



# VINNA for neonates: Orientation independence through latent augmentations

Leonie Henschel<sup>a</sup>, David Kügler<sup>a</sup>, Lilla Zöllei<sup>b,c,\*</sup>, Martin Reuter<sup>a,b,c,\*</sup>

<sup>a</sup>German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

<sup>b</sup>A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA, USA

<sup>c</sup>Department of Radiology, Harvard Medical School, Boston, MA, USA

\*Equal contribution

Corresponding Author: Martin Reuter ([martin.reuter@dzne.de](mailto:martin.reuter@dzne.de))

## ABSTRACT

A robust, fast, and accurate segmentation of neonatal brain images is highly desired to better understand and detect changes during development and disease, specifically considering the rise in imaging studies for this cohort. Yet, the limited availability of ground truth datasets, lack of standardized acquisition protocols, and wide variations of head positioning in the scanner pose challenges for method development. A few automated image analysis pipelines exist for newborn brain Magnetic Resonance Image (MRI) segmentation, but they often rely on time-consuming non-linear spatial registration procedures and require resampling to a common resolution, subject to loss of information due to interpolation and down-sampling. Without registration and image resampling, variations with respect to head positions and voxel resolutions have to be addressed differently. In deep learning, external augmentations such as rotation, translation, and scaling are traditionally used to artificially expand the representation of spatial variability, which subsequently increases both the training dataset size and robustness. However, these transformations in the image space still require resampling, reducing accuracy specifically in the context of label interpolation. We recently introduced the concept of resolution-independence with the Voxel-size Independent Neural Network framework, VINN. Here, we extend this concept by additionally shifting all rigid-transforms into the network architecture with a four degree of freedom (4-DOF) transform module, enabling resolution-aware internal augmentations (VINNA) for deep learning. In this work, we show that VINNA (i) significantly outperforms state-of-the-art external augmentation approaches, (ii) effectively addresses the head variations present specifically in newborn datasets, and (iii) retains high segmentation accuracy across a range of resolutions (0.5–1.0 mm). Furthermore, the 4-DOF transform module together with internal augmentations is a powerful, general approach to implement spatial augmentation without requiring image or label interpolation. The specific network application to newborns will be made publicly available as VINNA4neonates.

**Keywords:** computational neuroimaging, deep learning, structural MRI, artificial intelligence, high-resolution, newborn brain

## 1. INTRODUCTION

Collections of neonatal brain Magnetic Resonance Images (MRIs) are indispensable to understand brain development and to detect early signs of potential developmental disorders. One of the key tasks in MRI analysis

is automated segmentation, the labeling of anatomical regions of interest (ROIs) that can be used for quantitative modeling of healthy development, for analyses in population studies, for understanding disease effects as well as a starting point for further neuroimaging tasks. The segmentation of infant MRIs is a challenging and non-trivial

Received: 21 November 2023 Revision: 16 February 2024 Accepted: 19 April 2024 Available Online: 6 May 2024



© 2024 Massachusetts Institute of Technology.  
Published under a Creative Commons Attribution 4.0  
International (CC BY 4.0) license.

Imaging Neuroscience, Volume 2, 2024  
[https://doi.org/10.1162/imag\\_a\\_00180](https://doi.org/10.1162/imag_a_00180)

undertaking due to the rapid non-linear changes during the postnatal brain growth period, elevated levels of head motion, limited availability of congruent datasets, varying intensity profiles across scanners, protocols and modalities, as well as the inversion of gray-white contrast around the age of 5–9 months (Ajayi-Obe et al., 2000; Dubois et al., 2014; Gilmore et al., 2012; Prastawa et al., 2005). In this paper, we focus on a sub-group of the infant population—newborns—and present a four-degree of freedom (4-DOF) transform module to address two core challenges within this cohort: non-uniform image resolutions (scaling) and increased variability of head positions (rigid transformations = rotation and translation) during image acquisition. The 4-DOF transform module is directly integrated into the network architecture and addresses the variability of head positions internally. As such, it expands and generalizes the distinguishing feature, resolution independence, of the recently published Voxel-size Independent Neural Network (VINN) (Henschel et al., 2022) by rotation and translation transformations. We refer to our new framework as VINN with “internal augmentations” (VINNA).

In contrast to adults, newborn head positions in the scanner are far more diverse due to the scanning conditions (asleep) and overall smaller anatomy. While padding is often used to stabilize the child’s head and to occupy the space between head coil and participant (e.g., foam cushions, pads, mats, pillows, or visco-elastic matters) (Copeland et al., 2021), its standardization is difficult. This results in diverse head orientations within the scanner and potentially high variations among imaging studies.

In addition, there is no de-facto standard resolution for newborn imaging. In the case of research protocols, when more time is available, MRIs are often acquired at higher resolutions to address the small size of brain structures and stronger partial volume effects (Dubois et al., 2019, 2021; Liu et al., 2019). However, the range of recorded resolutions across research and clinical studies is relatively large and heterogeneous, ranging from 0.5 mm to 3.0 mm in past and recent research studies (e.g., NIH-PD (Evans, 2006), BCP (Howell et al., 2019), Developing Human Connectome Project (dHCP) (Bastiani et al., 2019; Makropoulos et al., 2018), HBCD (Volkow et al., 2021)). Similarly, resolutions are not standardized across atlases (UNC 0-1-2 Infant atlases (Shi et al., 2011), Imperial Brain Development atlases (Gousias et al., 2012; Kuklisova-Murgasova et al., 2011; Makropoulos et al., 2016; Serag et al., 2012)), that are often used to guide the automated labeling algorithms.

