Neuralizer: General Neuroimage Analysis without Re-Training

Steffen Czolbe
University of Copenhagen and MGH
per.sc@di.ku.dk

Adrian V. Dalca
MIT, and MGH, Harvard Medical School
adalca@mit.edu

Abstract

Neuroimage processing tasks like segmentation, reconstruction, and registration are central to the study of neuroscience. Robust 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 neuroscientists and clinical researchers who often lack the resources or machine-learning expertise to train deep learning models. In practice, this leads to a lack of adoption of deep learning, and neuroscience tools being dominated by classical frameworks.

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. 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. Our experiments on coronal slices show that when few annotated subjects are available, 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 enabled a deep understanding of the human brain. The field has also led to advanced patient care by facilitating non-invasive methods of diagnosis and treatment. Recent deep learning research promises to substantially increase the accuracy and speed of neuroimaging analysis methods.

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 type of data it has been trained on. Generalization to new tasks and domains, such as different acquisition protocols or new segmentation, is

Figure 1. Neuralizer can solve a broad range of image processing tasks, including new ones not seen during training, 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 main barrier to adoption [66]. Performing neuroimaging tasks like segmentation, registration, reconstruction, or motion correction requires different models for each processing step, despite operating on the same input data and methods exhibiting strong similarities in network architecture [13,47,90]. 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 often manually annotated, and the network, training, and data loading logic needs to be implemented. All these steps generally require machine learning and neuroimaging expertise. In addition, computational resources like specialized graphics processing hardware and software infrastructure needs to be available. 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

| Binary Segmentation | Modality Transfer | Super Resolution | Skull Stripping | Motion Correction | Undersampled Reconstruction | Denoising & Bias Correction | Inpainting |

Figure 2. Example neuroimaging tasks and modalities included in our dataset (top: input images, bottom: output images).

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, 36, 105]. 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 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 can solve new tasks, unseen during training, using a set of examples of the new task at inference (Fig. 1)

Neuralizer involves a convolutional architecture (Fig. 3), that takes as input a context set of examples that define the processing task, and thus does not require prior specification 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 limit our experiments to 2D. We evaluate our model by comparing the 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 ≤ 32 labeled examples are available, despite never training on the task. When generalizing to new segmentation protocols, Neuralizer matches the performance of baselines trained directly on the dataset.

2. 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 [64, 116], device, protocol, imaging sequence [60], and use of contrast agents [9, 38].

To analyze these images, a variety of processing tasks are most often combined in a processing pipeline. Common processing tasks include anatomical segmentation [12, 18, 22, 23, 32, 87, 103, 109], skull stripping [51, 51, 57, 92, 99, 115], defacing [2, 41], registration [5–7, 10, 20, 29, 47, 49, 58], modality transfer [84, 85, 102], in-painting [43, 72, 73, 83, 113], super-resolution [21, 62, 77, 78, 110], reconstruction, and de-noising [62, 74, 97], bias field removal [37, 63], surface fitting [50] and parcellation [96, 104].

Multiple toolboxes provide a suite of interoperable software components, most implementing classical optimization strategies. Widely used toolboxes include Freesurfer [32], FSL [55, 100, 112], SPM [35], CIVET [3], BrainSuite [95], HCP pipeline [106], and BrainIAK [61]. Deep-learning-based methods are starting to be included because of their improved accuracy and shorter runtime [13, 51]. While these methods provide solutions for common neuroimaging applications, most are limited to a single task and few modalities. Developers need to manually update the pipelines to include new processing tasks and to support a wider variety of image modalities. This process requires extensive technical expertiese and computational resources, often not available to the clinical neuroscientists focusing on scientific questions.

2.2. Multi-task learning

Multi-Task Learning (MTL) frameworks solve multiple tasks simultaneously by exploiting similarities between related tasks [17]. MTL can improve performance and reduce computational cost and development time compared to designing task-specific solutions [27, 93]. 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 [25, 40]. This strategy is challenging to scale as the number of tasks increases, requires prior determination of the set of tasks, and importantly 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 importantly with the ability to generalize to new tasks and modalities.

