOR IMAGE REGISTRATION

PAPER 1:

Steffen Czolbe, Paraskevas Pegios, Oswin Krause, and Aasa Feragen (2023). "Semantic similarity metrics for image registration". *Under revision for Medical Image Analysis.*

The version of this work presented here is currently under revision for the *Medical Image Analysis* journal. Two earlier versions of this work were published at a workshop and a conference, listed below. The paper won runner-up to the best paper award at MIDL 2021.

- **Steffen Czolbe**, Oswin Krause, and Aasa Feragen (Jul 2021). "Semantic similarity metrics for learned image registration". In: *Medical Imaging with Deep Learning (MIDL 2021).*
- **Steffen Czolbe**, Oswin Krause, and Aasa Feragen (Dec 2020). "DeepSim: Semantic similarity metrics for learned image registration". In: *NeurIPS workshop on Medical Imaging (MedNeurIPS 2020).*

Medical Image Analysis (2023)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Semantic similarity metrics for image registration

Steffen Czolbe[a,*], Paraskevas Pegios[b], Oswin Krause[a], Aasa Feragen[b]

[a]*Department of Computer Science, University of Copenhagen, Denmark*
[b]*DTU Compute, Technical University of Denmark, Denmark*

ARTICLE INFO

Article history:

Keywords: Image Registration, Deep Learning, Representation Learning

ABSTRACT

Image registration aims to find geometric transformations that align images. Most algorithmic and deep learning-based methods solve the registration problem by minimizing a loss function, consisting of a similarity metric comparing the aligned images, and a regularization term ensuring smoothness of the transformation. Existing similarity metrics like Euclidean Distance or Normalized Cross-Correlation focus on aligning pixel intensity values or correlations, giving difficulties with low intensity contrast, noise, and ambiguous matching. We propose a semantic similarity metric for image registration, focusing on aligning image areas based on semantic correspondence instead. Our approach learns dataset-specific features that drive the optimization of a learning-based registration model. We train both an unsupervised approach extracting features with an auto-encoder, and a semi-supervised approach using supplemental segmentation data. We validate the semantic similarity metric using both deep-learning-based and algorithmic image registration methods. Compared to existing methods across four different image modalities and applications, the method achieves consistently high registration accuracy and smooth transformation fields.

1. Introduction

Deformable registration, or nonlinear image alignment, is a fundamental tool in medical imaging to capture local deformations or changes between images. Applications include tracking disease progression (Yang et al., 2020; Castillo et al., 2013; Nielsen et al., 2019), population analysis (LaMontagne et al., 2019), co-registration of image modalities (Song et al., 2021; Lee et al., 2019a), object tracking (Ulman et al., 2017), and guiding of medical machinery (Trofimova et al., 2020). The registration model finds correspondences between a set of images and derives a geometric transformation to align them. Most algorithmic and deep learning-based methods solve the registration problem by the minimization of a loss function consisting of a similarity metric and a regularization term ensuring smoothness of the transformation. The similarity metric is essential to the optimization; it judges the quality of the match between registered images and has a strong influence on the result.

Pixel-based similarity metrics like euclidean distance and patch-wise cross-correlation are well explored within algorithmic and deep-learning-based image registration. These metrics assume that if the image intensities are aligned, or strongly correlated, the images are well aligned. Each choice of metric adds additional assumptions on the characteristics of the specific dataset. Thus, a common methodological approach is to trial registration models with multiple different pixel-based metrics, and choose the metric performing best on the dataset (Balakrishnan et al., 2019; Hu et al., 2019a).

The shortcomings of pixel-based similarity metrics have been studied substantially in the image generation community (Hou et al., 2017; Zhang et al., 2018), where they have been superseded by deep similarity metrics approximating human visual perception. Here, image representations are commonly ex-

*Corresponding author. *e-mail:* per.sc@di.ku.dk (Steffen Czolbe)

tracted by neural networks pre-trained on image-classification tasks (Deng et al., 2009). Performance can be further improved by fine-tuning the representation to human perception (Czolbe et al., 2020; Zhang et al., 2018). These representation-based deep similarity metrics have improved the visual quality of images generated with variational auto-encoders considerably. As image registration is a conditional generative problem (Dalca et al., 2018a; Czolbe et al., 2021b), we propose to apply deep similarity metrics within image registration to achieve a similar increase in performance for registration models.

Contributions. We propose a data-driven similarity metric for image registration based on the alignment of learned, task-specific semantic features. The experimental results illustrate that the method is robust toward image noise and achieves consistently favorable tradeoffs between registration accuracy and transformation smoothness. We evaluate the method using deep-learning based image registration with U-Nets (Ronneberger et al., 2015; Balakrishnan et al., 2019) and Transformers (Chen et al., 2022), and classical registration using the SyN algorithm of the ANTS package (Avants et al., 2008b).

To learn filters of semantic importance to the dataset, we present both an unsupervised approach using auto-encoders, and a semi-supervised approach using a segmentation model. We use the learned features to construct a similarity metric used for training a registration model, and validate our approach on four biomedical datasets of different image modalities and applications. For both methods and across all datasets, our method achieves consistently high registration accuracy and smooth transformation fields.

Finally, we perform an extensive ablation study to evaluate the influence of individual feature layers, model architectures, order of operations of the proposed loss, the possibility of using transfer learning in the absence of a dataset-specific semantic model, and the robustness toward noise in the images.

Previous prublications. Part of this work has been published at the Medical Imaging with Deep Learning (MIDL) conference (Czolbe et al., 2021c). This journal release contains an extended experimental evaluation, a fourth dataset, deeper discussion, and a broadened background section. We demonstrate the applicability of our method using the recently published state-of-the-art transformer TransMorph, and the well-established SyN algorithm. In addition, the popular similarity metrics of mutual information (Studholme et al., 1999) and MIND-SSC (Heinrich et al., 2013b), further called MIND, have been included as baselines in all experiments. A new ablation study section discusses multi-task pre-training, demonstrates transfer learning in the absence of a dataset-specific semantic feature extractor, and provides insights into multi-level feature learning by evaluating how different levels contribute to the registration accuracy. A new experiment confirms the robustness of our metric towards image noise.

2. Background & related work

2.1. Image registration

Intensity-based image registration frameworks model the problem as finding a transformation $\Phi\colon \Omega \to \Omega$ that aligns a moving image $\mathbf{I}\colon \Omega \to \mathbb{R}$ to a fixed image $\mathbf{J}\colon \Omega \to \mathbb{R}$. The morphed source image, obtained by applying the transformation, is expressed by function composition as $\mathbf{I} \circ \Phi$. The domain Ω denotes the set of all coordinates $\mathbf{x} \in \mathbb{R}^d$ within the image[1]. Images record intensity at discrete pixel-coordinates $\mathbf{p}$ but can be viewed as a continuous function by interpolation. The optimal transformation is found by minimization of a similarity metric D and a λ-weighted regularizer R, expressed via the loss function

$$L(\mathbf{I}, \mathbf{J}, \Phi) = D(\mathbf{I} \circ \Phi, \mathbf{J}) + \lambda R(\Phi) \ . \tag{1}$$

The choice of similarity metric D is the main objective of this paper, and common choices are discussed later. The regularizer R is necessary as many non-linear transformation models are over-parametrized, leading to many potential solutions. Smooth transformation fields, that avoid folds or gaps, are assumed to be physically plausible and encouraged by the regularizer (Leow et al., 2007; Kabus et al., 2009). Implicit regularizers achieve these properties by measuring the inverse consistency of the transformation (Greer et al., 2021; Shen et al., 2019b), while explicit regularizers operate on the displacement vector field directly (Balakrishnan et al., 2019). We use the explicit diffusion regularizer throughout this paper, which penalizes the spatial gradients of the displacement field. The displacement field $\mathbf{u}\colon \Omega \to \mathbb{R}^d$ of a discrete pixel-coordinate $\mathbf{p}$ is given by

$$\Phi(\mathbf{p}) = \mathbf{p} + \mathbf{u}(\mathbf{p}) \ , \tag{2}$$

and the diffusion regularizer thereon is defined as

$$R(\Phi) = \sum_{\mathbf{p} \in \Omega} \|\nabla \mathbf{u}(\mathbf{p})\|^2 \ , \tag{3}$$

with $\nabla \mathbf{u}(\mathbf{p})$ approximated via finite differences over the pixel coordinates.

2.2. Registration methods

Many methods of optimizing Eq. (1) have been proposed, and finding improved registration methods continues to be an active area of research. The field can be grouped into 1. algorithmic methods and 2. deep-learning-based methods.

1. Algorithmic methods optimize the objective for each pair of images individually, resulting in slow registration when many images have to be registered, for example in real-time applications or large population studies. Yet, this approach

[1] While the domain Ω is continuous in $\mathbb{R}^d$, recorded images and computations thereon are discrete. For simplicity of notation, we denote both the continuous and discrete domain as Ω. We implement $\sum_{\mathbf{p} \in \Omega}$ as a vectorized operation over the discrete pixel/voxel-coordinates and calculate $|\Omega|$ as the total count of discrete pixels/voxels of the image. The transformation Φ is implemented as a map from a discrete domain to a continuous one, and the sampling of continuous points from a discrete image is implemented via bi-/tri-linear interpolation.

does not require a large up-front investment into training datasets and resources. Most algorithms follow an iterative, gradient-descent-based approach. Some methods optimize the transformation directly, such as elastic models (Bajcsy and Kovačič, 1989; Davatzikos, 1997; Shen and Davatzikos, 2002), sparse parameterizations with b-splines (Rueckert et al., 1999), and Demons (Thirion, 1998; Vercauteren et al., 2007). Others parameterize intermediate transformation steps to offer diffeomorphic guarantees on the transformation field, such as the Large Diffeomorphic Distance Metric Mapping (LDDMM) algorithm (Faisal Beg et al., 2005) and standard symmetric normalization (SyN) (Avants et al., 2008b). Recent approaches follow a discrete optimization scheme (Heinrich et al., 2013a, 2015) while in Siebert et al. (2021), convex global optimization is combined with a local gradient-based instance refinement using an adaptive optimizer.

2. With the emergence of deep neural networks, model- and learning-based techniques for image registration are an area of active research. Compared to the algorithmic approach, deep-learning-based registration models are trained on a large dataset, necessitating a longer training time and a large collection of training images. However, after training is completed, inferring a transformation from the model is magnitudes faster than the algorithmic counterparts. Early works use supervised approaches, requiring ground-truth transformation fields (Yang et al., 2017; Krebs et al., 2017; Haskins et al., 2019). As these are often infeasible to attain, most contemporary works employ unsupervised or semi-supervised approaches by optimizing objective (1) directly. The dominant network architecture are fully-convolutional neural networks (CNNs), often in a U-Net configuration (Balakrishnan et al., 2019; Hu et al., 2019b; Hoopes et al., 2021). Various modifications, such as multi-level architectures (de Vos et al., 2019; Liu et al., 2019; Hu et al., 2019a; Mok and Chung, 2020, 2021; Shen et al., 2019a; Zhao et al., 2019), probabilistic models (Dalca et al., 2018b; Czolbe et al., 2021b), discretised architectures with a correlation layer (Dosovitskiy et al., 2015; Heinrich and Hansen, 2020), and fluid-diffeomorphism based transformations (Dalca et al., 2018b) have been proposed. Alternative approaches use vision transformers (Wang and Delingette, 2021; Chen et al., 2021, 2022; Mok and Chung, 2022; Shi et al., 2022; Song et al., 2022; Wang et al., 2022; Pegios and Czolbe, 2022) or graph-based networks (Hansen and Heinrich, 2021).

2.3. Similarity metrics for image registration

Similarity metric D measures the distance between the warped moving (*morphed*) image $\mathbf{I} \circ \Phi$ and the fixed image $\mathbf{J}$. Pixel-based metrics are well explored within algorithmic image registration, a comparative evaluation is given by Avants et al. (2011). We briefly recall four popular choices used as baselines in our evaluation: *mean squared error* (MSE), *normalized cross correlation* (NCC), *normalized mutual information* (NMI), and *modality independent neighborhood descriptor* (MIND), and discuss how these can be combined with supervised labels to obtain a *semi-supervised* similarity metric.

2.3.1. Mean Squared Error

The pixel-wise MSE is intuitive, computationally efficient, and easy to reason about. It is derived by maximizing the negative log-likelihood of a Gaussian normal distribution, making it an appropriate choice under the assumption of Gaussian noise. On a grid of discrete points $\mathbf{p}$ from the domain Ω, the MSE is defined as

$$\mathrm{MSE}(\mathbf{I} \circ \Phi, \mathbf{J}) = \frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} \|(\mathbf{I} \circ \Phi)(\mathbf{p}) - \mathbf{J}(\mathbf{p})\|^2 \quad . \tag{4}$$

2.3.2. Normalized Cross Correlation

Patch-wise NCC is robust to variations in brightness and contrast, making it a popular choice for images recorded with different acquisition tools and protocols, or even across image modalities. For two image patches $\mathbf{A}, \mathbf{B}$, represented as column-vectors of length N with patch-wise means $\bar{\mathbf{A}}, \bar{\mathbf{B}}$ and variance $\sigma^2_{\mathbf{A}}, \sigma^2_{\mathbf{B}}$, it is defined as

$$\mathrm{NCC}_{\mathrm{patch}}(\mathbf{A}, \mathbf{B}) = \sum_{n=1}^{N} \frac{(\mathbf{A}_n - \bar{\mathbf{A}})(\mathbf{B}_n - \bar{\mathbf{B}})}{\sigma_{\mathbf{A}} \sigma_{\mathbf{B}}} \quad . \tag{5}$$

The patch-wise similarities are then averaged over the image as

$$\mathrm{NCC}(\mathbf{I} \circ \Phi, \mathbf{J}) = \frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} \mathrm{NCC}_{\mathrm{patch}}(\mathbf{I}_{\mathbf{p}} \circ \Phi, \mathbf{J}_{\mathbf{p}}) \quad , \tag{6}$$

where $\mathbf{I}_{\mathbf{p}}, \mathbf{J}_{\mathbf{p}}$ denote the square image patch around pixel $\mathbf{p}$ (Gee et al., 1993; Avants et al., 2008b). Patches are centered around each pixel, leading to overlapping patches. Note that a slightly altered but computationally more efficient variant of NCC is used in some image registration works (Avants et al., 2011).

2.3.3. Normalized Mutual Information

Normalized mutual information (Studholme et al., 1999) models the probabilistic relation between the voxel intensities of the images. It is suitable for applications where no linear relation between the image intensities is present, making it the primary similarity metric used in multi-modal image registration. The relation between the morphed and fixed image is modeled as

$$\mathrm{NMI}(\mathbf{I} \circ \Phi, \mathbf{J}) = \frac{H(\mathbf{I} \circ \Phi) + H(\mathbf{J})}{I(\mathbf{I} \circ \Phi, \mathbf{J})} \quad , \tag{7}$$

with marginal entropy $H(\mathbf{I}) = -\int_{\mathbb{R}} p_{\mathbf{I}}(x) \log(p_{\mathbf{I}}(x))\, \mathrm{d}x$ and mutual information $I(x, y) = -\int_{\mathbb{R}} p_{\mathbf{IJ}}(x, y) \log(p_{\mathbf{IJ}}(x, y))\, \mathrm{d}x\, \mathrm{d}y$.

To calculate the joint probability $p_{\mathbf{IJ}}(x, y)$ and marginals $p_{\mathbf{I}}(x)$, $p_{\mathbf{J}}(y)$, the intensity distributions of both images are approximated by histograms, making them non-differentiable. For adaptation in gradient-based deep-learning frameworks, Parzen-window estimates with gaussian kernels are used to approximate the distributions. Given a Parzen window function w, the joint histogram for discrete bucket means $x, y \in \mathbb{R}$ is calculated as

$$h_{\mathbf{IJ}}(x, y) = \sum_{\mathbf{p} \in \Omega} w((\mathbf{I} \circ \Phi)(\mathbf{p}) - x)\, w(\mathbf{J}(\mathbf{p}) - y) \quad , \tag{8}$$

from which the joint $p_{\mathbf{IJ}}(x, y)$ is obtained by normalization and $p_{\mathbf{I}}(x)$, $p_{\mathbf{J}}(y)$ by marginalization thereof (de Vos et al., 2020; Qiu et al., 2021).

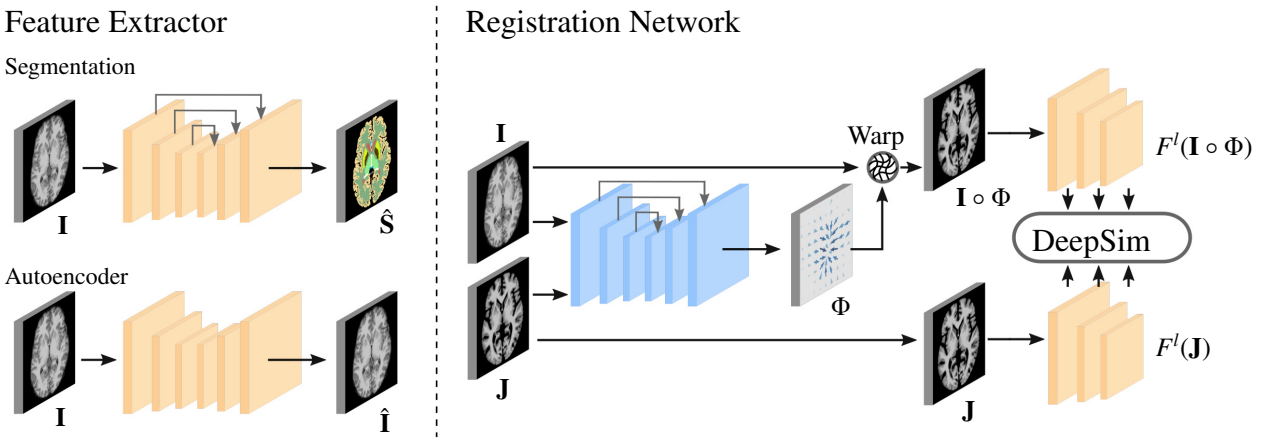


Fig. 1: Schematic overview of the method, using a U-Net model for registration. First, the feature extractor (yellow) is trained. We trial both a U-Net segmentation model trained on supervised segmentation masks (top left) and an unsupervised auto-encoder as the feature extractor (bottom left). The trained feature extractor is then used to drive the optimization of a registration model (blue, right). We trial both a U-Net (pictured) and transformer-based registration networks and algorithmic registration with SyN (Avants et al., 2008b). The registration model predicts the transformation Φ based on the moving and fixed images $\mathbf{I}, \mathbf{J}$. A spatial transformer module applies the transformation to obtain the morphed image $\mathbf{I} \circ \Phi$. Next, a pyramid of semantic representations $F^l(\cdot)$ is extracted by the frozen kernels of the encoding branch of the feature extractor. The DeepSim similarity metric compares the representations and calculates the similarity loss. It forms the training loss of the registration network together with the regularization of the transformation field.

2.3.4. *Modality Independent Neighborhood Descriptor*

The MIND-SSC image descriptor (Heinrich et al., 2012a, 2013b) extracts representations from images based on their self-similarity context (SSC). It is used as a loss function by comparing the extracted representations of images.

The self-similarity of two patches centered on $\mathbf{x}, \mathbf{y}$ with local variance σ^2 is caluclated as

$$S(\mathbf{I}, \mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{I}_\mathbf{x} - \mathbf{I}_\mathbf{y}\|^2}{\sigma^2}\right) . \tag{9}$$

Given this equation, the image descriptor of a pixel coordinate $\mathbf{p}$ is then calculated by evaluating the above equation on all pixels $\mathbf{x}, \mathbf{y} \in \mathcal{N}$ from the neighborhood of $\mathbf{p}$, only using pairs $\mathbf{x}, \mathbf{y}$ are adjacent to each other (euclidian distance of $\sqrt{2}$). Notably, the intensity of the center pixel $\mathbf{p}$ has no direct influence on the descriptor of $\mathbf{p}$. The descriptor is dependent on the choice of the patch size as well as the dilation and shape of the neighborhood $\mathcal{N}$, which have to be tuned for each application.

2.3.5. *Semi-Supervised Measures*

If additional information is available, the unsupervised similarity measures can be extended by a supervised component to align either ground-truth segmentation masks, pre-defined reference points, or reproduce a pre-determined reference transformation field. However, by adding a supervised component, the registration model is incentivized to be biased towards this component. Balakrishnan et al. (2019) study this in detail: as the strength of a supervised loss term is increased, the accuracy on unobserved regions and overall accuracy is decreasing. Thus, in the absence of perfect annotations, it is common practice to combine metrics operating on different representations of the image (Avants et al., 2008a). We compare to a semi-supervised metric by fusing an intensity-based loss $L_{\text{intensity}}$ with a semi-supervised metric L_{seg} operating on segmentation class annotations as

$$L = L_{\text{intensity}}(\mathbf{I} \circ \Phi, \mathbf{J}) + \gamma L_{\text{seg}}(\mathbf{S} \circ \Phi, \mathbf{T}) \tag{10}$$

for segmentation masks $\mathbf{S}, \mathbf{T}$ of images $\mathbf{I}, \mathbf{J}$ and weighting factor γ. For our expeirments, we use a supervised version of the NCC metric, formulated as $\text{NCC}_{\text{sup}}(\mathbf{I}, \mathbf{S}, \mathbf{J}, \mathbf{T}, \Phi) = \text{NCC}(\mathbf{I} \circ \Phi, \mathbf{J}) + \text{MSE}(\mathbf{S} \circ \Phi, \mathbf{T})$ for one-hot encoded $\mathbf{S}, \mathbf{T}$ (Balakrishnan et al., 2019).

2.4. *Deep similarity metrics in image registration*

While deep-learning-based image registration has received much interest recently, similarity metrics utilizing the compositional and data-driven advantages of neural networks remain under-explored. Some works explore how to incorporate scale-space into learned registration models, but similarity metrics remain intensity-based (Hu et al., 2019a; Li and Fan, 2018). Learned similarity metrics are proposed by Haskins et al. (2019) and Krebs et al. (2017), but both approaches require ground truth registration maps for training, which are either synthetically generated or manually created by a medical expert. Lee et al. (2019b) propose to learn annotated structures of interest as part of the registration model to aid alignment, but the method discards sub-regional and non-annotated structures.

Closest to our work is the approach by Wu et al. (2016), who learn a representation of the input images via a stacked autoencoder and use the resulting representations for the downstream task of algorithmic image registration. This is similar to our autoencoder-based approach combined with SyN registration. While their experimental evaluation has limitations, such as the patch-based training on small 21^3 patches, a model of 2 layers, and a small dataset of 66 images, their observations of increased accuracy and flexibility over handcrafted features are similar to ours. Majumdar et al. (2017) further investigates deep-learning-based features for algorithmic image registration. They find that on comparatively small datasets of less than 30 images, hand-crafted features can outperform learned ones.

2.5. Multi-modal image registration

Common data representations are frequently used as similarity metrics in multi-modal image registration (Heinrich et al., 2012b; Chen et al., 2016; Simonovsky et al., 2016; Pielawski et al., 2020; Blendowski et al., 2021). These approaches establish common representations across image modalities and are often learned from well-aligned images of multiple modalities. While our approach is similar, we instead aim to find a semantically augmented representation of images of a single modality, and show their applicability to mono-modality registration.

3. Method

We first discuss how the popular NCC metric assesses the similarity of image patches. We then modify the encoding of patches to include semantic information and finally outline how these semantic features are extracted from the image. A schematic overview of the method is given in Fig. 1.

3.1. A discussion of NCC

Our design of a semantic similarity metric starts by examining the popular NCC metric. We see that NCC between image patches $\mathbf{A}$ and $\mathbf{B}$ is equivalent to the cosine-similarity between the corresponding mean-centered vectors $f(\mathbf{A}) = \mathbf{A} - \bar{\mathbf{A}}$ and $f(\mathbf{B}) = \mathbf{B} - \bar{\mathbf{B}}$:

$$\mathrm{NCC}_{\mathrm{patch}}(\mathbf{A}, \mathbf{B}) = \frac{\langle f(\mathbf{A}), f(\mathbf{B}) \rangle}{\|f(\mathbf{A})\| \|f(\mathbf{B})\|} \quad , \tag{11}$$

with scalar product $\langle \cdot, \cdot \rangle$ and euclidean norm $\| \cdot \|$. Thus, an alternative interpretation of the NCC similarity measure is the cosine-similarity between two feature descriptors in a high-dimensional space. The descriptor is given by the intensity values of a centered image patch centered at a pixel $\mathbf{p}$. We will construct a similar metric, using semantic feature descriptors instead.

3.2. A semantic similarity metric for image registration

To align areas of similar semantic value, we propose a similarity metric based on the agreement of semantic feature representations of two images. Semantic feature maps are obtained by a *feature extractor*, which is pre-trained on a surrogate task. To capture alignment of both localized, concrete features, and global, abstract ones, we calculate the similarity at multiple layers of abstraction. Given a set of feature-extracting functions $F^l \colon \mathbb{R}^{\Omega \times C} \to \mathbb{R}^{\Omega_l \times C_l}$ for L layers, we define

$$\mathrm{DeepSim}(\mathbf{I} \circ \Phi, \mathbf{J}) = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{|\Omega_l|} \sum_{\mathbf{p} \in \Omega_l} \frac{\langle F^l_{\mathbf{p}}(\mathbf{I} \circ \Phi), F^l_{\mathbf{p}}(\mathbf{J}) \rangle}{\|F^l_{\mathbf{p}}(\mathbf{I} \circ \Phi)\| \|F^l_{\mathbf{p}}(\mathbf{J})\|} \quad , \tag{12}$$

where $F^l_{\mathbf{p}}(\mathbf{J})$ denotes the l^{th} layer feature extractor applied to image $\mathbf{J}$, at the spatial coordinate $\mathbf{p}$. It is a vector of C_l output channels, and the spatial size of the l^{th} feature map is denoted by $|\Omega_l|$.

Just as for NCC, the neighborhood of the pixel is considered by the similarity metric, as F^l is composed of convolutional filters with increasingly large receptive field sizes. In contrast to NCC, it is not necessary to zero-mean the feature descriptors, as the semantic feature representations are trained to be robust to variances in image brightness present in the training data.

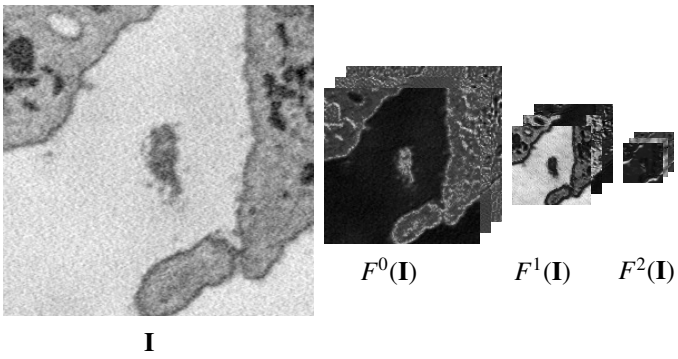


Fig. 2: The DeepSim similarity metric aligns a pyramid of semantic feature representations of an image. Left: Image $\mathbf{I}$. Right: Examples of feature maps $F^l(\mathbf{I})$ extracted at layers $l \in \{0, 1, 2\}$. Feature maps extracted from deeper layers of the feature extraction network encompass more global information, and are of lower spatial resolution. Each feature maps is C_l channels deep, with $C_l = 64, 128, 256$ in our experiments.

3.3. Feature extraction

To aid registration, the functions $F^l(\cdot)$ should extract features of semantic relevance for the registration task, while ignoring noise and artifacts inherent in image acquisition methods. To achieve these properties we extract features from the encoding branch of networks trained on two surrogate tasks:

1. Semi-Supervised measure: If segmentation masks are available, we can learn features on a supplementary segmentation task. Segmentation models excel at learning relevant kernels for the data while attaining invariance towards non-predictive features like noise, but require an annotated dataset for training. We denote the proposed similarity metric with feature extractors conditioned on this task as $\mathrm{DeepSim}_{\mathrm{seg}}$.
2. Unsupervised measure: We can learn an abstract feature representation of the dataset in an unsupervised setting with auto-encoders. Auto-encoders learn an efficient data encoding by training the network to ignore signal noise. A benefit of this approach is that no additional annotations are required. While variational methods for encoding tasks have several advantages, we choose a deterministic auto-encoder for its simplicity and lack of hyperparameters. We denote the similarity metric with feature extractors conditioned on this task as $\mathrm{DeepSim}_{\mathrm{ae}}$.

The choice of depth and receptive field size of the feature extracting functions has further impact on the metric. Deeper feature extractors can model more complex datasets, but increase computation time and memory requirements during training. Exclusively using high-level features, such as the last layer of a segmentation network, might only align the borders of anatomical regions and has the potential to ignore finer structures within those regions. On the contrary, too shallow features can behave similarly to intensity-based metrics. We evaluate different depth configurations as an ablation study, and use kernels up to the bottleneck of the segmentation network for our main experiments, effectively building a feature pyramid as visualized in Fig. 2.

4. Experimental setup

We evaluate our method using both deep-learning-based and algorithmic image registration. We train deep registration models with the proposed unsupervised $\text{DeepSim}_{\text{ae}}$ and semi-supervised $\text{DeepSim}_{\text{seg}}$, and compare to baselines MSE, NCC, NCC_{sup} (NCC with supervised information), NMI, and MIND. Our implementation of the baseline metrics follows Avants et al. (2011), Balakrishnan et al. (2019), Qiu et al. (2021), Hou et al. (2017), and (Heinrich et al., 2013b). To show that our method is also applicable to algorithmic image registration, we compare intensity-based registration using SyN (Avants et al., 2008b) to SyN registration of images augmented with semantic features learned by DeepSim. To ensure reproducibility, all code and experiments are available at github.com/SteffenCzolbe/DeepSimRegistration.

4.1. Data

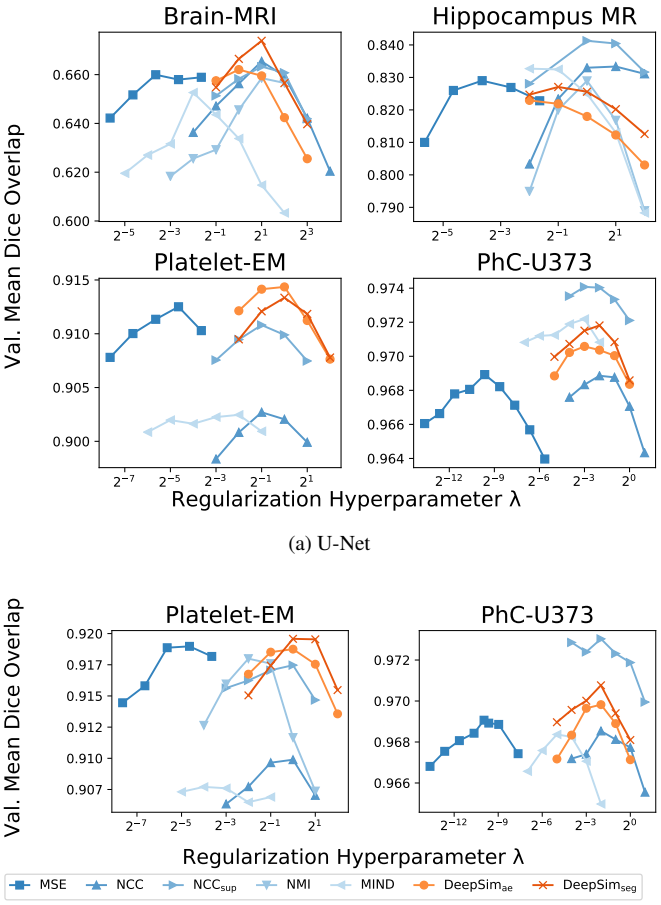
To show that our approach applies to a variety of registration tasks, we validate it on four 2D and 3D datasets of different modalities:

(1) T1 weighted Brain-MRI scans from the ABIDE-I, ABIDE-II (Di Martino et al., 2014) and OASIS3 (LaMontagne et al., 2019) studies for atlas-based alignment of *Brain-MRI* scans. Acquisition details, subject age ranges, and health conditions differ for each dataset, but no large anatomical anomalies are present. We perform standard pre-processing as in Balakrishnan et al. (2019), including intensity normalization, affine spatial alignment, skull-stripping and segmentation for each scan using FreeSurfer (Fischl, 2012) and crop the resulting images to $160 \times 192 \times 224$ voxels. Anatomical regions labeled separately on each hemisphere and smaller regions such as on the sub-structures of the cingulate cortex are combined, resulting in 24 distinct segmentation classes. After scans with preprocessing errors are discarded, we split the data 3665/250/250 for train-, validation-, and test-set, and register images to an atlas.

(2) T1 weighted MR scans of the hippocampus from the 2022 Learn2Reg challenge (Hering et al., 2022). The dataset was originally introduced in (Jafari-Khouzani et al., 2011) and included in the Medical Segmentation Decathlon Antonelli et al. (2022). It contains images from 90 healthy adults and 105 adults with a non affective psychotic disorder. Images are cropped to $64 \times 64 \times 64$ voxels. We split the data into 156 train-, 52 validation-, and 52 test-images, and perform inter-subject registration, giving 24000 unique training pairs.

(3) Slices of human blood cells from the Platelet-EM dataset (Quay et al., 2018). Images are recorded using serial block-face scanning electron microscopy. The dataset contains 74 slices manually annotated with three classes (Cytoplasm, Organelle, Background). Images are affinely pre-aligned and the dataset is split 50/12/12 for train-, validation-, and test-set. We register neighboring 2d slices.

(4) Cell tracking video of the PhC-U373 dataset from the ISBR cell tracing challenge (Maška et al., 2014; Ulman et al., 2017). The video sequence contains 230 2d images and is annotated with two classes (Cells, Background). We split the data 115/68/67 for train-, validation-, and test-set and register images of adjacent time steps.



(a) U-Net

(b) Transformer

Fig. 3: Hyperparameter tuning for a) U-Net and b) Transformer-based registration models. We trial regularizer strength parameter $\lambda = 2^n$ for some $n \in \mathbb{Z}$ for each model, similarity metric, and dataset independently. Parameter λ on the log-scaled x-axis, validation mean dice overlap on the y-axis. For each model, we select the λ with the highest validation mean dice overlap for further evaluation.

4.2. Deep learning models

For the registration model, we trial both well-established 2D and 3D U-Net (Ronneberger et al., 2015) architectures as popularized through VoxelMorph (Balakrishnan et al., 2019), and the recent state-of-the-art transformer model TransMorph (Chen et al., 2022).

We use the same U-Net architecture for the image registration model and segmentation-based feature extraction networks. We use a similar architecture for the auto-encoder feature extractor but without the shortcut connections. Each network consists of three encoder and decoder stages. Each stage consists of one batch normalization (Ioffe and Szegedy, 2015), two convolutional, and one dropout layer (Gal, Yarin and Ghahramani, 2016). After the final decoder step, we smooth the model output with three more convolutional layers. We experimented with deeper architectures but found they do not increase performance. The activation function is LeakyReLu throughout the network, Softmax for the final layer of the segmentation network, Sigmoid for the final layer of the auto-encoder, and linear for the final layer of the registration network. The stages have 64, 128, 256

Table 1: Parameters of the baseline similarity metrics used in our experiments.

Baseline	Parameter	Value			
		Brain-MRI	Hippocampus MR	Platelet-EM	PhC-U373
NCC / NCC_{sup}	window size	9	9	9	9
NMI	number of bins	64	64	8	32
MIND	radius	2	2	6	4
MIND	dilation	2	2	6	4
MIND	neighborhood size	6	6	4	4

channels for 2d datasets, and 32, 64, 128 channels for 3d.

In our experiments with TransMorph, we tried different model variants and sizes. We use the original TransMorph version which consists of 4 stages with $\{2, 2, 4, 2\}$ number of Swin Transformer (Liu et al., 2021) blocks and $\{4, 4, 8, 8\}$ number of heads in each stage respectively but set the embedding dimension to $C = 64$ because it performed better. As suggested we set the window size to be the same as at the input size after 32-fold downsampling while we zero-pad the images for the PhC-U373 dataset to make the spatial dimensions divisible by 32.

The segmentation model is trained with a cross-entropy loss function, the auto-encoder with the mean squared error. Both U-Net and transformer-based registration networks are trained with the loss given by Eq. 1. The optimization algorithm for all models is ADAM (Kingma and Ba, 2015), the initial learning rate is 10^{-4}, decreasing by a factor of 10 each time the validation loss plateaus. All models are trained until convergence. Training images are augmented with random affine transformations. Due to the large 3D volumes involved, the choice of batch-size is often limited by available memory. We sum gradients over multiple passes to arrive at effective batch-sizes of 3-5 samples.