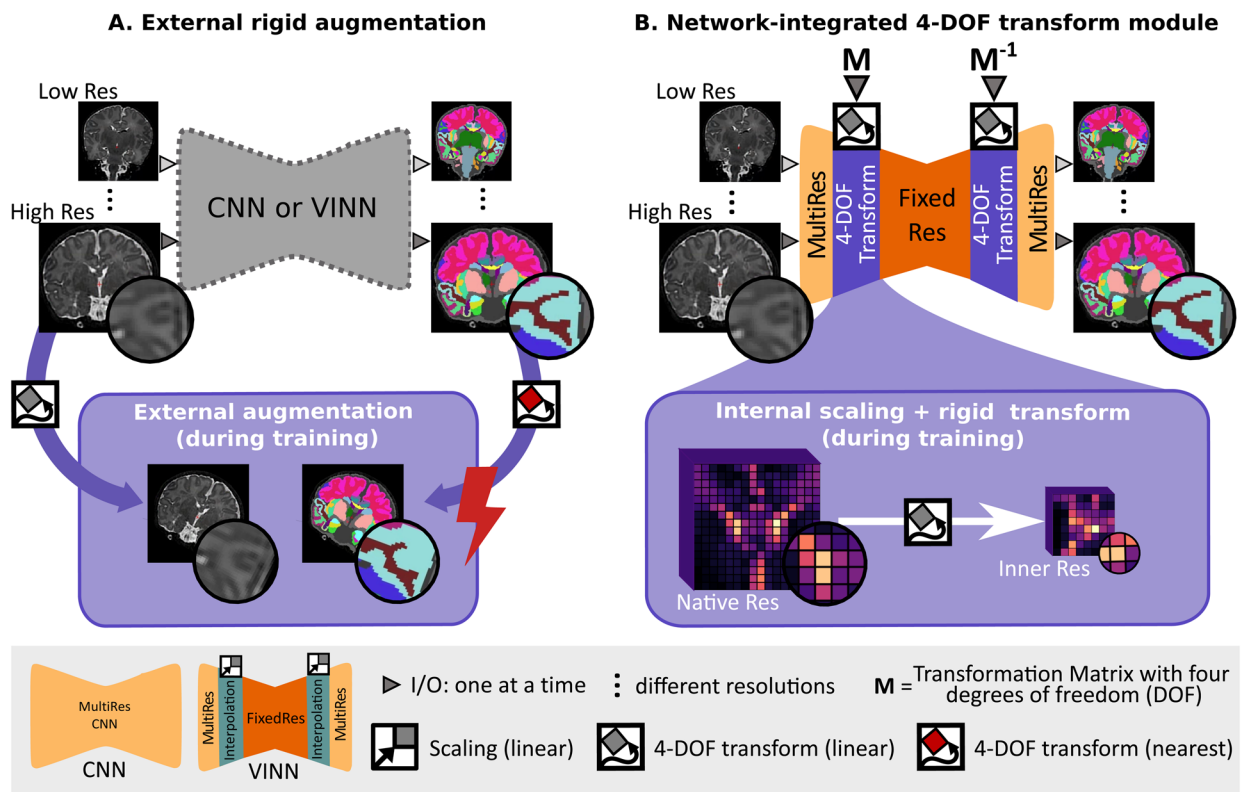
Traditional tools for newborn or infant segmentation predominantly interpolate images to a single chosen resolution and harmonize the spatial head position via atlas registrations (Cherel et al., 2015; Dai et al., 2013;

Makropoulos et al., 2018; Prieto et al., 2019; Shi et al., 2010; Zöllei et al., 2020). Resampling of images can, however, result in loss of information, especially in the context of high-resolution label maps. Furthermore, atlas-guided approaches are usually highly dependent on registration accuracy. For newborns, registration is particularly challenging due to lower tissue contrast, specifically in the T1w scans. Errors in the process are hence common and improvement of the registration, for example, with spatio-temporal information, anatomical constraints, and surface models, is an active field of research (Ahmad et al., 2019; Bozek et al., 2018; Dubois et al., 2021; Garcia et al., 2018; Kuklisova-Murgasova et al., 2011; Lebenberg et al., 2018; Makropoulos et al., 2014; Shi et al., 2010).

The explicit definition of spatial and intensity features can be avoided by using Convolutional Neural Networks (CNNs). In fact, fast deep-learning methods for semantic segmentation are becoming increasingly popular for infant segmentation (Bui et al., 2019; Dolz et al., 2020; Kumar et al., 2018; Moeskops et al., 2015; Nie et al., 2016; Qamar et al., 2020; Wang et al., 2022, 2023; Zeng & Zheng, 2018; Zeng et al., 2023; Zhang et al., 2015). Applicability of deep-learning approaches, however, is generally restricted to domains where sufficiently large training datasets exist. While there have been several initiatives to collect larger neuroimaging cohorts of newborns and infants in recent years (Bastiani et al., 2019; Howell et al., 2019; Makropoulos et al., 2018; Volkow et al., 2021), their size is still relatively small compared to equivalent cohorts in the adult population. Additionally, accompanying manual labels are sparse, due to high annotation costs (time and money) and non-uniform labeling protocols, limiting the pool for supervised training options further. Considering the newborn cohort, the variability in resolution and head positioning is likely underrepresented in the publicly available datasets, questioning whether a network trained on the available pairs of scans and labels can be robust enough without additional augmentation.