2.3. 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 [45] trained on ImageNet [24] and fine-tuning part of the network for a new task [52, 59, 107]. For medical imaging, networks pre-trained on large sets of medical images are available [19], and fine-tuning them to new tasks results in shortened training time and higher accuracy [4, 76]. However, fine-tuning also requires machine learning expertise and computational resources, most often not available in clinical research. Additionally, in scenarios with small datasets, fine-tuning models trained on large vision datasets can be harmful [88].

2.4. Few-shot learning

Few-shot models generate predictions from just a few labeled examples [70, 89, 91, 111], or in the case of zero-shot methods [14], none at all. Many of these methods require training or fine-tuning. In computer vision, several methods pass a query image, along with a set of support images and labels as input to the model [70, 94, 101, 108]. Natural image segmentation methods [71, 118] use single image-label pairs [65, 117] as support or aggregate information from a larger support set [67]. Recent few-shot learning methods in the medical image segmentation setting [11, 30, 31] operate on a specific anatomical region in a single image modality [44, 120]. Similar Prior-Data Fitted Networks (PFNs) are fitted to multiple datasets at once to learn the training and prediction algorithm [82]. During training, this strategy draws a dataset, a set of data points and their labels from it, masks one of the labels and predicts it. The resulting model aims to generalize to new datasets. PFNs have only been applied to low-dimensional and tabular data [48]. Our solution builds on ideas from these methods, but aims to solve a much larger range of diverse image-to-image tasks on neuroimages of many modalities.

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. NN-UNet [53] automates the design and training of models for biomedical image segmentation, and has been successfully applied to brain segmentation [26, 54, 75]. While AutoML effectively reduces 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 [90, 119]. It improves model 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 [13, 16, 47, 51, 98]. In meta-learning, data augmentation can further be used to generate entirely new tasks [15, 69, 114]. 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. We then introduce the model architecture, training, and inference strategies.

3.1. Generalizabe multi-task 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 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-output image pairs passed to the model alongside the prediction task. The context set defines the desired task, and can vary in size $|C_t| = N$ and is re-sampled from the underlying task-datasets for each input. Fig. 1 gives an example for a modality transfer task.

We employ a neural network $g_\theta(x_t, C_t) = y_t$ with weights θ that applies the task defined by context set C_t

Neuralizer

Target Input

Target Output

Embedding

Pairwise-Conv-Avg Block

Output Convs

Context Set (varying size N)

Pairwise-Conv-Avg Block

Target Representation

Residual Unit

AVG

Context Representations

Residual Unit (shared weights)

Concat

1x1 conv (s. weights)

Residual Unit (shared weights)

Concat

1x1 conv (s. weights)

Residual Unit (shared weights)

$\times N$

Figure 3. Neuralizer consists of 7 Pairwise-Conv-Avg blocks (right), arranged in a U-Net-like [81, 90] 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. The architecture is invariant to context size N.

to the input neuroimage x_t. We optimize the network using supervised training with the loss

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

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, neuroimage encodings, and generation of the training set for each task type.

Loss function. Neuralizer solves both segmentation tasks (e.g. anatomical segmentation and skull-stripping via a brain mask), more general and image-to-image tasks (e.g. denoising). We use the Soft Dice Loss [81] for segmentation-like tasks, and the pixel-wise Mean Squared Error $MSE(y_t, g(x_t, C_t)) = \frac{1}{2\sigma^2} \sum_p [y_{t_p} - g(x_t, C_t)_p]^2$ with balancing hyperparameter σ^2 for other tasks. As the network optimizes multiple tasks during training, the balance of the loss terms can dramatically affect the optimization and resulting performance.

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 any channels unnecessary for a specific task. The output space $\mathcal{Y}$ follows the same design but uses only one channel.