4.3. Hyperparameter selection

The characteristics of deformable image registration methods are strongly influenced by the strength of regularization that is applied. Additionally, some baseline metrics have further hyperparameters, e.g, the number of bins in NMI or radius and dilation in MIND. For a fair comparison, we tune all parameters on the validation split of each dataset. The parameter choices used in our experiments can be found in Table 1. For hyperparameter λ, we trial values $\lambda = 2^n$ for some $n \in \mathbb{Z}$ for each U-Net model and hyperparameter selection, and plot the validation mean dice overlap in Fig. 3. We selected the parameter choices scoring the highest for further evaluation.

4.4. Algorithmic image registration

We further investigate whether algorithmic image registration benefits from the semantic image representations used for DeepSim. As a baseline, we choose to register the intensity images using the well-establish SyN algorithm (Avants et al., 2008b) from the ANTS software package (Avants et al., 2009), using the default registration parameters. For the semantic similarity metrics, we augment the intensity images by registering semantic feature maps obtained from either the auto-encoder or the segmentation feature extractor as additional modalities. We use channel-wise normalization, so that $\|F_c^l(\cdot)\|^2 = 1$ for each channel and layer, and up-scale all feature maps to image size using bi-/tri-linear interpolation. All modalities contribute equally to the objective function.

5. Results

5.1. Qualitative results

We plot the fixed and moving images $\mathbf{I}, \mathbf{J}$ and the morphed image $\mathbf{I} \circ \Phi$ warped by transformations obtained from the U-Net registration models trained with each similarity metric model in Fig. 4. The transformation is visualized by grid-lines and segmentation classes are overlaid for guidance.

5.2. Registration accuracy

We measure registration accuracy by the mean Sørensen Dice overlap of the annotated segmentation masks on the unseen test-set of each dataset. Results are presented in Fig. 6. U-Net registration models trained with our proposed $DeepSim_{ae}$ and $DeepSim_{seg}$ metrics achieve higher accuracy than all baselines on the Brain-MRI and Platelet-EM datasets. On the PhC-U373 dataset, only the NCC_{sup} baseline performs better. On the Hippocampus MR dataset, $DeepSim_{ae}$ and $DeepSim_{seg}$ outperform MSE, but fall behind the other baselines.

In Fig. 7, we contrast registration accuracy with transformation regularity (Leow et al., 2007; Kabus et al., 2009). We see that $DeepSim_{ae}$ and $DeepSim_{seg}$ are placed in the bottom right corner for three out of four datasets, indicating very smooth transformation fields combined with high registration accuracy.

Using algorithmic registration with SyN, the semantic features of $DeepSim_{seg}$ improve the registration accuracy over the baseline on all four datasets. The auto-encoder-based features of $DeepSim_{ae}$ fall short of just intensity-based registration.

We perform statistical significance testing of the model's results with the Wilcoxon signed rank test for paired samples. A significance level of 5% gives a Bonferroni-adjusted significance threshold $p = 0.002$. We further measure the effect size with Cohen's d and show the results in Table 2. We see that most results are statistically significant. On the Platelet-EM dataset, the performance difference between models trained with MSE and our proposed metrics falls below the statistical threshold, yet our method outperforms the baselines with at least small effect sizes. On the PhC-U373 dataset, the baseline NCC_{sup} outperform DeepSim with very small effect sizes.

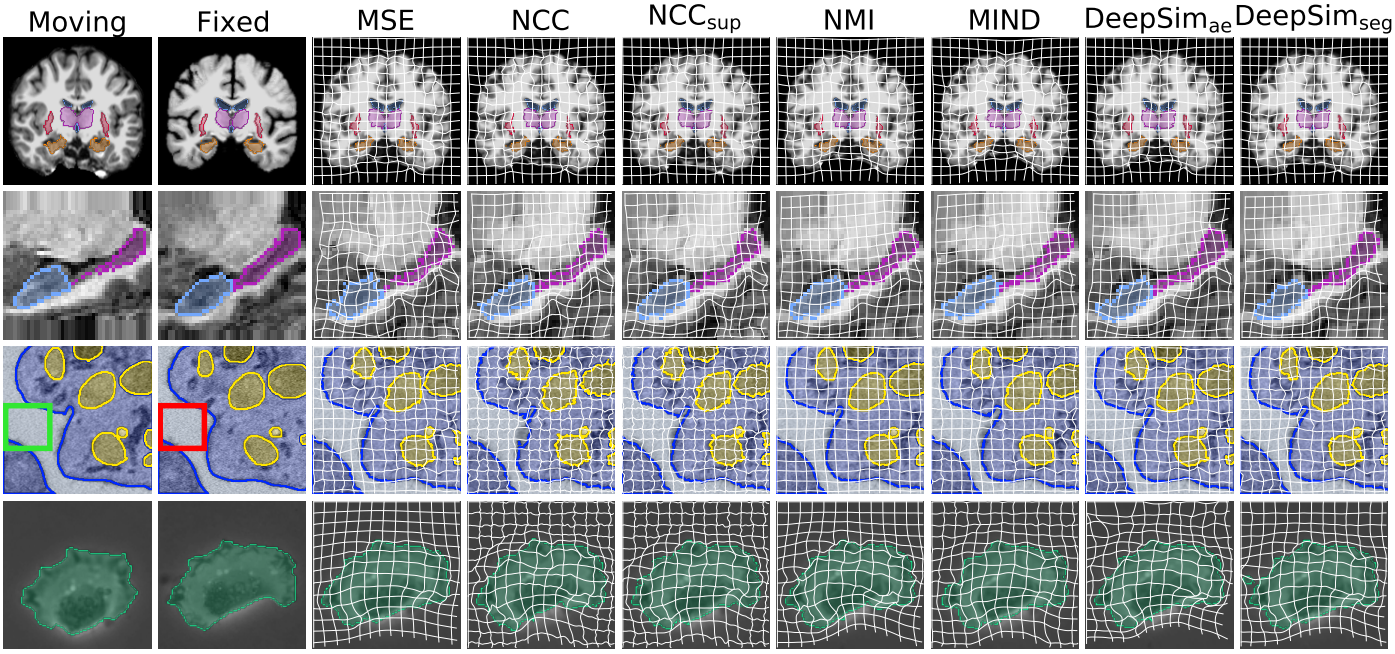


Fig. 4: Qualitative comparison of U-Net based deep-learning registration models. We register the moving image $\mathbf{I}$ (1st column) to the fixed image $\mathbf{J}$ (2nd column). Morphed images $\mathbf{I} \circ \Phi$ obtained from registration models trained with baseline similarity metrics MSE, NCC, NCC_{sup}, NMI, and MIND in columns 3–7. Morphed images obtained from our methods $\text{DeepSim}_{\text{ae}}$ and $\text{DeepSim}_{\text{seg}}$ in columns 8 and 9. Rows: Datasets Brain-MRI, Hippocampus MR, Platelet-EM, PhC-U373. Select segmentation classes annotated in color. The transformation is visualized by morphed grid-lines.

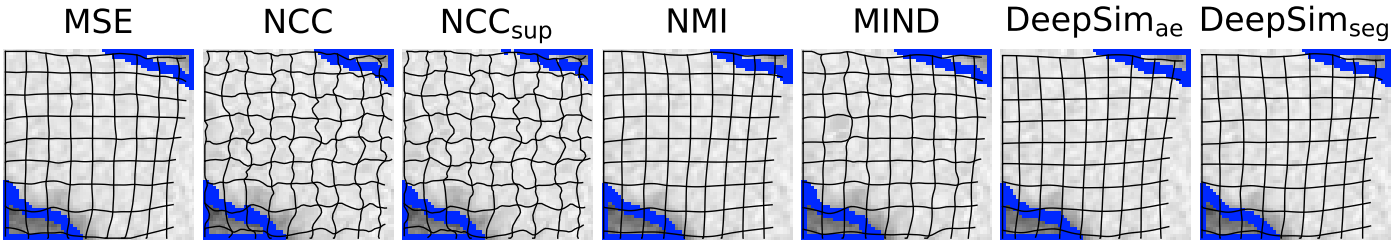


Fig. 5: Detail view of transformation grids on the highlighted spot of the Platelet-EM dataset in Fig. 4. The regularity of the transformation fields on this noisy image patch varies considerably between methods. Models trained with NCC, NCC_{sup}, and MIND exhibit the most irregular transformation fields. Models trained with $\text{DeepSim}_{\text{ae}}$ and $\text{DeepSim}_{\text{seg}}$ show the smoothest transformation fields. The cell-boundary is annotated in blue. The transformation is visualized by morphed grid-lines.

5.3. Regularity of the Transformation

To highlight the differences in transformation fields between methods, we display a noisy background patch of the Platelet-EM dataset in Fig. 5. The patch has been registered with transformations obtained from the U-Net registration models trained with each similarity metric model. Black grid-lines visualize the transformation. On this patch, models trained with NCC, NCC_{sup}, and MIND produce highly irregular transformation fields. Transformations obtained from $\text{DeepSim}_{\text{ae}}$, $\text{DeepSim}_{\text{seg}}$ and NMI are the most smooth on this dataset.

We perform a quantitative analysis of the regularity of the transformations produced by the U-Net models in Table 3, measuring transformation irregularity by the variance of the log-determinant of the Jacobian of the transformation field $\sigma^2(\log|J_\Phi|)$, and domain folding by the percentage of transformation voxels with a negative determinant.

5.4. Noise Resistance

We further evaluate registration performance in the presence of noise in the input data. Without retraining the models, we measure the mean dice overlap on the test set of the Platelet-EM dataset with added Gaussian noise. We sample the noise from $\mathcal{N}(0, \sigma^2)$, and test noise levels of $\sigma = 0, 0.05, 0.1, ..., 0.35$. We show results and examples of the noisy image patches in Fig. 8. The performance of all models decreases as noise is added. However, the models trained with the baselines loose performance quicker then model trained with Deepsim.

5.5. Convergence and speed

We monitor the mean training and validation dice overlap of the U-Net based deep-learning models during training in Fig. 9. The training accuracy is, with few exceptions, similar to the test accuracy, indicating that results generalize well. The relative time per epoch of models trained with each loss function is given in Table 4. Training models with DeepSim adds between $4-52\%$ time per epoch.

5.6. Anatomical regions

The Brain-MRI dataset contains annotations of the brain's anatomical regions. We plot the dice overlap per region in a

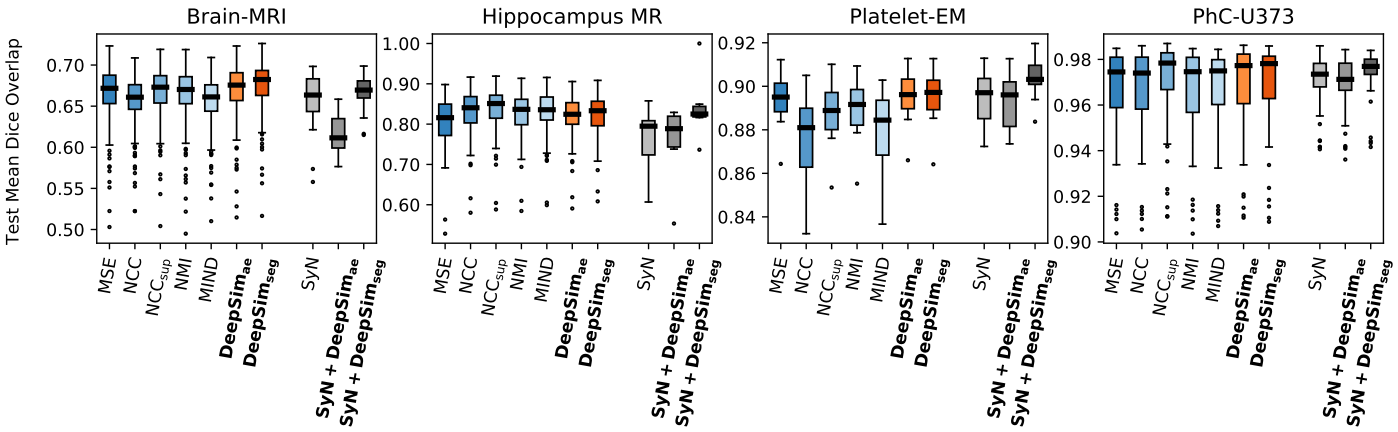


Fig. 6: For the datasets Brain-MRI, Hippocampus MR, Platelet-EM and PhC-U373, we trial multiple registration models and algorithms and record their test mean dice overlap. U-Net based deep-learning models trained with similarity metrics MSE, NCC (Gee et al., 1993), NCC_{sup} (Balakrishnan et al., 2019), NMI (Studholme et al., 1999), MIND (Heinrich et al., 2012a), $DeepSim_{ae}$ (ours), $DeepSim_{seg}$ (ours) on the left side of each plot. On the right side of each plot is algorithmic registration with the SyN algorithm (Avants et al., 2008b), and the SyN algorithm augmented with semantic features from $DeepSim_{ae}$ and $DeepSim_{seg}$. Boxplot with median, quartiles, deciles and outliers. Labels of our methods in bold.

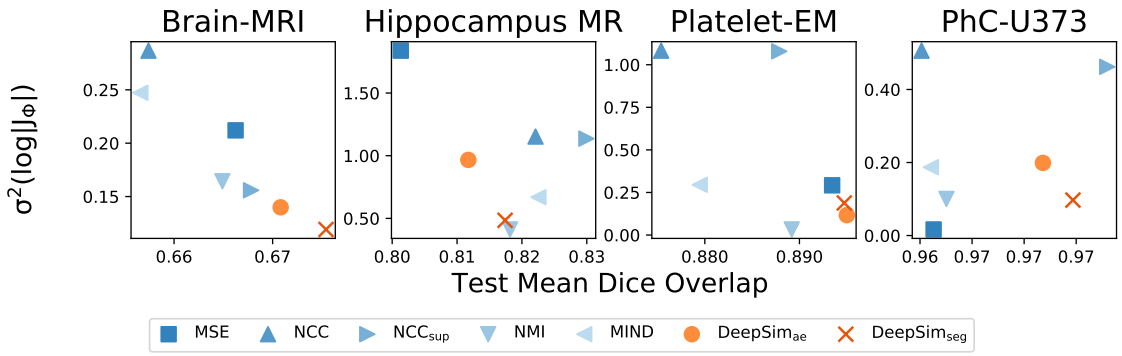


Fig. 7: Registration accuracy and irregularity of the transformation fields. Test mean dice overlap from Fig. 6 on the x-axis, variance of the log Jacobian determinant of the transformation $\sigma^2(\log|J_\Phi|)$ on the y-axis. Higher dice overlaps indicate a better alignment, lower variance indicates smoother transformation fields and fewer deformations.

Table 2: Significance testing of the results, performed with the Wilcoxon signed rank test for paired samples. Effect size measured with Cohen's d. Statistically insignificant results (significance level 0.05, Bonferroni-corrected to $p > 0.002$) and very small effect sizes ($|d| < 0.1$) in grey.

Dataset	Method	Baseline									
		MSE		NCC		NCC_{sup}		NMI		MIND	
		p	d	p	d	p	d	p	d	p	d
Brain-MRI	$DeepSim_{ae}$	<0.001	0.14	<0.001	0.43	<0.001	0.10	<0.001	0.18	<0.001	0.46
	$DeepSim_{seg}$	<0.001	0.30	<0.001	0.60	<0.001	0.25	<0.001	0.34	<0.001	0.63
Hippocampus MR	$DeepSim_{ae}$	0.003	0.15	<0.001	−0.16	<0.001	−0.28	<0.001	−0.10	<0.001	−0.17
	$DeepSim_{seg}$	<0.001	0.24	0.006	−0.07	<0.001	−0.19	0.238	−0.01	0.020	−0.08
Platelet-EM	$DeepSim_{ae}$	0.016	0.12	<0.001	1.14	<0.001	0.51	<0.001	0.43	<0.001	0.96
	$DeepSim_{seg}$	0.034	0.10	<0.001	1.12	<0.001	0.49	<0.001	0.40	<0.001	0.94
PhC-U373	$DeepSim_{ae}$	<0.001	0.10	<0.001	0.11	<0.001	−0.06	<0.001	0.08	<0.001	0.10
	$DeepSim_{seg}$	<0.001	0.12	<0.001	0.13	0.002	−0.03	<0.001	0.11	<0.001	0.12

Table 3: Regularity of the transformation. The determinant of the Jacobian of the transformation $|J_\Phi|$ is a measure of how the image volume is compressed or stretched by the transformation. We assess transformation smoothness by the variance of the voxel-wise log Jacobian determinant $\sigma^2(\log|J_\Phi|)$, a lower variance indicates a more volume-preserving transformation. Additionally, we assess the regularity of the transformation by measuring the percentage of voxels for which the determinant is < 0, which indicates domain folding.

Method	Dataset							
	Brain-MRI		Hippocampus MR		Platelet-EM		PhC-U373	
	$\sigma^2(\log\|J_\Phi\|)$	$\|J_\Phi\| < 0\,(\%)$	$\sigma^2(\log\|J_\Phi\|)$	$\|J_\Phi\| < 0\,(\%)$	$\sigma^2(\log\|J_\Phi\|)$	$\|J_\Phi\| < 0\,(\%)$	$\sigma^2(\log\|J_\Phi\|)$	$\|J_\Phi\| < 0\,(\%)$
MSE	0.21	0.42	1.84	8.75	0.29	0.40	0.02	0.02
NCC	0.29	0.93	1.15	4.08	1.08	4.15	0.51	0.71
NCC_{sup}	0.16	0.28	1.14	3.99	1.08	4.03	0.46	0.57
NMI	0.16	0.24	0.41	0.51	0.03	0.00	0.10	0.30
MIND	0.25	0.77	0.67	1.62	0.29	0.23	0.19	0.20
$DeepSim_{ae}$	0.14	0.20	0.97	3.02	0.12	0.04	0.20	0.35
$DeepSim_{seg}$	0.12	0.12	0.48	0.89	0.19	0.14	0.10	0.32

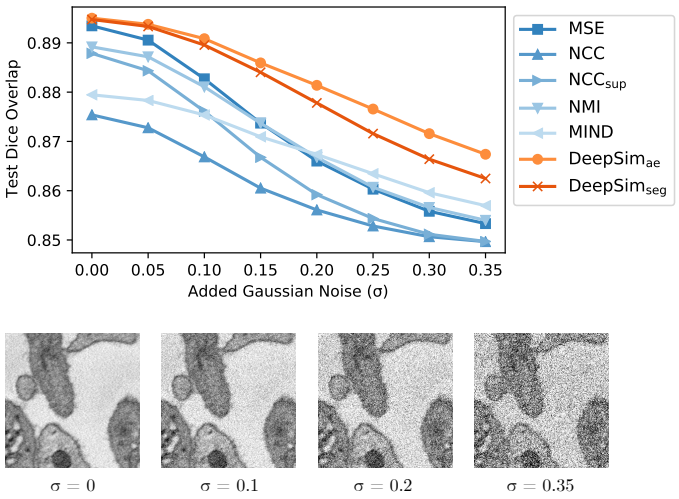


Fig. 8: Model performance on noisy data. We add Gaussian noise sampled from $\mathcal{N}(0, \sigma^2)$ to the input data, and measure the dice overlap on the test set. The x-axis shows the noise level, and the y-axis the test Dice overlap. Images below the plot show image patches under the noise levels.

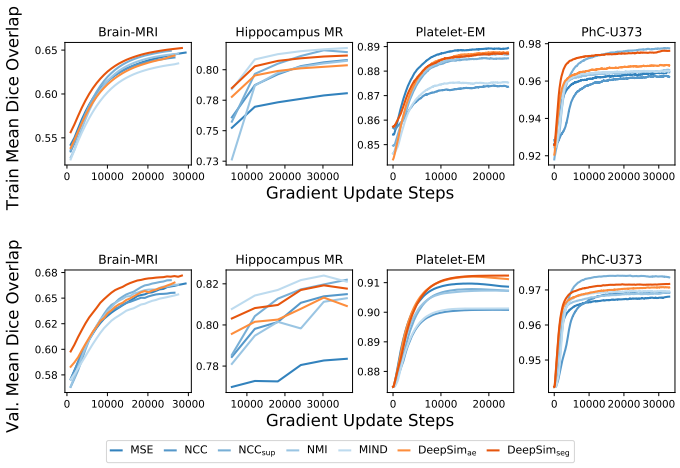


Fig. 9: Convergence during training and validation of U-Net models. Gradient update steps on the x-axis, train and validation mean dice overlap on the y-axis. The training duration per model on a single RTX 2080 GPU is approximately seven days for the Brain-MRI dataset and one day for the Platelet and PhC-U373 datasets.

boxplot in Fig. 10, and highlight regions where both of our metrics perform better than all baselines bold. Baseline methods (blue) perform very similar, despite NCC_{sup} as a supervised metric requiring more information over the unsupervised MSE and NCC.

5.7. Image registration using Transformers

We further evaluate the flexibility of the proposed method using the recent state-of-the-art transformer-based model TransMorph on the 2d datasets. As in the previous experiments, we perform hyperparameter tuning both for DeepSim and baseline loss functions and select the transformer model with the highest validation dice overlap.

Given the best parameter choices, we evaluate the tradeoff between dice-overlap and transformation smoothness on the test sets in Fig. 11. Results are similar to ones obtained with the U-Net model in Fig. 7, albeit slightly better overall. TransMorph registration networks trained with DeepSim achieve favorable

Table 4: Relative training time of U-Net registration models, based on measurements of 1 epoch of training. MSE = 1.00. Time measurement includes feed-forward and back-propagation through the model and loss function, as well as weight update for the model.

Method	Dataset			
	Brain-MRI	Hip. MR	Plat.-EM	PhC-U373
MSE	1.00	1.00	1.00	1.00
NCC	1.09	1.22	1.05	1.02
NCC_{sup}	1.12	1.24	1.05	1.03
NMI	1.07	1.05	1.08	1.04
MIND	1.02	1.05	1.03	1.02
$DeepSim_{ae}$	1.04	1.52	1.17	1.15
$DeepSim_{seg}$	1.03	1.49	1.16	1.15

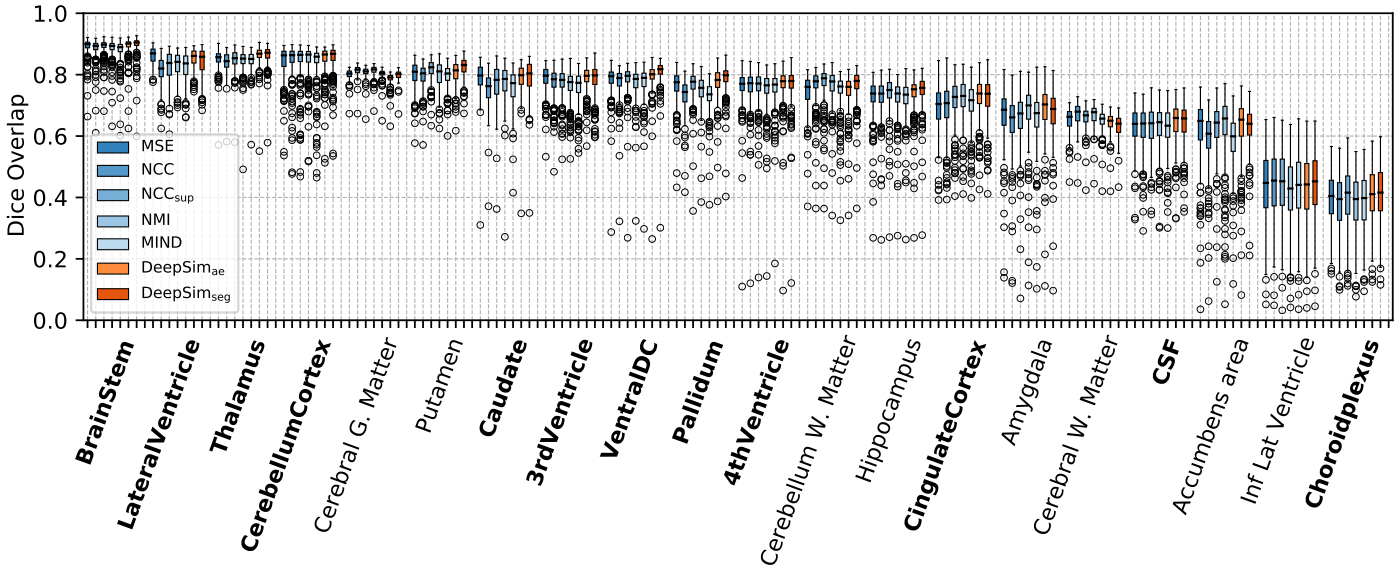


Fig. 10: Dice overlaps of the anatomical regions of the Brain-MRI dataset. Baselines in shades of blue, our methods in red. Bold labels for regions where *both* of our methods score higher than all baselines. We combined labels of the left and right brain hemispheres into a single class. The boxplot shows median, quartiles, deciles and outliers.

accuracy-smoothness tradeoffs for both the 2d datasets, placing in the bottom right corner of the plot on both datasets.

6. Ablation studies

After establishing the DeepSim similarity metric and comparing it to established choices, we now focus on investigating decisions made in the design of the metric. We investigate the effect of different levels of extracted features, assess if a dedicated feature extractor needs to be trained for each dataset, and inquire about the order of operations within the metric. These experiments are performed on the 2d dataset only.

6.1. *Levels of extracted features*

The abstraction levels at which semantic features are extracted can have an impact on the proposed metric. We investigate how different levels of features contribute to the registration accuracy. We trial DeepSim loss functions using deep features extracted from multiple combinations of layers, using only features extracted feature extraction layers 1, 2, 3, 1+2, 1+3, 2+3, and all layers. The level of the feature extraction layer is denoted with superscript, e.g, DeepSim^{1} compares shallow features extracted only from the first layer of a deep feature extractor, while DeepSim^{12} combines features only from the two first layers.

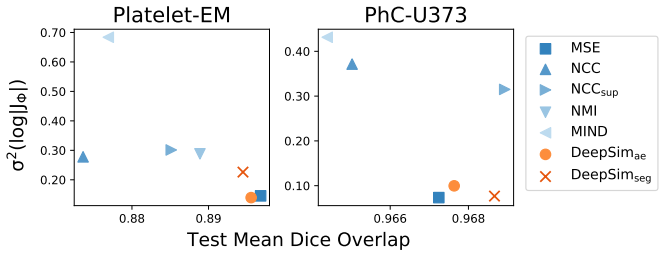
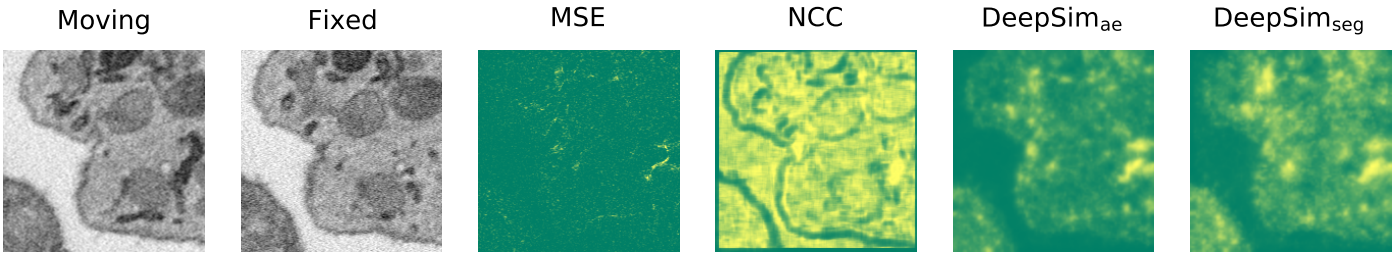


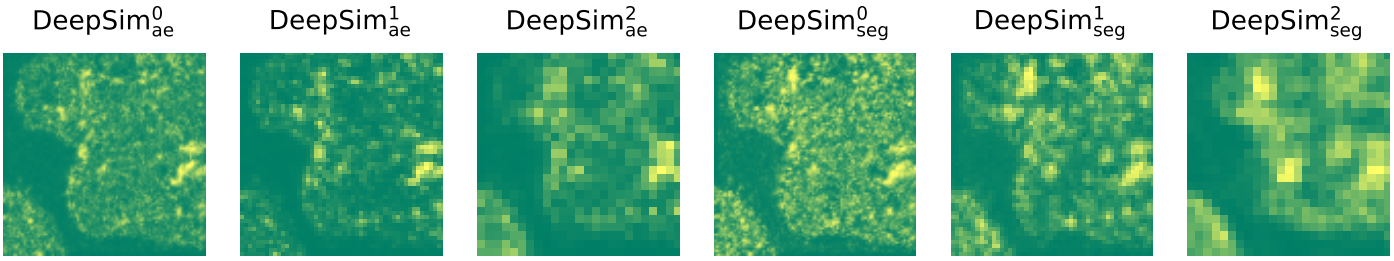
Fig. 11: Evaluation of the TransMorph registration network trained with different loss functions. Registration accuracy and irregularity of the transformation fields. Test mean dice overlap on the x-axis, variance of the log Jacobian determinant of the transformation $\sigma^2(\log|J_\Phi|)$ on the y-axis. Higher dice overlaps indicate a better alignment, lower variance indicates smoother transformation fields that are often considered more realistic.

We re-tune the regularization hyperparameter λ on the 2d datasets for the different layer configurations of DeepSim both for our unsupervised and semi-supervised approach, using U-Net based registration models. We plot the registration accuracy on the validation sets for $\text{DeepSim}_{\text{ae}}$ and $\text{DeepSim}_{\text{seg}}$ in Fig. 13a and Fig. 13b respectively. We observe that for the Platelet-EM dataset, which contains noisy images, using high-level features such as level 3 of the deep feature extractors improves accuracy. Most notably, disregarding the shallowest layer, $\text{DeepSim}^{23}_{\text{seg}}$ achieves slightly better performance than $\text{DeepSim}_{\text{seg}}$. On the other hand, on the non-noisy PhC-U373 dataset, level 1 contributes the most to the registration accuracy. This is in line with previous results, where the intensity-based baselines performed competitively on the PhC-U373 dataset. In general, it is evident that including both low-level, concrete features and high-level, more abstract ones in the loss, is beneficial to the performance of the registration model in almost all of the cases.

We further plot heatmaps of the loss occurred under different loss functions in Fig. 12a, and a per-layer view of the loss occurred under DeepSim in Fig. 12b. The moving and fixed images (top left) have been registered using U-Net models trained with the presented similarity metrics, and the occurred loss at each spatial coordinate is plotted on a color scale normalized to each model. The results show that, compared to the baselines, the DeepSim loss is more evenly distributed around non-matching image parts. It also doesn't occur a loss in the noisy background area between the cells.



(a) Loss-heatmaps of selected loss functions



(b) Losses at layers 1,2,3 of DeepSim$_{ae}$ and DeepSim$_{seg}$

Fig. 12: Heatmaps of the loss occurred after registration of the Moving and Fixed image (top left). Top Row: Registration and loss under models trained with MSE, NCC, DeepSim$_{ae}$, and DeepSim$_{seg}$. Bottom row: Heatmaps of loss occurred at layers 1,2,3 of DeepSim$_{ae}$, and DeepSim$_{seg}$. Brighter colors indicate a higher loss. Loss values have been normalized to one color scale.

6.2. Transfer Learning

One drawback of DeepSim is that a feature extractor has to be trained for each dataset. We investigate whether this is necessary, or if we can use feature-extractors trained on related or similar data instead. This approach is commonly referred to as Transfer Learning.

We trial three separate configurations: for the Platelet-EM dataset, we use the auto-encoder and segmentation model trained on the PhC-U373 dataset as the feature extractor. Vice versa, for the PhC-U373 dataset, we use the auto-encoder and segmentation model trained on the Platelet-EM dataset. We denote these similarity metrics with transferred features as DeepSim$_{ae}$ – TL and DeepSim$_{seg}$ – TL. Finally, we investigate the performance of our method using a universal feature extractor. To this end, we extract features from a VGG (Simonyan and Zisserman, 2014) classification network trained on ImageNet (Deng et al., 2009), and we denote this variant of the method as DeepSim$_{VGG}$.

For each configuration, we train U-Net based registration networks with different regularization parameters λ and plot the mean validation dice overlap in Fig. 14. We observe that for PhC-U373 all transfer learning approaches (DeepSim$_{ae}$ – TL, DeepSim$_{seg}$ – TL, DeepSim$_{VGG}$) not only improve the performance compared to the default setup of our method, but also surpass in performance all the baselines from the previous experiments. This might indicate that the PhC-U373 dataset does not have sufficient complexity and size to train a good feature extractor on it. For the Platelet-EM dataset, the performance of DeepSim transfer learning variants falls short of the original method but is still comparable with other baseline loss functions.

6.3. Feature extraction and transformation

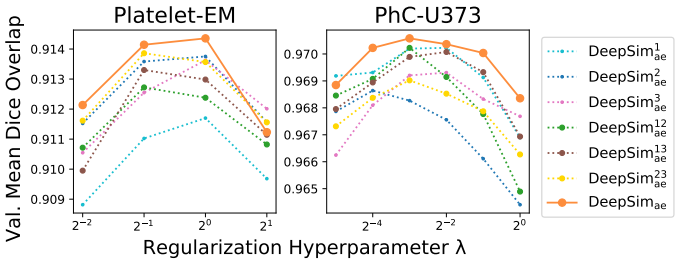
As defined in Eq. (12), the DeepSim metric first applies the transformation to the moving image, and then extracts a semantic representation from the morphed image (Transform before Extraction, TbE). A recently used alternative approach (Czolbe et al., 2021a) first extracts a semantic representation from the moving image, and then transforms the semantic representation (Extraction before Transformation, EbT). We empirically compare both variants, using both an auto-encoder and a segmentation model as the feature extractor.

We re-tune the regularization hyperparameter λ for the alternative implementation DeepSim (EbT). The necessary transformation of lower resolution feature maps is implemented by down-sampling and -scaling the transformation before warping the feature map. Registration accuracy on the validation sets is displayed in Fig. 15. We observe that the optimal choice for λ differs between the variations of the loss function, with the optimal value for the EbT version being consistently lower than for the unaltered TbE version across all datasets and feature extractors. The loss functions achieving the highest dice overlap are inconsistent, with the TbE version performing better on the Brain-MRI dataset, both versions achieving similar scores on the Platelet-EM dataset, and the EbT version performing better on the PhC-U373 dataset.

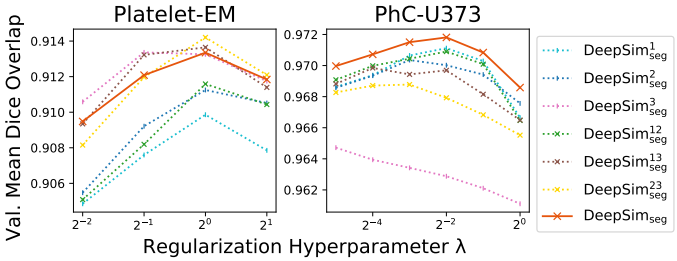
7. Discussion

The experimental results show that registration methods optimized with the proposed semantic similarity metric achieve small improvements in accuracy. Additionally, they are rebust to noise and produce smoother transformations, resulting in consistent improvements in the accuracy-smoothness tradeoff and more plausible transformations. The trend holds true across four diverse datasets and registration with SyN, U-Nets, and Transformers, showing the general applicability of the results.

We see the largest performance increase on the Platelet-EM dataset, which we hypothesize is caused by the significant noise



(a) DeepSim$_{ae}$



(b) DeepSim$_{seg}$

Fig. 13: Effect of different layer configurations on registration accuracy with (a) DeepSim$_{ae}$ and (b) DeepSim$_{seg}$. We trial loss functions using only features from feature extraction layers 1, 2, 3, 1+2, 1+3, 2+3 and all layers. For each configuration, we train U-Nets with different regularization parameters λ (x-axis) and observe the validation dice overlap (y-axis).

present in the dataset. The intensity-based metrics incentivize the model to align the noise, producing the observed unsmooth transformations and high loss values across the image, overall hindering registration. The proposed semantic similarity metric has instead learned that the noise is of no semantic importance, thus ignoring it in the registration.

Similarity metrics are independent of the registration method used. To show the general applicability of our metric and its independence from the underlying registration framework, we conducted experiments with U-Nets, transformer-based architectures, and algorithmic registration using the SyN algorithm from the ANTS package. The observed results are similar, especially between the two deep-learning based approaches. This is an indication that the choice of the registration model matters less compared to the metric used during training. Our method is robust and behaves consistently across registration methods.