The most widely used solution to artificially increase the training set size, robustness, and generalizability of deep-learning methods has been traditional data augmentation, such as rotation, scaling, or translation (Fig. 1A). In this case, both images and their labelmaps are interpolated to a new random position during training. Interpolation, however, in the native image space requires resampling of the discrete ground truth segmentations, resulting in information loss (e.g., from lossy nearest-neighbor (NN) interpolation) and reduction in accuracy (Henschel et al., 2022).

With the VINN architecture (Henschel et al., 2022), we recently established the first network for resolution-independent deep learning, which effectively circumvents



**Fig. 1.** Spatial augmentations in deep-learning networks: (A) One single resolution-ignorant CNN or resolution-independent voxel size independent network (VINN) can learn to segment multiple resolutions and head positions by training on a diverse dataset. External scale, rotation, and translation augmentation (+ external augmentation, A. bottom) diversifies the existing training samples by resampling the images and the reference segmentation maps. Here, however, lossy interpolation and resulting artefacts, especially from nearest-neighbor interpolation of discrete label maps, may result in a loss of structural details and sub-optimal performance. (B) Our 4-DOF transform module (VINNA) completely avoids interpolation of the images and discrete labels by integrating the interpolation step into the network architecture itself (B. bottom). Rotations, translations, and scalings applied in the first interpolation block are later reversed, assuring segmentations to be in the original image orientation. Furthermore, the explicit transition from the native resolution to a normalized internal resolution facilitates an understanding of the difference between image features (MultiRes blocks with distances measured in voxels) and anatomical features (FixedRes inner blocks with normalized distances).

scaling augmentation and subsequent external resampling, while leveraging information across datasets of varying resolutions. In VINN, the classic fixed-factor integer down- and up-scale transitions, often implemented via pooling operations in UNet-like architectures (Ronneberger et al., 2015), are replaced with a flexible re-scaling for the first and last scale transitions. This network-integrated resolution-normalization allows for segmentation in the native space during both, training and inference. In adults, this approach has been shown to outperform fixed-resolution CNNs as well as resolution-ignorant CNNs trained with external scaling augmentation, and to improve performance both for sub-millimeter and one-millimeter scans.

Since newborn datasets are often acquired at various native resolutions and are particularly subject to partial volume effects, the resolution-normalization feature offers a basis to improve segmentation performance here as

well. As is, VINN only addresses scaling and would still require external augmentations, and hence label interpolation, to address the increased variability of head positions and limited availability of training data for newborns. With our VINNA and its 4-DOF transform module, we now close this gap and propose to shift away from classical data augmentation towards a detail-preserving internal augmentation scheme (Fig. 1B). While avoiding any type of label interpolation, we extend VINN’s network-integrated resolution-normalization with spatial augmentations (i.e., rotation and translation). At the first layer scale transition, the feature maps are hence not only flexibly rescaled, but also randomly transformed to imitate the position variations commonly found in newborns, subsequently increasing the training distribution.

In conclusion, VINNA and its 4-DOF transform module effectively address the challenges associated with newborn segmentation, namely variation in head positions and

resolutions in the context of limited data availability. The four key contributions of VINNA presented in this work are as follows:

- (i) We provide the first publicly available open-source deep-learning pipeline for a combined subcortical segmentation as well as cortical, and white matter parcellation for newborn T1w or T2w MRIs.
- (ii) We introduce a novel augmentation concept, which for the first time moves spatial augmentations into the network (instead of being performed outside). Our experimental results compare various spatial augmentation approaches side-by-side to isolate their effects.
- (iii) We ensure fair comparisons throughout, for example, by fixed dataset splits, retraining methods under equal data and parameter settings, comparing architectures and setups with minimal differences, and quantifying real-world performance.
- (iv) We, further, provide extensive comparison with state-of-the-art deep-learning methods (2D and 3D nnUNet) adapted for newborn segmentation (retrained on the same data) and present an extensive comparison to the publicly available newborn segmentation pipelines iBEAT and infantFS.

The specific application of VINNA to newborns (approximately 24–44 weeks post-menstrual age) will be made available as VINNA4neonates within our open source repository<sup>1</sup> including Docker containers offering easy accessibility for the community.

### 1.1. Related work

While various reliable and sensitive traditional (Fischl et al., 2002; Friston et al., 2007; Jenkinson et al., 2012; Zhang et al., 2001) and fast deep-learning solutions exist (Billot et al., 2020; Chen et al., 2018; Coupé et al., 2020; Henschel et al., 2020, 2022; Huo et al., 2019; Iglesias et al., 2021; Ito et al., 2019; McClure et al., 2019; Mehta et al., 2017; Roy et al., 2019; Sun et al., 2019; Wachinger et al., 2018) for adult whole-brain segmentation, application of these methods to younger ages is hampered by the significant differences in size, MRI contrast profiles, and rapidly changing postnatal anatomy that is challenging to model with static templates.

### 1.2. Traditional tools for infant segmentation

Infant-specific traditional atlas-guided tools (Beare et al., 2016; Chérel et al., 2015; Makropoulos et al., 2018; Prieto

et al., 2019; Shi et al., 2010; Zöllei et al., 2020) are predominantly optimized for a specific age range, resolution, and modality. Further, they differ significantly in the number of segmented classes and structure definitions.