Training dataset creation. At each training iteration, we first sample a task t from T_{seen}, selecting the task-specific dataset (Tab. 1). From this dataset, we sample the input image, ground truth output, and image pairs for the context set.

Table 1. Tasks, Modalities, Datasets, and Segmentation classes used in this paper, and involved in 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 [49, 79]	Freesurfer protocol, 31
BRATS [8, 9, 80]	classes [13, 32]
IXI [1]	
ATLAS R2.0 [68]	Manually-annotated Hammers
Hammers Atlas [42]	Atlas, 96 classes [28, 39, 42]
WMH Challenge [60]	
ISLES2022 [86]	Brainmasks [32, 51]

To increase the range of images that can be used to condition the trained model, the image modalities and acquisition protocols of entries 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.

3.3. Model architecture

Fig. 3 shows the Neuralizer network architecture, adapted with the concurrently developed [15] – a method

that focuses on solving broad segmentation tasks. 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 using multiple Pairwise-Conv-Avg Blocks (explained below), arranged as a U-Net-like configuration [81,90] to exploit multiple scales. The output r_x^{out} of the final Pairwise-Conv-Avg Block is processed by a residual unit [45] 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 [46].

Compared to standard CNNs, Neuralizer uses a mechanism to enable knowledge transfer from the context set to the input image. We design 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^{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, which involve two convolutions, operate on the context representations and have shared weights. Second, we pairwise concatenate the context representations with the target representation on the channel dimension: $p_i = \text{cat}(r_x^{\text{int}}, r_{C_i}^{\text{int}})$. We combine the pairwise representations and reduce the channel size back to c using 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 did not lead to an increased generalization to unseen tasks.

3.4. Task augmentations

To further diversify the training dataset, we employ task augmentations [15], a group of transformations applied at random to the input, output, and context images. The objective is 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 summarized 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 desired 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 [15]. We also increase the context set by sampling with replacement from the context set, and add small affine augmentations.

4. Experiments

We first compare Neuralizer with task-specific networks, which require substantial expertise and compute. We then analyze the effect of the size of the context set, and the multi-task generalization to unseen segmentation protocols 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 [49, 79], BRATS [8, 9, 80], Atlas R2.0 [68], Hammers Atlas [42], IXI [1], ISLES2022 [86], and the White Matter Hyperintensities Challenge [60]. We segment all subjects with Synthseg [13, 32]. Based on the segmentation, we affinely align the images to the MNI 152 template space [33, 34], and resample to 1mm isometric resolution at a size of $192 \times 224 \times 192$mm. We perform manual quality control of the segmentation and registration by ensuring no segmented areas fall outside of the cropped volume and discard subjects failing this check (4 subjects). 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 [32, 51]. The final dataset contains 2,282 subjects with 15,911 images and segmentation masks across 8 modalities. Subjects of the seven original datasets are split into 80% for 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

5

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 standard deviation across modalities. We extract results for T1 scans in Supplement D.

4.2. Models

Neuralizer-seen. This Neuralizer model includes all tasks available during training. We use this model to evaluate the performance 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 Neuralizer performance 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 [81, 90]. However, training one model for each task and modality requires overwhelming computational resources. To reduce the computational requirement, we follow previous modality-agnostic models [13, 51] 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 given an equal amount of data, we train baselines with training set sizes of $\{1, 2, 4, 8, 16, 32, \text{all}\}$ and employ standard 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 when all data is available for training, and 64 channels otherwise. Using larger U-Nets resulted in overfitting and lower performance. Supplement H summarizes model parameter counts and inference costs.

4.3. Training

We use supervised training, task-specific loss functions, and weigh the MSE loss by selecting $\sigma^2 = 0.05$, resulting in the 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 [56]. To speed up training, we undersample 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

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.

deformations, and random flips along the sagittal plane. For Baseline-seen, we reuse the augmentations but remove those that introduce uncertainty in the desired output. 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 exhibiting no decrease in validation loss. The model with the lowest validation loss is used for further evaluation on the test set.