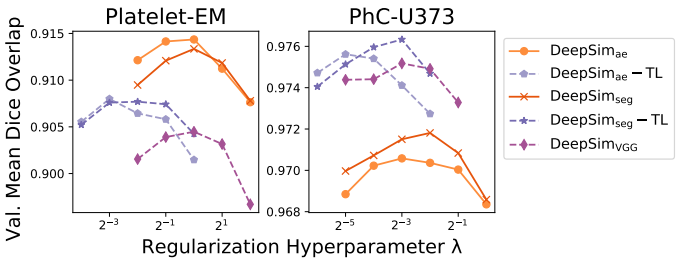


Fig. 14: We trial transferring feature extractors between datasets. Models trained with DeepSim$_{ae}$ and DeepSim$_{seg}$ use a feature extractor trained on their dataset. Models trained with DeepSim$_{ae}$ – TL and DeepSim – TL use a feature extractor trained on the opposite dataset. The model trained with DeepSim$_{VGG}$ uses features form the VGG image classification network in the feature extractor. For each loss function, we train models with a range of regularization parameters λ (x-axis) and observe the mean dice overlap on the validation set (y-axis).

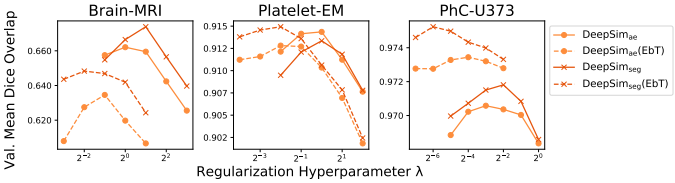


Fig. 15: We test an alternative version of DeepSim, where semantic features are extracted from the images before the transformation is applied (EbT: Extract before Transform). We train models with both DeepSim and DeepSim (EbT), using both segmentation models and auto-encoders as the feature extractor. For each of the three datasets Brain-MRI, Platelet-EM, and PhC-U373 we trial multiple choices for the regularization hyperparameter λ (x-axis) and observe the mean dice overlap on the validation set (y-axis).

A drawback of our method is that it requires a feature extractor to obtain features of semantic importance to the dataset. The experimental evaluation shows that the availability of annotated anatomical regions can help with learning semantic features, particularly if the dataset is large enough to support the training of such models. However, labeling a dataset is expensive and time-consuming, especially in biomedical settings.

To alleviate this issue, we investigated two alternatives that do not require labeled data or the training of a dedicated feature extractor. The need for labeled data can be removed by using semantic features extracted from an auto-encoder. This metric outperformed the baselines in both registration accuracy and transformation smoothness when registering images with the deep-learning-based models. However, using algorithmic registration, the unsupervised approach underperformed the baselines, particularly on the 3D Brain-MRI dataset. This could be caused by shortcomings of the auto-encoder, which yielded blurry reconstructions on the large brain volumes. On the other hand, for the 3D Hippocampus MR dataset, our proposed similarity metric provides a noticeable improvement when using the SyN registration algorithm.

To remove the requirement of having to train a dedicated feature extractor completely, one can use transfer learning to use an extractor trained on a different dataset. We used extractors trained on other medical datasets, and a general computer vision feature extractor, trained on ImageNet. Both worked especially well when the target dataset was small, even outperforming feature extractors trained directly on the data for the PhC-U373 dataset. This approach could be further expanded by using networks pre-trained on a large range of medical imaging tasks (Chen et al., 2019).

We focused on mono-modal image registration. The presented method could be extended to multi-modal registration in two different ways: (1) Through the use of modality-specific feature extractors to map each input modality to a common semantic representation (Pielawski et al., 2020), followed by alignment thereof. (2) Alternatively, separate feature extractors can be trained on each modality, and their semantic representations compared with a multi-modal metric such as MIND, NMI, or NCC.

7.1. Weaknesses

A weakness of our method is the need to include a separate model for the extraction of semantic features. While we have

shown that no dedicated model is required – other models can be reused with only slight decreases in performance – the design, training, and testing of a second model takes additional resources.

In the absence of ground-truth transformation fields, evaluation of deformable image registration is performed through proxy tasks. We measured accuracy by segmentation dice-overlap, but this evaluation technique only measures the overlap of larger areas, while discarding the alignment of sub-structures inside the annotated regions and does not evaluate point-to-point matches. We further evaluated the smoothness of the transformation fields and balanced this with the dice overlap in our evaluation, but no conclusive way of combining these metrics exists in the image registration literature. We welcome that recent registration challenges focus increasingly on measures besides segmentation dice overlap.

As our similarity metric depends on an auxiliary task, there is also a risk that the metric is biased by this choice of task, as well as by the segmentation masks that are used to train the auxiliary segmentation network. This label bias is perhaps most of all a problem in that its potential downstream effects are hard to foresee. However, we also note that the annotations often used to validate registration algorithms come with similar risks. Registration algorithms are often validated using annotated landmarks or Dice overlap of segmentation masks. We argue that these validation methods, which also affect which models are eventually chosen and published as state-of-the-art, come with a similar risk of label bias.

Any method is based on a large number of choices, decisions, and hyperparameters. While we did extensive trials of some of them in the ablation studies, there is always more that can be tested. We weighted all semantic features evenly in our method, and only considered features extracted from the encoding branch of the feature extraction networks. Tuning the individual weight of each feature is computationally expensive, but can be achieved in the presence of dedicated datasets, as Zhang et al. (2018) show for perceptual similarity metrics in image generation, or through hyperparameter learning strategies as shown by Hoopes et al. (2021) and Mok and Chung (2021) for image registration. While we did tune the regularization hyperparameter for the deep-learning-based models, we did not tune the parameters of SyN, instead using the same default parameters for each method. Due to technical constraints, we did not use the exact formulation of DeepSim for the SyN registration experiment but instead treated the semantic representations from DeepSim as additional modalities during registration with SyN. Because of practical issues, NMI is not used for TransMorph on the PhC-U373 dataset. Due to limited hardware availability, we do not include the 3d datasets in some of the ablation studies and our experiments with TransMorph.

8. Conclusion

We designed a semantic similarity metric for image registration. The new metric measures image similarity via the agreement of semantic and hierarchical image representations. The semantic representations can be extracted either in an unsupervised approach using an auto-encoder or in a supervised approach using supplemental segmentation data. In the absence of both, we have shown that features trained on related datasets can also be used.

The proposed metric achieves robust performance across four diverse datasets and three different registration models, using both deep-learning-based and algorithmic image registration. Image registration optimized with our method shows improved accuracy and smoother transformation fields compared to metrics such as MSE, NCC, NMI, and MIND.

The method is applicable to image registration tasks of all modalities and anatomies. Beyond the diverse range of datasets presented here, our good results in the presence of noise let us hope that our method will improve registration accuracy in domains such as low-dose CT, ultrasound, or microscopy, where details are often hard to identify, and image quality is poor.

We further emphasize that the application of semantic similarity metrics is not limited to the image registration domain. Semantic similarity metrics have the potential to improve methods in other image regression tasks, such as image synthesis, -translation, and -reconstruction.

Acknowledgments

We thank Matthew Quay, the Cell Tracking Challenge, and the Cancer Imaging Archive for the provision of the datasets, and Huaqi Qui for providing the implementation of the NMI metric. This work was funded in part by the Lundbeck Foundation (grant no. R218-2016-883), in part by the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (grant no. 0062606), and in part through the Danish National Research Foundation through the Danish Pioneer Centre for AI (grant no. P1).

References

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al., 2022. The medical segmentation decathlon. Nature communications 13, 1–13.

Avants, B., Duda, J.T., Kim, J., Zhang, H., Pluta, J., Gee, J.C., Whyte, J., 2008a. Multivariate analysis of structural and diffusion imaging in traumatic brain injury. Academic radiology 15, 1360–1375.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008b. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis 12, 26–41.

Avants, B.B., Tustison, N., Song, G., 2009. Advanced normalization tools (ANTS). Insight j 2, 1–35.

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage 54, 2033–2044.

Bajcsy, R., Kovačič, S., 1989. Multiresolution elastic matching. Computer Vision, Graphics, and Image Processing 46, 1–21.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. IEEE Transactions on Medical Imaging 38, 1788–1800.

Blendowski, M., Hansen, L., Heinrich, M.P., 2021. Weakly-supervised learning of multi-modal features for regularised iterative descent in 3d image registration. Medical image analysis 67, 101822.

Castillo, R., Castillo, E., Fuentes, D., Ahmad, M., Wood, A.M., Ludwig, M.S., Guerrero, T., 2013. A reference dataset for deformable image registration

spatial accuracy evaluation using the COPDgene study archive. Physics in Medicine and Biology 58, 2861–2877.

Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y., 2022. Transmorph: Transformer for unsupervised medical image registration. Medical image analysis 82, 102615.

Chen, J., He, Y., Frey, E.C., Li, Y., Du, Y., 2021. ViT-V-Net: Vision Transformer for Unsupervised Volumetric Medical Image Registration. Medical Imaging with Deep Learning , 2020–2022.

Chen, S., Ma, K., Zheng, Y., 2019. Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 .

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P., 2016. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, in: Proceedings of the 30th International Conference on Neural Information Processing Systems.

Czolbe, S., Feragen, A., Krause, O., 2021a. Spot the Difference: Detection of Topological Changes via Geometric Alignment. Advances in Neural Information Processing Systems 34.

Czolbe, S., Feragen, A., Krause, O., 2021b. Spot the Difference: Topological Anomaly Detection via Geometric Alignment .

Czolbe, S., Krause, O., Cox, I., Igel, C., 2020. A Loss Function for Generative Neural Networks Based on Watson's Perceptual Model. Advances in Neural Information Processing Systems .

Czolbe, S., Krause, O., Feragen, A., 2021c. Semantic similarity metrics for learned image registration. Proceedings of Machine Learning Research .

Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R., 2018a. Unsupervised Learning for Fast Probabilistic Diffeomorphic Registration. Medical Image Computing and Computer Assisted Intervention , 729–738.

Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R., 2018b. Unsupervised Learning for Fast Probabilistic Diffeomorphic Registration. Medical Image Computing and Computer Assisted Intervention , 729–738.

Davatzikos, C., 1997. Spatial Transformation and Registration of Brain Images Using Elastically Deformable Models. Computer Vision and Image Understanding 66, 207–222.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: Conference on Computer Vision and Pattern Recognition, pp. 248–255.

Di Martino, A., Yan, C.G., Li, Q., Others, 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry 19, 659–667.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T., 2015. Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE international conference on computer vision, pp. 2758–2766.

Faisal Beg, M., Miller, M.I., Trouvétrouv, A., Younes, L., 2005. Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms. International Journal of Computer Vision 61, 139–157.

Fischl, B., 2012. FreeSurfer. NeuroImage, 62 (2), 774–781.

Gal, Yarin and Ghahramani, Z., 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, in: International Conference on Machine Learning, pp. 1050—-1059.

Gee, J.C., Reivich, M., Bajcsy, R., 1993. Elastically Deforming a Three-Dimensional Atlas to Match Anatomical Brain Images. Technical Report.

Greer, H., Kwitt, R., Vialard, F.X., Niethammer, M., 2021. Icon: Learning regular maps through inverse consistency, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3396–3405.

Hansen, L., Heinrich, M.P., 2021. Deep learning based geometric registration for medical images: How accurate can we get without visual features? Information Processing in Medical Imaging .

Haskins, G., Kruecker, J., Kruger, U., Xu, S., Pinto, P.A., Wood, B.J., Yan, P., 2019. Learning deep similarity metric for 3D MR–TRUS image registration. International Journal of Computer Assisted Radiology and Surgery 14, 417–425.

Heinrich, M.P., Hansen, L., 2020. Highly accurate and memory efficient unsupervised learning-based discrete ct registration using 2.5 d displacement search, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 190–200.

Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.A., 2012a. Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. Medical image analysis 16, 1423–1435.

Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, S.M., Schnabel, J.A., 2012b. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. Medical Image Analysis 16, 1423–1435.

Heinrich, M.P., Jenkinson, M., Brady, M., Schnabel, J.A., 2013a. Mrf-based deformable registration and ventilation estimation of lung ct. IEEE transactions on medical imaging 32, 1239–1248.

Heinrich, M.P., Jenkinson, M., Papież, B.W., Brady, S.M., Schnabel, J.A., 2013b. Towards realtime multimodal fusion for image-guided interventions using self-similarities, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 187–194.

Heinrich, M.P., Maier, O., Handels, H., 2015. Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities. VISCERAL Challenge@ ISBI 1390, 27.

Hering, A., Hansen, L., Mok, T.C., Chung, A.C., Siebert, H., Häger, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., et al., 2022. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. IEEE Transactions on Medical Imaging .

Hoopes, A., Hoffmann, M., Fischl, B., Guttag, J., Dalca, A.V., 2021. HyperMorph: Amortized Hyperparameter Learning for Image Registration. Information Processing in Medical Imaging .

Hou, X., Shen, L., Sun, K., Qiu, G., 2017. Deep feature consistent variational autoencoder, in: Winter Conference on Applications of Computer Vision, IEEE. pp. 1133–1141.

Hu, X., Kang, M., Huang, W., Scott, M.R., Wiest, R., Reyes, M., 2019a. Dual-Stream Pyramid Registration Network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 382–390.

Hu, Y., Gibson, E., Barratt, D.C., Emberton, M., Noble, J.A., Vercauteren, T., 2019b. Conditional Segmentation in Lieu of Image Registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 401–409.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, International Machine Learning Society (IMLS). pp. 448–456.

Jafari-Khouzani, K., Elisevich, K.V., Patel, S., Soltanian-Zadeh, H., 2011. Dataset of magnetic resonance images of nonepileptic subjects and temporal lobe epilepsy patients for validation of hippocampal segmentation techniques. Neuroinformatics 9, 335–346.

Kabus, S., Klinder, T., Murphy, K., van Ginneken, B., Lorenz, C., Pluim, J.P., 2009. Evaluation of 4d-ct lung registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 747–754.

Kingma, D.P., Ba, J.L., 2015. Adam: A method for stochastic optimization, in: nternational Conference on Learning Representations, International Conference on Learning Representations, ICLR.

Krebs, J., Mansi, T., Delingette, H., Zhang, L., Ghesu, F.C., Miao, S., Maier, A.K., Ayache, N., Liao, R., Kamen, A., 2017. Robust non-rigid registration through agent-based action learning, in: Lecture Notes in Computer Science, Springer Verlag. pp. 344–352.

LaMontagne, P.J., Benzinger, T.L.S., Morris, J.C., Others, 2019. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. medRxiv .

Lee, D.S., Sahib, A., Wade, B., Narr, K.L., Hellemann, G., Woods, R.P., Joshi, S.H., 2019a. Multimodal Data Registration for Brain Structural Association Networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 373–381.

Lee, M.C.H., Oktay, O., Schuh, A., Schaap, M., Glocker, B., 2019b. Image-and-Spatial Transformer Networks for Structure-Guided Image Registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 337–345.

Leow, A.D., Yanovsky, I., Chiang, M.C., Lee, A.D., Klunder, A.D., Lu, A., Becker, J.T., Davis, S.W., Toga, A.W., Thompson, P.M., 2007. Statistical properties of jacobian maps and the realization of unbiased large-deformation nonlinear image registration. IEEE transactions on medical imaging 26, 822–832.

Li, H., Fan, Y., 2018. Non-rigid image registration using self-supervised fully convolutional networks without training data, in: Proceedings - International Symposium on Biomedical Imaging, IEEE Computer Society. pp. 1075–1078.

Liu, L., Hu, X., Zhu, L., Heng, P.A., 2019. Probabilistic Multilayer Regularization Network for Unsupervised 3D Brain Image Registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 346–354.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021.

Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

Majumdar, A., Mehta, R., Sivaswamy, J., 2017. To Learn or Not to Learn Features for Deformable Registration? Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11038 LNCS, 52–60.

Maška, M., Ulman, V., Svoboda, D., Matula, P., Matula, P., Ederra, C., Urbiola, A., España, T., Venkatesan, S., Balak, D.M.W., Karas, P., Bolcková, T., Štreitová, M., Carthel, C., Coraluppi, S., Harder, N., Rohr, K., Magnusson, K.E.G., Jaldén, J., Blau, H.M., Dzyubachyk, O., Krížek, P., Hagen, G.M., Pastor-Escuredo, D., Jimenez-Carretero, D., Ledesma-Carbayo, M.J., Muñoz-Barrutia, A., Meijering, E., Kozubek, M., Ortiz-de Solorzano, C., 2014. A benchmark for comparison of cell tracking algorithms. Bioinformatics 30, 1609–1617.

Mok, T.C., Chung, A., 2020. Large deformation diffeomorphic image registration with laplacian pyramid networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 211–221.

Mok, T.C., Chung, A., 2021. Conditional deformable image registration with convolutional neural network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 35–45.

Mok, T.C., Chung, A., 2022. Affine medical image registration with coarse-to-fine vision transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20835–20844.

Nielsen, R.K., Darkner, S., Feragen, A., 2019. TopAwaRe: Topology-Aware Registration. International Conference on Medical Image Computing and Computer-Assisted Intervention , 364—-372.

Pegios, P., Czolbe, S., 2022. Can transformers capture long-range displacements better than cnns?, in: Medical Imaging with Deep Learning.

Pielawski, N., Wetzer, E., Öfverstedt, J., Lu, J., Wählby, C., Lindblad, J., Sladoje, N., 2020. CoMIR: Contrastive Multimodal Image Representation for Registration. Proceedings of the 33rd International Conference on Neural Information Processing Systems .

Qiu, H., Qin, C., Schuh, A., Hammernik, K., Rueckert, D., 2021. Learning Diffeomorphic and Modality-invariant Registration using B-splines, in: Proceedings of Machine Learning Research.

Quay, M., Emam, Z., Anderson, A., Leapman, R., 2018. Designing deep neural networks to automate segmentation for serial block-face electron microscopy, in: International Symposium on Biomedical Imaging, IEEE Computer Society. pp. 405–408.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer Verlag. pp. 234–241.

Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: Application to breast mr images. IEEE Transactions on Medical Imaging 18, 712–721.

Shen, D., Davatzikos, C., 2002. HAMMER: Hierarchical attribute matching mechanism for elastic registration. IEEE Transactions on Medical Imaging 21, 1421–1439.

Shen, Z., Han, X., Xu, Z., Niethammer, M., 2019a. Networks for joint affine and non-parametric image registration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4224–4233.

Shen, Z., Vialard, F.X., Niethammer, M., 2019b. Region-specific diffeomorphic metric mapping. Advances in Neural Information Processing Systems 32.

Shi, J., He, Y., Kong, Y., Coatrieux, J.L., Shu, H., Yang, G., Li, S., 2022. Xmorpher: Full transformer for deformable medical image registration via cross attention, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 217–226.

Siebert, H., Hansen, L., Heinrich, M.P., 2021. Fast 3d registration with accurate optimisation and little learning for learn2reg 2021. arXiv propteprint arXiv:2112.03053 .

Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N., Komodakis, N., 2016. A deep metric for multimodal registration, in: Lecture Notes in Computer Science, Springer Verlag. pp. 10–18.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv propteprint arXiv:1409.1556 .

Song, L., Liu, G., Ma, M., 2022. Td-net: unsupervised medical image registration network based on transformer and cnn. Applied Intelligence , 1–9.

Song, X., Guo, H., Xu, X., Chao, H., Xu, S., Turkbey, B., Wood, B.J., Wang, G., Yan, P., 2021. Cross-modal Attention for MRI and Ultrasound Volume Registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 66–75.

Studholme, C., Hill, D.L., Hawkes, D.J., 1999. An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognition 32, 71–86.

Thirion, J.P., 1998. Image matching as a diffusion process: an analogy with maxwell's demons. Medical image analysis 2, 243–260.

Trofimova, D., Adler, T., Kausch, L., Ardizzone, L., Maier-Hein, K., Köthe, U., Rother, C., Maier-Hein, L., 2020. Representing Ambiguity in Registration Problems with Conditional Invertible Neural Networks, in: Medical Imaging Meets NeurIPS Workshop at Neural Information Processing Systems 2020.

Ulman, V., Maška, M., Magnusson, K.E., Ronneberger, O., Haubold, C., Harder, N., Matula, P., Matula, P., Svoboda, D., Radojevic, M., Smal, I., Rohr, K., Jaldén, J., Blau, H.M., Dzyubachyk, O., Lelieveldt, B., Xiao, P., Li, Y., Cho, S.Y., Dufour, A.C., Olivo-Marin, J.C., Reyes-Aldasoro, C.C., Solis-Lemus, J.A., Bensch, R., Brox, T., Stegmaier, J., Mikut, R., Wolf, S., Hamprecht, F.A., Esteves, T., Quelhas, P., Demirel, Ö., Malmström, L., Jug, F., Tomancak, P., Meijering, E., Muñoz-Barrutia, A., Kozubek, M., Ortiz-De-Solorzano, C., 2017. An objective comparison of cell-tracking algorithms. Nature Methods 14, 1141–1152.

Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., 2007. Non-parametric diffeomorphic image registration with the demons algorithm, in: Lecture Notes in Computer Science, pp. 319–326.

de Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I., 2019. A deep learning framework for unsupervised affine and deformable image registration. Medical Image Analysis 52, 128–143.

de Vos, B.D., van der Velden, B.H.M., Sander, J., Gilhuijs, K.G.A., Staring, M., Išgum, I., 2020. Mutual information for unsupervised deep learning image registration, in: Medical Imaging 2020: Image Processing, International Society for Optics and Photonics. p. 113130R.

Wang, Y., Qian, W., Li, M., Zhang, X., 2022. A transformer-based network for deformable medical image registration, in: CAAI International Conference on Artificial Intelligence, Springer. pp. 502–513.

Wang, Z., Delingette, H., 2021. Attention for Image Registration (AiR): an unsupervised Transformer approach .

Wu, G., Kim, M., Wang, Q., Munsell, B.C., Shen, D., 2016. Scalable High-Performance Image Registration Framework by Unsupervised Deep Feature Representations Learning. IEEE Transactions on Biomedical Engineering 63, 1505–1516.

Yang, Q., Fu, Y., Giganti, F., Ghavami, N., Chen, Q., Noble, J.A., Vercauteren, T., Barratt, D., Hu, Y., 2020. Longitudinal Image Registration with Temporal-order and Subject-specificity Discrimination, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 243–252.

Yang, X., Kwitt, R., Styner, M., Niethammer, M., 2017. Quicksilver: Fast predictive image registration - A deep learning approach. NeuroImage 158, 378–396.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. Conference on Computer Vision and Pattern Recognition , 586–595.

Zhao, S., Lau, T., Luo, J., Eric, I., Chang, C., Xu, Y., 2019. Unsupervised 3d end-to-end medical image registration with volume tweening network. IEEE journal of biomedical and health informatics 24, 1394–1404.

6

CAN TRANSFORMERS CAPTURE LONG-RANGE DISPLACEMENTS BETTER THAN CNNS?

PAPER 2:

Paraskevas Pegios and **Steffen Czolbe** (Jul 2022). "Can Transformers capture long-range displacements better than CNNs?". In: *Medical Imaging with Deep Learning (MIDL 2022).*

This is the first time I took the role of the senior author. My thank goes out to Paraskevas Pegios, who delivered this outstanding master's thesis under my supervision. The work has been published at MIDL 2022.

Can Transformers capture long-range displacements better than CNNs?

Paraskevas Pegios PPEGIOSK@GMAIL.COM
DTU Compute, Technical University of Denmark, Denmark
Steffen Czolbe PER.SC@DI.KU.DK.COM
Department of Computer Science, University of Copenhagen, Denmark

Abstract

Convolutional Neural Networks (CNNs) are well-established in medical imaging tackling various tasks. However, their performance is limited due to their incapacity to capture long spatial correspondences within images. Recently proposed deep-learning-based registration methods try to overcome this limitation by assuming that transformers are better at modeling long-range displacements thanks to the nature of the self-attention mechanism. Even though existing transformers are already considered state-of-the-art in image registration, there is no extensive validation of the key premise. In this work, we test this hypothesis by evaluating the target registration error as a function of the displacement. Our findings show that transformers outperform CNNs on a public dataset of lung 3D CT images with large displacements. Yet, the performance difference stems from transformers registering small displacements with higher accuracy. Contrary to previous beliefs, we find no evidence to support the hypothesis that transformers register long displacements better than CNNs. Additionally, our experiments provide insights on how to train vision transformers effectively for image registration on small datasets with less than 50 image pairs.

Keywords: Image Registration, Vision Transformers, Convolutional Neural Networks

1. Introduction

Image registration aims to find geometric transformations that align images. During the last years, CNN-based methods, such as VoxelMorph (Balakrishnan et al., 2019), have attracted wide attention in the field of deformable registration. After training, these methods can significantly speed up medical image processing pipelines while achieving comparable registration accuracy with traditional optimization approaches. The main limitation of CNNs is that they tend to focus on local aspects of images, which is problematic especially when the displacements between the moving and the fixed images become larger than the effective receptive field. Vision transformers lack the inductive biases of CNNs, such as translation invariance and locally restricted receptive fields and their success is usually ascribed to their ability to capture long-range dependencies within an image, even from the shallowest layers. Very recently, (Park and Kim, 2022) questioned this explanation by revealing new intuitions on how vision transformers work. Following the current trend in computer vision and medical imaging, transformer-based models such as ViT-V-Net (Chen et al., 2021b) and TransMorph (Chen et al., 2021a) have been proposed as strong candidates for better modeling of long-range displacements. Even though these models can have a global view of the entire image (Chen et al., 2021a) achieving state-of-the-art results in image registration, there is no extensive validation of the main hypothesis that transformers can capture long-range displacements better than CNNs.

2. Experimental Setup

Given a fixed volume $\mathbf{F}$, and a moving volume $\mathbf{M}$, we seek to predict a transformation $\Phi = Id + \mathbf{u}$, where $\mathbf{u}$ is the displacement field and Id is the identity transformation. The warping is applied using a spatial transformation function, i.e, $\mathbf{M} \circ \Phi$. We model $\mathbf{u}$ using a deep network which can be either a convolutional (VoxelMorph) or a transformer-based (Vit-V-Net, TransMorph) network. In that sense, a network is used to generate the transformation between the images, i.e, $g_{\boldsymbol{\theta}}(\mathbf{F}, \mathbf{M}) = \mathbf{u}$, where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{enc}, \boldsymbol{\theta}_{dec}\}$ and subscripts enc and dec denote the parameters of the encoder and decoder part of network respectively. Instead of naive random initialization, we leverage IXI[1] *pre-trained* weights to initialize $\boldsymbol{\theta}_{enc}$, while $\boldsymbol{\theta}_{dec}$ (task specific) are initialized randomly. To the best of our knowledge, *transfer learning* has not been used for image registration because established CNN-based methods can achieve good performance even for small datasets (Balakrishnan et al., 2019). During training, normalized cross correlation is used as distance metric between $\mathbf{M} \circ \Phi$ and $\mathbf{F}$, together with diffusion regularization, weighted by a hyper-parameter λ. A warm-up phase is evaluated by gradually increasing the learning rate up to a specific point and then using standard schedulers. This is a common technique for fine-tuning transformers because layer normalization in multi-head self-attention (MSA) layers can lead to high gradients at early iterations. Intuitively, by taking small steps we prevent adaptive optimizers from going towards wrong directions. Previous studies focused on evaluating transformers mainly in terms of DICE using large datasets. We conduct the evaluation in terms of Target Registration Error (TRE) since our aim is test to the ability of the models to capture long-range displacements. For evaluation, we use the "Learn2Reg: CT Lung Registration" dataset which contains 30 cases of inhaling and exhaling image pairs. Since there are no available landmarks for the original test pairs, we reorganize the dataset and split it (20/5/5), in order to use available keypoints for our validation (6-10) and test cases (1-5).

3. Results & Discussion

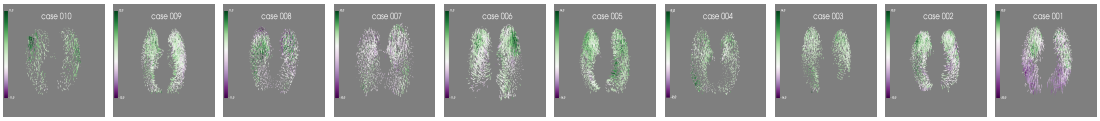


Figure 1: **Comparison of displacement fields between TransMorph^{++} and VoxelMorph-2^{++}.**

For a fair comparison, we tuned λ for baseline (random initialized) and fine-tuned (initialized with encoder pre-trained weights) models using the validation set. Transformers because of the lack of inductive bias required stronger regularization ($\lambda = 1$) than VoxelMorph ($\lambda = 0.5$). We evaluated the irregularity and the smoothness of the transformations and the results are reported in Table 1. Transfer learning proved beneficial not only for transformers but also for VoxelMorph. This can be very useful in practice when working with limited hardware and datasets. As expected, TransMorph benefited the most from transfer learning since its encoder is completely composed of transformer layers. Furthermore, TransMorph outperformed both VoxelMorph and ViT-V-Net, but apart from transfer learning, it required a warm-up phase to improve the smoothness of transformations.

1. https://github.com/junyuchen245/TransMorph_Transformer_for_Medical_Image_Registration/blob/main/IXI/TransMorph_on_IXI.md

A qualitative comparison of the displacements produced by the *fine-tuned* VoxelMorph and TransMorph models is shown in Fig.1. The displacement vectors are colored based on the difference in TRE. The greener the vector the better TransMorph^{++} is while as a vector gets more purple the better VoxelMorph^{++} is. *Diagnostic plots* to measure TRE by the length of the displacement are illustrated in Fig.2. The displacements were binned into approximately evenly-sized bins, in order to determine the mean and a confidence interval for each bin. In this way, diagnostic curves were produced for each *fine-tuned* model aiming to inspect TRE as the displacement length increases. Bin-wise statistical t-tests with Benjamini/Hochberg correction were used to highlight the significant bins (p-value ≤ 0.05) for the pair-wise model comparisons. It is evident that transformers were better at small to medium lengths while for larger displacements there is no such difference.

Warm-up	Model	**TRE** ↓	$\lvert J_\Phi\rvert_{<0}(\%)$ ↓	$\sigma(\lvert J_\Phi\rvert)$
-	Affine	15.34	-	-
-	VoxelMorph-2	11.74	0.61	0.133
-	**VoxelMorph-2^{++}**	**10.58**	**0.45**	**0.120**
-	ViT-V-Net	10.80	0.16	0.090
-	**ViT-V-Net^{++}**	**10.12**	**0.84**	**0.122**
-	TransMorph	11.88	2.29	0.188
✓	**TransMorph^{++}**	**9.91**	**0.34**	**0.105**

Table 1: **Evaluation metrics on our test set (cases 1-5). Transfer learning is denoted with a ++ superscript. TRE is measured in mm.**

Conclusion Overall, transformers outperformed VoxelMorph but the performance gain came from better registering small displacements. To answer the question posed in the title of the paper: *contrary to previous assumptions, we found no evidence to support the claim that transformers register long displacements better than CNNs.* This finding seems to be supported by (Park and Kim, 2022) where it is shown that "the success of MSAs for computer vision is NOT due to their weak inductive bias and capturing long-range dependency".

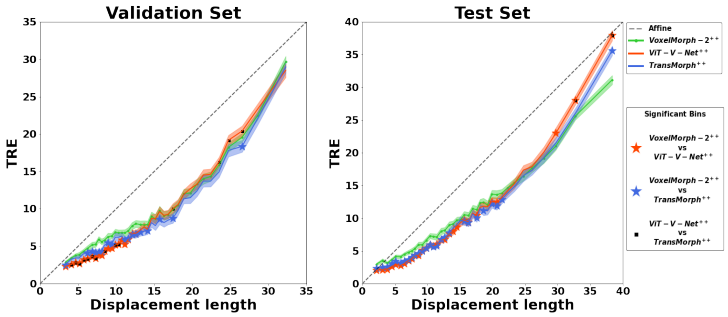


Figure 2: **TRE across the displacement length domain.**

References

Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

Junyu Chen, Yong Du, Yufan He, William P Segars, Ye Li, and Eirc C Frey. Transmorph: Transformer for unsupervised medical image registration. *arXiv preprint arXiv:2111.10480*, 2021a.

Junyu Chen, Yufan He, Eric C Frey, Ye Li, and Yong Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. *arXiv preprint arXiv:2104.06468*, 2021b.

Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations*, 2022.

7

SPOT THE DIFFERENCE: DETECTION OF TOPOLOGICAL CHANGES VIA GEOMETRIC ALIGNMENT

PAPER 3:

Steffen Czolbe, Aasa Feragen, and Oswin Krause (Dec 2021). "Spot the Difference: Detection of Topological Changes via Geometric Alignment". In: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021).*

This paper has been published at NeurIPS 2021. My special thanks go out to Oswin Krause, who sat with me in front of a whiteboard for many days to derive and debug the learnable prior of the VAE.

Spot the Difference: Detection of Topological Changes via Geometric Alignment

Steffen Czolbe
Department of Computer Science
University of Copenhagen
per.sc@di.ku.dk

Aasa Feragen
DTU Compute
Technical University of Denmark
afhar@dtu.dk

Oswin Krause
Department of Computer Science
University of Copenhagen
oswin.krause@di.ku.dk

Abstract

Geometric alignment appears in a variety of applications, ranging from domain adaptation, optimal transport, and normalizing flows in machine learning; optical flow and learned augmentation in computer vision and deformable registration within biomedical imaging. A recurring challenge is the alignment of domains whose topology is not the same; a problem that is routinely ignored, potentially introducing bias in downstream analysis. As a first step towards solving such alignment problems, we propose an unsupervised algorithm for the detection of changes in image topology. The model is based on a conditional variational auto-encoder and detects topological changes between two images during the registration step. We account for both topological changes in the image under spatial variation and unexpected transformations. Our approach is validated on two tasks and datasets: detection of topological changes in microscopy images of cells, and unsupervised anomaly detection brain imaging.

1 Introduction

Geometric alignment is a fundamental component of widely different algorithms, ranging from domain adaptation [7], optimal transport [40] and normalizing flows [35, 42] in machine learning; optical flow [21, 51] and learned augmentation [20] in computer vision, and deformable registration within biomedical imaging [5, 15, 19, 39, 53]. A recurring challenge is the alignment of domains whose topology is not the same. When the objects to be aligned are probability distributions [35], this appears when distributions have different numbers of modes whose support is separated into separate connected components. When the objects to be aligned are scenes or natural images, the problem occurs with occlusion or temporal changes [51]. In biomedical image registration, the problem is very common and happens when the studied anatomy differs from "standard" anatomy [36]. Despite being extremely common, this problem is routinely ignored or accepted as inevitable, potentially introducing bias in downstream analysis.

We study two cases from biomedical image registration. One is the alignment of image slices to reconstruct a 3d volume, where changes in topology between slices introduce challenges in post-processing (Figure 1). The other is the registration of brain MRI scans, where tumors give common examples of anatomies that are topologically different from healthy brains. In deformable image registration, a "moving image" is mapped via a nonlinear transformation to make it as similar as possible to a "target" image, enabling matching local features or transferring information from one

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

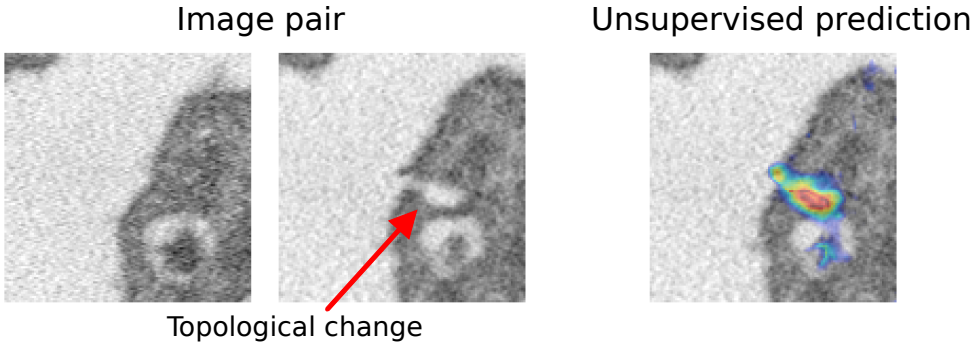


Figure 1: Left: Example of topological changes between two adjacent slices of human blood cells imaged via serial block-face scanning electron microscopy [41]. We aim to detect the change of topology caused by an emerging organelle within the cell (highlighted by the red arrow) while accounting for non-linear deformations of the image introduced by natural shape changes between slices. Right: Heatmap of the likelihood of topological changes predicted by our unsupervised model.

image to another. It is common to numerically stabilize the estimation of the transformation by constraining the predicted transformation to be diffeomorphic, that is, bijective and continuously differentiable in both directions. In particular, diffeomorphic transformations are homeomorphic, or topology-preserving, which implies that a common topology is assumed across all images [13, 15]. This topology is often provided by a common template image $\mathbf{I}_{\text{template}}$, from which all other images are obtained via the transformation Φ from the group of diffeomorphisms $\mathcal{G}$. Under this common topology assumption, the set of all images is given by

$$\mathcal{I} = \{\mathbf{I}_{\text{template}} \circ \Phi | \Phi \in \mathcal{G}\} \ .$$

Topological differences in biomedical images can be caused by a variety of processes. For instance, image slices obtained from a volume do not all contain the same elements. Tumor growth or the removal of surgical tissue can alter the topology of an image. Various processes can lead to the replacement or deformation of organic tissue, which cannot be mapped to the original image. We choose to model these topological differences as the inability to obtain one image from the other via a homeomorphic transformation of the image domain. Since, within image registration, transformations are assumed to be continuously differentiable, we are effectively modelling topological differences between pairs of images via the failures of diffeomorphic image registration in aligning them.