The more recent Infant Brain Extraction and Analysis Toolbox (iBEAT) V2.0 (Wang et al., 2023) is a combination of age-specific CNNs for tissue segmentation, traditional surface generation, and parcellation, based on atlas registration, into 34 regions following the Desikan-Killiany protocol (Desikan et al., 2006). It supports a large age range (0–6 years), and allows segmentation of both, T1w and T2w MRI. While multiple input resolutions are supported, iBEAT internally reorients and resamples each image to a standard format (RAS orientation and 0.8 mm isotropic resolution). Hence, it does not support native resolution segmentation and image interpolation is required to map segmentations back to the original input space. The resampling step is automatically included in the pipeline such that in- and output resolutions are flexible. Furthermore, in its publicly available docker pipeline,<sup>2</sup> segmentation is limited to white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF).

infantFreeSurfer (infantFS) (Zöllei et al., 2020), on the other hand, mimics the FreeSurfer (Fischl, 2012) processing pipeline for adults and processes images from the first 2 years postnatally. It supports anatomical segmentation into 32 classes based on multi-atlas label fusion strategy, including registration to the infantFreeSurfer training data set (de Macedo Rodrigues et al., 2015). The entire pipeline is publicly available<sup>3</sup> and allows processing of T1w images at a resolution of 1.0 mm, where the atlas training data are defined. For newborns, T1w images often suffer from poor tissue contrast due to the underlying myelination process, aggravating accurate registration from the atlases onto individual brains. This age group can therefore be a challenge for infantFS's mono-modality approach.

The dHCP minimal-processing-pipeline (Makropoulos et al., 2018) is an optimized framework for cortical and subcortical volume segmentation, cortical surface extraction, and cortical surface inflation, which has been specifically designed for high-resolution T2w MRIs of newborns (Hughes et al., 2017). Here, an Expectation-Maximization algorithm, including an atlas-based spatial prior term, labels 87 classes based on a modified version of the ALBERTs atlas (Gousias et al., 2012; Makropoulos et al., 2014). The segmentations include subcortical structures, cortical and WM parcellations. Due to the cubic increase in voxel size for high-resolution images, processing times are in the order of hours to days for a single subject. This is a common limitation among traditional methods.

<sup>1</sup> [github.com/Deep-MI/VINNA4neonates](https://github.com/Deep-MI/VINNA4neonates) upon publication

<sup>2</sup> <https://github.com/iBEAT-V2/iBEAT-V2.0-Docker>

<sup>3</sup> <https://surfer.nmr.mgh.harvard.edu/fswiki/infantFS>

### 1.3. Deep-learning for infant segmentation

#### 1.3.1. Newborns

Overall, networks for cortical parcellations and subcortical structure segmentations in newborns are limited. The few existing CNNs support a single modality (T2w), fixed resolution, and segment a single (Rajchl et al., 2017) or eight (Moeskops et al., 2015) tissue classes. One recent exception is the deep-learning based neuroimaging pipeline by Shen et al. (2023), which is trained with the dHCP data. Here, the authors propose a 3D multi-task deep learning model with a U-Net like architecture to segment structural T1w and T2w images on both thin and thick sliced images. Unfortunately, the network follows a fixed-resolution scheme (0.8 mm), it does not support native segmentation across resolutions commonly encountered in newborn cohorts, and it is not readily available online.

#### 1.3.2. Isointense phase

The vast majority of deep-learning tools focus on processing of images at the isointense phase around 6 months after birth (Ding et al., 2022; Dolz et al., 2020; Kumar et al., 2018; Nie et al., 2016; Pasban et al., 2020; Qamar et al., 2020; Zeng & Zheng, 2018; Zeng et al., 2023; Zhang et al., 2015). Via the iSeg-challenge (Sun et al., 2021; Wang et al., 2019); data for training and validation are conveniently available, partly explaining this predominance. While many interesting architectural solutions have arisen, the main focus of the works is the effective combination of information from both T1w and T2w images to generate a broad segmentation into CSF, GM, and WM. This modality combination is specifically important in the isointense phase, which is characterized by changes in the myelination strongly effecting the appearance of the recorded MRIs (Gui et al., 2012; Wang et al., 2015; Weisenfeld & Warfield, 2009). The inversion of the WM-GM signal results in extremely low tissue contrast. The newborn cohort, on the other hand, demonstrates good contrast between GM and WM, specifically on the T2w images. While the age difference is small, the two cohorts as well as the associated challenges are distinct and networks trained on the one cannot easily be applied to the other. Subsequently, neither resolution-independence nor the stronger variation of head positions is specifically accounted for in network solutions for the isointense phase.

#### 1.3.3. Cross-age generalizability

To address generalizability across different age groups, recent research has suggested optimized training strategies for neonates, such as multi-task learning of tissue

segmentation and geodesic distances (Bui et al., 2019) or the use of inductive biases in the form of pre-trained weights (i.e., fine-tuning to the target domain) (Wang et al., 2022). Both approaches improve segmentation accuracy, but they are still limited in their generalizability. They require retraining and hence a sufficient amount of labeled data; additionally, they rely on private datasets, limiting their reproducibility. Recently, a contrast agnostic segmentation via synthetic images, originally proposed for adult brain segmentation (Billot et al., 2020; Iglesias et al., 2021), has also been adopted for infant segmentation (Shang et al., 2022). Unfortunately, the output resolution is fixed for the network, and native resolution segmentations are not supported. Furthermore, while the model was able to generalize across a broader age range, the synthetic images still differ considerably from real data and the network, therefore they underperformed compared to age-specific models trained on existing MRIs. It should be noted that the authors did not aim to generate realistic images but rather a better segmentation tool.