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 low-data regime 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: Baseline Comparison

To assess if the proposed multi-task approach is competitive with 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, averaged across modalities in Fig. 4. We also provide an evaluation of using just the T1 modality in Supplement D, since many task-specific networks in neuroimage analysis literature focus on T1 images. We further aggregate performance across all tasks in Fig. 5, and provide tabular results in Supplement E. Both Neuralizer models outperform most 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 analysis

We assess the few-shot setting that is prevalent in neuroimage analysis, where few annotated images are often available for a new task. 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. For most tasks, 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 a new segmentation protocol

The Hammers Atlas dataset [28,39,42] 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 differences in visual characteristics. We use the Hammers Atlas dataset to evaluate Neuralizer-

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 protocol is withheld from training. Evaluation of major labels located in the center and right of the coronal slice. See Supplement F for class abbreviations.

unseen by entirely withholding the dataset and its annotations 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 neuroimaging 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). Neuralizer provides a performance advantage on smaller datasets likely by exploiting neuroimaging similarities across the many other 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 and is an interesting direction of study. 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 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 compute cluster. Entire volumetric data also impose a challenging memory requirement on Neuralizer models. To tackle 3D data in the future, we plan to process multiple slices at a time.

We affinely aligned the neuroimages 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.

We originally experimented with tumor and lesion segmentation tasks but found this to be a more challenging scenario. Lesions are spatially heterogeneous, making learning from the context set much harder. We excluded tumor and lesion segmentation masks from the experiments, but plan to study this setting in the future.

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 plan to extend Neuralizer to tackle these in the future.

6. Conclusion

Neuralizer performs accurate rapid single-pass, multi-task generalization, and even 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 prohibitive requirement of retraining or fine-tuning a task-specific model. We believe this will facilitate neuroscience analyses not currently possible.

Acknowledgments

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. 2

[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 Greve, Koen Van Leemput, Bruce Fischl, Juan Eugenio Iglesias*, and Adrian V Dalca*. A learning strategy for contrast-agnostic mri segmentation. In *Medical Imaging with Deep Learning*, pages 75–93. PMLR, 2020. 2

[13] 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. 1, 2, 3, 4, 5, 6

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

[15] Victor Ion Butoi*, Jose Javier Gonzalez Ortiz*, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. *arXiv preprint arXiv:2304.06131*, 2023. 3, 4, 5

[16] 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

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

[18] 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

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

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

[21] Adrian V Dalca, Katherine L Bouman, William T Freeman, Natalia S Rost, Mert R Sabuncu, and Polina Golland. Medical image imputation from image collections. *IEEE transactions on medical imaging*, 38(2):504–514, 2018. 2

[22] Adrian V Dalca, John Guttag, and Mert R Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9290–9299, 2018. 2

[23] Adrian V Dalca, Evan Yu, Polina Golland, Bruce Fischl, Mert R Sabuncu, and Juan Eugenio Iglesias. Unsupervised deep learning for bayesian brain mri segmentation. In *Medical Image Computing and Computer Assisted Intervention– MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pages 356–365. Springer, 2019. 2

[24] 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

[25] 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

[26] 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

[27] 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

[28] 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, 24

[29] 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

[30] 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

[31] 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

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

[33] 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

[34] 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

[35] Richard SJ Frackowiak. *Human Brain Function*. 2004. 2

[36] 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

[37] 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. 2

[38] 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

[39] 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, 24

[40] 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

[41] 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

[42] 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, 24

[43] 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

[44] 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

[45] 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

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

[47] 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. 1, 2, 3, 19

[48] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *NeurIPS 2022 First Table Representation Workshop*, jul 2022. 3