As most registration algorithms align images based on intensity, e.g. minimizing mean squared error (MSE), these tissue changes make it difficult to map images correctly. The strong local deformations required to deal with the non-diffeomorphic part of the image inevitably also deform the surrounding area, leading to distorted transformation fields in topologically matching parts of the image [36]. These transformation fields adversely affect downstream tasks, for example indicating false size changes in adjacent regions.

Previous work on aligning topologically inconsistent domains. Attempting to relax the same-image assumption induced by fully diffeomorphic transformations is not new. In the context of organs sliding against each other, several approaches exist, most of which rely on pre-annotating the sliding boundary using organ segmentation [6, 10, 22, 37, 43, 46], with a few extensions to un-annotated images [38, 45].

When topological holes are created or removed in the domain, for example through tumors, pathologies, or surgical resections, the loss function used for registration can be locally weighted or masked [26, 29, 30], or an artificial insection can be grown to correct anatomies [36]. These approaches rely on annotation of the topological differences, which have to be provided manually or by segmentation. An exception is given by Li and Wyatt [30], which detects changes in topology from the difference between the aligned images. This depends crucially on the ability to find a good diffeomorphic registration *outside* the anomaly, which is difficult all the while the applied transformation is still diffeomorphic.

An alternative approach to registering topologically inconsistent images is to inpaint the difference in the source images to obtain a topologically consistent quasi-normal image. Then standard registration methods can be used on the altered images. Quasi-normal images can be obtained through low-rank and sparse matrix decomposition [32, 33], principle component analysis [16, 18], denoising VAEs [52], or learning of a blended representation [17]. Registration with the quasi-normal approach retains the diffeomorphic properties of the transformation but does not register the topologically inconsistent areas of the images.

Our contribution. We propose an unsupervised algorithm for the detection of changes in image topology. To this end, we train a conditional variational autoencoder for predicting image-to-image alignment, obtaining a per-target-pixel probability of being obtained from the moving image via diffeomorphic transformation. We combine a semantic loss function trained to extract contextual information [8], with a learnable prior of transformations [9], allowing us to incorporate both the reconstruction error, as well as knowledge about the expected transformation strength.

We test the validity of our approach on a novel dataset of cell slices with annotated topological changes and on the proxy task of unsupervised brain-tumor detection. We also validate our approach by investigating a spatial "topological inconsistency likelihood", and showing that this likelihood is higher in regions where topological inconsistencies are known to be common. Our model is able to detect topological inconsistencies with a purely registration-driven framework, and thus provides the first step towards an end-to-end registration model for images with topological discrepancies. The implementation is available at github.com/SteffenCzolbe/TopologicalChangeDetection.

2 Background

2.1 Notation of images and transformations

We view an image $\mathbf{I}$ interchangeably as two different structures. First, it is a continuous function $\mathbf{I} : \Omega_\mathbf{I} \to \mathbb{R}^C$, where $\Omega_\mathbf{I} = [0,1]^D$ is the domain of the image, and C the number of channels. This function can be approximated by a grid of n pixels with positions $x_k \in \Omega_\mathbf{I}$ leading to the image representation $\mathbf{I}_k^{(c)}$, where c is an index over the channels and $\mathbf{I}_k = (\mathbf{I}_k^{(1)}, \dots, \mathbf{I}_k^{(C)})^T = \mathbf{I}(x_k)$. Second, this pixel grid is accompanied by a graph structure that encodes the neighbourhood of each pixel. In this view, the set of neighbours of a pixel with index k (for example the 4-neighbourhood of a pixel on the image grid) is referred to as $N(k)$ and $|N(k)|$ is the number of neighbours. The neighborhoods of a pixel gives rise to a graph which can be described via the graph laplacian $\Lambda \in \mathbb{R}^{n \times n}$ with $\Lambda_{k,k} = |N(k)|$ and $\Lambda_{k,k'} = -1$ when pixel $k' \in N(k)$, and zero otherwise.

Applying a spatial transformation $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ to an image is written as $\mathbf{J} = \mathbf{I} \circ \Phi$, which can be seen as its own image with domain $\Omega_\mathbf{J} = [0,1]^D$ with pixel coordinates $y_k \in \Omega_\mathbf{J}$ and $\mathbf{J}_k = \mathbf{I}(\Phi(y_k))$. The transformation Φ can be seen as a vector field on the image domain which assigns each pixel in $\mathbf{J}$ a position on $\mathbf{I}$ and thus it can be parameterized as a pixel grid $\Phi_k^{(d)}$, $d = 1, \dots, D$ at the pixel coordinates of $\mathbf{J}$ using $\Phi(y_k) = y_k + \Phi_k$. To make this choice of coordinate system clear, we will refer to a transformation that moves a pixel position from the domain $\Omega_\mathbf{J}$ to the corresponding pixel in domain $\Omega_\mathbf{I}$ as $\Phi_{\mathbf{J}\to\mathbf{I}}$, whenever it is not clear from the context. If Φ is a diffeomorphism, it can alternatively be parameterized by a vector field V on the tangent space around the identity, where the mapping between the tangent space and the transformation is given by $\Phi = \exp(V)$, which amounts to integration over the vector field [2].

2.2 Variational registration framework

It is possible to phrase the problem of fitting a registration model in terms of variational inference, using an approach similar to conditional variational autoencoders [47]. Here, we summarize the approach taken by [9, 31]. For a D-dimensional image pair $(\mathbf{I}, \mathbf{J})$, we assume that $\mathbf{J}$ is generated from $\mathbf{I}$ by drawing a transformation Φ from a prior distribution $p(\Phi|\mathbf{I})$, apply it to $\mathbf{I}$ and then add pixel-wise noise:

$$p(\mathbf{J}|\mathbf{I}) = \int p_{\text{noise}}(\mathbf{J}|\mathbf{I} \circ \Phi) p(\Phi|\mathbf{I}) \, d\Phi$$

This includes the common topology assumption implicitly via $p(\Phi|\mathbf{I})$, which is typically chosen to produce invertible transformations depending only on the topology of $\mathbf{I}$, as well as the noise

model which does not assume systematic changes between $\mathbf{J}$ and $\mathbf{I}$. This model can be learned using variational inference using a proposal distribution $q(\Phi|\mathbf{I},\mathbf{J})$ with evidence lower bound (ELBO)

$$\log p(\mathbf{J}|\mathbf{I}) \geq E_{q(\Phi|\mathbf{I},\mathbf{J})}\left[\log p_{\text{noise}}(\mathbf{J}|\mathbf{I}\circ\Phi)\right] - KL(q(\Phi|\mathbf{I},\mathbf{J})\|p(\Phi|\mathbf{I}))\ . \tag{1}$$

In contrast to variational autoencoders, the decoder is given by the known application of Φ to $\mathbf{I}$. Thus, the degrees of freedom in this model are in the choice of the encoder, prior, and the noise distribution. Dalca et al. [9] proposed to parameterize Φ as a vector field $V_k^{(d)}$ on the tangent space, which turns application of $\Phi = \exp(V)$ into sampling an image with a spatial transformer module [24]. As a prior for this parameterization, they chose a prior independent of $\mathbf{I}$

$$p(\Phi) = \prod_{d=1}^{D} \mathcal{N}\left(V^{(d)} \mid 0, \Lambda^{-1}\right)\ ,$$

where we used the implicit identification of Φ and V and the precision matrix Λ is chosen as the Graph Laplacian over the neighbourhood graph (see notation). Using an encoder that for each pixel proposes $q(V_k^{(d)}|\mathbf{I},\mathbf{J}) = \mathcal{N}(\mu_k^{(d)}, v_k^{(d)})$, the KL divergence is derived as

$$\mathrm{KL}\left(q(\Phi|\mathbf{I},\mathbf{J})\|p(\Phi|\mathbf{I})\right) = \frac{1}{2}\sum_{d=1}^{D}\sum_{k=1}^{n} -\log v_k^{(d)} + |N(k)|v_k^{(d)} + \sum_{l\in N(k)}\left(\mu_k^{(d)} - \mu_l^{(d)}\right)^2 + \text{const}\,. \tag{2}$$

It is worth noting that this equation is invariant under translations of μ. This invariance manifests in rank-deficiency of Λ and as a result, const is infinite. Thus, sampling from the prior and bounding the objective is impossible. Still training with this term works in practice as images are usually pre-aligned with an affine transformation and thus translations are close to zero. We will present a slightly modified approach, rectifying the missing eigenvalue.

3 Detection of topological differences

The variational approach for learning the distribution of transformations introduced before optimizes an ELBO on $\log p(\mathbf{J}|\mathbf{I})$. This information is enough to detect images that contain topological differences under the assumption that these images will overall have a lower likelihood. However, in our application, we need not only to detect the existence but also the position of outliers in the image. For this, we have to ensure that $\log p(\mathbf{J}|\mathbf{I})$ can be decomposed into a likelihood for each pixel of the image. It is immediately obvious by inspection of the ELBO (1) together with the KL-Divergence (2), that the lower bound on $\log p(\mathbf{J}|\mathbf{I})$ can be decomposed into pixel-wise terms if $\log p_{\text{noise}}(\mathbf{J}|\mathbf{I}\circ\Phi)$ can be decomposed as such. To enforce this, we will introduce a general form of error function, which can be decomposed and includes the MSE as a special case. For this, we first map the images $\mathbf{I}$ and $\mathbf{J}$ to feature maps over the pixel positions k via a mapping $f_k(\mathbf{I}) \in \mathbb{R}^F$ and define the loss as:

$$p_{\text{noise}}(\mathbf{J}|\mathbf{I}\circ\Phi) = \prod_{k=1}^{n} \mathcal{N}(f_k(\mathbf{J})|f_k(\mathbf{I})\circ\Phi, \Sigma_f)\ , \tag{3}$$

where $\Sigma_f \in \mathbb{R}^{F\times F}$ is a diagonal covariance matrix with variances learned during training.

The ability to decompose the likelihood is not enough for a *meaningful* metric, as we have to ensure that each term is calculated in the correct coordinate system. This depends on the parameterisation and regularisation of Φ. In the approach by Dalca et al. [9] the parameterization V of Φ is defined on the tangent space and consequently the prior is also on this space. Since the connection between Φ and V is given by integration of the vector field, decomposing (2) for a single pixel k will produce estimates based on the local differential of the transformation, but will not take the full path with starting and endpoints into account. Thus, correct cost assignments require integration of (2) over the computed path, which is expensive and suffers from severe integration inaccuracies. Instead, we will use an alternative approach, where we parameterize Φ directly as a vector field on the image domain. Transformations parameterized this way are not necessarily invertible anymore, yet smoothness is still encouraged by the prior.

Learnable prior Using this parameterization, we extend the approach by Dalca et al. [9] and introduce a parameterized prior on Φ_k that is learned simultaneously with the model:

$$p(\Phi) = \prod_{d=1}^{D} \mathcal{N}\left(\Phi^{(d)} \mid 0, \Lambda_{\alpha\beta}^{-1}\right),\ \ \Lambda_{\alpha\beta} = \alpha\Lambda + \frac{\beta}{n^2}\mathbb{1}\mathbb{1}^T \tag{4}$$

The expected variations and translations between transformation vectors are governed by α and β. Unlike most works in image registration, we do not treat these as tuneable hyperparameters, but instead view them as unknowns to be fitted to the data during training similar to [28, 49]. For efficient learning, we use an estimate for the optimal values for α, β over a batch of samples during training, and use a running average at test time. A detailed explanation is given in supplementary material A.

The second term of (4) ensures that $\Lambda_{\alpha\beta}$ is invertible, by adding a multiple of the eigenvector $\mathbb{1} = (1, \dots, 1)^T$. It can be verified easily that $\Lambda\mathbb{1} = 0$. Unlike adding a multiple of the identity matrix to Λ, adding the missing eigenvalue does not modify the prior in any other way than regularizing the translations. Further, it ensures that the KL divergence of the resulting matrix can be quickly computed up to a constant as α and β do not modify the same eigenvalues. Recomputing the KL-divergence for n transformation vectors in D dimensions leads to

$$\begin{aligned} 2\,\mathrm{KL}\left(q(\Phi|\mathbf{I},\mathbf{J})\|p_{\alpha\beta}(\Phi)\right) &= -(n-1)D\log\alpha - D\log\beta + \beta\sum_{d=1}^{D}\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i^{(d)}\right)^2 \\ &+ \sum_{d=1}^{D}\sum_{k=1}^{n} -\log v_k^{(d)} + \left(\alpha|N(k)| + \frac{\beta}{n^2}\right)v_k^{(d)} + \alpha\sum_{l\in N(k)}\left(\mu_k^{(d)} - \mu_l^{(d)}\right)^2 + \mathrm{const} \end{aligned} \tag{5}$$

Decomposed error metric We define our pixel-wise error measure for topological change detection based on the ELBO (1) with KL-divergence (5) as follows, where we compute $\mu_k^{(d)}$ and $v_k^{(d)}$ via the proposal distribution $q(\Phi|\mathbf{I},\mathbf{J})$ and pick $\Phi_k^{(d)} = \mu_k^{(d)}$:

$$\begin{aligned} L_k(\mathbf{J}|\mathbf{I}) &= -\log\mathcal{N}(f_k(\mathbf{J})|f_k(\mathbf{I})\circ\Phi,\Sigma_f) + \frac{\beta\mu_k^{(d)}}{n^2}\sum_{d=1}^{D}\sum_{i=1}^{n}\mu_i^{(d)} \\ &+ \sum_{d=1}^{D} -\log v_k^{(d)} + \left(\alpha|N(k)| + \frac{\beta}{n^2}\right)v_k^{(d)} + \alpha\sum_{l\in N(k)}\left(\mu_k^{(d)} - \mu_l^{(d)}\right)^2 . \end{aligned} \tag{6}$$

We will treat the loss over all pixels $L(\mathbf{J}|\mathbf{I}) = (L_1(\mathbf{J}|\mathbf{I}), \dots, L_n(\mathbf{J}|\mathbf{I}))$ as another image with domain and pixel coordinates the same as $\mathbf{J}$. This measure is not symmetric. The prior distribution does not treat the distributions $q(\Phi|\mathbf{I},\mathbf{J})$ and $q(\Phi|\mathbf{J},\mathbf{I})$ equally. If $\Phi_{\mathbf{J}\to\mathbf{I}}$ maps a line in $\mathbf{J}$ to an area in $\mathbf{I}$, this will incur a large visible feature along the line due to violating the smoothness assumption encoded in the prior. On the other hand, if an area in $\mathbf{J}$ gets mapped to a line in $\mathbf{I}$, the overall error contribution is smoothed out over the area. To rectify this issue, we will compute a bidirectional measure $L_{\text{sym}}(\mathbf{J}|\mathbf{I}) = L(\mathbf{J}|\mathbf{I}) + L(\mathbf{I}|\mathbf{J})\circ\Phi_{\mathbf{I}\to\mathbf{J}}$, where $\Phi_{\mathbf{I}\to\mathbf{J}}$ is the same as the one used to compute $L(\mathbf{J}|\mathbf{I})$. For this measure it holds that if $\Phi_{\mathbf{J}\to\mathbf{I}} = \Phi_{\mathbf{I}\to\mathbf{J}}^{-1}$, we have $L_{\text{sym}}(\mathbf{I}|\mathbf{J}) = L_{\text{sym}}(\mathbf{J}|\mathbf{I})\circ\Phi_{\mathbf{J}\to\mathbf{I}}$ up to interpolation errors caused by the finite coordinate grid.

Topological outlier detection L_{sym} detects topological changes between two images. However, for evaluation on the Brain dataset, we are interested in topological outliers. Outliers can be detected using L_{sym} by contrasting the observed deviations with the observed deviations within a larger set of control images $\mathcal{C}$. This leads to the score

$$Q(\mathbf{J}) = \mathbb{E}_{\mathbf{I}\in\mathcal{C}}\left[L_{\text{sym}}(\mathbf{J}|\mathbf{I}) - \mathbb{E}_{\mathbf{K}\in\mathcal{C}}\left[L_{\text{sym}}(\mathbf{I}|\mathbf{K})\right]\circ\Phi_{\mathbf{I}\to\mathbf{J}}\right]. \tag{7}$$

4 Evaluation

We evaluate our approach on two tasks. In the first, we measure prediction agreement with annotated topological changes on a dataset of cell slices. For this, we introduce the first dataset with annotated topological differences for image registration (see Section 4.1), which allows us to significantly expand on the evaluation strategies of prior work [26, 29, 30]. In the second task, we adapt our approach to anomaly detection in order to detect brain tumors on slices of MRI images.

On the change detection task, we use our model prediction of L_{sym} directly. On the anomaly detection task, we use the score (7), which subtracts the average scores over healthy patients for each pixel.

We compare our model to the following baselines:

1. Two unsupervised approaches for topological change detection:
   - Li and Wyatt's [30] intensity difference and image gradient-based approach using a deterministic registration model [5] to obtain the transformations.
   - Using the same model, we devise a method based on the Jacobian Determinant of the transformation field $|J_\Phi|$. We expect strong stretching or shrinkage in areas of topological mismatch, which we measure using use the score $\log(|\det J_\Phi|)^2$.

   We adapt both approaches to the task of tumor detection by subtracting the average scores over healthy patients, analogous to (7).
2. The approach by An and Cho [1] for unsupervised anomaly detection in images is based on the local reconstruction error of a variational autoencoder. The error score is $\|\mathbf{J} - \text{dec}(\text{enc}(\mathbf{J}))\|^2$, where $\text{enc}(\mathbf{J})$ maps $\mathbf{J}$ to the mean of the variational proposal distribution and dec is the corresponding learned decoder. As the score does not use registration, we cannot use equation (7).
3. A supervised segmentation model trained for segmenting topological changes based on two input images on the cell dataset, and tumor segmentation based on a single input image on the brain dataset. Since this model requires annotated data, we withhold 75% of the annotated volumes for training and evaluate the segmentation model only on the remaining samples.

In both tasks, we measure the pixel-wise agreement of the models with the annotated ground-truth using the receiver operating characteristic curve (ROC curve) and compare the area under the curve (AUC) between the models. AUC estimates are bootstrapped on the subject level to obtain error estimates.

As additional evaluations, we present qualitative examples and investigate whether brain regions with known topological variability get assigned higher scores in our model. For this we compute the pairwise average score L_{sym} over multiple healthy subjects and register them all to a brain atlas using $\mathbb{E}_{\mathbf{I},\mathbf{K}}\left[L_{\text{sym}}(\mathbf{I}|\mathbf{K}) \circ \Phi_{\mathbf{I}\to\text{Atlas}}\right]$. We group the scores by their position on the brain atlas into partitions: cortical surfaces, subcortical regions, and ventricles.

4.1 Tasks and Data

Topology change detection in Cells Serial block-face scanning electron microscopy (SBEM) is a method to obtain three-dimensional images from small biological samples. An image is taken from the face of the block, after which a thin section is cut from the sample to expose the next slice. A challenge is the accurate reconstruction of the volume, as neighboring slices differ by both natural deformations and changes in topology. Natural deformations can be introduced by shape-changes of objects between the slices, and deformations of the sample due to the physical cutting. Changes in topology occur due to objects present in one slice but not the other, and tears of the physical sample induced by the cutting.

We evaluate our method on the detection of topological changes between neighboring slices of human platelet cells recorded with SBEM. We use the pre-segmented dataset by Quay et al. [41] as a base. In the dataset, image slices are affinely pre-aligned and manually segmented into 7 classes. Afterwards, for the validation and test set, we annotated changes in the topology of the segmentation masks. Using this approach, not all instances of topological changes in the image can be annotated as the segmentation maps merge several types of cell components into a single class. The data is cropped into patches of 256×256 pixels and we use 9 patches of 50 slices for training, 4 patches of 24 slices for validation, and 5 patches of 24 slices for test (3 patches for the supervised approach due to the training-test split of annotated data).

Brain tumor detection Individual brains offer a range of topological differences, especially in the presence of tumors. Further, inter-subject differences are found at the cortical surface, where the sulci vary significantly [48], and near ventricles, which can either be open cavities, or partially closed [36]. We quantitatively evaluate our method on the proxy task of detecting brain tumors. Tumors change the morphology of the brain and can thus be detected indirectly via the large transformations they cause. For this, we first train our model using a dataset of healthy images from the control group and then use (7) to obtain a score for topological outlier detection. For the control set, we combine T1 weighted MRI scans of the healthy subjects from the ABIDE I [11][1], ABIDE II [12] and OASIS3 [27] studies.

[1] CC BY-NC-SA 3.0, `https://creativecommons.org/licenses/by-nc-sa/3.0/`

For the tumor set we use MRI scans from the BraTS2020 brain tumor segmentation challenge [3, 4, 34], which have expert-annotated tumors. We use the T1 weighted MRI scans, and combine labels of the classes necrotic/cystic and enhancing tumor core into a single tumor class. All datasets are anonymized, with no protected health information included and participants gave informed consent to data collection.

We perform standard pre-processing on both brain datasets, including intensity normalization, affine spatial alignment, and skull-stripping using FreeSurfer [14]. From each 3D volume, we extract a center slice of 160×224 pixels. Scans with preprocessing errors are discarded, and the remaining images of the control dataset are split 2381/149/162 for train/validation/test. Of the tumor dataset, 84 annotated images with tumors larger than 5cm^2 along the slice are used for evaluation (17 for the supervised approach due to the training-test split of subjects).

4.2 Model and training

All models evaluated are based on a U-Net [44] architecture, except An and Cho [1], which we implement using as a spatial VAE following the previously published adaptation to Brain-scans by Venkatakrishnan et al. [50]. The networks consist of encoder and decoder stages of $64, 128, 256$ channels for all registration models, and $32, 64, 128, 256$ channels for the segmentation and VAE models. Each stage consists of a batch normalization [23] and a convolutional layer.

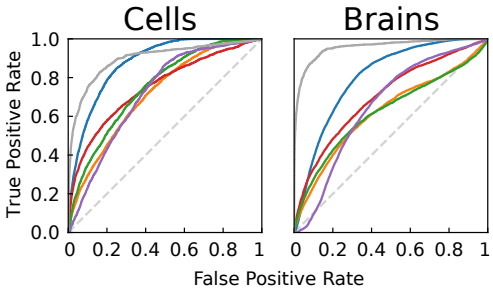
In our approach, we use a U-Net to model $p(\Phi|\mathbf{I}, \mathbf{J})$. The output of the last decoder stage is fed through separate convolution layers with linear activation functions to predict the transformation mean and log-scaled variance. Throughout the network, we use LeakyReLu activation functions. The generator step $\mathbf{I} \circ \Phi$ is implemented by a parameterless spatial transformer layer [24]. During training of our model, we use the analytical solution for prior parameters α, β (supplementary material, Eq. 8), averaged over the mini-batch of 32 image pairs. For validation and test, we use the running mean recorded during training. The diagonal covariance of the reconstruction loss Σ_f is treated as a trainable parameter.

For all datasets, we use data augmentation with random affine transformations of the training images. For training, the optimization algorithm is ADAM [25] with a learning rate of 10^{-4}. Regularization of all models is performed by applying an L_2-penalty to the weights with a factor of 0.01 for the cell dataset and 0.0005 for the brains. We train each model on a single TitanRTX GPU, with maximum training times of 1 day for the cells and 4 days for the brains. Hyperparameters: The network by Venkatakrishnan et al. [50] has $\sigma = 1$ chosen from $\{0.1, 1, 10\}$, based on reconstruction loss on validation set. The deterministic registration model was trained using $\lambda = 0.1$ as in [8]. For [30], the parameters σ of the Gaussian derivative kernel and hyper-parameter K where chosen to maximize the AUC score, selecting $\sigma = 6$, $K = 2$ out of $\{1, \ldots, 9\}^2$.

For the reconstruction loss, we compare two different loss functions. The first is using the MSE as in [9, 30]. The second is a semantic similarity metric similar to [8]. To obtain the semantic image descriptors, we train a U-net with $32, 64, 64$ channels for image segmentation, using the manual annotations of the cell set and automatically created labels obtained with FreeSurfer [14] for the brain control images. Notably, the segmentation models used for the loss have not been trained on images or pairs containing topological changes or tumors. From this network, we extract the features of the first three stages and use them as a 160-channel feature map in the loss (3). For both the MSE and the semantic loss, we learn the variance parameters while training the variational autoencoder.

4.3 Results

The ROC curves of all trained models on the cell and brain tasks can be seen in Figure 2. For both tasks, the supervised model performed best (AUC 0.90, 0.95), while our proposed approach with semantic loss performed best among the unsupervised models (AUC 0.88, 0.80). The unsupervised approach for topological change detection by Li and Wyatt [30] (AUC 0.75, 0.70) performed overall best among the baselines, but worse than our method. The unsupervised anomaly detection method by An and Cho [1] (AUC 0.72, 0.67) performed well at detecting brain tumors, but worse at detecting topological changes in the cell images. Using the Jacobian determinant (AUC 0.75, 0.62) performed well on the cell images but worse on the brain tumor detection task. Our approach using MSE (AUC 0.72, 0.61) performed worse than the other methods on both tasks.



Method	AUC (Cells)	AUC (Brains)
Ours (Sem. Loss)	$.877 \pm .003$	$.797 \pm .005$
Ours (MSE)	$.712 \pm .003$	$.611 \pm .009$
Li and Wyatt	$.755 \pm .009$	$.697 \pm .006$
Jac. Det.	$.752 \pm .005$	$.616 \pm .011$
An and Cho	$.718 \pm .004$	$.670 \pm .011$
Supervised Seg.	$.899 \pm .004$	$.946 \pm .025$

Figure 2: Receiver operating characteristic curves (ROC) and area under the curve (AUC) for detecting topological changes on the cell and brain datasets. We test models of our method for unsupervised topological change detection, trained with a semantic loss function and the MSE in the reconstruction term, and compare against unsupervised baselines from image registration (Li and Wyatt [30], Jacobian Determinant) and unsupervised anomaly detection (An and Cho [1]). For reference, we also include a supervised segmentation model, which has been trained on the ground truth annotations.

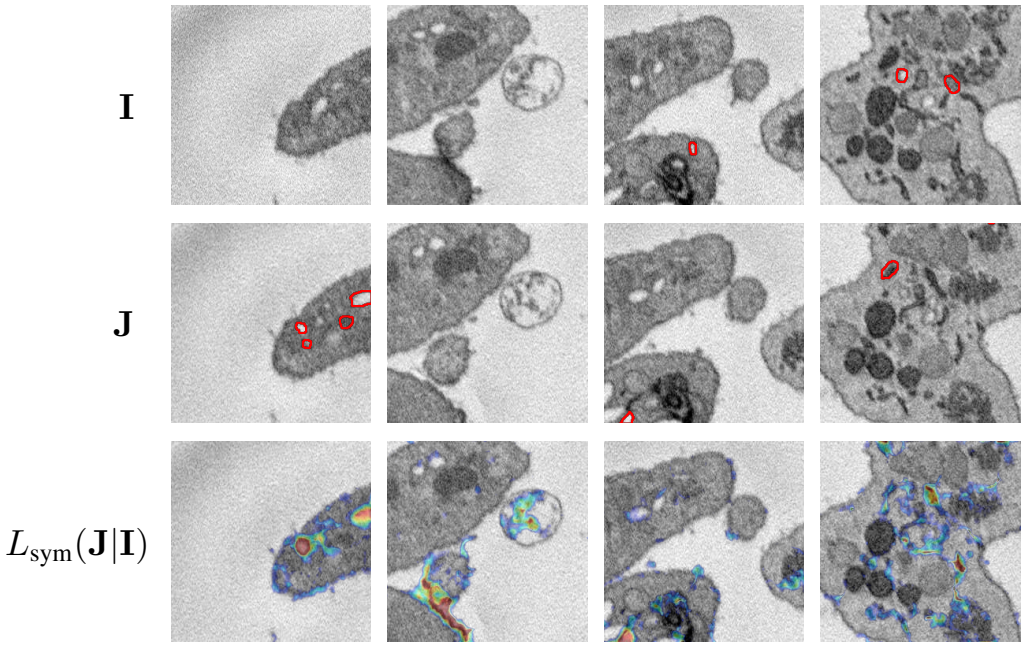


Figure 3: Topological differences detected by our method, cell dataset. Neighboring slices **I**, **J** in rows 1 and 2. Heatmaps of the likelihood of topological differences detected with L_{sym} in row 3. Heatmaps are overlayed on image **J** to ease comparison. Annotated topological differences used for evaluation outlined in red. Note that only a subset of topological anomalies present is annotated in our dataset.

When analyzing the ROC curves, our model performed best among the unsupervised models for all false positive rates, while the supervised model is the best overall. Finally, even though both models share the same trained model, the score used by Li and Wyatt [30] performed better than scoring using the Jacobian determinant on the brain tumor detection task, while on the cell dataset, both approaches performed the same.

We show qualitative results on the cell dataset in Figure 3. In row 3, we see that L_{sym} detected annotated areas of topological change (contoured in red), but is more certain at detecting changes in areas with high intensity difference. In many cases, the model assigns a likelihood of topological changes to areas that have not been annotated in the dataset, such as the merging cell boundary in column 2 or many small changes in the cell interior in column 4.

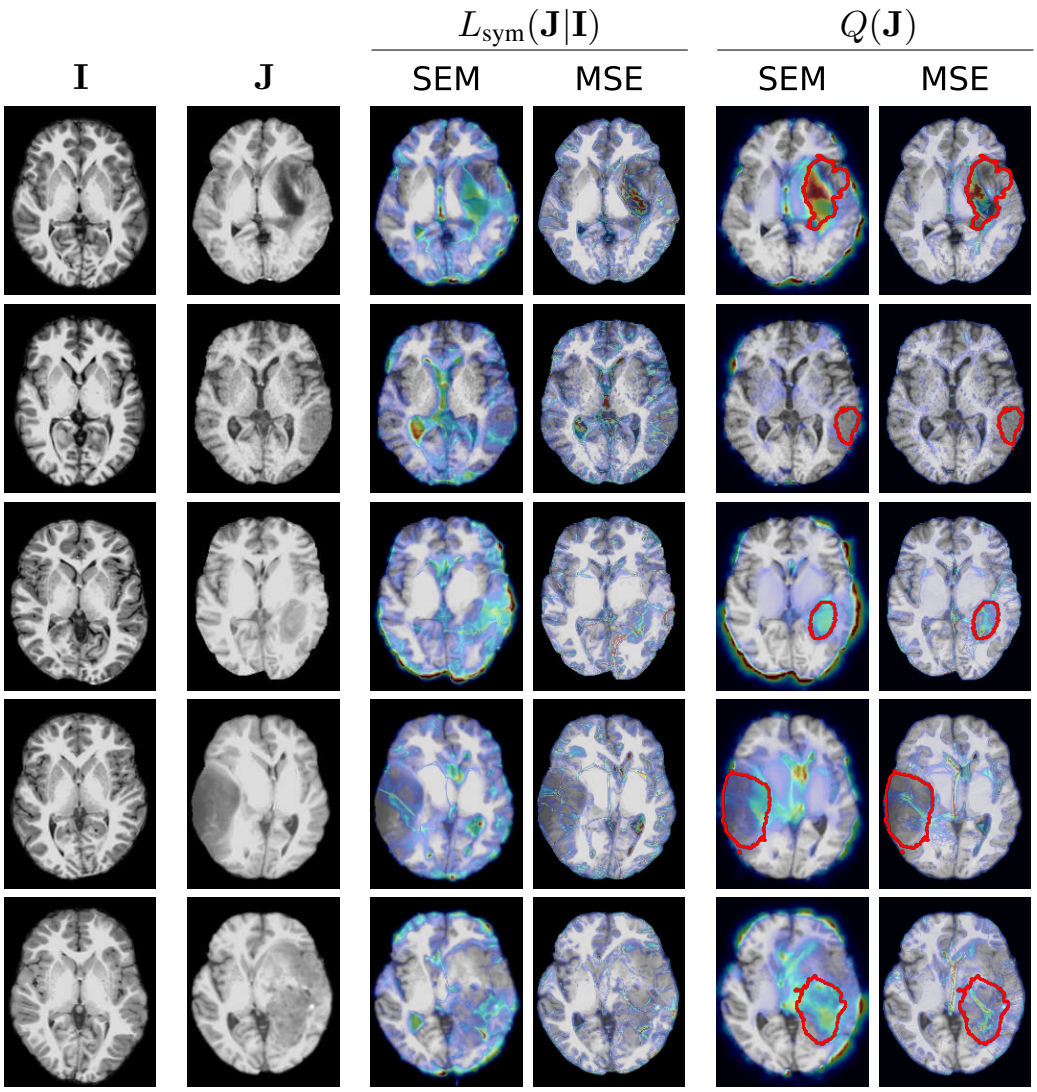


Figure 4: Topological differences detected by our method, brain dataset. Structurally normal brain $\mathbf{I}$ in column 1, brain with tumor $\mathbf{J}$ in column 2. Heatmaps of the likelihood of topological differences detected with L_{sym} in columns 3, 4 . Likelihood of topological differences caused by the structural anomaly filtered by Eq. 7 in columns 5, 6. Contour of the ground truth brain tumor in red. Heatmaps are overlayed on image $\mathbf{J}$ to ease comparison.

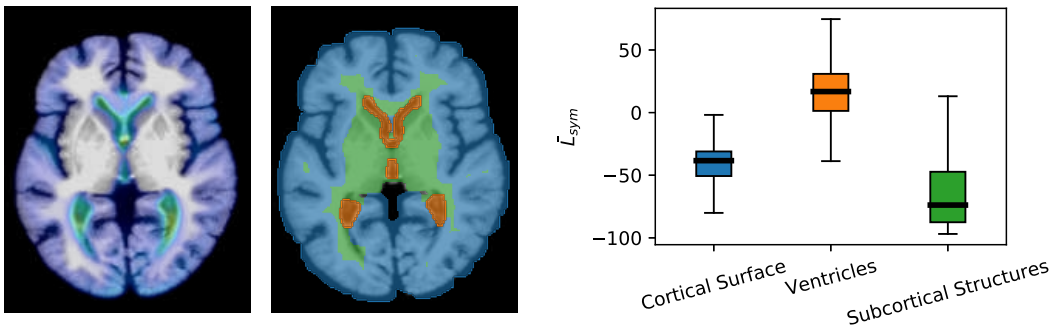


Figure 5: Left: Heatmap of average location of topological differences among the control group, predicted by the semantic model, averaged with $\mathbb{E}_{\mathbf{I},\mathbf{K}}\left[L_{\text{sym}}(\mathbf{I}|\mathbf{K}) \circ \Phi_{\mathbf{I}\to\text{Atlas}}\right]$ using a brain atlas as reference image. Center: We use morphological operations to split the atlas into cortical surface (blue), ventricles (orange) and sub-cortical structures (green). Right: Likelihood of topological differences occurring in each region. Boxplot with median, quartiles, deciles.

Qualitative results on the brain data are presented in Figure 4. When looking at columns 3 and 4, we see that L_{sym} detected notable areas with high changes in topology compared to the reference image $\mathbf{I}$. This includes the ventricles (rows 2,3), the cortical areas with the sulci (all rows) as well as tumor areas (rows 1,3,5). There was a clear difference in the behaviour between semantic loss and MSE as the semantic loss highlights broader regions of the surface. When comparing the outlier-detection measure $Q(\mathbf{J})$ in columns 5 and 6, we can see that our approach filtered most of the ventricles and sulci leaving an area around most tumor regions. Notable exceptions are rows 2 and 4, where the tumor area was not highlighted, as well as row 1 where only part of the tumor was detected.

In Figure 5, we show the average topological change score on healthy subjects. We see on the brain image and the box plot, that the cortical surfaces and ventricles get assigned higher scores than the subcortical structures.

5 Discussion and conclusion

In this work, we have introduced a novel approach for the detection of topological changes. We evaluated our approach qualitatively and compared it quantitatively to previous approaches using both a novel dataset with purpose-made annotations and on an unsupervised segmentation proxy task. On both tasks, our approach performed best among the unsupervised methods, but could not reach the performance of the supervised method.