### 1.4. Resolution-independence and position transforms in deep learning

A general resolution-ignorant framework addressing position transforms via external augmentations is nnUNet (Isensee et al., 2021). This network has successfully been applied for a variety of segmentation tasks due to its inherent ability to construct optimal parameter settings based on the input data itself. It provides different network set-ups (2D, 3D, and a cascaded 3D approach for large images) as well as a number of external image augmentations, including random rotation, scaling, mirroring, and gamma transformation. Interestingly, while the trained network also follows a fixed-resolution scheme, pre- and post-processing automatically resamples between original image and network resolution. While native resolution segmentation is not supported, in- and output resolutions are not fixed and the method is therefore a valid alternative to our VINNA. Both, the 2D and 3D nnUNet with external augmentation (exA) therefore serve as a state-of-the-art baseline for the newborn segmentation task.

A siamese network for semi-supervised training of a network to become equivariant to elastic transformation has been proposed in a single-resolution setting (Bortsova et al., 2019). A dedicated loss function assures that segmentations are consistent under a given class of transformation applied first to the image, and second to the output. The approach therefore relies on external augmentation and applies the transformation in the image space (before and after a UNet). The proposed VINNA, on the other hand, is fundamentally different. It shifts this

step into the network itself, hence creating an internal augmentation. Overall, the approach by [Bortsova et al. \(2019\)](#) does therefore assure consistency across transformations in the labeling space, while VINNA targets spatial consistency of the feature maps.

In spatial transformers ([Jaderberg et al., 2015](#)), transformations attempt to harmonize or re-orient the image into a better position. To this end, an affine transformation is implicitly learned via a dedicated localisation network. Subsequent application of the calculated coordinate grid resamples the source feature maps via bi-linear interpolation to the new position. While our approach shares grid calculation and interpolation within the network with spatial transformers, our internal augmentation approach is inherently different. First, spatial transformers do not diversify or augment feature maps, but rather try to reach a harmonized position with respect to the data seen during training. External augmentations are still necessary to expose the network to a wide data variety and approximate equivariance. Second, instead of a localization network, we directly determine the sampling-grid based on a transformation matrix, which allows for an explicit integration of knowledge about the image, such as the resolution. As a result, computational complexity is reduced while achieving the desired position diversification and resolution-independence.

## 2. MATERIAL AND METHODS

### 2.1. Datasets

As the largest publicly available newborn dataset with intensity images, accompanying subcortical segmentations as well as cortical and WM parcellations at the time of our experiments, we randomly assign participants from the dHCP cohort with corresponding T1w and T2w MRIs to the training, testing, and validation sets, while ensuring equal distribution of age and gender. Additionally, the Melbourne Children's Regional Infant Brain (M-CRIB) atlas cohort serves as an independent testing set for external validation of the final method. Written informed consent to participate in this study was provided by the participants' legal guardian or next of kin in accordance with the Institutional Review Board. Complete ethic statements are available at the respective study webpages.

#### 2.1.1. dHCP

The developing Human Connectome Project ([Makropoulos et al., 2018](#)) includes T1w and T2w MRIs of newborns imaged without sedation on a 3 T Philips Achieva scanner. It provides 0.5 mm isotropic de-faced scans of individuals

imaged postnatally between 24 to 45 weeks post-menstrual age. Imaging data for 578 participants with matching T2w and T1w are selected. The original images were acquired in sagittal and axial slice stacks with in-plane resolution of 0.8 mm  $\times$  0.8 mm and 1.6 mm slices overlapped by 0.8 mm. Motion correction and super-resolution reconstruction techniques ([Cordero-Grande et al., 2018](#); [Kuklisova-Murgasova et al., 2012](#)) created isotropic volumes of resolution 0.5 mm. All T1w scans follow the same inversion recovery multi-slice fast spin-echo protocol with TR 4.795 s, TE 8.7 ms, TI 1.740 s, and SENSE factor 2.27 (axial) / 2.56 (sagittal). The parameters for the T2w scans are TR 12 s, TE 156 ms, and SENSE factor 2.11 (axial) / 2.66 (sagittal). The full dataset is available online.<sup>4</sup> In the present study, 318 quality-checked images are used for network training and 90 for validation. A total of 170 images are used in the final test set.

Even though the dHCP follows a well-defined protocol, standardization of positioning during scanning is still a challenge. As shown in [Figure 2](#), inter-subject head position diversity in the newborn cohort is larger than an equally standardized adult cohort (HCP).