[49] Andrew Hoopes, Malte Hoffmann, Douglas N Greve, Bruce Fischl, John Guttag, and Adrian V Dalca. Learning the effect of registration hyperparameters with hypermorph. *Machine Learning for Biomedical Imaging*, 1:1–30, 2022. 2, 4, 5

[50] 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. 2

[51] 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

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

[53] 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

[54] 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

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

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

[57] 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

[58] 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

[59] 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

[60] 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

[61] 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. 2

[62] 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

[63] Erik G Learned-Miller and Parvez Ahammad. Joint MRI Bias Removal Using Entropy Minimization Across Images. *Advances in Neural Information Processing Systems*, 17, 2004. 2

[64] 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

[65] 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

[66] 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

[67] 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

[68] 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

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

[70] 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

[71] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. CR-Net: 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. 3

[72] 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

[73] 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

[74] 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. 2

[75] 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

[76] 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

[77] 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

[78] 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

[79] 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

[80] 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

[81] 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

[82] Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers Can Do Bayesian Inference. In *International Conference on Learning Representations*, mar 2022. 3

[83] 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

[84] 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

[85] 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

[86] 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

[87] Kilian M Pohl, Sylvain Bouix, Motoaki Nakamura, Torsten Rohlfing, Robert W McCarley, Ron Kikinis, W Eric L Grimson, Martha E Shenton, and William M Wells. A hierarchical algorithm for mr brain image parcellation. *IEEE transactions on medical imaging*, 26(9):1201–1212, 2007. 2

[88] 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

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

[90] 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. 1, 3, 4, 5, 6

[91] 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. 3

[92] 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

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

[94] 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

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

[96] 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. 2

[97] 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. 2, 25

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

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

[100] 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. 2

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

[102] 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

[103] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Octave Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc 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, 2002. 2

[104] 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. 2

[105] 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

[106] 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. 2

[107] 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

[108] 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

[109] Christian Wachinger, Martin Reuter, and Tassilo Klein. Deepnat: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage*, 170:434–445, 2018. 2

[110] 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

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

[112] 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. 2

[113] 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

[114] 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

[115] 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

[116] 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

[117] 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, 2019. 3

[118] 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

[119] 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

[120] 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

Task augmentations are randomized data augmentations applied to both input and target images or segmentation maps. These change not only the appearance of the input image, but also the target and members of the context set, essentially altering the task itself. We apply task augmentations not to create plausible neuroimaging tasks, but instead to expand the set of tasks the model is exposed to during training. This prevents memorization of the training tasks, and aids generalization to unseen tasks during inference. We first describe the task augmentations in C.1, then discuss their composition in C.2, and finally provide 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 [47] 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 [47]. 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.

Mask Augmentations

Compose (p=1)
- OneOf (p=2/3)
  - MaskDilation (p=1)
  - Compose (p=1)
    - MaskContour (p=1)
    - MaskDilation (p=1)
- MaskInvert (p=0.4)

Intensity Augmentations

OneOf (p=1)
- SobelFilter (p=0.5)
- IntensityMapping (p=0.75)
- SyntheticModality (p=1)

Channel Augmentations

Compose (p=1)
- PermuteChannels (p=1)
- DupliateChannels (p=0.2)

Spatial Augmentations

Compose (p=1)
- AffineTransform (p=1)
- ElasticDeformation (p=1)
- RandomFlip (p=0.5)

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.

C.3. Examples of task augmentations

Fig. 11 provides visual examples of task augmentations applied to a segmentation and bias correction task.

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.

21

D. Evaluation on T1 modality

We aggregated scores across all modalities in Fig. 4. To aid comparison to existing literature, which most often focuses on T1 images, we provide the same evaluation, performed on just the T1 modality here. Some tasks are easier on T1 data, thus improving scores. For small dataset sizes of 1 or 2 subjects, the baselines sometimes underperform on the T1 modality. This is often because images of the T1 modality may 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.

E. 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.

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 [28, 39, 42].

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 [97]. 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 model parameter counts and inference costs. 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