An unsupervised method is useful in practice, as annotations of topological changes are rarely available. While our results are not pixel exact, they indicate where a registration algorithm must be used more carefully to obtain a valid registration. The results on the cell dataset align well with the annotations, and many of the false positives appear to be caused by incomplete annotation of the data. This is also reflected in the reported ROC-curves, which show that our model outperforms the supervised segmentation model at false positive rates larger than 0.5. The results obtained on the tumor segmentation proxy task are reinforced by the distribution of scores obtained on healthy patients in different parts of the brain. The high likelihood of topological differences in ventricles found agrees with previous work [36] and the higher scores in cortical surfaces reflect the fact, that the sulci of the cortical surface exhibit high variability between subjects [7], which was previously difficult to quantify.

Our results also show that using a semantic loss function is advantageous compared to the MSE in this task, as all MSE based methods performed worse than our approach using the semantic loss. This is likely because the contrast between some anatomical areas is quite small and thus missed by the MSE. In contrast, the semantic loss incorporates more texture information and thus is capable of differentiating between areas of similar intensity but different semantics. However, particularly on the brain example, even the semantic approach misses tumors close to the cortex. We hypothesize, that this is in part caused by the similar appearance of tumors and grey matter, in part by the semantic model not being trained on tumors, and in part due to the cortical area containing high topological variation among the control group as well.

On the brain dataset, our unsupervised results for the method by An and Cho [1] are in line with previously reported results on a comparable dataset [50]. However, our supervised results are not comparable to the results published for the BRATS challenge, as we selected a subset of data for training and only used structural MRI images, discarding the other modalities. On the cell dataset, no other work on topology change or outlier detection is available.

Our study has several limitations. We only investigate registrations in 2D and topological differences might vanish if the whole 3D volume is considered. The transformations obtained by our unsupervised method differ from strongly regularised methods, as the hyperparameter-less learned prior under-regularises in order to maximize the likelihood of a topological match during training. Conversely, the poor performance of the Jacobian determinant might be due to a strong regularisation for good performance in image registration as we used the hyperparameters as found in [8].

In conclusion, our approach serves as the first step for unsupervised annotation of topological changes in image registration. Our approach is fully unsupervised and hyperparameter-free, making it a prospective building block in an end-to-end topology-aware image registration model.

Acknowledgements

This work was funded by the Novo Nordisk Foundation (grants no. NNF20OC0062606 and NNF17OC0028360) and the Lundbeck Foundation (grant no. R218-2016-883).

The human platelet SBEM data and segmentations were provided by Matthew Quay, the topological change annotations are ours. The Brain tumor data was provided by the BraTS challenge. The Brain control data was provided in part by OASIS Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly-owned subsidiary of Eli Lilly.

References

[1] Jinwon An and Sungzoon Cho. *Variational Autoencoder based Anomaly Detection using Reconstruction Probability*. Tech. rep. 2015.

[2] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. "A log-euclidean framework for statistics on diffeomorphisms". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer Verlag, 2006, pp. 924–931.

[3] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features". In: *Scientific Data* 4 (2017).

[4] Spyridon Bakas et al. "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge". In: *arXiv* 124 (2018).

[5] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. "VoxelMorph: A Learning Framework for Deformable Medical Image Registration". In: *IEEE Transactions on Medical Imaging* 38.8 (2019), pp. 1788–1800.

[6] Xiang Chen, Nishant Ravikumar, Yan Xia, and Alejandro F Frangi. "A Deep Discontinuity-Preserving Image Registration Network". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2021.

[7] Nicolas Courty, Remi Flamary, Devis Tuia, and Alain Rakotomamonjy. "Optimal Transport for Domain Adaptation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.9 (2017), pp. 1853–1865.

[8] Steffen Czolbe, Oswin Krause, and Aasa Feragen. "Semantic similarity metrics for learned image registration". In: *Proceedings of Machine Learning Research* (2021).

[9] Adrian V. Dalca, Guha Balakrishnan, John Guttag, and Mert R. Sabuncu. "Unsupervised Learning for Fast Probabilistic Diffeomorphic Registration". In: *Medical Image Computing and Computer Assisted Intervention* (2018), pp. 729–738.

[10] V Delmon, S Rit, R Pinho, and D Sarrut. "Registration of sliding objects using direction dependent B-splines decomposition Registration of sliding objects using direction dependent B-splines decomposition *". In: *Phys. Med. Bio* 58.5 (2013), pp. 1303–1314.

[11] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, et al. "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism". In: *Molecular psychiatry* 19.6 (2014), pp. 659–667.

[12] Adriana Di Martino et al. "Enhancing studies of the connectome in autism using the autism brain imaging data exchange II". In: *Scientific Data* 4.1 (2017), pp. 1–15.

[13] Mirza Faisal Beg, Michael I Miller, Alain Trouvétrouv, and Laurent Younes. "Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms". In: *International Journal of Computer Vision* 61.2 (2005), pp. 139–157.

[14] B Fischl. *FreeSurfer. Neurolmage, 62 (2), 774–781*. 2012.

[15] Ulf Grenander and Michael I Miller. "Computational anatomy: An emerging discipline". In: *Quarterly of applied mathematics* 56 (1998), pp. 617–694.

[16] Xu Han, Roland Kwitt, Stephen Aylward, Spyridon Bakas, Bjoern Menze, Alexander Asturias, Paul Vespa, John Van Horn, and Marc Niethammer. "Brain extraction from normal and pathological images: A joint PCA/Image-Reconstruction approach". In: *NeuroImage* 176 (2018), pp. 431–445.

[17] Xu Han, Zhengyang Shen, Zhenlin Xu, Spyridon Bakas, Hamed Akbari, Michel Bilello, Christos Davatzikos, and Marc Niethammer. "A Deep Network for Joint Registration and Reconstruction of Images with Pathologies". In: *11th International Workshop on Machine Learning in Medical Imaging*. Springer Nature, 2020, pp. 342–352.

[18] Xu Han, Xiao Yang, Stephen Aylward, Roland Kwitt, and Marc Niethammer. "Efficient registration of pathological images: A joint PCA/image-reconstruction approach". In: *4th International Symposium on Biomedical Imaging* 2017 (2017), pp. 10–14.

[19] Lasse Hansen and Mattias P. Heinrich. "Tackling the Problem of Large Deformations in Deep Learning Based Medical Image Registration Using Displacement Embeddings". In: *Medical Imaging with Deep Learning* (2020).

[20] Søren Hauberg, Oren Freifeld, Anders Boesen, Lindbo Larsen, John W Fisher, Iii Lars, and Kai Hansen. "Dreaming More Data: Class-dependent Distributions over Diffeomorphisms for Learned Data Augmentation". In: *Artificial Intelligence and Statistics*. Vol. 41. PMLR, 2016, pp. 342–350.

[21] Berthold KP Horn and Brian G Schunck. "Determining optical flow". In: *Artificial intelligence* 1 (1981), pp. 185–203.

[22] Rui Hua, Jose M. Pozo, Zeike A. Taylor, and Alejandro F. Frangi. "Multiresolution eXtended Free-Form Deformations (XFFD) for non-rigid registration with discontinuous transforms". In: *Medical Image Analysis* 36 (2017), pp. 113–122.

[23] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International Conference on Machine Learning*. International Machine Learning Society (IMLS), 2015, pp. 448–456.

[24] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. "Spatial Transformer Networks". In: *Advances in neural information processing systems* (2015).

[25] Diederik P. Kingma and Jimmy Lei Ba. "Adam: A method for stochastic optimization". In: *International Conference on Learning Representations*. 2015.

[26] Dongjin Kwon, Marc Niethammer, Hamed Akbari, Michel Bilello, Christos Davatzikos, and Kilian M. Pohl. "PORTR: Pre-operative and post-recurrence brain tumor registration". In: *IEEE Transactions on Medical Imaging* 33.3 (2014), pp. 651–667.

[27] Pamela J LaMontagne, Tammie L S Benzinger, John C Morris, et al. "OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease". In: *medRxiv* (2019).

[28] Dong Bok Lee, Dongchan Min, Seanie Lee, and Sung Ju Hwang. "Meta-GMVAE: Mixture of Gaussian VAE for Unsupervised Meta-Learning". In: *International Conference on Learning Representations*. 2020.

[29] Xiaoxing Li, Xiaojing Long, Christopher Wyatt, and Paul Laurienti. "Registration of Images with Varying Topology Using Embedded Maps". In: *IEEE Transactions on Medical Imaging* 31.3 (2012), pp. 749–765.

[30] Xiaoxing Li and Chritopher Wyatt. "Modeling topological changes in deformable registration". In: *2010 7th IEEE International Symposium on Biomedical Imaging*. 2010, pp. 360–363.

[31] Lihao Liu, Xiaowei Hu, Lei Zhu, and Pheng-Ann Heng. "Probabilistic Multilayer Regularization Network for Unsupervised 3D Brain Image Registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019, pp. 346–354.

[32] Xiaoxiao Liu, Marc Niethammer, Roland Kwitt, Matthew McCormick, and Stephen Aylward. "Low-Rank to the Rescue – Atlas-Based Analyses in the Presence of Pathologies". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8675 LNCS.PART 3 (2014), pp. 97–104.

[33] Xiaoxiao Liu, Marc Niethammer, Roland Kwitt, Nikhil Singh, Matt McCormick, and Stephen Aylward. "Low-Rank Atlas Image Analyses in the Presence of Pathologies". In: *IEEE Transactions on Medical Imaging* 34.12 (2015), pp. 2583–2591.

[34] Bjoern H. Menze et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)". In: *IEEE Transactions on Medical Imaging* 34.10 (2015), pp. 1993–2024.

[35] Didrik Nielsen, Priyank Jaini, Emiel Hoogeboom, Ole Winther, and Max Welling. *SurVAE Flows: Surjections to Bridge the Gap between VAEs and Flows*. Tech. rep. 2020, pp. 12685–12696.

[36] Rune Kok Nielsen, Sune Darkner, and Aasa Feragen. "TopAwaRe: Topology-Aware Registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), pp. 364–372.

[37] Danielle F. Pace, Stephen R. Aylward, and Marc Niethammer. "A locally adaptive regularization based on anisotropic diffusion for deformable image registration of sliding organs". In: *IEEE Transactions on Medical Imaging* 32.11 (2013), pp. 2114–2126.

[38] Bartłomiej W. Papiez, Mattias P. Heinrich, Jérome Fehrenbach, Laurent Risser, and Julia A. Schnabel. "An implicit sliding-motion preserving regularisation via bilateral filtering for deformable image registration". In: *Medical Image Analysis* 18.8 (2014), pp. 1299–1311.

[39] Sarah Parisot, William Wells, Stéphane Chemouny, Hugues Duffau, and Nikos Paragios. "Concurrent tumor segmentation and registration with uncertainty-based sparse non-uniform graphs". In: *Medical Image Analysis* 18.4 (2014), pp. 647–659.

[40] Gabriel Peyré and Marco Cuturi. "Computational optimal transport". In: *Foundations and Trends in Machine Learning* 11.5-6 (2019), pp. 1–257.

[41] Matthew Quay, Zeyad Emam, Adam Anderson, and Richard Leapman. "Designing deep neural networks to automate segmentation for serial block-face electron microscopy". In: *International Symposium on Biomedical Imaging*. Vol. 2018-April. IEEE Computer Society, 2018, pp. 405–408.

[42] Danilo Jimenez Rezende and Shakir Mohamed. "Variational Inference with Normalizing Flows". In: *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.

[43] Laurent Risser, François Xavier Vialard, Habib Y. Baluwala, and Julia A. Schnabel. "Piecewise-diffeomorphic image registration: Application to the motion estimation between 3D CT lung images with sliding conditions". In: *Medical Image Analysis* 17.2 (2013), pp. 182–193.

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Vol. 9351. Springer Verlag, 2015, pp. 234–241.

[45] Dan Ruan, Selim Esedoğlu, and Jeffrey A. Fessler. "Discriminative sliding preserving regularization in medical image registration". In: *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009*. 2009, pp. 430–433.

[46] Alexander Schmidt-Richberg, Jan Ehrhardt, René Werner, and Heinz Handels. "Fast explicit diffusion for registration with direction-dependent regularization". In: *Biomedical Image Registration* 7359 (2012), pp. 220–228.

[47] Kihyuk Sohn, Xinchen Yan, and Honglak Lee. "Learning Structured Output Representation using Deep Conditional Generative Models". In: *Advances in Neural Information Processing Systems*. Vol. 28. 2015, pp. 3483–3491.

[48] Elizabeth R. Sowell, Paul M. Thompson, David Rex, David Kornsand, Kevin D. Tessner, Terry L. Jernigan, and Arthur W. Toga. "Mapping sulcal pattern asymmetry and local cortical surface gray matter distribution in vivo: Maturation in perisylvian cortices". In: *Cerebral Cortex* 12.1 (2002), pp. 17–26.

[49] Arash Vahdat and Jan Kautz. "NVAE: A Deep Hierarchical Variational Autoencoder". In: *Advances in Neural Information Processing Systems* (2020).

[50] Abinav Ravi Venkatakrishnan, Seong Tae Kim, Rami Eisawy, Franz Pfister, and Nassir Navab. "Self-Supervised Out-of-Distribution Detection in Brain CT Scans". In: *Medical Imaging Meets NeurIPS Workshop* (2020).

[51] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. "Occlusion Aware Unsupervised Learning of Optical Flow". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4884–4893.

[52] Xiao Yang, Xu Han, Eunbyung Park, Stephen Aylward, Roland Kwitt, and Marc Niethammer. "Registration of Pathological Images". In: *Simulation and synthesis in medical imaging (Workshop)* 9968 (2016), p. 97.

[53] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. "Quicksilver: Fast predictive image registration - A deep learning approach". In: *NeuroImage* 158 (2017), pp. 378–396.

8

IS SEGMENTATION UNCERTAINTY USEFUL?

PAPER 4:

Steffen Czolbe, Kasra Arnavaz, Oswin Krause, and Aasa Feragen (Jun 2020). "Is segmentation uncertainty useful?". In: *International Conference on Information Processing in Medical Imaging (IPMI 2021).*

I share the first authorship of this work evenly with my collaborator and fellow Ph.D. student Kasra Arnavaz. I also want to thank Aasa Feragen for initiating the project and being very involved throughout. The paper was published at IPMI 2021.

Is segmentation uncertainty useful?

Steffen Czolbe[†1], Kasra Arnavaz[†1], Oswin Krause[1], and Aasa Feragen[2]

[1] University of Copenhagen, Department of Computer Science, Denmark,
per.sc,kasra,oswin.krause@di.ku.dk
[2] Technical University of Denmark, DTU Compute, Denmark,
afhar@dtu.dk
† Authors contributed equally

Abstract. Probabilistic image segmentation encodes varying prediction confidence and inherent ambiguity in the segmentation problem. While different probabilistic segmentation models are designed to capture different aspects of segmentation uncertainty and ambiguity, these modelling differences are rarely discussed in the context of applications of uncertainty. We consider two common use cases of segmentation uncertainty, namely assessment of segmentation quality and active learning. We consider four established strategies for probabilistic segmentation, discuss their modelling capabilities, and investigate their performance in these two tasks. We find that for all models and both tasks, returned uncertainty correlates positively with segmentation error, but does not prove to be useful for active learning.

Keywords: Image segmentation · Uncertainty quantification · Active learning.

1 Introduction

Image segmentation – the task of delineating objects in images – is one of the most crucial tasks in image analysis. As image acquisition methods can introduce noise, and experts disagree on ground truth segmentations in ambiguous cases, predicting a single segmentation mask can give a false impression of certainty. Uncertainty estimates inferred from the segmentation model can give some insight into the confidence of any particular segmentation mask, and highlight areas of likely segmentation error to the practitioner. It adds transparency to the segmentation algorithm and communicates this uncertainty to the user. This is particularly important in medical imaging, where segmentation is often used to understand and treat disease. Consequently, quantification of segmentation uncertainty has become a popular topic in biomedical imaging [6, 11].

Training segmentation networks requires large amounts of annotated data, which are costly and cumbersome to attain. Active learning aims to save the annotator's time by employing an optimal data gathering strategy. Some active

Code available at github.com/SteffenCzolbe/probabilistic_segmentation

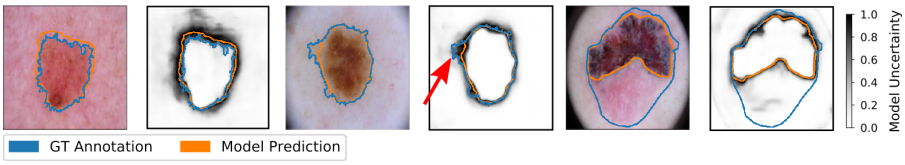


Fig. 1: Segmentation uncertainty is often interpreted as probable segmentation error, as seen near the lesion boundary in the first two examples. In the third example, however, model bias leads to a very certain, yet incorrect segmentation.

learning methods use uncertainty estimates to select the next sample to annotate [7,9,10]. While several potential such data gathering strategies exist [13,16], a consistent solution remains to be found [8].

While several methods have been proposed to quantify segmentation uncertainty [4, 6, 11], it is rarely discussed what this uncertainty represents, whether it matches the user's interpretation, and if it can be used to formulate a data-gathering strategy. We compare the performance of several well-known probabilistic segmentation algorithms, assessing the quality and use cases of their uncertainty estimates. We consider two segmentation scenarios: An unambiguous one, where annotators agree on one underlying true segmentation, and an ambiguous one, where a set of annotators provide potentially strongly different segmentation maps, introducing variability in the ground truth annotation.

We investigate the degree to which the inferred uncertainty correlates with segmentation error, as this is how reported segmentation uncertainty would typically be interpreted by practitioners. We find that uncertainty estimates of the models coincide with likely segmentation errors and strongly correlate with the uncertainty of a set of expert annotators. Surprisingly, the model architecture used does not have a strong influence on the quality of estimates, with even a deterministic U-Net [12] giving good pixel-level uncertainty estimates.

Second, we study the potential for uncertainty estimates to be used for selecting samples for annotation in active learning. Reducing the cost of data annotation is of utmost importance in biomedical imaging, where data availability is fast-growing, while annotation availability is not. We find that there are many pitfalls to an uncertainty-based data selection strategy. In our experiment with multiple annotators, the images with the highest model uncertainty were precisely those images where the annotators were also uncertain. Labeling these ambiguous images by a group of expert annotators yielded conflicting ground truth annotations, providing little certain evidence for the model to learn from.

2 Modelling segmentation uncertainty

Image segmentation seeks to estimate a well-defined binary[3] segmentation $g\colon \Omega \to \{0,1\}$ for a discrete image domain Ω. Typically, a predictive model $h(\mathbf{x}, \mathbf{w})$ with

[3] For simplicity, we consider binary segmentation; the generalization to multi-class segmentation is straightforward.

parameters $\mathbf{w}$, such as a neural network, is fitted to binary annotation data $a\colon \Omega \to \{0,1\}$ by minimizing a loss $\mathcal{L}(a, h(\mathbf{x}, \mathbf{w}))$. Here, $\mathbf{x} \in \mathbb{R}^{\Omega}$ is the image, and $\mathbf{y} = h(\mathbf{x}, \mathbf{w})$ defines an image of pixel-wise segmentation probabilities, such as the un-thresholded softmax output of a segmentation network h.

Typically, the annotation is assumed to be error-free, that is $a = g$, and predictors are typically trained on a single annotation per image. We assume that the trained neural network $h(\mathbf{x}, \mathbf{w})$ satisfies

$$h(\mathbf{x}, \mathbf{w}) = g(\mathbf{x}) + b + err \ ,$$

where b and err denote bias and segmentation error. Segmentation uncertainty is often interpreted as correlating with this error, although this is primarily realistic for small bias. Such segmentation tasks are called *unambiguous*; we consider a running example of skin lesion segmentation from dermoscopic images [3, 15], where the lesion boundary is clearly visible in the image (Fig. 1).

Recent work has considered *ambiguous* segmentation tasks [6,11], where there is no accessible "ground truth" segmentation, either because the data is not sufficient to estimate the segmentation, or because there is subjective disagreement. Examples include lesions in medical imaging, where the boundary can be fuzzy due to gradual infiltration of tissue, or where experts disagree on whether a tissue region is abnormal or not.

In such tasks, we make no assumption on the underlying segmentation g or the errors err, but regard the observed annotations as samples from an unknown "ground truth" distribution $p(a|\mathbf{x})$ over annotations a conditioned on the image $\mathbf{x}$. The goal of segmentation is to estimate the distribution $p(a|\mathbf{x})$, or its proxy distribution $p(\mathbf{y}|\mathbf{x})$ over pixel-wise class probabilities $\mathbf{y}\colon \Omega \to [0,1]$, as accurately as possible for a given image $\mathbf{x}$. If successful, such a model can sample coherent, realistic segmentations from the distribution, and estimate their variance and significance. As a running example of an ambiguous segmentation task, we consider lung lesions [1, 2, 6]. For such tasks, predictors are typically trained on multiple annotators, who may disagree both on the segmentation boundary and on whether there is even an object to segment.

From the uncertainty modelling viewpoint, these two segmentation scenarios are rather different. Below, we discuss differences in uncertainty modelling for the two scenarios and four well-known uncertainty quantification methods.

3 Probabilistic Segmentation Networks

A probabilistic segmentation model seeks to model the distribution $p(\mathbf{y}|\mathbf{x})$ over segmentations given an input image $\mathbf{x}$. Here, our annotated dataset $(\mathbf{X}, \mathbf{A})$ consists of the set $\mathbf{X}$ of N images $\left\{\mathbf{x}_n \mid n = 1, ..., N\right\}$, and L annotations are available per image, so that $\mathbf{A} = \left\{a_n^{(l)} \sim p(\mathbf{y}|\mathbf{x}_n) \mid (n,l) = (1,1), ..., (N,L)\right\}$.

Taking a Bayesian view, we seek the distribution

$$p(\mathbf{y}|\mathbf{x}, \mathbf{X}, \mathbf{A}) = \int p(\mathbf{y}|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathbf{X}, \mathbf{A}, h)\, d\mathbf{w} \ , \tag{1}$$

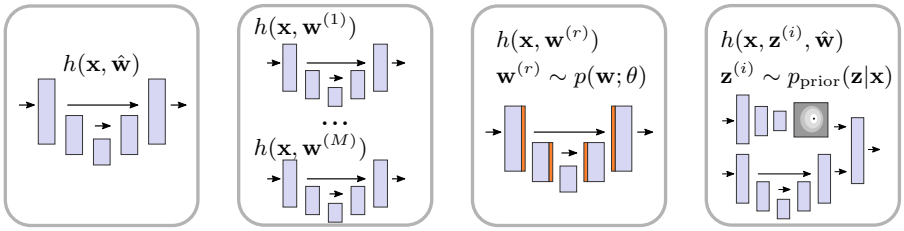


Fig. 2: Schematic overview (adapted from [6]) of the evaluated models. Blue: residual blocks . Orange: Dropout layers essential to the networks' functionality.

over segmentations $\mathbf{y}$ given image $\mathbf{x}$ and data $(\mathbf{X}, \mathbf{A})$, which can be obtained by marginalization with respect to the weights $\mathbf{w}$ of the model h.

In most deep learning applications, our prior belief over the model h, denoted $p(h)$, is modelled by a Dirac delta distribution indicating a single architecture with no uncertainty. In the context of uncertain segmentation models, however, we would like to model uncertainty in the parameters $\mathbf{w}$. Denoting our prior belief over the parameters $\mathbf{w}$ by $p(\mathbf{w}|h)$, Bayes' theorem gives

$$p(\mathbf{w}|\mathbf{X}, \mathbf{A}, h) = \frac{p(\mathbf{w}|h)p(\mathbf{A}|\mathbf{X}, \mathbf{w}, h)}{p(\mathbf{A}|\mathbf{X}, h)}, \tag{2}$$

where the likelihood update function is given by

$$p(\mathbf{A}|\mathbf{X}, \mathbf{w}, h) = \exp\left(\sum_{n=1}^{N}\sum_{l=1}^{L} \mathbf{A}^{(l)} \log\left(h(\mathbf{x}_n, \mathbf{w})\right) + (1 - \mathbf{A}^{(l)}) \log\left(1 - h(\mathbf{x}_n, \mathbf{w})\right)\right)$$

and normalizing constant

$$p(\mathbf{A}|\mathbf{X}, h) = \int p(\mathbf{w}|h)p(\mathbf{A}|\mathbf{X}, \mathbf{w}, h)\, d\mathbf{w} \ .$$

This integral is generally intractable, making it impossible to obtain the proper posterior (2). Below, we discuss how empirical approximations $\hat{p}(\mathbf{y}|\mathbf{x}, \mathbf{X}, \mathbf{A})$ to the distribution $p(\mathbf{y}|\mathbf{x}, \mathbf{X}, \mathbf{A})$ found in (1) are performed in four common segmentation models. Note that both p and $\hat{p}$ can be degenerate, depending on the number of annotations available and models used.

U-Net with softmax output. The well established U-Net [12] architecture with a softmax output layer yields class-likelihood estimates. As the model is deterministic, $p(h|\mathbf{X}, \mathbf{A})$ is degenerate. Parameters are selected by a maximum a posteriori (MAP) estimate i.e. $p(\mathbf{w}|\mathbf{X}, \mathbf{A}, h) \approx \delta(\mathbf{w} - \hat{\mathbf{w}})$ in which $\hat{\mathbf{w}} = \text{argmax}\ p(\mathbf{w}|\mathbf{X}, \mathbf{A}, h)$. The model output (1) is approximated by the degenerate distribution $\hat{p}(\mathbf{y}|\mathbf{x}, \mathbf{X}, \mathbf{A}) \approx p(\mathbf{y}|\mathbf{x}, \hat{\mathbf{w}})$. The softmax output layer predicts

a pixel-wise class probability distribution $p(\mathbf{y}_{(i,j)}|\mathbf{x}, \mathbf{X}, \mathbf{A})$. As no co-variance or dependencies between pixel-wise estimates are available, segmentation masks sampled from the pixel-wise probability distributions are often noisy [11]. An alternative approach followed by our implementation is the thresholding of pixel-wise probability values, which leads to a single, coherent segmentation map.

Ensemble methods combine multiple models to obtain better predictive performance than that obtained by the constituent models alone, while also allowing the sampling of distinct segmentation maps from the ensemble. We combine M U-Net models $h(\mathbf{x}, \mathbf{w}^{(m)})$ where, if labels from multiple annotators are available, each constituent model is trained on a disjoint label set $\mathbf{A}^{(m)}$. When trained on datasets with a single label, all constituent models are trained on the same data and their differences stem from randomized initialization and training. Treating the models as samples, we obtain an empirical distribution approximating (1) by drawing from the constituent models at random.

Monte-Carlo Dropout [4] is a Bayesian approximation technique based on dropout, where samples from the posterior over dropout weights give a better approximation of the true posterior than a MAP estimation. Given a selected model h, one can approximate (1) as $\hat{p}(\mathbf{y}|\mathbf{x}, \mathbf{X}, \mathbf{A}) \approx 1/R \sum_{r=1}^{R} p(\mathbf{y}|\mathbf{x}, \mathbf{w}^{(r)})$ when $\mathbf{w}^{(r)} \sim p(\mathbf{w}|\mathbf{X}, \mathbf{A}, h)$. Since $p(\mathbf{w}|\mathbf{X}, \mathbf{A}, h)$ is intractable, it is approximated [4] by a variational distribution $p(\theta)$ as $\theta_i = w_i \cdot z_i, z_i \sim \text{Bernoulli}(p_i)$, where p_i is the probability of keeping the weight w_i in a standard dropout scheme.

The Probabilistic U-Net [6] fuses the output of a deterministic U-Net with latent samples from a conditional variational auto-encoder modelling the variation over multiple annotators. Test-time segmentations are formed by sampling a latent $\mathbf{z}$, which is propagated with the image through the U-Net. Predictions are made as $\hat{p}(\mathbf{y}|\mathbf{x}, \mathbf{X}, \mathbf{A}) \approx p(\mathbf{y}|\mathbf{x}, \mathbf{z}^{(i)}, \hat{\mathbf{w}})$, with $\mathbf{z}^{(i)} \sim p_{\text{prior}}(\mathbf{z}|\mathbf{x})$.

4 Experiments

4.1 Data

Practical applications of uncertainty in segmentation tasks differ both in the type of ambiguity, and the availability of expert annotations. We select two representative datasets for our evaluation.

The **ISIC18** dataset consists of skin lesion images with a single annotation available [3, 15], and is used as an example of unambiguous image segmentation. We rescale the images to 256×256 pixels and split the dataset into 1500 samples for the train-set and 547 each for the validation and test sets.

The **LIDC-IDRI** lung cancer dataset [1, 2] contains 1018 lung CT scans from 1010 patients. For each scan, 4 radiologists (out of 12) annotated abnormal lesions. Anonymized annotations were shown to the other annotators, who were

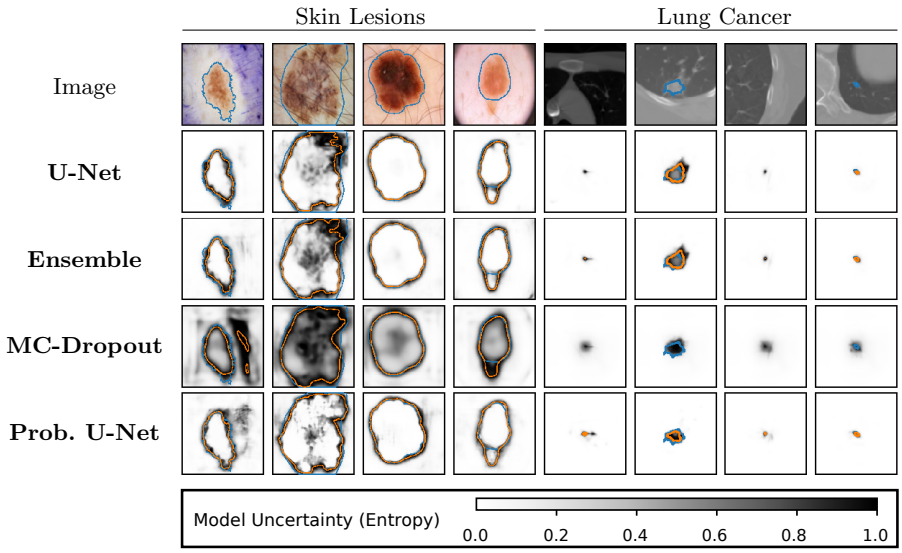


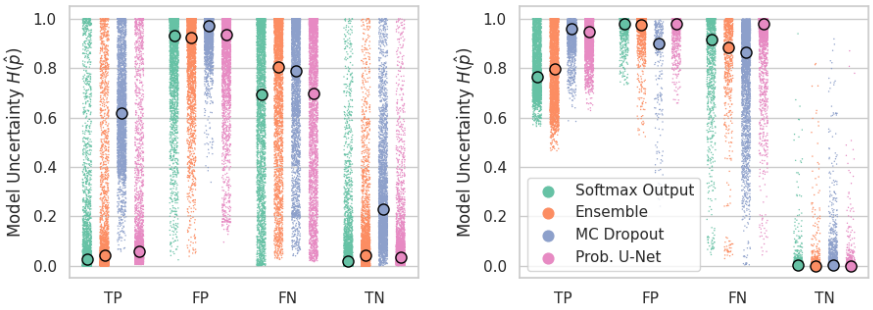
Fig. 3: Segmentation Uncertainty. Samples from the test set of the two datasets. Images in row one, model uncertainty (entropy) heat-maps in rows 2-5. Outline of mean ground truth annotations in Blue, mean model predictions in Orange.

allowed to adjust their own masks. Significant disagreement remains between the annotators: Among the extracted patches where at least one annotator marked a lesion, an average of 50% of the annotations are blank. We pre-processed the images as in [6], resampled to 0.5mm resolution, and cropped the CT-slices with lesions present to 128×128 pixels. The dataset is split patient-wise into three groups, 722 for the training-set and 144 each for the validation and test sets.

4.2 Model tuning and training

To allow for a fair evaluation, we use the same U-Net backbone of four encoder and decoder blocks for all models. Each block contains an up-/down-sampling layer, three convolution layers, and a residual skip-connection. The ensemble consists of four identical U-Nets. The latent-space encoders of the probabilistic U-Net are similar to the encoding branch of the U-Nets, and we choose a six-dimensional latent space size, following the original paper's recommendation.

All models were trained with binary cross-entropy. The probabilistic U-Net has an additional β-weighted KL-divergence loss to align the prior and posterior distributions, as per [6]. The optimization algorithm was Adam, with a learning rate of 10^{-4} for most models, except the probabilistic U-Net and MC-Dropout models on the skin lesion dataset, where a lower learning rate of 10^{-5} gave better results. We utilized early stopping to prevent over-fitting, and define the stopping criteria as 10 epochs without improvement of the validation loss, 100 epochs for models trained with the reduced learning rate. For the MC-Dropout



(a) Skin Lesion Dataset (b) Lung Cancer Dataset

Fig. 4: Pixelwise uncertainty by prediction correctness (True Positive, False Positive, False Negative, True Negative). The scatter plot shows individual pixels, with the median circled. For the lung cancer dataset, we discarded pixels with annotator disagreement.

and probabilistic U-Net models we performed a hyper-parameter search over the dropout probability p and the loss function weighting factor β, selecting the configuration with the lowest generalized energy distance on the validation set. We arrived at $p = 0.5$, $\beta = 0.0005$.

4.3 Uncertainty Estimation

For all models, our uncertainty estimates are based on non-thresholded pixel-wise predictions. For the U-Net, we take the final softmax predictions; for the remaining models we average across 16 non-thresholded samples. We quantify the *pixel-wise* uncertainty of the model by the entropy

$$\mathrm{H}(p(\mathbf{y}_{(i,j)}|\mathbf{x}, \mathbf{X}, \mathbf{A})) = \sum_{c \in C} p(\mathbf{y}_{(i,j)} = c|\mathbf{x}, \mathbf{X}, \mathbf{A}) \log_2 \frac{1}{p(\mathbf{y}_{(i,j)} = c|\mathbf{x}, \mathbf{X}, \mathbf{A})}$$

with $p(\mathbf{y}_{(i,j)} = c|\mathbf{x})$ as the pixel-wise probability to predict class $c \in C$. We plot the resulting uncertainty map for random images $\mathbf{x}$ from both datasets in Fig. 3. For visual reference, we overlay the mean expert annotation in Blue, and the mean model prediction in Orange. Darker shades indicate higher uncertainty.

We quantitatively assess the quality of uncertainty estimates by examining their relation to segmentation error in Fig. 4. On both datasets, models are more certain when they are correct (true positive, true negative) compared to when they are incorrect (false positive, false negative). A repeated measure correlation test finds a significant ($\alpha = 0.01$) correlation between segmentation error and model uncertainty on both datasets, for all methods. The relation holds, but is less strong, for MC-dropout on the skin dataset, which retains high uncertainty

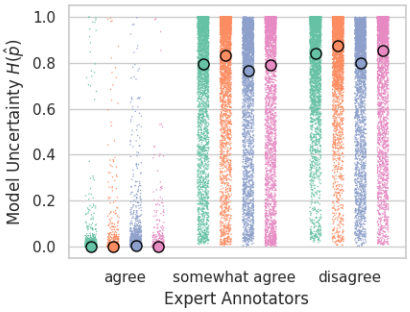


Fig. 5: Pixel-wise model uncertainty on the lung cancer dataset, grouped by agreement of expert annotations. Experts agree: $H(p) = 0$, somewhat agree $0 < H(p) < 1$, disagree $H(p) = 1$.

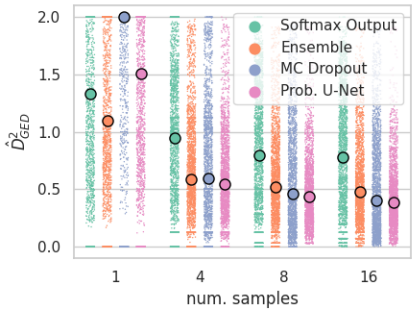


Fig. 6: Generalized Energy Distance of models on the lung cancer dataset, approximation by 1 to 16 samples, median highlighted. Lower distances are better.

even when it is correct. On the lung cancer dataset, all models have high uncertainty on true positive predictions. This might be caused by the imbalance of the dataset, where the positive class is strongly outweighed by the background and annotators often disagree. We tried training the models with a class-occurrence weighted loss function, which did produce true positive predictions with higher certainty but suffered an overall higher segmentation error.

We assess the correlation of model uncertainty with the uncertainty of the annotators on the lung cancer dataset in Fig. 5. For all models, this correlation is significant ($\alpha = 0.01$) . The median model uncertainty is very low (< 0.1) when all annotators agree, but high (> 0.7) when they disagree. There is a minor difference in model uncertainty between partial agreement (annotators split 3 – 1) and full disagreement (annotators split 2 – 2).