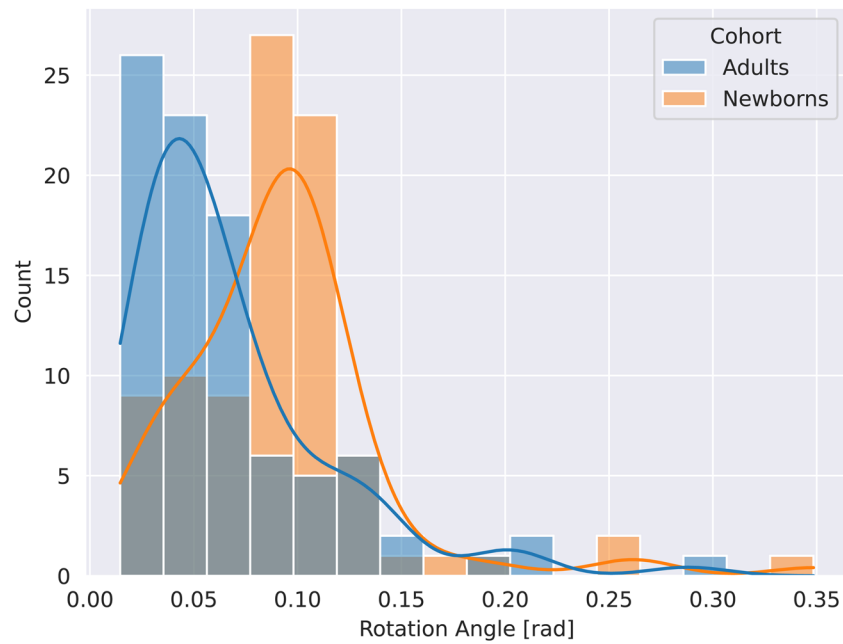
#### 2.1.2. M-CRIB

The Melbourne Children's Regional Infant Brain (M-CRIB) atlas ([Alexander et al., 2019](#)) is constructed from 10 T2w MRI and corresponding manual segmentations of healthy term-born neonates (four females, six males) with post-menstrual age-at-scan between 40.29–43.00 weeks. The atlas comprises 94 neonatal brain regions compatible with the widely-used Desikan-Killiany-Tourville adult cortical atlas ([Klein & Tourville, 2012](#)). The T2w MRIs scanning protocols include the usage of a transverse T2 restore turbo spin echo sequence with 1.0 mm axial slices, a TR of 8.910 s, TE of 152 ms, flip angle of 120 degrees, Field Of View of 192 mm  $\times$  192 mm, and in-plane resolution of 1 mm (zero-filled interpolated to 0.5 mm  $\times$  0.5 mm  $\times$  1 mm). The T2w images are bias-corrected, skull-stripped, and resampled to 0.63 mm  $\times$  0.63 mm isotropic voxels. All 10 participants are used as an independent testing set for our external validation experiments.

### 2.2. Generation of reference segmentation with the dHCP-minimal-processing-pipeline

To imitate various resolutions and create the desired reference segmentations for training, we processed all raw dHCP MRIs with the dHCP-minimal-processing-pipeline ([Makropoulos et al., 2018](#)) at 1.0 mm, 0.8 mm, and

<sup>4</sup> <https://data.developingconnectome.org/>



**Fig. 2.** Comparison of variation in head position in newborns (orange) and adults (blue). Newborns show greater variation with respect to rotation angles. Rotation transformation is based on alignment of each individual subject to their midspace.  $N = 90$  for both cohorts.

0.5 mm. The structure definitions follow the ALBERTs atlas (Gousias et al., 2012) with the subdivision of the WM and cortex proposed by Makropoulos et al. (2014), resulting in a total of 87 structures (3 background labels, 20 subcortical regions, 32 cortical parcels, 32 WM parcels). We further lateralized the intracranial background based on the average Euclidean distance to neighboring labels, resulting in a final count of 88 labels. We provide a list of all segmentation labels used for training in the Appendix (see Appendix Table B). As the dHCP cohort includes both, T2w and a co-registered T1w MRIs, we trained dedicated networks for each modality. A manual quality check for all selected scans assured good overlap after the registration. Note that the dHCP-minimal-processing-pipeline relies on the original T2w images to create its segmentations, which are generally of higher quality in this collection.

### 2.3. Traditional infant segmentation tools

To evaluate VINNA against state-of-the-art traditional segmentation methods, we further process the testing set with the docker version of the iBEAT V2.0 pipeline (Wang et al., 2023) and infantFS (Zöllei et al., 2020).

#### 2.3.1. iBEAT

The iBEAT V2.0 pipeline (Wang et al., 2023) combines both traditional and deep-learning models to create tis-

sue segmentations into three classes (GM, WM, and CSF), surface models, and cortical parcellations of the pial surface into 34 regions based on the Desikan-Killiany protocol (Desikan et al., 2006). For tissue segmentation, iBEAT uses seven age-specific CNNs trained on data for the representative age group ( $\leq 1, 3, 6, 9, 12, 18,$  and  $24+$  months of age). Neither the source code nor the training data and labels are publicly available. Hence, retraining of the models is not possible and comparisons are limited to the iBEAT pipeline output as is. For processing with iBEAT, submissions via a webserver<sup>5</sup> or processing with a docker image<sup>6</sup> are possible. The docker version does not currently support the cortical parcellations of the surface models. Due to the large number of participants, privacy concerns, and longer processing times when submitting via the webserver, we decided to use the docker version to process the T2w images of the testing set at the original 0.5 mm resolution. The resulting 3-label tissue segmentations form the basis for comparison to the other tools in this paper.

#### 2.3.2. infantFS

To allow comparison of segmentation performance to VINNA, all available T1w images from the dHCP testing set are processed with infantFS with default settings. The

<sup>5</sup> <https://ibeat.wildapricot.org/>

<sup>6</sup> <https://github.com/iBEAT-V2/iBEAT-V2.0-Docker>

neuroimaging pipeline infantFS creates surface models, anatomical segmentations, and cortical parcellations based on the Desikan-Killiany-Tourville atlas (Klein & Tourville, 2012) for 0- to 2-year-old infants akin to the version for adults (FreeSurfer (Fischl, 2012)). The tool runs on T1w MRIs at a resolution of 1.0 mm (non-conforming images are resampled). infantFS relies on a registration-based multi-atlas label fusion strategy and returns an anatomical segmentation into 32 classes, including two labels for GM and WM.