4.4 Sampling Segmentation Masks

Fig. 7 shows segmentation masks $\mathbf{y}$ sampled from the trained models $\hat{p}(\mathbf{y}|\mathbf{x})$. The U-Net model is fully deterministic and does not offer any variation in samples. The sample diversity of the ensemble is limited by the number of constituent models (four in our experiment). The MC-Dropout and probabilistic U-Net allow fully random sampling and achieve a visually higher diversity. On the skin lesion dataset, where only one export annotation per image is available, models still produce diverse predictions. On the lung cancer dataset, samples from the MC-Dropout and probabilistic U-Net represent the annotator distribution well.

We measure the distance between the model distribution $\hat{p}(\mathbf{y}|\mathbf{x}, \mathbf{X}, \mathbf{A})$ and the annotator distribution $p(\mathbf{y}|\mathbf{x})$ with the Generalized Energy Distance [6, 11, 14]. The distance measure is calculated as

$$D^2_{GED}(p, \hat{p}) = 2\mathbb{E}_{y\sim p, \hat{y}\sim\hat{p}}\left[d(y, \hat{y})\right] - \mathbb{E}_{y,y'\sim p}\left[d(y, y')\right] - \mathbb{E}_{\hat{y},\hat{y}'\sim\hat{p}}\left[d(\hat{y}, \hat{y}')\right] \quad . \quad (3)$$

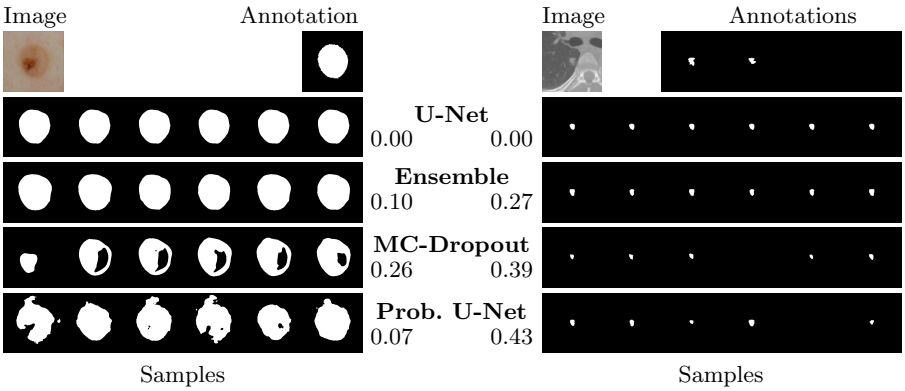


Fig. 7: Samples from the probabilistic models. First row: Image and ground truth annotations from the skin dataset (left) and lung nodule dataset (right). Following rows: samples $\mathbf{y} \sim \hat{p}(\mathbf{y}|\mathbf{x}, \mathbf{X}, \mathbf{A})$ drawn from the various models. Sample diversity over the entire dataset shown next to the model name.

We use $1 - \text{IoU}(\cdot, \cdot)$ as the distance d. A low D^2_{GED} indicates similar distributions of segmentations. We approximate the metric by drawing up to 16 samples from both distributions, and sample with replacement. The results are shown in Fig. 6. We observe that the annotator distribution is best approximated by the probabilistic U-Net, with MC-dropout and Ensemble closely behind; these pairwise ranks are significant ($\alpha = 0.01$) with left-tailed t-tests. A deterministic U-Net architecture is not able to reproduce the output distribution. Our results are consistent with [6], verifying our implementation. Following [11], we use the last term of (3) to assess the diversity of samples drawn from the model and note them in Fig. 7. They reinforce the qualitative observations of sample diversity.

4.5 Uncertainty estimates for active learning

Instead of training the models with all available data $\{\mathbf{X}, \mathbf{A}\}$, we now start with a small random subset $\{\mathbf{X}_0, \mathbf{A}_0\}$. We train the model with this subset at iteration $t = 0$, and then add a set of k images from $\{\mathbf{X}, \mathbf{A}\}$ to form $\{\mathbf{X}_{t+1}, \mathbf{A}_{t+1}\}$. Samples are selected based on the sum of pixel-wise entropies [7]. We repeat for T iterations, benchmarking against a random sample selection strategy.

For both skin lesion and lung cancer datasets, we start with a training size of 50 images, add $k = 25$ images at each iteration, and repeat $T = 10$ times. The models are trained for 5000 gradient updates with a batch size of 16 and 32 for the respective datasets. Since annotations are costly and to speed up computations, no validation-loss based early stopping is used. The experimental setup has been picked to ensure meaningful model uncertainties for the data selection policy and to ensure convergence within each active learning iteration.

The learning curves in Fig. 8 show that random-based sampling leads to a faster reduction in test loss over the uncertainty-based sampling strategy for

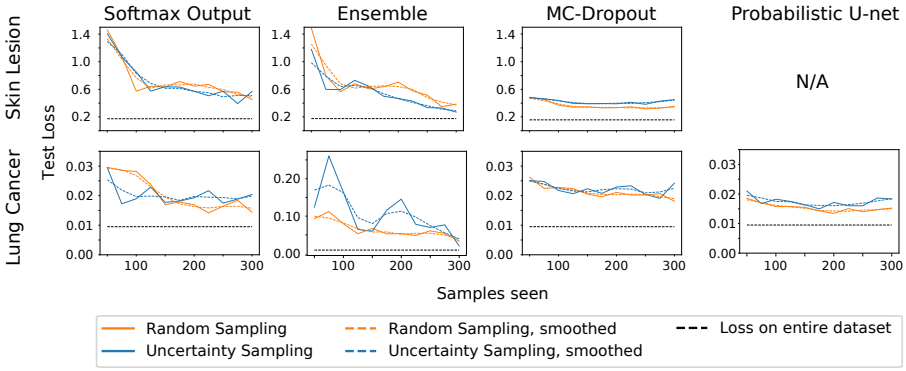


Fig. 8: Learning curves for the four algorithms on both datasets. Note that the Probabilisitic U-net only applies to the ambiguous segmentation.

Fig. 9: An example of the ambiguous samples frequently selected for inclusion into the training set under the uncertainty-based data gathering strategy. This unseen sample was selected when 150 annotations were revealed. The group of expert annotators provided disagreeing segmentation masks, confirming the model uncertainty but providing little additional information to learn from.

both datasets. We further investigated the samples selected by the uncertainty-based strategy by looking at the images which caused a large increase in the test error. One such image is shown in Fig. 9.

5 Discussion & Conclusion

Our results in Fig. 4 show that there is a clear relation between uncertainty estimates and segmentation error. The examples in Fig. 3 further highlight that areas of high uncertainty are not merely distributed around class boundaries, but also encompass areas with ambiguous labels. Fig. 5 shows that the uncertainty estimates obtained from the model are a good representation of the uncertainty of a group of expert annotators. We conclude that pixel-wise model uncertainty estimates give the practitioner a good indication of possible errors in the presented segmentation mask, allowing those predictions to be examined with care.

The learning curves in Fig. 8 show that estimated uncertainty is not generally useful for selecting active learning samples, for any model or dataset. Our results

depend on using the sum of pixel-wise entropies as a per-image entropy, which is correct for the softmax model, but only an approximation for the other models. This might impact our results. For the Lung Cancer dataset, all models estimate high uncertainty for the positive class, and the active learner thus selects images with a large foreground, skewing the proportion of classes represented in the training set. Furthermore, the selected images often have high annotator disagreement, illustrated in Fig. 9. If the active learner prefers sampling ambiguous images, it will be presented with inconsistent labels leading to harder learning conditions and poor generalisation. This may stem from an incorrect active learning assumption that annotations are noise-free and unambiguous, which is often not true. In conclusion, for a fixed budget of annotated images, we find no advantage in uncertainty-based active learning.

We observed similar behaviour of pixel-wise uncertainty estimates across all four segmentation models. The models differ in their ability to generate a distribution of distinct and coherent segmentation masks, with only the MC-dropout and probabilistic U-Net offering near unlimited diversity (see Fig. 7). But these models are harder to implement, more resource-intensive to train, and require hyperparameter tuning. The choice of model is ultimately application dependent, but our experiments show that even a simple U-net is competitive for the common task of assessing segmentation error. This agrees with [5], which compared uncertainty quantification models for unambiguous segmentation.

Our division of segmentation tasks into ambiguous and unambiguous considers it as "unambiguous" when a fundamentally ambiguous segmentation task is covered by a single annotator – or potentially several annotators, but with only one annotator per image, as for the Skin Lesion dataset. Even if the underlying task *is* ambiguous, the models considered in this paper inherently assume that it is *not*, as there is no mechanism to detect annotator variance when every image is only annotated once. More fundamental modelling of segmentation ambiguity and uncertainty thus remains a highly relevant open problem.

To conclude – is segmentation uncertainty useful? We find that uncertainty, even in the simplest models, reliably gives practitioners an indication of areas of an image that might be ambiguous, or wrongly segmented. Using uncertainty estimates to reduce the annotation load has proven challenging, with no significant advantage over a random strategy.

Acknowledgements. Our data was extracted from the "ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection" grand challenge datasets [3, 15]. The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used here. This work was funded in part by the Novo Nordisk Foundation (grants no. NNF20OC0062606 and NNF17OC0028360) and the Lundbeck Foundation (grant no. R218-2016-883).

References

1. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical physics **38**(2), 915–931 (2011)
2. Clark, K., et al.: The cancer imaging archive (TCIA): maintaining and operating a public information repository. Journal of digital imaging **26**(6), 1045–1057 (2013)
3. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 IEEE 15th International Symposium on Biomedical Imaging. pp. 168–172
4. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059 (2016)
5. Jungo, A., Reyes, M.: Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 48–56. Springer (2019)
6. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. Advances in neural information processing systems **31**, 6965–6975 (2018)
7. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Machine learning proceedings 1994, pp. 148–156. Elsevier (1994)
8. Loog, M., Yang, Y.: An empirical investigation into the inconsistency of sequential active learning. In: 2016 23rd international conference on pattern recognition (ICPR). pp. 210–215. IEEE (2016)
9. MacKay, D.J.: The evidence framework applied to classification networks. Neural computation **4**(5), 720–736 (1992)
10. MacKay, D.J.: Information-based objective functions for active data selection. Neural computation **4**(4), 590–604 (1992)
11. Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., Glocker, B.: Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 12756–12767 (2020)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
13. Settles, B.: Active learning literature survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences (2009)
14. Székely, G.J., Rizzo, M.L.: Energy statistics: A class of statistics based on distances. Journal of statistical planning and inference **143**(8), 1249–1272 (2013)
15. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**, 180161 (2018)
16. Yang, Y., Loog, M.: A benchmark and comparison of active learning for logistic regression. Pattern Recognition **83**, 401–415 (2018)

9

NEURALIZER: NEUROIMAGE ANALYSIS WITHOUT RE-TRAINING

PAPER 5:

Steffen Czolbe and Adrian Dalca (2023). "Neuralizer: Neuroimage Analysis without Re-Training". *Under Review for CVPR*.

This work, in large parts inspired by and performed during a 6-month stay at the Athinoula A. Martinos Center for Biomedical Imaging in Boston, is currently under review for CVPR 2023.

Neuralizer: Neuroimage Analysis without Re-Training

Steffen Czolbe
University of Copenhagen
Copenhagen, Denmark
per.sc@di.ku.dk

Adrian V. Dalca
A. A. Martinos Center, Massachusetts General Hospital, Boston, MA, USA
Harvard Medical School, Cambridge, MA, USA
Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA
adalca@mit.edu

Abstract

Neuroimage processing tasks like segmentation, reconstruction, and registration are central to the study of neuroscience. Deep learning strategies and architectures used to solve these tasks are often similar. Yet, when presented with a new task or a dataset with different visual characteristics, practitioners most often need to train a new model, or fine-tune an existing one. This is a time-consuming process that poses a substantial barrier for the thousands of scientists and clinical researchers who often lack the resources or expertise to train deep learning models. In practice, this leads to a lack of adoption of deep learning and neuroscience being dominated by classical frameworks. In this paper, we introduce Neuralizer, a single model that generalizes to previously unseen neuroimaging tasks and modalities without the need for re-training or fine-tuning. Tasks do not have to be known a priori, and generalization happens in a single forward pass during inference. We show experimentally that the model can solve processing tasks across multiple image modalities, acquisition methods, and datasets, and generalize to tasks and modalities it has not been trained on. When few annotated subjects are available ($\leq$ 32 in our experiments), our multi-task network outperforms task-specific baselines without training on the task.

1. Introduction

Computational methods for the processing and analysis of neuroimages have deepened our understanding of the human brain. The field has also led to advanced patient care by developing non-invasive methods of diagnosis and treatment, and has attracted large interest from the medical community and funding bodies. Recent deep learning research promises to further increase the accuracy and speed of neuroimaging methods.

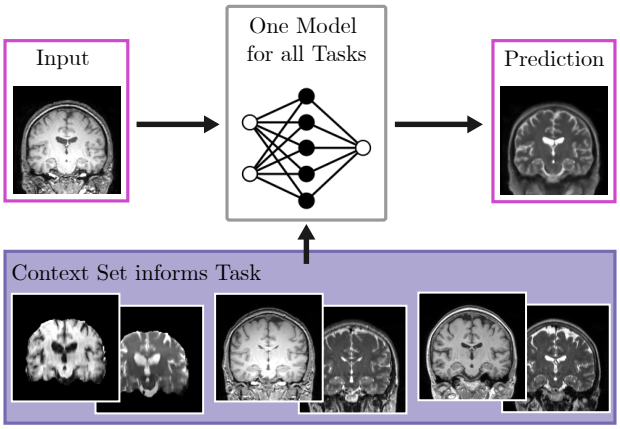


Figure 1. We solve a broad range of image processing tasks with a single model by conditioning the prediction on a context set of examples. After training on a diverse set of tasks, the model can generalize to new tasks in a single forward pass without re-training or fine-tuning. The model is highly flexible, requiring no prior definition of the set of tasks, and can be conditioned with context sets of any length.

A drawback of most current deep-learning-based approaches is that each model is limited to solving the task it has been trained on, on the data it has been trained on. Generalization to new task and domains, such as different acquisition protocols or new segmentation targets, remains a barrier to adoption [60]. Performing neuroimaging tasks like segmentation, registration, reconstruction, or motion correction requires different models for each pro-

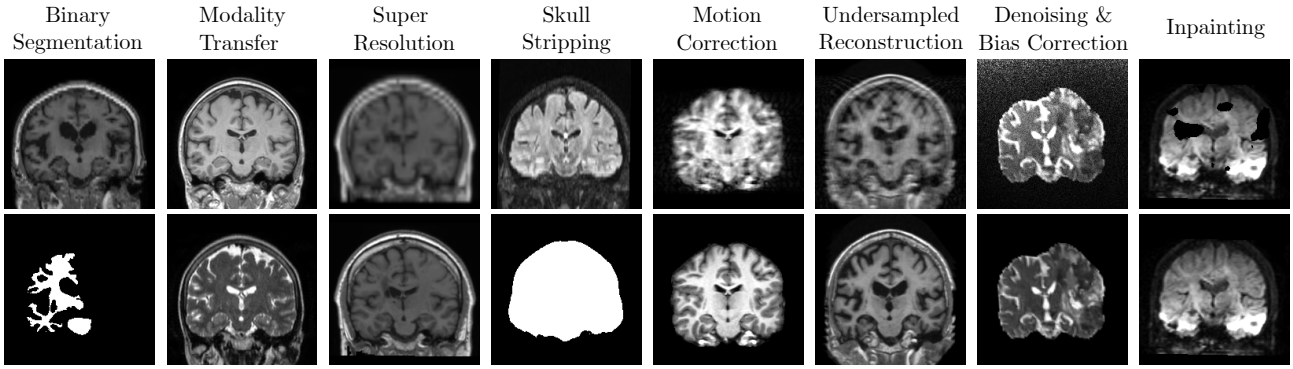


Figure 2. Examples of the neuroimaging tasks and modalities included in our dataset. Input images in the top row, output images in the bottom row.

cessing step, despite operating on the same input data and methods exhibiting strong similarities in network architecture [12, 42, 82]. Yet, designing and training models to solve these tasks on each dataset is prohibitively expensive. To train a deep learning model, a dataset needs to be compiled and manually annotated, and the network, training and data loading logic needs to be implemented, all of which generally require machine learning and neuroimaging expertise. In addition, computational resources like specialized graphics processing hardware needs to be available for running the optimization. These requirements are particularly problematic in clinical research settings due to a high cost of annotation and a lack of machine learning expertise and hardware. The many closely related neuroimaging tasks and image modalities and acquisition characteristics require custom solutions, many of which are not available. As a consequence, many works forgo using methods adapted to their task and data characteristics, and instead use existing methods even when their data acquisition falls outside of the protocols used for building the tool [10, 31, 96]. As neuroimaging tasks have much in common, generalization is a promising proposal to reduce the number of models that have to be trained.

Contribution

We introduce Neuralizer, a general-purpose neuroimaging model that, given a set of examples at inference (Fig. 1), can solve a broad range of neuroimaging tasks on diverse image modalities (Fig. 2), without the need for task-specific training or fine-tuning.

Neuralizer involves a novel architecture (Fig. 3), that takes as input a context set of examples to inform the processing task, and thus does not require prior definition of the tasks. The method enables single-pass generalization during inference and can process any number of reference images in a single pass to inform the prediction.

As a first method tackling task generalization in neuroimaging, we focus on analyzing the capabilities of such system and presenting general insights, and thus focus on 2D experiments allowed by our compute environment. We evaluate our model by comparing it's single-pass generalization performance to task-specific baselines conditioned on an equivalent amount of data. We find that Neuralizer outperforms the baselines on tasks where $\leq$ 32 labeled examples are available, despite never training on the task. When generalizing across segmentation protocols, Neuralizer matches the performance of baselines trained directly on the dataset.

2. Background & related work

We give a short introduction to neuroimaging tasks, terminology, and methods. We then provide an overview of fundamental methods for adapting a model to multiple domains, including multi-task learning, few-shot learning, fine-tuning, and data synthesis.

2.1. Neuroimage analysis

Neuroimage analysis employs computational techniques to study the structure and function of the human brain. Common imaging techniques are structural magnetic resonance imaging (MRI), functional MRI, diffusion tensor imaging (DTI), computed tomography (CT), and Positron emission tomography (PET). Each imaging method can create diverse images with different characteristics and contrasts, which are further diversified depending on the properties of the acquisition site [58, 107], device, protocol, imaging sequence [54], and use of contrast agents [9, 33].

To analyze these images, a variety of processing tasks are most often combined in a processing pipeline. Common processing tasks include anatomical segmentation [12, 16, 27], skull stripping [45, 51, 84, 91, 106], defacing [2, 36], registration [5–7, 10, 18, 24, 42, 52], modality transfer [77, 78, 94], in-painting [38, 66, 67, 76, 104], super-resolution [56, 71, 72, 100], compressed sensing, reconstruc-

tion, and de-noising [56, 68, 89], bias field removal [32, 57], surface fitting [44] and parcellation [88, 95].

Multiple toolboxes provide a suite of interoperable software components, most implementing classical optimization strategies. Widely used toolboxes are Freesurfer [27], FSL [49, 92, 103], SPM [30], CIVET [3], BrainSuite [87], HCP pipeline [97], and BrainIAK [55]. Deep-learning-based methods are just now starting to be included because of their improved accuracy and shorter runtime [12, 45]. While these toolboxes provide solutions for common neuroimaging applications, most tasks are limited to a single modality. Manual updates by the authors are required to include new processing tasks and to support a wider variety of image modalities.

2.2. Multi-task learning

Multi-Task Learning (MTL) frameworks solve multiple tasks simultaneously by exploiting similarities between related tasks [15]. MTL can improve performance and reduce computational cost and development time compared to designing task-specific solutions [22, 85]. In neuroimaging, MTL networks were recently proposed for the simultaneous segmentation and classification of brain tumors by training a single network with separate prediction heads associated with the different tasks [20, 35]. This strategy does not scale as the number of tasks increases, requires prior determination of the set of tasks, and does not enable generalization of the model to new tasks. With Neuralizer, we build on these methods to achieve scalable MTL, without the need for multiple network heads, and impotently the ability to generalize to new tasks and modalities.

2.3. Few-shot learning

Few-shot models generate predictions from just a few labeled examples [64, 81, 83, 101], or in the case of zero-shot methods [13], none at all. In computer vision, several methods pass a query image, along with a set of support images and labels as input to the model [64, 86, 93, 98]. Natural image segmentation methods [65, 109] use single image-label pairs [59, 108] as support or aggregate information from a larger support set [61]. Recent n-shot learning methods in the medical imaging setting [11, 25, 26] operate on a specific anatomical region in a single image modality [39, 111]. We build on ideas from these methods but aim to solve a much larger range of image-to-image tasks on images of many modalities, leading to unique challenges.

2.4. Fine-tuning

To tackle problems in the limited data scenarios frequent in medical imaging, neural networks can be pre-trained on a related task with high data availability and then fine-tuned for specific tasks. For example, a common approach involves taking a Res-Net [40] trained on ImageNet [19] and fine-tuning part of the network for a new task [46, 53, 99]. For medical imaging, networks pre-trained on large sets of medical images are available [17], and fine-tuning them to new tasks results in shortened training time and higher accuracy [4, 70]. However, as with training from scratch, fine-tuning requires machine learning expertise and computational resources not always available in clinical research. Additionally, in scenarios with small datasets, fine-tuning models trained on large vision datasets can be harmful [80].

2.5. AutoML methods

AutoML tools can be used to automate the steps of implementation, training, and tuning deep learning models, reducing the technical knowledge required of the user. For example, NN-UNet [47] is a software package that automates the design and training of U-Nets [82] for biomedical image segmentation, and has been successfully applied to brain segmentation [21, 48, 69]. While AutoML methods effectively reduce the technical requirements for the implementation, massively parallel hardware is still required for performing the internal hyper-parameter search and training the model. Additionally, AutoML methods reduce the flexibility in solution design, as they are often specific to a type of task or data structure.

2.6. Data augmentation and synthesis

Data augmentation increases the diversity of training data by augmenting or modifying existing data [82, 110]. It improves the model's robustness to input variability that may not be available in the original training data. In neuroimaging, arbitrary image modalities can be simulated by synthesis of images without requiring any real data [12, 14, 42, 45, 90]. In meta-learning, data augmentation can further be used to generate entirely new tasks [63, 102, 105]. We use data augmentations and further expand existing methods by developing rich neuroimaging task augmentations for generalization to unseen neuroimaging tasks.

3. Neuralizer

We introduce Neuralizer, a multi-task model for neuroimage analysis tasks. In this section, we first define the training framework and adaptations necessary to operate on a diverse range of tasks and input types. Then, we introduce the model architecture, training, and inference strategies.

3.1. Generalizabe multi-task model

Fig. 1 gives a high-level overview of our model. Let T represent a set of tasks, with a subset of tasks T_{seen} seen during training. Each task consists of input-output image pairs (x_t, y_t) from potentially multiple underlying datasets with joint input and output spaces $x_t \in \mathcal{X}, y_t \in \mathcal{Y}$.

To enable generalization to unseen tasks, we condition the model on a context set $C_t = \{(x_{t,i}, y_{t,i})\}_{i=1}^{N}$ of input-

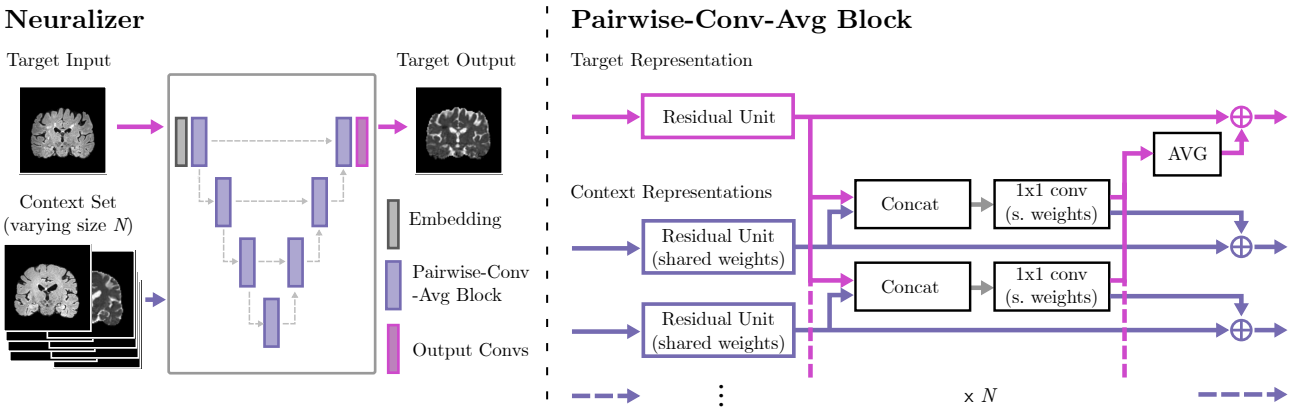


Figure 3. Neuralizer consists of 7 Pairwise-Conv-Avg blocks (right), arranged in a U-Net-like [75, 82] configuration (left). Each Pairwise-Conv-Avg block enables interaction between the input image and the image pairs present in the context set. The block consists of a residual unit, pairwise convolution of each context member with the target, and an averaging of results across the context set to update the representation. To extract information from context sets of any size, the architecture is invariant in the length of the context size N.

output image pairs that exemplify the task to be performed. The context set can vary in size $|C_t| = N$.

We employ a neural network $g_\theta(x_t, C_t) = y_t$ with weights θ that aims to apply the task defined by context set C_t to the input neuroimage x_t. We optimize the network using supervised training with the loss

$$\mathcal{L}(T_{\text{seen}}; \theta) = \mathbb{E}_{t \in T_{\text{seen}}} \left[\mathbb{E}_{(x_t, y_t, C_t)} [\mathcal{L}_t(y_t, g_\theta(x_t, C_t))] \right], \quad (1)$$

where $\mathcal{L}_t$ is a task-specific loss function.

3.2. Design for diverse tasks

To process different tasks with a single model, we carefully select the loss function, image encoding, and generation of the training set for each task type.

Loss function. Neuralizer needs to solve both segmentation tasks (e.g. anatomical segmentation and skull-stripping via a brain mask), and image-to-image tasks (e.g. denoising). We use the Soft Dice Loss [75] for the segmentation-like tasks, and the pixel-wise Mean Squared Error $MSE(y_t, g(x_t, C_t)) = \frac{1}{2\sigma^2} \sum_p [y_{tp} - g(x_t, C_t)_p]^2$ with balancing hyperparameter σ^2 for all other tasks. As the network optimizes multiple tasks during training, the balance of the loss terms can dramatically affect the result.

Input and output encoding. For Neuralizer to work on both segmentation and image-to-image tasks, we facilitate simultaneous input of multiple image modalities and masks. We design the input space $\mathcal{X}$ to accept floating point value images with three channels and zero-pad empty channels. The output space $\mathcal{Y}$ follows the same design but uses only one channel.

Training dataset creation. On each training iteration, we first sample a task t from T_{seen}. Given the task, the task-specific dataset is created from the underlying training data (Tab. 1). From this dataset, we sample the input image, ground truth output, and image pairs for the context set. To increase the range of images that can be used to condition the trained model, the image modalities and acquisition protocols of members of the context set can differ from the input image for some tasks. Supplemental section G contains a detailed description of the training data generator.

Table 1. Tasks, Modalities, Datasets, and Segmentation classes used for training Neuralizer.

Tasks	Modalities
Binary Segmentation	T1-w.
Modality Transfer	T2-w.
Super Resolution	MRA
Skull Stripping	PD
Motion Correction	FLAIR
Undersampled Reconstruction	ADC
Denoising & Bias correction	DWI
Inpainting	DTI (17 dir.)

Datasets	Segmentation Classes
OASIS 3 [43, 73]	Freesufer Aseg, 31
BRATS [8, 9, 74]	classes [12, 27]
IXI [1]	
ATLAS R2.0 [62]	Manually-annotated Hammers
Hammers Atlas [37]	Atlas, 96 classes [23, 34, 37]
WMH Challenge [54]	
ISLES2022 [79]	Brainmasks [27, 45]

3.3. Model architecture

Fig. 3 shows the Neuralizer network architecture, developed jointly with [102]. As the architecture is independent of the task, we omit the task subscript in this section.

The input image x and the image pairs of the context set $C_i = (x_i, y_i), i = 1, ..., N$ are first passed through an embedding layer consisting of a single 1×1 convolution with learnable kernels e_x, e_C, to obtain the representations $r_x = x * e_x$, $r_{C_i} = \text{cat}(x_i, y_i) * e_C$ where $*$ is the convolution operator. This combines each context image pair to a joint representation r_{C_i} and maps all representations to a uniform channel width c, which is constant throughout the model. Next, we process the representations by multiple Pairwise-Conv-Avg Blocks (explained later), arranged as a U-Net-like configuration [75, 82] to include multiple scales. The output r_x^{out} of the final Pairwise-Conv-Avg Block is processed by a residual unit [40] and a final 1×1 conv layer to map to one output channel. All residual units consist of two 3×3 conv layers, a shortcut connection, and GELU activation functions [41].

Compared to standard CNNs, Neuralizer needs a mechanism to enable knowledge transfer from the context set to the input image. We introduce the Pairwise-Conv-Avg Block (Fig. 3, right) to model this interaction. The block maps from representations of the target input r_x^{in} and context pairs $r_{C_i}^{\text{in}}$ to output representations $r_x^{\text{out}}, r_{C_i}^{\text{out}}$ of the same size. First, we process each input separately with a residual unit to obtain $r_x^{\text{int}} = \text{ResUnit}_x(r_x^{\text{in}})$ and $r_{C_i}^{\text{int}} = \text{ResUnit}_C(r_{C_i}^{\text{in}})$. The residual units operating on the context representations have shared weights. Second, we pairwise concatenate the context representations with the target representation on the channel dimension so that $p_i = \text{cat}(r_x^{\text{int}}, r_{C_i}^{\text{int}})$. We combine the pairwise representations and reduce the channel size back to c with a 1×1 convolution with learnable kernel k_x, and update the target representation by averaging across context members $r_x^{\text{out}} = r_x^{\text{int}} + \frac{1}{N} \sum_{i=1}^{N} p_i * k_x$. The context representations are updated with a separate kernel $r_{C_i}^{\text{out}} = r_{C_i}^{\text{int}} + p_i * k_C$. We then re-size the outputs of a Pairwise-Conv-Avg Block before feeding them as input the next block. We experimented with attention-based and weighted average approaches but found that they do not lead to an increased generalization to unseen tasks.

3.4. Task augmentations

To further diversify the training dataset, we employ task augmentations [102], a group of transformations applied at random to the input and output images, and the images of their context set. The objective is not to create plausible neuroimaging tasks, but instead to increase the diversity of tasks to discourage the model from merely memorizing the tasks in the training data. A list of all task augmentations is given in Tab. 2, with more detailed descriptions and visual examples in Supplement C.

3.5. Inference

During inference, we supply an input image x_i and a context set C_i from the same task. Given these inputs, a simple feed-forward pass through the model provides the prediction $\hat{y} = g(x, C)$. To further increase accuracy at test-time, we use context-set bootstrapping [102]. If less than the maximum computationally feasible number of context images is provided, it is padded to full size by sampling with replacement from the provided set, with small affine movements applied to the padded samples.

4. Experiments

We first compare Neuralizer with task-specific networks, which require substantial expertise and compute. We then analyze the effect of the length of context set, and the multi-task generalization to unseen segmentation schemes and image modalities. For this first method of large-scale multi-task generalization in neuroimaging, we conduct the experiments on 2D image slices.

4.1. Data

To create a diverse dataset encompassing a multitude of different modalities, acquisition protocols, devices, and tasks, we pool neuroimages from the public datasets OASIS3 [43, 73], BRATS [8, 9, 74], Atlas R2.0 [62], Hammers Atlas [37], IXI [1], ISLES2022 [79], and the White Matter Hyperintensities Challenge [54]. We segment all subjects with Synthseg [12, 27]. Based on the segmentation, we affinely align the images to the MNI 152 template space [28, 29], and resample to 1mm isometric resolution at a size of $192 \times 224 \times 192$mm. We perform quality assurance of the segmentation and registration by ensuring no segmented areas fall outside of the cropped volume and discard subjects failing this check (4 subjects discarded). We extract a coronal slice of 192×192mm, bisecting the frontal Brain stem, Hippocampus, Thalamus, and Lateral ventricles. We rescale image intensities to the $[0, 1]$ interval using dataset-specific percentiles. For full head images, we create a brain mask with Synthstrip [27, 45]. The final dataset contains 2,282 subjects with 15,911 images and segmentation masks across 8 modalities and 17 DTI directions. Subjects of the seven original datasets are split into 80% training and validation, 20% test, with a minimum of 15 test subjects per dataset.

Table 2. Task Augmentations

Task Augmentations	
IntensityMapping	SyntheticModality
SobelFilter	MaskInvert
MaskContour	MaskDilation
PermuteChannels	DuplicateChannels

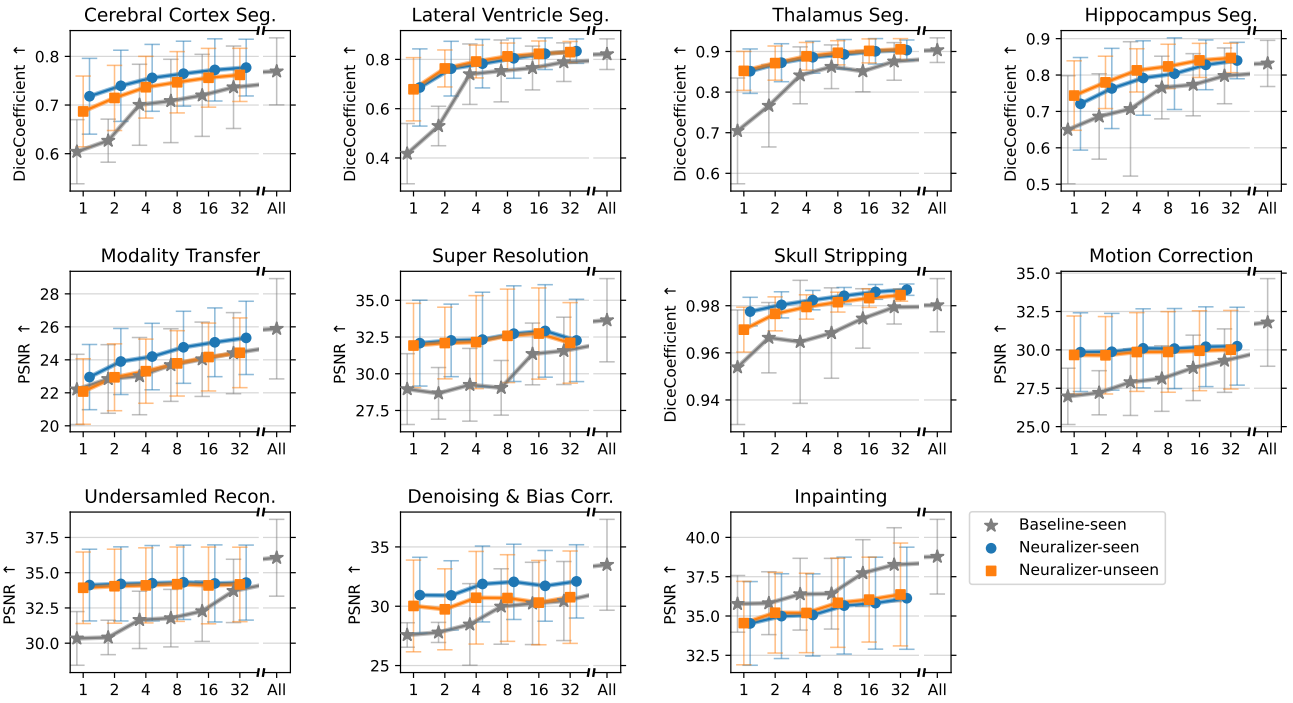


Figure 4. Performance of multi-task Neuralizer and the task-specific baselines on each task, averaged across all modalities in the test set. The tasks being evaluated were included in the training of Neuralizer-seen (orange), held out in Neuralizer-unseen (blue), and specifically trained on by each task-specific baseline (gray). The x-axis is the size of the train/context set, and the y-axis is the Dice/PSNR score. Some points on the x-axis are omitted for better visibility. 'All' refers to all available train data for the task, ranging from 249 to 2,282 subjects depending on the task. The bars denote the standard deviation across modalities. We provide an evaluation on just the T1 modality in Supplement E.

4.2. Models

Neuralizer-seen. This Neuralizer model has seen all tasks we have available during training. We evaluate the performance of Neuralizer on unseen scans from tasks and modalities that have been included in the training. The model uses the 4-stage architecture shown in Fig. 3 with 64 channels per layer. During training, the context size $|C_i|$ is sampled from $\mathcal{U}_{\{1,32\}}$ at each iteration.

Neuralizer-unseen. To evaluate the performance of Neuralizer on tasks and modalities it has not been trained on, we train a family of Neuralizer models where a single task or modality is excluded from the training set. The model architecture of Neuralizer-unseen is identical to Neuralizer-seen.

Baseline-seen. As no established baseline for multi-task and multi-modality models in neuroimaging can tackle the number of tasks we aim for, we compare Neuralizer to an ensemble of task-specific U-Nets [75, 82]. However, training one model for each task and modality requires overwhelming computational resources. To reduce the computational requirement, we follow previous modality-agnostic models [12, 45] and train each model on multiple input modalities. This lowers the number of models to be trained to one per task, segmentation class, and modality-transfer output modality. To compare Baseline-seen with Neuralizer-unseen conditioned on an equal amount of data, we train U-Net models with training set sizes of $\{1, 2, 4, 8, 16, 32, \text{all}\}$ and use data augmentation.

We use a 4-stage U-Net architecture with one residual block per layer. The channel width is tuned experimentally for each training dataset size. We select 256 channels for all data, and 64 channels otherwise. Using larger U-Nets resulted in overfitting and lower performance. An overview of model parameter count and inference cost is given in Supplement H.

4.3. Training

We use supervised training, task-specific loss functions, and weigh the MSE loss by selecting $\sigma^2 = 0.05$, resulting in both loss terms being of similar magnitude. All models are trained with a batch size of 8, a learning rate of 10^{-4}, and the ADAM optimizer [50]. To speed up training, we under-

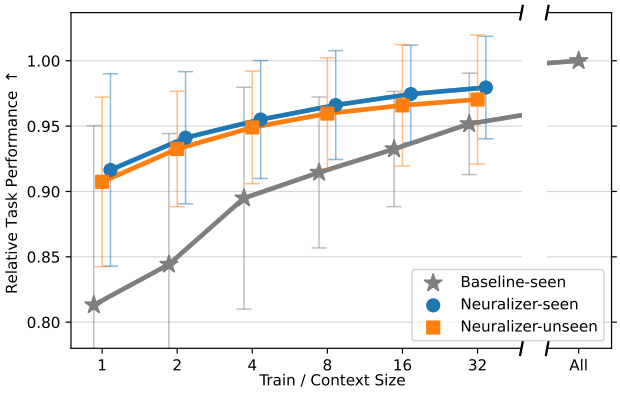


Figure 5. Results averaged across tasks, expressed as relative performance compared to the baseline trained on all data. The tasks being evaluated were included in the training of Neuralizer-seen (orange), held out in Neuralizer-unseen (blue), and specifically trained on by each task-specific baseline (gray). The x-axis is the size of the train/context set, and the y-axis is the relative score. Some points on the x-axis are omitted for better visibility. 'All' refers to all available data for the task, ranging from 249 to 2,282 subjects depending on the task. Bars: standard deviation across tasks/modalities.

sample tasks that the model learns quickly, with sampling weights given in Supplement I.

In addition to the task augmentations, we use data augmentations via random affine movements, random elastic deformations, and random flips along the sagittal plane. For Baseline-seen, we reuse the task augmentations but remove augmentations that introduce uncertainty over the model output due to not having access to the context set. The training time for Neuralizer is 7 days on a single A100 GPU. The training time of the baseline models is capped at 5 days. All models use early stopping, ending the training after 25 epochs without a decrease in validation loss. The model with the lowest validation loss is used for further evaluation.

4.4. Evaluation

We evaluate the Dice coefficient for the segmentation and skull stripping tasks, and the Peak Signal-to-Noise Ratio (PSNR) for the image-to-image tasks on the test set. As the few-shot capability is of particular interest, we measure performance as a function of context set size for the Neuralizer models and use training set size as an analog for the U-Net models. We evaluate context sets of up to 32 subjects. Larger context sets are possible but come at a linear cost in memory.

4.5. Experiment 1: Comparison to task-specific networks

To assess if Neuralizer's multi-task approach is competitive compared to task-specific models, we evaluate the performance and runtime of Neuralizer-seen, Neuralizer-unseen, and Baseline-seen on the test-set of each task. For Neuralizer-unseen, we withhold image modalities using a leave-one-out strategy during training and evaluate on the unseen modalities at test time.

Results. We display the results by task in Fig. 4, aggregated across all tasks in Fig. 5, and in tabular form in Supplement D. Both Neuralizer models outperform task-specific baselines trained on up to 32 samples. When training the baselines on all available data, the baselines outperform Neuralizer-seen by 2 percentage points in relative performance, and Neuralizer-unseen by 3 percentage points. The loss in performance when generalizing to an unseen modality (between Neuralizer-seen and Neuralizer-unseen) is less than 2 percentage points for all context set sizes.

Training the baseline model to convergence on 32 samples took on average 28.2 ± 16.6 hours per task, using one A100 GPU. Since Neuralizer only requires inference for a new task, it is orders of magnitude faster, requiring less than 0.1 seconds on a GPU and less than 3 seconds on a CPU.

We provide qualitative samples of the predictions from Neuralizer-seen model in Supplement A, Figures 6-8.

4.6. Experiment 2: Context set size

To assess the few-shot setting, we evaluate performance as a function of the number of labeled samples. For Neuralizer, we evaluate the model with context-set sizes of $\{1, 2, 4, 8, 16, 32\}$ unique subjects from the test set. For the baseline, we trained models with reduced training set sizes of the same amount of subjects. To reduce the effect of random training subject selection, we train three separate baselines with $n = 1$, two baselines with $n = 2$, and average results of models with the same n.

Results. Tab. 4 and Figs. 4, 5 illustrate the results. For all models, prediction accuracy increases with the availability of labeled data, with diminishing returns. For both Neuralizer models, a context set size of one achieves more than 90% of the performance attainable with all data. The baseline performs overall worse than both Neuralizer models when ≤ 32 labeled samples are available but achieves the best overall performance on larger datasets.

4.7. Experiment 3: Generalization to unseen segmentation protocol

The Hammers Atlas dataset [23, 34, 37] provides an alternative anatomical segmentation protocol to the widely-used Freesurfer segmentation available for most subjects in the dataset. The shape, size, and amount of annotated regions in the protocols differ drastically. A different image acquisition site also leads to images of different visual characteristics. We use the Hammers Atlas dataset to evaluate Neuralizer-unseen by withholding the dataset and its

Model	Task Seen	Segmentation Class (Hammers Atlas)														Mean *(std)*
		Hip	PAG	STG	MIG	FuG	Stm	Ins	PCG	Tha	CC	3V	PrG	PoG	ALG	
Baseline-seen	✓	.88	.86	.93	.92	.79	.87	.82	.87	.90	.80	.68	.83	.77	.82	.84 *(.07)*
Neuralizer-seen	✓	.88	.86	.92	.92	.76	.88	.83	.85	.90	.82	.73	.86	.77	.81	.84 *(.07)*
Neuralizer-unseen	✗	.88	.87	.93	.91	.78	.87	.82	.85	.90	.81	.72	.85	.78	.81	.84 *(.06)*

Table 3. Segmentation of the Hammers Atlas dataset. For Neuralizer-unseen, this dataset and segmentation scheme is withheld from training, allowing comparative evaluation of the Dice overlap to models that have been trained using this dataset. The Hammers Atlas dataset contains region of interest segmentations significantly different from the Freesurfer protocol available for most of the training data. Evaluation of major segmentation classes located in the center and right of the coronal slice. See Supplement F for class abbreviations.

annotations entirely from training. We evaluate the Dice coefficient of the 14 major anatomical segmentation classes present in the center and right half of the coronal slice.

Results. Tab. 3 illustrates the results. Neuralizer-unseen performs similarly to Neuralizer-seen and the baseline, while not requiring lengthy re-training or fine-tuning on the Hammers Atlas dataset, and not having seen the segmentation protocol. All three models achieve a mean Dice coefficient of 0.84. The largest performance difference is in the third ventricle class, where both Neuralizer models outperform the baseline by at least 0.04 Dice. The Freesurfer segmentation protocol included in the training set of the Neuralizer models also contains a third ventricle class.

5. Discussion

Our experiments using modality and segmentation class hold-outs show that Neuralizer can generalize well to unseen neuroimaging tasks. Across all context set sizes, the generalization loss between seen and unseen modalities and segmentation classes is less than 2 percentage points across experiments 1 and 2. On the smaller held-out Hammers-Atlas segmentation dataset, we find that Neuralizer can generalize to unseen tasks with similar performance. These results show promise that a single Neuralizer model can perform multiple tasks including generalization to new inference tasks not seen during training.

In settings with 32 or fewer labeled example images, Neuralizer-unseen outperforms task-specific baselines despite never having seen the task or modality at train time, and taking nearly no effort or compute compared to the baselines which require substantial expertise, manual labor, and compute resources. The performance difference is largest when only one labeled subject is available, but still present at 32 subjects (Fig. 5). The Neuralizer few-shot approach provides a performance advantage on smaller datasets likely by exploiting similarities across the many neuroimaging tasks and datasets available in training. When training the baselines on all available data, they can outperform Neuralizer-seen and Neuralizer-unseen by at most 3 percentage points. The inflection point of identical performance between Neuralizer and the baselines is not covered by the range of context set sizes chosen for training and evaluation due to prohibitive computational costs.

When large annotated datasets are available, the baselines performed best on most tasks. However, training task-specific models comes at a significant cost. As a first step in the proposed problem formulation, Neuralizer offers an alternative with near equal performance, while only requiering seconds to infer any task from the context set.

Limitations

We made simplifying assumptions in this first paper demonstrating the potential of multi-task generalization in neuroimaging. The presented experiments are conducted on 2D data slices. In large part, we did this since running the hundreds of baselines in 3D would be infeasible on our available compute resources. However, entire volumetric data would also impose prohibitive memory requirements on Neuralizer models. To tackle 3D data in the future, we plan to process multiple slices at a time.

We affinely aligned the images of the context set to the target image. Early in Neuralizer development, we tried training on non-aligned inputs but found that it deteriorated performance. The need for affine alignment provides an obstacle to adoption. While existing affine-alignment tools are fast and can be employed, we also believe that this requirement can be removed with further development.

Early on, we experimented with lesion segmentation tasks but found the results to be unsatisfactory. Lesions are spatially heterogeneous, making learning from the context set much harder for convolutional architectures. As with 3D data and affine alignment, we believe this to be an interesting future research challenge. While we demonstrate the proposed ideas on a broad range of tasks and modalities, Neuroimage analysis can involve more domains, tasks, and populations, like image registration, surface-based tasks, CT image domains, and pediatric data. We look further to extend Neuralizer to tackle these in the future.

6. Conclusion

Neuralizer performs rapid few-shot, single-pass, multi-task generalization, and outperforms task-specific baselines in limited data scenarios. Even when a large amount of annotated data is available, Neuralizer often matches baseline performance despite not training on the data. Neuralizer provides clinical researchers and scientists with a single model to solve a wide range of neuroimaging tasks on images of many modalities and can be easily adapted to new tasks without the substantial investment of retraining or fine-tuning a task-specific model.

References

[1] IXI Dataset. 4, 5

[2] David Abramian and Anders Eklund. Refacing: Reconstructing anonymized facial features using GANS. *International Symposium on Biomedical Imaging*, 2019-April:1104–1108, apr 2019. 2

[3] Yasser Ad-Dab'bagh, O Lyttelton, JS Muehlboeck, C Lepage, D Einarson, K Mok, O Ivanov, RD Vincent, J Lerch, and E Fombonne. The CIVET image-processing environment: a fully automated comprehensive pipeline for anatomical neuroimaging research. In *Proceedings of the 12th annual meeting of the organization for human brain mapping*, 2006. 3

[4] Laith Alzubaidi, Muthana Al-Amidie, Ahmed Al-Asadi, Amjad J. Humaidi, Omran Al-Shamma, Mohammed A. Fadhel, Jinglan Zhang, J. Santamaría, and Ye Duan. Novel Transfer Learning Approach for Medical Imaging with Limited Labeled Data. *Cancers 2021, Vol. 13, Page 1590*, 13(7):1590, mar 2021. 3

[5] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 924–931. Springer Verlag, 2006. 2

[6] John Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95–113, oct 2007. 2

[7] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, feb 2008. 2

[8] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4, sep 2017. 4, 5

[9] Spyridon Bakas, Bjoern Menze, and Others. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *arXiv*, 124, nov 2018. 2, 4, 5

[10] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, feb 2019. 2

[11] Cheng Bian, Chenglang Yuan, Kai Ma, Shuang Yu, Dong Wei, and Yefeng Zheng. Domain Adaptation Meets Zero-Shot Learning: An Annotation-Efficient Approach to Multi-Modality Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 41(5):1043–1056, may 2022. 3

[12] Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, and Juan Eugenio Iglesias. SynthSeg: Domain Randomisation for Segmentation of Brain Scans of any Contrast and Resolution. jul 2021. 2, 3, 4, 5, 6

[13] Maxime Bucher, Tuan-Hung VU, Matthieu Cord, and Patrick Pérez. Zero-Shot Semantic Segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[14] Víctor M. Campello, Carlos Martín-Isla, Cristian Izquierdo, Steffen E. Petersen, Miguel A.González Ballester, and Karim Lekadir. Combining Multi-Sequence and Synthetic Images for Improved Segmentation of Late Gadolinium Enhancement Cardiac MRI. *Lecture Notes in Computer Science*, 12009 LNCS:290–299, 2020. 3

[15] Rich Caruana, Lorien Pratt, and Sebastian Thrun. Multitask Learning. *Machine Learning 1997 28:1*, 28(1):41–75, 1997. 3

[16] Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng Ann Heng. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage*, 170:446–455, apr 2018. 2

[17] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer Learning for 3D Medical Image Analysis. apr 2019. 3

[18] Steffen Czolbe, Oswin Krause, and Aasa Feragen. Semantic similarity metrics for learned image registration. *Proceedings of Machine Learning Research*, 2021. 2

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[20] Francisco Javier Díaz-Pernas, Mario Martínez-Zarzuela, David González-Ortega, and Míriam Antón-Rodríguez. A Deep Learning Approach for Brain Tumor Classification and Segmentation Using a Multiscale Convolutional Neural Network. *Healthcare*, 9(2):153, feb 2021. 3

[21] Houssam El-Hariri, Luis A. Souto Maior Neto, Petra Cimflova, Fouzi Bala, Rotem Golan, Alireza Sojoudi, Chris Duszynski, Ibukun Elebute, Seyed Hossein Mousavi, Wu Qiu, and Bijoy K. Menon. Evaluating nnU-Net for early ischemic change segmentation on non-contrast computed tomography in patients with Acute Ischemic Stroke. *Computers in Biology and Medicine*, 141:105033, feb 2022. 3

[22] Theodoras Evgeniou and Massimiliano Pontil. Regularized multi-task learning. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117, 2004. 3

[23] Isabelle Faillenot, Rolf A. Heckemann, Maud Frot, and Alexander Hammers. Macroanatomy and 3D probabilistic atlas of the human insula. *NeuroImage*, 150:88–98, apr 2017. 4, 7, 23

[24] Mirza Faisal Beg, Michael I Miller, Alain Trouvétrouv, and Laurent Younes. Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms. *International Journal of Computer Vision*, 61(2):139–157, 2005. 2

[25] Ruiwei Feng, Xiangshang Zheng, Tianxiang Gao, Jintai Chen, Wenzhe Wang, Danny Z. Chen, and Jian Wu. Interactive Few-Shot Learning: Limited Supervision, Better Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 40(10):2575–2588, oct 2021. 3

[26] Abdur R Feyjie, Reza Azad, Marco Pedersoli, Claude Kauffman, Ismail Ben Ayed, and Jose Dolz. Semi-supervised few-shot learning for medical image segmentation. mar 2020. 3

[27] B Fischl. FreeSurfer. NeuroImage, 62 (2), 774–781, 2012. 2, 3, 4, 5

[28] VS Fonov, AC Evans, RC McKinstry, CR Almli, and DL Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102, jul 2009. 5

[29] Vladimir Fonov, Alan C. Evans, Kelly Botteron, C. Robert Almli, Robert C. McKinstry, and D. Louis Collins. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1):313–327, jan 2011. 5

[30] Richard SJ Frackowiak. *Human Brain Function*. 2004. 3

[31] Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, jul 2016. 2

[32] Tal Goldfryd, Shiri Gordon, and Tammy Riklin Raviv. Deep semi-supervised bias field correction of mr images. *Proceedings - International Symposium on Biomedical Imaging*, 2021-April:1836–1840, apr 2021. 3

[33] Enhao Gong, John M. Pauly, Max Wintermark, and Greg Zaharchuk. Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI. *Journal of Magnetic Resonance Imaging*, 48(2):330–340, aug 2018. 2

[34] Ioannis S. Gousias, Daniel Rueckert, Rolf A. Heckemann, Leigh E. Dyet, James P. Boardman, A. David Edwards, and Alexander Hammers. Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. *NeuroImage*, 40(2):672–684, apr 2008. 4, 7, 23

[35] Sachin Gupta, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. MAG-Net: Multi-task Attention Guided Network for Brain Tumor Segmentation and Classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13147 LNCS:3–15, 2021. 3

[36] J Hale, Nakeisha Schimke, and John Hale. Quickshear defacing for neuroimages Attack Graph Generation on HPC Clusters View project Quickshear Defacing for Neuroimages. *Frontiers in psychiatry*, 21, 2011. 2

[37] Alexander Hammers, Richard Allom, Matthias J. Koepp, Samantha L. Free, Ralph Myers, Louis Lemieux, Tejal N. Mitchell, David J. Brooks, and John S. Duncan. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Human brain mapping*, 19(4):224–247, aug 2003. 4, 5, 7, 23

[38] Xu Han, Roland Kwitt, Stephen Aylward, Spyridon Bakas, Bjoern Menze, Alexander Asturias, Paul Vespa, John Van Horn, and Marc Niethammer. Brain extraction from normal and pathological images: A joint PCA/Image-Reconstruction approach. *NeuroImage*, 176:431–445, aug 2018. 2

[39] Stine Hansen, Srishti Gautam, Robert Jenssen, and Michael Kampffmeyer. Anomaly Detection-Inspired Few-Shot Medical Image Segmentation Through Self-Supervision With Supervoxels. *Medical Image Analysis*, 78, mar 2022. 3

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3, 5

[41] Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). jun 2016. 5

[42] Malte Hoffmann, Benjamin Billot, Douglas N. Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V. Dalca. SynthMorph: learning contrast-invariant registration without acquired images. *IEEE Transactions on Medical Imaging*, apr 2020. 2, 3, 19

[43] Andrew Hoopes, Malte Hoffmann, Bruce Fischl, John Guttag, and Adrian V. Dalca. HyperMorph: Amortized Hyperparameter Learning for Image Registration. *Information Processing in Medical Imaging*, jan 2021. 4, 5

[44] Andrew Hoopes, Juan Eugenio Iglesias, Bruce Fischl, Douglas Greve, and Adrian V Dalca. TopoFit: Rapid Reconstruction of Topologically-Correct Cortical Surfaces. *Proceedings of Machine Learning Research-Under Review*, pages 1–13, 2022. 3

[45] Andrew Hoopes, Jocelyn S. Mora, Adrian V. Dalca, Bruce Fischl, and Malte Hoffmann. SynthStrip: Skull-Stripping for Any Brain Image. mar 2022. 2, 3, 4, 5, 6

[46] Minyoung Huh, Pulkit Agrawal, and Alexei A. Efros. What makes ImageNet good for transfer learning? aug 2016. 3

[47] Fabian Isensee, Paul F. Jaeger, Simon A.A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods 2020 18:2*, 18(2):203–211, dec 2020. 3

[48] Fabian Isensee, Paul F. Jäger, Peter M. Full, Philipp Vollmuth, and Klaus H. Maier-Hein. nnU-Net for Brain Tumor Segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12659 LNCS:118–132, 2021. 3

[49] Mark Jenkinson, Christian F. Beckmann, Timothy E.J. Behrens, Mark W. Woolrich, and Stephen M. Smith. FSL. *NeuroImage*, 62(2):782–790, aug 2012. 3

[50] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, dec 2015. 6

[51] Jens Kleesiek, Gregor Urban, Alexander Hubert, Daniel Schwarz, Klaus Maier-Hein, Martin Bendszus, and Armin Biller. Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. *NeuroImage*, 129:460–469, apr 2016. 2

[52] Arno Klein, Jesper Andersson, Babak A. Ardekani, John Ashburner, Brian Avants, Ming Chang Chiang, Gary E. Christensen, D. Louis Collins, James Gee, Pierre Hellier, Joo Hyun Song, Mark Jenkinson, Claude Lepage, Daniel Rueckert, Paul Thompson, Tom Vercauteren, Roger P. Woods, J. John Mann, and Ramin V. Parsey. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, 46(3):786–802, jul 2009. 2

[53] Simon Kornblith, Jonathon Shlens, and Quoc V Le Google Brain. Do Better ImageNet Models Transfer Better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2661–2671, 2019. 3

[54] Hugo J. Kuijf, Adrià Casamitjana, D. Louis Collins, Mahsa Dadar, Achilleas Georgiou, Mohsen Ghafoorian, Dakai Jin, April Khademi, Jesse Knight, Hongwei Li, Xavier Lladó, J. Matthijs Biesbroek, Miguel Luna, Qaiser Mahmood, Richard Mckinley, Alireza Mehrtash, Sebastien Ourselin, Bo Yong Park, Hyunjin Park, Sang Hyun Park, Simon Pezold, Elodie Puybareau, Jeroen De Bresser, Leticia Rittner, Carole H. Sudre, Sergi Valverde, Veronica Vilaplana, Roland Wiest, Yongchao Xu, Ziyue Xu, Guodong Zeng, Jianguo Zhang, Guoyan Zheng, Rutger Heinen, Christopher Chen, Wiesje Van Der Flier, Frederik Barkhof, Max A. Viergever, Geert Jan Biessels, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, and M. Jorge Cardoso. Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge. *IEEE Transactions on Medical Imaging*, 38(11):2556–2568, nov 2019. 2, 4, 5

[55] Manoj Kumar, Michael J. Anderson, James W. Antony, Christopher Baldassano, Paula P. Brooks, Ming Bo Cai, Po-Hsuan Cameron Chen, Cameron T. Ellis, Gregory Henselman-Petrusek, David Huberdeau, J. Benjamin Hutchinson, Y. Peeta Li, Qihong Lu, Jeremy R. Manning, Anne C. Mennen, Samuel A. Nastase, Hugo Richard, Anna C. Schapiro, Nicolas W. Schuck, Michael Shvartsman, Narayanan Sundaram, Daniel Suo, Javier S. Turek, David Turner, Vy A. Vo, Grant Wallace, Yida Wang, Jamal A. Williams, Hejia Zhang, Xia Zhu, Mihai Capotă, Jonathan D. Cohen, Uri Hasson, Kai Li, Peter J. Ramadge, Nicholas B. Turk-Browne, Theodore L. Willke, and Kenneth A. Norman. BrainIAK: The Brain Imaging Analysis Kit. *Aperture Neuro*, 2021(4), jan 2021. 3

[56] Sonia Laguna, Riana Schleicher, Benjamin Billot, Pamela Schaefer, Brenna Mckaig, Joshua N Goldstein, Kevin N Sheth, Matthew S Rosen, W Taylor Kimberly, and Juan Eugenio Iglesias. Super-resolution of portable low-field MRI in real scenarios: integration with denoising and domain adaptation. *Medical Imaging with Deep Learning (MIDL)*, 2022. 2, 3

[57] Erik G Learned-Miller and Parvez Ahammad. Joint MRI Bias Removal Using Entropy Minimization Across Images. *Advances in Neural Information Processing Systems*, 17, 2004. 3

[58] Jaein Lee, Eunsong Kang, Eunjin Jeon, and Heung Il Suk. Meta-modulation Network for Domain Generalization in Multi-site fMRI Classification. *Lecture Notes in Computer Science*, 12905 LNCS:500–509, 2021. 2

[59] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive Prototype Learning and Allocation for Few-Shot Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8334–8343, 2021. 3

[60] Xiaoxiao Li, Yufeng Gu, Nicha Dvornek, Lawrence H. Staib, Pamela Ventola, and James S. Duncan. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Medical Image Analysis*, 65:101765, oct 2020. 1

[61] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. FSS-1000: A 1000-Class Dataset for Few-Shot Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2869–2878, 2020. 3

[62] Sook-Lei Liew, Bethany P. Lo, Miranda R. Donnelly, Artemis Zavaliangos-Petropulu, Jessica N. Jeong, Giuseppe Barisano, Alexandre Hutton, Julia P. Simon, Julia M. Juliano, Anisha Suri, Zhizhuo Wang, Aisha Abdullah, Jun Kim, Tyler Ard, Nerisa Banaj, Michael R. Borich, Lara A. Boyd, Amy Brodtmann, Cathrin M. Buetefisch, Lei Cao, Jessica M. Cassidy, Valentina Ciullo, Adriana B. Conforto, Steven C. Cramer, Rosalia Dacosta-Aguayo, Ezequiel de la Rosa, Martin Domin, Adrienne N. Dula, Wuwei Feng, Alexandre R. Franco, Fatemeh Geranmayeh, Alexandre Gramfort, Chris M. Gregory, Colleen A. Hanlon, Brenton G. Hordacre, Steven A. Kautz, Mohamed Salah Khlif, Hosung Kim, Jan S. Kirschke, Jingchun Liu, Martin Lotze, Bradley J. MacIntosh, Maria Mataró, Feroze B. Mohamed, Jan E. Nordvik, Gilsoon Park, Amy Pienta, Fabrizio Piras, Shane M. Redman, Kate P. Revill, Mauricio Reyes, Andrew D. Robertson, Na Jin Seo, Surjo R. Soekadar, Gianfranco Spalletta, Alison Sweet, Maria Telenczuk, Gregory Thielman, Lars T. Westlye, Carolee J. Winstein, George F. Wittenberg, Kristin A. Wong, and Chunshui Yu. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific Data*, 9(1):1–12, jun 2022. 4, 5

[63] Jialin Liu, Fei Chao, and Chih-Min Lin. Task Augmentation by Rotating for Meta-Learning. feb 2020. 3

[64] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-Shot Unsupervised Image-to-Image Translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10551–10560, 2019. 3

[65] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. CRNet: Cross-Reference Networks for Few-Shot Segmenta-

tion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4165–4173, 2020. 3

[66] Xiaoxiao Liu, Marc Niethammer, Roland Kwitt, Nikhil Singh, Matt McCormick, and Stephen Aylward. Low-Rank Atlas Image Analyses in the Presence of Pathologies. *IEEE Transactions on Medical Imaging*, 34(12):2583–2591, dec 2015. 2

[67] Xiaofeng Liu, Fangxu Xing, Chao Yang, C. C.Jay Kuo, Georges El Fakhri, and Jonghye Woo. Symmetric-Constrained Irregular Structure Inpainting for Brain MRI Registration with Tumor Pathology. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12658 LNCS:80–91, 2021. 2

[68] Michael Lustig, David L. Donoho, Juan M. Santos, and John M. Pauly. Compressed sensing MRI: A look at how CS can improve on current imaging techniques. *IEEE Signal Processing Magazine*, 25(2):72–82, 2008. 3

[69] Huan Minh Luu and Sung Hong Park. Extending nn-UNet for Brain Tumor Segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12963 LNCS:173–186, 2022. 3

[70] Kushagra Mahajan, Monika Sharma, and Lovekesh Vig. Meta-DermDiagnosis: Few-Shot Skin Disease Identification Using Meta-Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 730–731, 2020. 3

[71] José V. Manjón, Pierrick Coup, Antonio Buades, Vladimir Fonov, D. Louis Collins, and Montserrat Robles. Non-local MRI upsampling. *Medical Image Analysis*, 14(6):784–792, dec 2010. 2

[72] José V. Manjón, Pierrick Coupé, Antonio Buades, D. Louis Collins, and Montserrat Robles. MRI superresolution using self-similarity and image priors. *International Journal of Biomedical Imaging*, 2010, 2010. 2

[73] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. *Journal of Cognitive Neuroscience*, 19:1498–1507, 2007. 4, 5

[74] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M.S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, oct 2015. 4, 5

[75] Fausto Milletari, Nassir Navab, and Seyed Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, pages 565–571, dec 2016. 4, 5, 6

[76] Bao Nguyen, Adam Feldman, Sarath Bethapudi, Andrew Jennings, and Chris G. Willcocks. Unsupervised region-based anomaly detection in brain mri with adversarial image inpainting. *Proceedings - International Symposium on Biomedical Imaging*, 2021-April:1127–1131, apr 2021. 2

[77] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10435 LNCS:417–425, 2017. 2

[78] Alexander F.I. Osman and Nissren M. Tamam. Deep learning-based convolutional neural network for intramodality brain MRI synthesis. *Journal of Applied Clinical Medical Physics*, 23(4):e13530, apr 2022. 2

[79] Moritz Roman Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Enrique Valenzuela Pinilla, Mauricio Reyes, Maria Ines Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, David Robben, Alexander Hutton, Tassilo Friedrich, Teresa Zarth, Johannes Bürkle, The Anh Baran, Bjoern Menze, Gabriel Broocks, Lukas Meyer, Claus Zimmer, Tobias Boeckh-Behrens, Maria Berndt, Benno Ikenberg, Benedikt Wiestler, and Jan S. Kirschke. ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. jun 2022. 4, 5

[80] Maithra Raghu, Chiyuan Zhang, Google Brain, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding Transfer Learning for Medical Imaging. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[81] Sachin Ravi and Hugo Larochelle. Optimization as a Model for Few-Shot Learning. In *International Conference on Learning Representations (ICLR)*, jul 2017. 3

[82] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 9351, pages 234–241. Springer Verlag, 2015. 2, 3, 4, 5, 6

[83] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders, 2019. 3

[84] F. Ségonne, A. M. Dale, E. Busa, M. Glessner, D. Salat, H. K. Hahn, and B. Fischl. A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, 22(3):1060–1075, jul 2004. 2

[85] Ozan Sener and Vladlen Koltun. Multi-Task Learning as Multi-Objective Optimization. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[86] Jun Seo, Young-Hyun Park, Sung Whan Yoon, and Jaekyun Moon. Task-Adaptive Feature Transformer with Semantic Enrichment for Few-Shot Segmentation. feb 2022. 3

[87] David W. Shattuck and Richard M. Leahy. BrainSuite: An automated cortical surface identification tool. *Medical Image Analysis*, 6(2):129–142, jun 2002. 3

[88] X. Shen, F. Tokoglu, X. Papademetris, and R. T. Constable. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *NeuroImage*, 82:403–415, nov 2013. 3

[89] Nalini M Singh, Juan Eugenio Iglesias, Elfar Adalsteinsson, Adrian V Dalca, and Polina Golland. Joint Frequency and Image Space Learning for MRI Reconstruction and Analysis. *Journal of Machine Learning for Biomedical Imaging*, 2022:1–28, 2022. 3, 24

[90] Youssef Skandarani, Nathan Painchaud, Pierre-Marc Jodoin, and Alain Lalande. On the effectiveness of GAN generated cardiac MRIs for segmentation. may 2020. 3

[91] Stephen M. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, nov 2002. 2

[92] Stephen M. Smith, Mark Jenkinson, Mark W. Woolrich, Christian F. Beckmann, Timothy E.J. Behrens, Heidi Johansen-Berg, Peter R. Bannister, Marilena De Luca, Ivana Drobnjak, David E. Flitney, Rami K. Niazy, James Saunders, John Vickers, Yongyue Zhang, Nicola De Stefano, J. Michael Brady, and Paul M. Matthews. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23(1), 2004. 3

[93] Jake Snell, Kevin Swersky, and Twitter Richard Zemel. Prototypical Networks for Few-shot Learning. *Advances in Neural Information Processing Systems*, 30, 2017. 3

[94] Haoliang Sun, Ronak Mehta, Hao H. Zhou, Zhichun Huang, Sterling C. Johnson, Vivek Prabhakaran, and Vikas Singh. DUAL-GLOW: Conditional Flow-Based Generative Model for Modality Transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10611–10620, 2019. 2

[95] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain. *NeuroImage*, 15(1):273–289, jan 2002. 3

[96] T. G.M. Van Erp, D. P. Hibar, J. M. Rasmussen, D. C. Glahn, G. D. Pearlson, O. A. Andreassen, I. Agartz, L. T. Westlye, U. K. Haukvik, A. M. Dale, I. Melle, C. B. Hartberg, O. Gruber, B. Kraemer, D. Zilles, G. Donohoe, S. Kelly, C. McDonald, D. W. Morris, D. M. Cannon, A. Corvin, M. W.J. Machielsen, L. Koenders, L. De Haan, D. J. Veltman, T. D. Satterthwaite, D. H. Wolf, R. C. Gur, R. E. Gur, S. G. Potkin, D. H. Mathalon, B. A. Mueller, A. Preda, F. Macciardi, S. Ehrlich, E. Walton, J. Hass, V. D. Calhoun, H. J. Bockholt, S. R. Sponheim, J. M. Shoemaker, N. E.M. Van Haren, H. E.H. Pol, R. A. Ophoff, R. S. Kahn, R. Roiz-Santiaez, B. Crespo-Facorro, L. Wang, K. I. Alpert, E. G. Jönsson, R. Dimitrova, C. Bois, H. C. Whalley, A. M. McIntosh, S. M. Lawrie, R. Hashimoto, P. M. Thompson, and J. A. Turner. Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular Psychiatry*, 21(4):547–553, jun 2015. 2

[97] David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, and Kamil Ugurbil. The WU-Minn Human Connectome Project: An overview. *NeuroImage*, 80:62–79, oct 2013. 3

[98] Oriol Vinyals, Google Deepmind, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. *Advances in Neural Information Processing Systems*, 29, 2016. 3

[99] Oriol Vinyals, Google Deepmind, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. *Advances in Neural Information Processing Systems*, 29, 2016. 3

[100] Qi Wang, Julius Steiglechner, Tobias Lindig, Benjamin Bender, Klaus Scheffler, and Gabriele Lohmann. Super-Resolution for Ultra High-Field MR Images. In *Medical Imaging with Deep Learning (MIDL)*, 2022. 2

[101] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a Few Examples. *ACM Computing Surveys*, 53(3), jun 2020. 3

[102] Authors withheld for Review. Universeg: Universal Medical Image Segmentation. *Draft Document*, 2023. 3, 5

[103] Mark W. Woolrich, Saad Jbabdi, Brian Patenaude, Michael Chappell, Salima Makni, Timothy Behrens, Christian Beckmann, Mark Jenkinson, and Stephen M. Smith. Bayesian analysis of neuroimaging data in FSL. *NeuroImage*, 45(1), 2009. 3

[104] Xiao Yang, Xu Han, Eunbyung Park, Stephen Aylward, Roland Kwitt, and Marc Niethammer. Registration of Pathological Images. *Simulation and synthesis in medical imaging (Workshop)*, 9968:97, 2016. 2

[105] Huaxiu Yao, Long-Kai Huang, Linjun Zhang, Ying Wei, Li Tian, James Zou, Junzhou Huang, and Zhenhui Li. Improving Generalization in Meta-learning via Task Augmentation. In *Proceedings of the 38th International Conference on Machine Learning, PMLR*, pages 11887–11897. PMLR, jul 2021. 3

[106] Chandan Ganesh Bangalore Yogananda, Benjamin C. Wagner, Gowtham K. Murugesan, Ananth Madhuranthakam, and Joseph A. Maldjian. A deep learning pipeline for automatic skull stripping and brain segmentation. *Proceedings - International Symposium on Biomedical Imaging*, pages 727–731, apr 2019. 2

[107] Lin Yuan, Xue Wei, Hui Shen, Ling Li Zeng, and Dewen Hu. Multi-center brain imaging classification using a novel 3d cnn approach. *IEEE Access*, 6:49925–49934, sep 2018. 2

[108] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid Graph Networks With Connection Attentions for Region-Based One-Shot Semantic Segmentation. In *Proceedings of the IEEE/CVF In-*

ternational Conference on Computer Vision (ICCV), pages 9587–9595, 2019. 3

[109] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. CANet: Class-Agnostic Segmentation Networks With Iterative Refinement and Attentive Few-Shot Learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5217–5226, 2019. 3

[110] Amy Zhao, Guha Balakrishnan, Frédo Durand, John V Guttag, and Adrian V Dalca. Data Augmentation Using Learned Transformations for One-Shot Medical Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8543–8553, 2019. 3

[111] Guoyan Zheng, Chengwen Chu, Daniel L. Belavý, Bulat Ibragimov, Robert Korez, Tomaž Vrtovec, Hugo Hutt, Richard Everson, Judith Meakin, Isabel Lŏpez Andrade, Ben Glocker, Hao Chen, Qi Dou, Pheng Ann Heng, Chunliang Wang, Daniel Forsberg, Aleš Neubert, Jurgen Fripp, Martin Urschler, Darko Stern, Maria Wimmer, Alexey A. Novikov, Hui Cheng, Gabriele Armbrecht, Dieter Felsenberg, and Shuo Li. Evaluation and comparison of 3D intervertebral disc localization and segmentation methods for 3D T2 MR data: A grand challenge. *Medical Image Analysis*, 35:327–344, jan 2017. 3

Supplementary Material for Neuralizer: Neuroimage Analysis without Re-Training

A. Samples

We provide examples of model inputs – target image and context set – and Neuralizer-seen predicted outputs. The inputs are sampled at random from the test dataset. The context set length is sampled from the discrete random uniform distribution $\mathcal{U}_{\{1,32\}}$. To reduce visual clutter, we display up to eight context image pairs and omit the rest in the visualization. We also only show one channel, excluding additional inputs like multiple modalities, or the binary mask for in-painting tasks. We provide a collection of images from the first 50 samples from the test dataset. We only excluded examples to avoid duplication of tasks.