### 2.3.3. Label harmonization

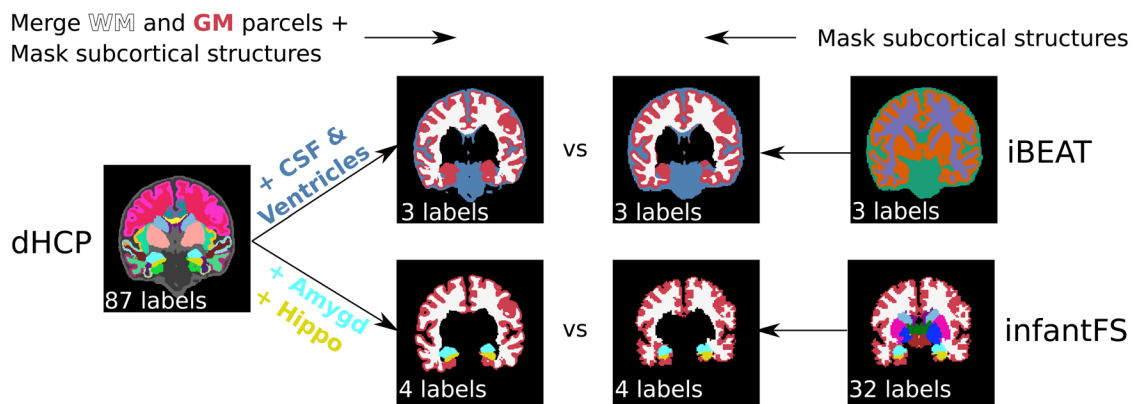
As the dHCP-ALBERTs atlas differs from the resulting segmentations of both iBEAT and infantFS, we merge, remove, and mask classes to reach an approximate consensus across predictions. Figure 3 shows the merging protocol on a representative participant with the original and mapped ground truth dHCP labels (left side) together with iBEAT (top right side) and infantFS (bottom right side). First, the 32 cortical and WM parcels from the dHCP ground truth segmentation are reduced to two labels (cortex and WM) (top left in Fig. 3). For iBEAT, the WM additionally includes the corpus callosum while GM also encompasses the hippocampus and amygdala. The CSF label corresponds to the union of lateral-ventricles and CSF in the dHCP-ALBERTs atlas. In the dHCP ground truth, these labels are consequently merged to create the final three classes (GM, WM, and CSF; top second to left image). All other subcortical structures without a single possible assignment to GM, WM, or CSF are masked in

the iBEAT prediction using the dHCP ground truth (Fig. 3, top right two images). For infantFS, the hippocampus and amygdala label remain, while individual cortex and WM parcels of the dHCP ground truth are merged (Fig. 3, bottom left images). Hence, in the infantFS predictions the following four labels remain: cortex, WM, hippocampus, and amygdala (Fig. 3, bottom two images to the right).