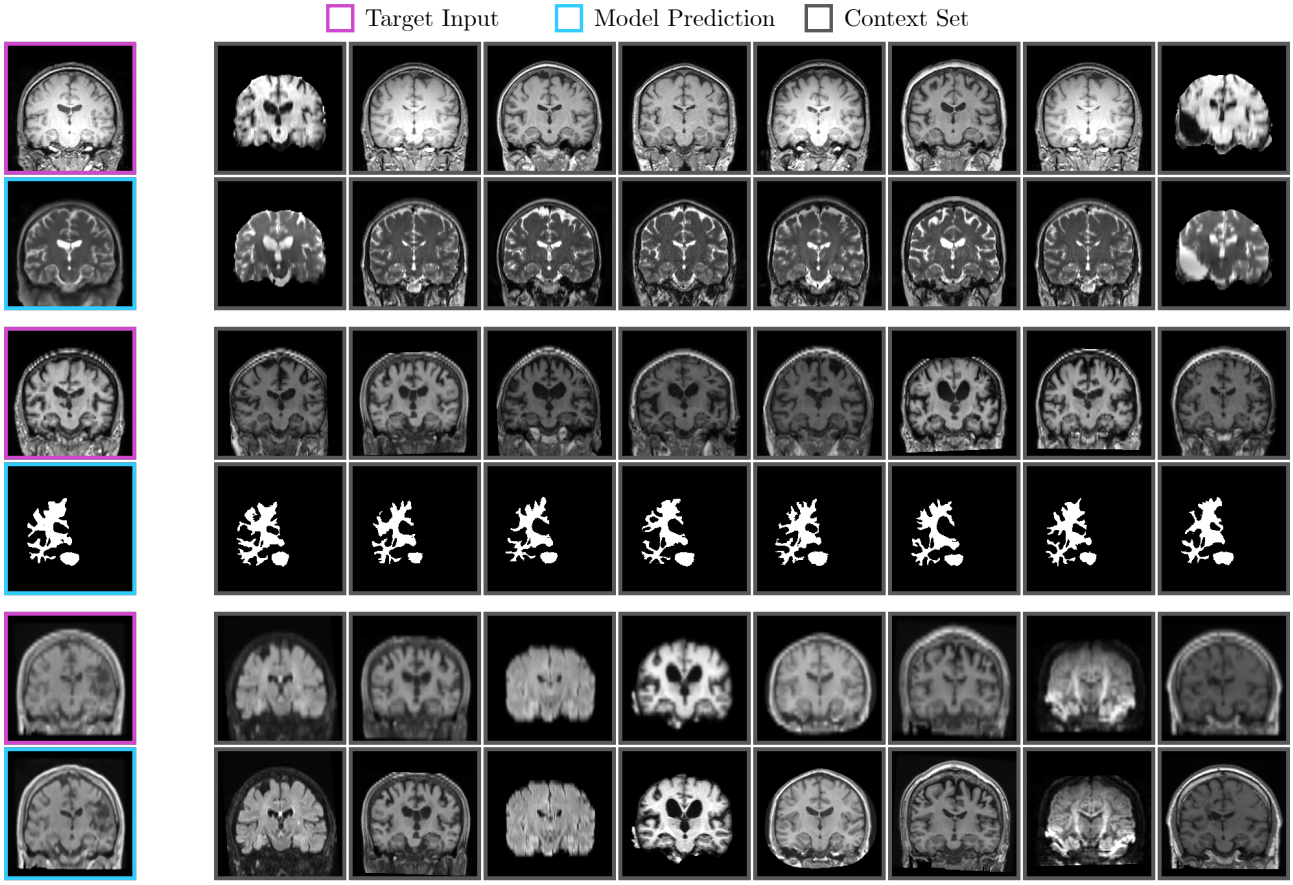


Figure 6. Sample Neuralizer-seen predictions. Left: Target input (magenta frame) and model prediction (blue frame). Right: context set supplied to inform the task (grey frame). We provide more samples on the next pages.

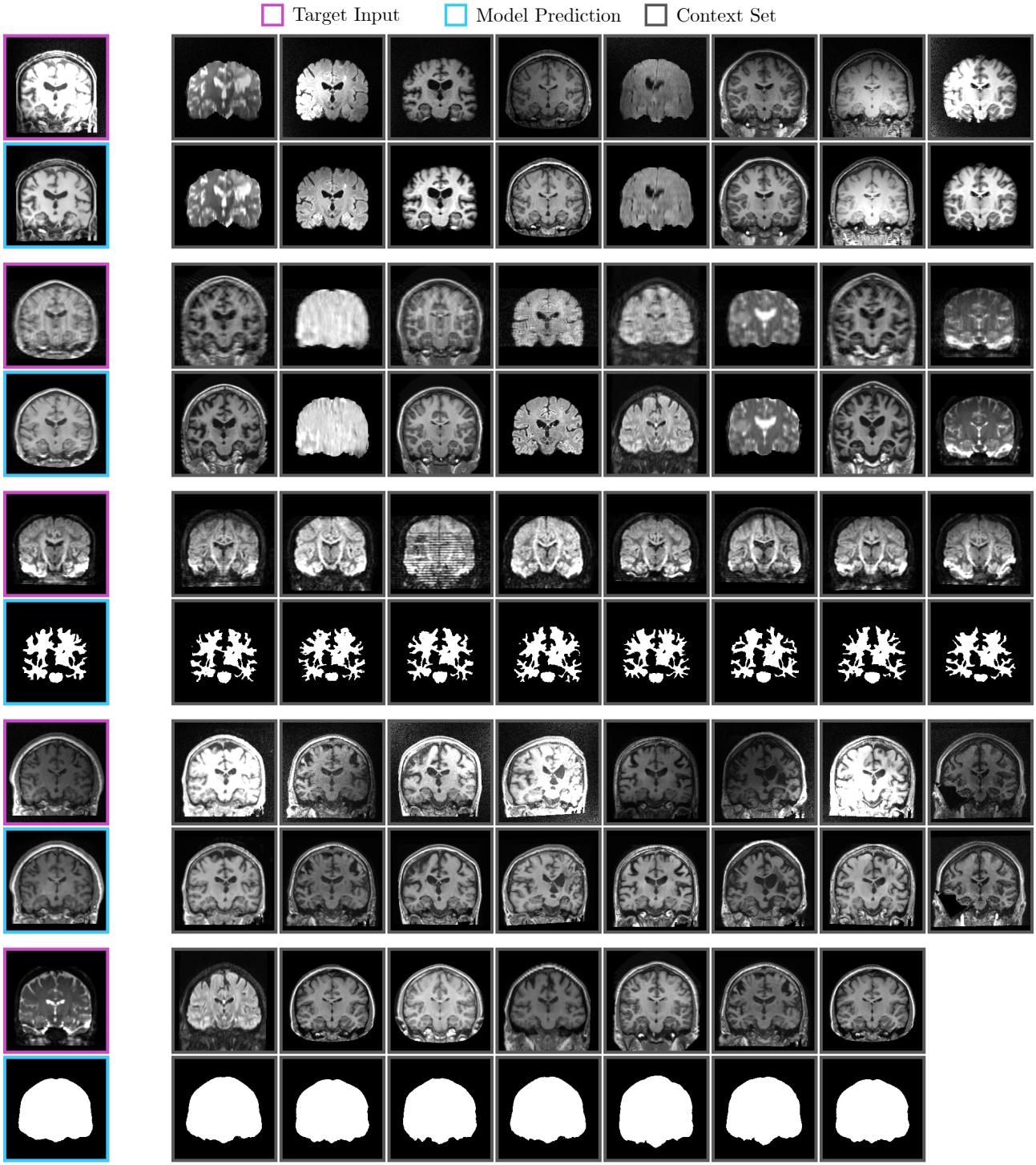


Figure 7. Sample Neuralizer-seen predictions (continued). Left: Target input (magenta frame) and model prediction (blue frame). Right: context set supplied to inform the task (grey frame).

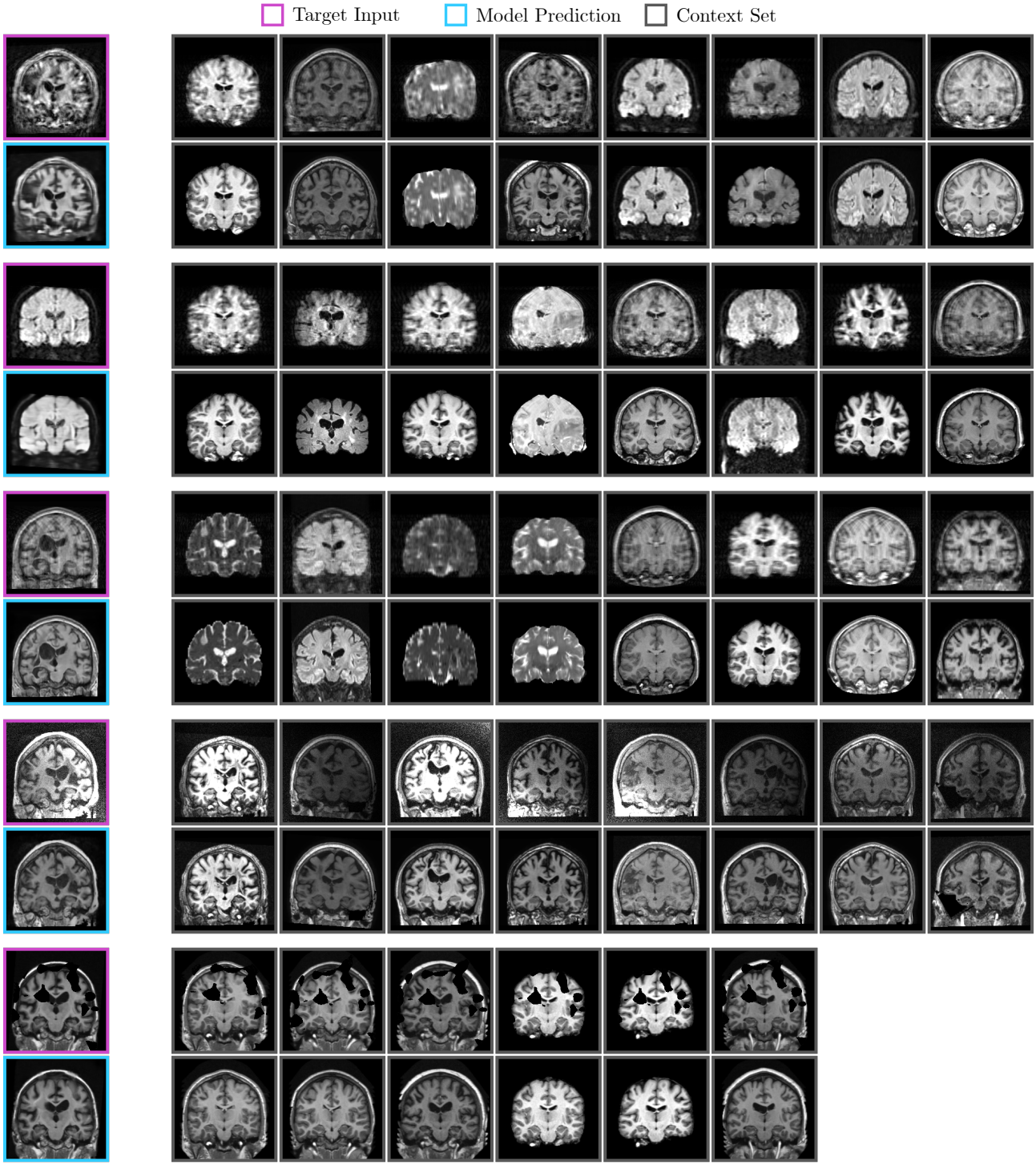


Figure 8. Sample Neuralizer-seen predictions (continued). Left: Target input (magenta frame) and model prediction (blue frame). Right: context set supplied to inform the task (grey frame).

B. Train samples

We provide samples from the train set, including data and task augmentations, and show all three input channels. Further examples of the visual diversity possible with task augmentations are shown in Fig. 11.

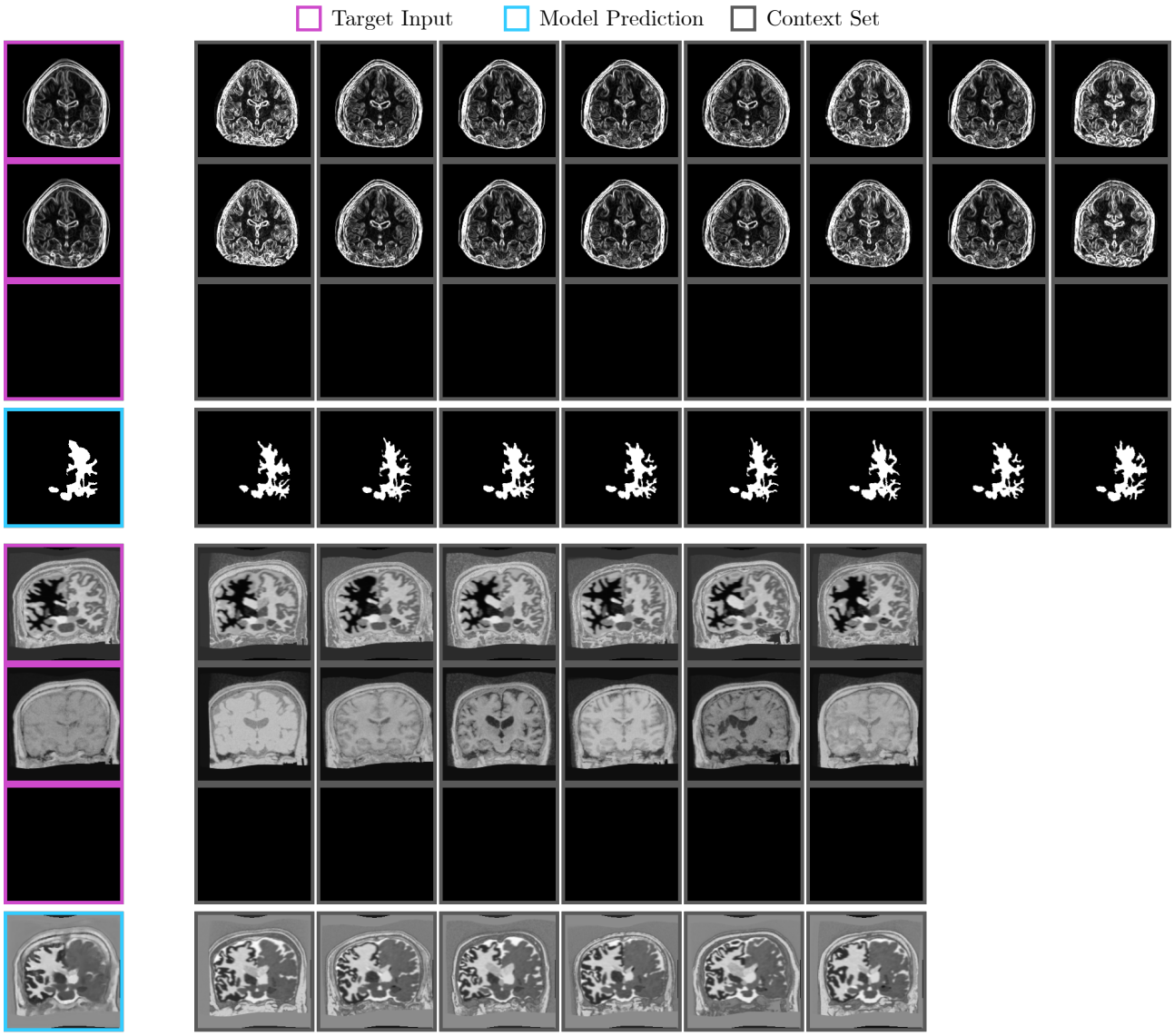


Figure 9. Sample Neuralizer-seen predictions from the train set, with data and task augmentations. All three channels of the input are shown. Left: Target input (magenta frame) and model prediction (blue frame). Right: context set supplied to inform the task (grey frame).

C. Task augmentations

We provide a more detailed description of the task augmentation strategies we employ. First, we describe each task augmentation implemented by us. Then, we show the composition and likelihood of the augmentations during training. Finally, we show examples in Fig. 11. Hyper-parameters for all augmentations are selected by visual inspection.

C.1. Task augmentations

We provide a description of each task augmentation. In addition to the task augmentations, we use data augmentations via random affine movements, random elastic deformations, and random flips along the sagittal plane.

SobelFilter. A Sobel filter is applied to an intensity image.

IntensityMapping. The intensity of an image is remapped. To perform this operation, the image intensity values are split into histogram bins, and each bin is assigned a new intensity difference value. To obtain new intensity values, we compute a distance from the original intensity value to the two neighboring bin centers, using linear interpolation.

SyntheticModality. An intensity image is replaced with a synthetic one generated from an anatomical segmentation map of the subject, using previous work [42]. Each anatomical segmentation class is randomly assigned an intensity mean and standard deviation and the new synthetic modality image of the brain is generated according to these distributions. As our anatomical segmentations do not cover the skull, we take an extra step to ensure skulls are present in the synthetic data: If the original intensity image had a skull, the generated brain is overlaid onto the original image, thus keeping the skull.

MaskContour. We extract a contour of the binary mask in a segmentation task, which then represents the new target segmentation mask. Contoured Masks are always dilated to a width of 3 voxels.

MaskDilation. The binary segmentation mask is dilated by 1 voxel.

MaskInvert. The binary segmentation mask is inverted.

PermuteChannels. The input images are represented by three channels. On each input during training, we permute the input channels. This encourages the network to ignore the specific channel order.

DuplicateChannels. We overwrite empty input channels with the duplication of a non-zero channel. The augmentation is applied to each empty channel with a probability p.

C.2. Composition and likelihood of task augmentations

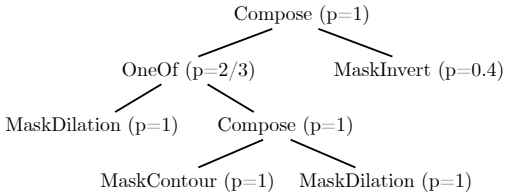
We compose task and data augmentations during training. Some task augmentations can be combined (e.g. MaskDilation and MaskInvert), while others are exclusive to each other (e.g. SobelFilter and SyntheticModality). To model these dependencies, we define the default composition tree used for most tasks in Fig. 10. The augmentation groups "Mask Augmentations", "Intensity Augmentations", "Channel Augmentations", and "Spatial Augmentations" are applied in this order. Augmentations in child nodes of "Compose" are applied left to right, while "OneOf" selects a single child augmentation to apply. A node is applied with probability p stated on the node.

Some tasks use modified versions of this composition tree. As a safety feature, we do not use RandomFlip for segmentation-related tasks, as this can lead to information leakage when evaluating on non-symmetric class-holdouts (in our experiments presented here we always hold out the same anatomical class on both sides of the brain, but this has not always been the case during development). To simplify other tasks, we omit MaskContour and MaskDilate from the inpainting task, and SobelFilter and SyntheticModality form the modality transfer task.

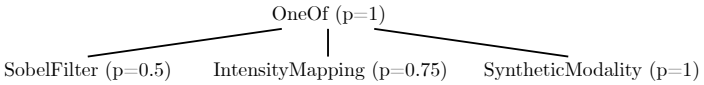
C.3. Examples of task augmentations

Fig. 11 provides visual examples of task augmentations applied to a segmentation and bias correction task.

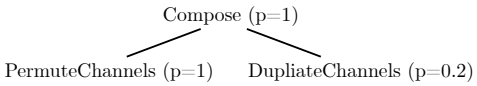
Mask Augmentations



Intensity Augmentations



Channel Augmentations



Spatial Augmentations

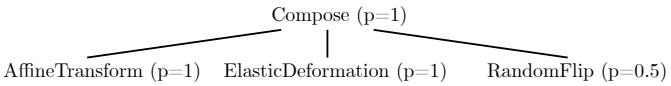


Figure 10. Default composition of augmentations used for most tasks during training. We use "Compose" and "OneOf" nodes to model these restrictions. Augmentations in child nodes of "Compose" are applied left to right, while "OneOf" selects a single child augmentation to apply. A node is applied with probability p.

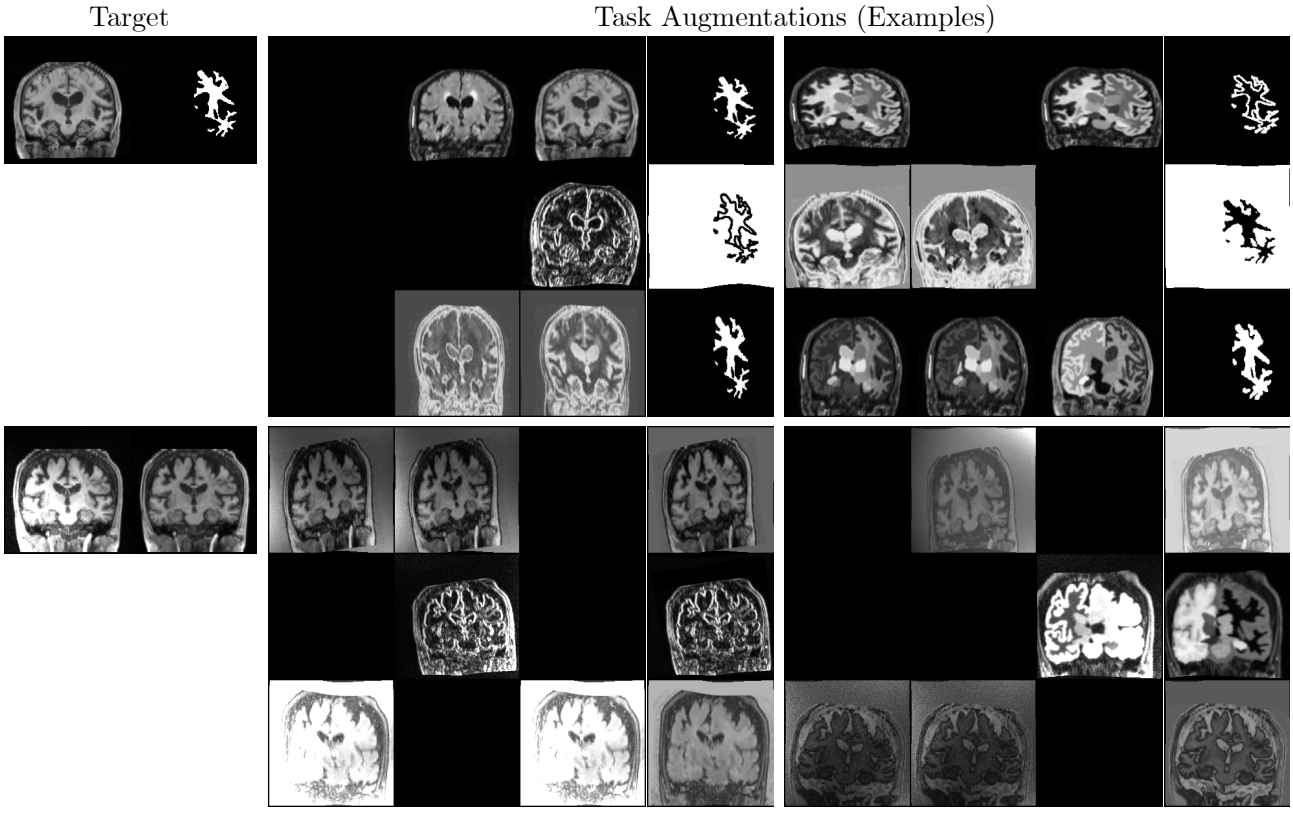


Figure 11. Examples of task augmentations, designed to increase the diversity of neuroimaging tasks seen by the model during training. We show non-augmented target input and output image of T1 modality on the left. We show examples of random data- and task-augmentations applied to the target during training on the right. The augmented target input is represented by up to three channels of real and synthetic modalities of the subject. The target output is augmented with synthetic image modalities and alterations to the segmentation mask. The same augmentations are applied to the context set.

D. Experiments 1 and 2 tabular results

Model	Trained	Subjects	Segmentation	Mod. Transfer	Super Res.	Skull Strip.	Motion Recon.	Undersamp. Recon.	Noise Recon.	Inpainting
Baseline-seen	✓	all	$.83 \pm .08$	25.9 ± 3.0	33.6 ± 2.8	$.98 \pm .01$	31.8 ± 2.9	36.1 ± 2.7	33.5 ± 3.8	38.8 ± 2.4
		32	$.80 \pm .09$	24.4 ± 2.5	31.6 ± 2.3	$.98 \pm .01$	29.3 ± 2.1	33.7 ± 2.2	30.4 ± 3.3	38.3 ± 2.4
		16	$.78 \pm .09$	24.0 ± 2.3	31.3 ± 2.1	$.97 \pm .01$	28.8 ± 2.1	32.3 ± 2.1	30.2 ± 3.5	37.7 ± 2.1
		8	$.77 \pm .11$	23.7 ± 2.2	29.0 ± 1.9	$.97 \pm .02$	28.1 ± 2.1	31.8 ± 2.0	30.0 ± 3.2	36.4 ± 2.2
		4	$.75 \pm .14$	23.0 ± 2.3	29.2 ± 2.5	$.96 \pm .03$	27.9 ± 2.2	31.7 ± 2.0	28.5 ± 3.4	36.4 ± 2.3
		2	$.65 \pm .12$	22.8 ± 2.1	28.7 ± 1.8	$.97 \pm .01$	27.2 ± 1.4	30.4 ± 1.2	27.8 ± 0.8	35.8 ± 2.0
		1	$.59 \pm .16$	22.2 ± 2.1	29.0 ± 2.4	$.95 \pm .02$	27.0 ± 1.8	30.3 ± 1.9	27.6 ± 1.0	35.8 ± 1.8
Neuralizer-seen	✓	32	$.84 \pm .07$	25.3 ± 2.2	32.3 ± 2.8	$.99 \pm .00$	30.2 ± 2.5	34.3 ± 2.7	32.1 ± 3.1	36.1 ± 3.2
		16	$.83 \pm .07$	25.1 ± 2.1	32.9 ± 3.1	$.99 \pm .00$	30.2 ± 2.6	34.2 ± 2.7	31.7 ± 3.0	35.8 ± 2.9
		8	$.82 \pm .09$	24.8 ± 2.2	32.7 ± 3.3	$.98 \pm .00$	30.1 ± 2.6	34.3 ± 2.6	32.1 ± 3.2	35.7 ± 3.1
		4	$.80 \pm .09$	24.2 ± 2.0	32.3 ± 3.2	$.98 \pm .00$	30.1 ± 2.6	34.3 ± 2.7	31.9 ± 3.2	35.1 ± 2.6
		2	$.78 \pm .10$	23.9 ± 2.0	32.3 ± 2.5	$.98 \pm .01$	29.9 ± 2.5	34.2 ± 2.6	30.9 ± 2.9	35.0 ± 2.7
		1	$.74 \pm .13$	23.0 ± 2.0	32.1 ± 2.9	$.98 \pm .01$	29.9 ± 2.6	34.1 ± 2.5	30.9 ± 3.2	34.5 ± 2.7
Neuralizer-unseen	✗	32	$.84 \pm .07$	24.4 ± 2.1	32.1 ± 2.7	$.98 \pm .00$	30.0 ± 2.6	34.2 ± 2.6	30.8 ± 3.9	36.4 ± 3.3
		16	$.83 \pm .07$	24.2 ± 2.1	32.7 ± 3.1	$.98 \pm .00$	29.9 ± 2.6	34.1 ± 2.7	30.3 ± 3.6	36.0 ± 2.7
		8	$.82 \pm .08$	23.8 ± 2.0	32.6 ± 3.2	$.98 \pm .00$	29.9 ± 2.6	34.2 ± 2.6	30.7 ± 3.7	35.8 ± 2.8
		4	$.81 \pm .08$	23.3 ± 1.9	32.2 ± 3.2	$.98 \pm .01$	29.9 ± 2.6	34.1 ± 2.7	30.7 ± 3.9	35.2 ± 2.5
		2	$.78 \pm .09$	22.9 ± 2.0	32.1 ± 2.4	$.98 \pm .01$	29.6 ± 2.5	34.0 ± 2.6	29.7 ± 3.4	35.2 ± 2.6
		1	$.74 \pm .11$	22.1 ± 2.0	31.9 ± 2.9	$.97 \pm .01$	29.7 ± 2.5	33.9 ± 2.5	30.0 ± 3.9	34.5 ± 2.7

Table 4. Model scores (Dice for segmentation and skull-stripping, PSNR for other tasks) for each model and task as a function of the available subjects for training (U-Net) or context set (Neuralizer). Higher values are better. We average scores across all test subjects, eight modalities, and four segmentation classes (Cerebal cortex, Lateral ventricle, Thalamus, Hippocampus). Standard deviation across modalities and segmentation classes.

E. Evaluation on T1 modality

We aggregated scores across all modalities in Fig. 4. To aid comparison to other works, we provide the same evaluation, performed on just the T1 modality here. Some tasks are easier on T1 data, thus improving scores. Note that for small dataset sizes of 1 or 2 subjects, the baselines sometimes underperform on the T1 modality. This is due to a limitation in our implementation, where images of the T1 modality are not always present in small training sets. For sizes of 4 subjects and larger, the t1 modality is always included in the training set.

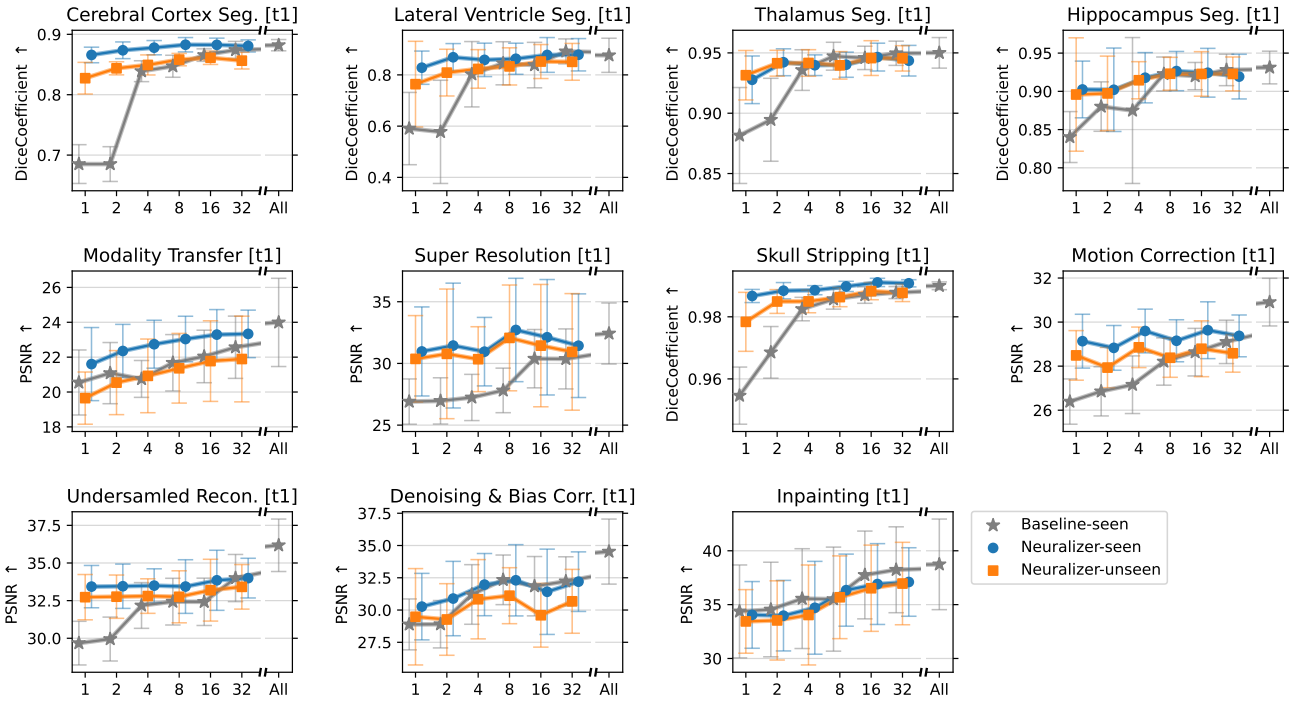


Figure 12. Performance of multi-task Neuralizer and the task-specific baselines on each task, t1 modality only. The tasks being evaluated were included in the training of Neuralizer-seen (orange), held out in Neuralizer-unseen (blue), and specifically trained on by each task-specific baseline (gray). The x-axis is the size of the train/context set, and the y-axis is the Dice/PSNR score. Some points on the x-axis are omitted for better visibility. 'All' refers to all available train data for the task, ranging from 249 to 2,282 subjects depending on the task. The bars denote the standard deviation across subjects.

F. Class names for Hammers Atlas dataset (experiment 3)

We provide label names and indices for the tissue classes in Tab. 3, re-compiled from [23, 34, 37].

Abbreviation	Class Index	Class Name
Hip	2	Hippocampus
PAG	10	Parahippocampal and ambient gyri
STG	12	Superior temporal gyrus
MIG	14	Middle and inferior temporal gyri
FuG	16	Lateral occipitotemporal gyrus (fusiform gyrus)
Stm	19	Brainstem
Ins	20	Insula
PCG	26	Gyrus cinguli, posterior part
Tha	40	Thalamus
CC	44	Corpus callosum
3V	49	Third ventricle
PrG	50	Precentral gyrus
PoG	60	Postcentral gyrus
ALG	94	Anterior long gyrus

Table 5. Hammers Atlas label abbreviations.

G. Training dataset creation

We dynamically generate input image x_t, ground truth output y_t, and context set $\{(x_{t,j}, y_{t,j})\}_{j=1}^{N}$ from a collection of underlying datasets (Tab. 1) during training.

In every training iteration, we first sample a task t from T_{seen}. Next, one of the underlying datasets is selected to generate the sample (x, y). Due to the makeup of the datasets, not every task can be performed on every dataset. For example, a dataset involving a single modality can not naturally be used to generate a modality transfer task. From the list of valid datasets, we sample the datasets for the input and context images independently, with a $1/3$rd chance of all context images coming from the same dataset as the input, $1/3$rd chance that context datasets are sampled at random from the valid datasets, and $1/3$rd chance that the context does not contain any subjects of the input dataset.

After the selection of task and dataset, we create the input and output images. This creation varies by task. We draw the subjects from each dataset at random, but exclude the input subject to re-occur as a context set member. For most tasks, we sample a subset of between one to three image modalities from the subject. For the segmentation task, we join a random subset of available segmentation classes into a binary target mask. For reconstruction and denoising tasks, noise and artifacts in the input images are simulated according to [89]. For the modality transfer task, we select a separate target modality. For the inpainting task, we create a random binary mask from Perlin noise mask these areas from the input image. For skull stripping, the target is a binary brain mask. For tasks other than segmentation and modality transfer, the modality of context images can vary from the input image.

H. Inference cost and model size

We provide an overview of model parameter count and inference cost here. we use a Baseline U-net with 64 channels for experiments with limited data set sizes, and a U-Net with 256 channels for experiments on all data. For Neuralizer, we use the same model in all experiments, but the inference cost increases linearly with the size of the context set.

Table 6. Model size and inference cost.

Model	inference FLOP (g)	Parameters (m)
Baseline, 64 channels	20.7	0.62
Baseline, 256 channels	329.7	9.84
Neuralizer, 1 ctx image	39.1	1.27
Neuralizer, 32 ctx images	610.5	1.27

I. Task weights

To speed up training, we use weighted sampling of tasks during training. Task weights are shown in Tab. 7. These values have been tuned experimentally. Tasks that converge fast and achieve high-quality results are given a lower weight. Tasks that take longer to converge or are given a higher weight.

Table 7. Task weights during training.

Task	Weight
Binary Segmentation	2.0
Modality Transfer	2.0
Superresolution	1.0
Skull Stripping	.5
Motioncorrection Reconstruction	.5
Denoising & Bias correction	.5
k-space Undersampling Recon.	1.0
Inpainting	1.0
Simulated Modality Transfer	1.0
Masking	.5