## 2.4. Network architectures

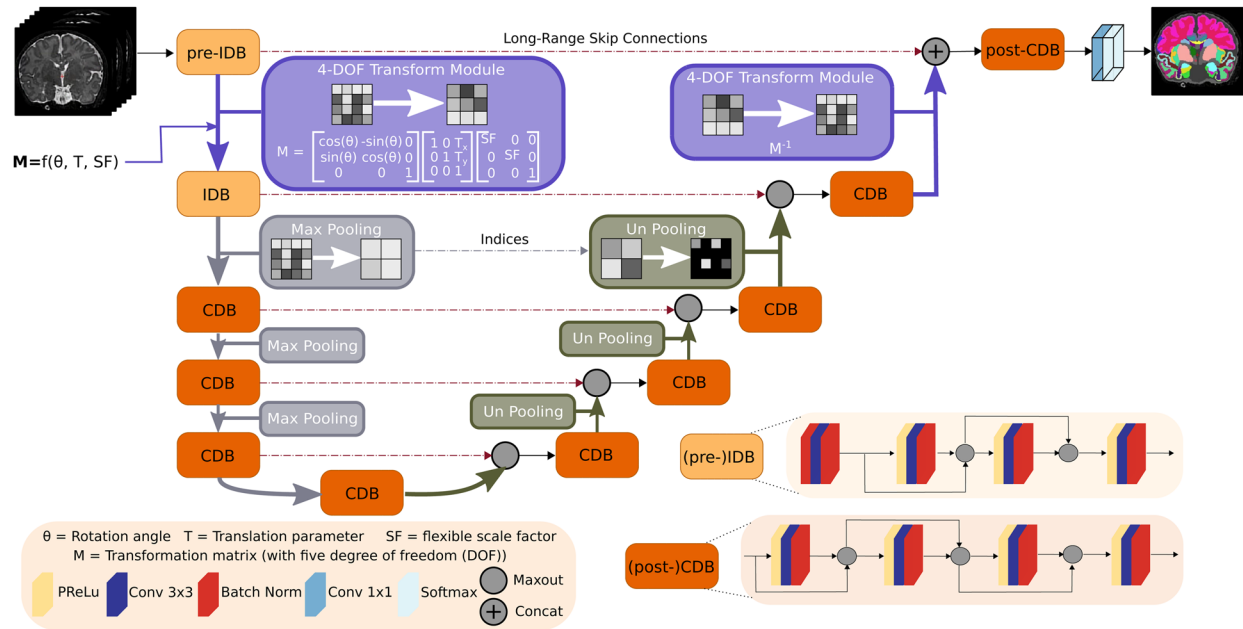
### 2.4.1. Macro architecture

Figure 4 shows the macro architecture for VINNA. While the proposed 4-DOF transform module (purple) can, in theory, be included in any UNet-like architecture, we use the same basic setup for all trained models to assure maximum comparability (i.e., same number of parameters, same kernel sizes, etc.). The parameter-equal CNN is referred to as CNN\*. CNN\*, VINN, and VINNA, all contain an encoder and decoder consisting of five competitive dense blocks, respectively, which are separated by a bottleneck layer. In the encoder, maxpooling operations rescale the feature maps at each level by one half between the blocks using a pooling kernel of size  $2 \times 2$  and stride 2. In contrast, index-unpooling doubles the feature map size in the decoder. Skip connections between the blocks at each level allow the gradient to flow efficiently. In CNN\* (Henschel et al., 2022), pooling and unpooling operations transition between all levels (i.e., the purple block in Figure 4 is substituted with the gray maxpooling/unpooling operation). In VINN (Henschel et al., 2022), the first



**Fig. 3.** Harmonization of inconsistent label protocols between iBEAT, infantFS, and dHCP. Reduction of the original dHCP segmentation from 88 labels (left) by first merging all cortical parcels to cortex, WM parcels to WM and removal of subcortical structures, is followed by addition of CSF and ventricles (top) or amygdala and hippocampus (bottom) for comparison to iBEAT (3 labels: CSF (blue), WM (white), and GM (red), top second from left top) or infantFS (4 labels: WM (white), cortex (red), hippocampus (yellow), and amygdala (blue), bottom second from left), respectively. For iBEAT (right top), all GM and WM are modified using the dHCP segmentation for the subcortical structures, resulting in three labels for the final mapped version (3 labels, top second from right). For infantFS (right bottom), all structures except WM, cortex, hippocampus, and amygdala are removed (4 labels, bottom second from right).





**Fig. 4.** Network-integrated position variation and scaling normalization in VINNA. Flexible transitions between resolutions and head positions become possible by replacing (un)pooling with our network-integrated 4-DOF transform module (purple) after the first input dense block in the encoder (pre-IDB) and before the last competitive dense block in the decoder (post-CDB). A transformation matrix composed of rotation angle  $\theta$ , translation parameter  $T$ , and the scaling factor  $SF$  defines the feature alterations. Scale transitions between the other competitive dense blocks (CDB) remain standard MaxPool and UnPool operations. Each CDB is composed of four sequences of parametric rectified linear unit (PReLU), convolution (Conv), and batch normalization (BN). In the first two encoder blocks ((pre-)IDB), the PReLU is replaced with a BN to normalize the inputs.

layer pooling and unpooling operation is replaced with a resolution-normalization module. This network-integrated flexible interpolation step allows transitions between resolutions without restrictions to pre-defined fixed voxel sizes, both during training and inference. Hence, images can be processed at their native resolution without prior resampling. Similar to spatial transformers (Jaderberg et al., 2015), the interpolation-based transition is divided into two parts: (i) calculation of the sampling coordinates (*grid generator*) and (ii) interpolation operation (*sampler*) to retrieve the spatially transferred output feature maps. Here, the sampling calculation relies only on the scale factor  $SF$ : the quotient of the resolution information of the inner normalized scale  $Res_{inner}$ , a tune-able hyperparameter set to 0.8 mm throughout our experiments, and the input image  $Res_{native}$ . The addition of parameter  $\alpha$  sampled from a Gaussian distribution with parameters sigma = 0.1 and mean = 0 slightly augments the scale factor  $SF$  ( $SF = Res_{inner} / Res_{native} + \alpha$ ), introduces small resolution variations to the sampling, and increases the robustness of the latent space interpolation. Specifically, the presence of alpha allows for augmentations at the actual anatomical rather than the voxel size as the normalization of the resolution inside VINN disentangles perceived voxel versus actual structure size differences.

#### 2.4.2. Network-integrated 4-DOF transform module

In VINNA, the transition is implemented via the new network-integrated 4-DOF transform module shown in purple in Figure 4. Here, the sampling coordinate calculation is based on a transformation matrix  $M \in \mathbb{R}^{3 \times 3}$  with four degrees of freedom, encoding not only scaling, but also in-plane rotation, and translation. Parameters for the rotation angle  $\theta \in \mathbb{R}$  and translation  $T \in \mathbb{R}^2$  are randomly sampled on the fly during training. The scale factor  $SF$  is calculated as in VINNs resolution-normalization module, that is, by dividing the inner normalized scale by the image resolution with augmentation by the parameter  $\alpha$ . In the first transition step (Fig. 4, pre-IDB to IDB), the affine per-channel mapping  $M: \mathcal{U} \rightarrow \mathcal{V}$  samples the input feature maps  $U \in \mathbb{R}^{H_{native} \times W_{native}}$  to the output feature maps  $V \in \mathbb{R}^{H_{inner} \times W_{inner}}$ . In the final transition step (Fig. 4, competitive dense block (CDB) to post-CDB), this spatial transformation is reversed by using the inverse transformation matrix  $M^{-1}: \mathcal{V} \rightarrow \mathcal{U}$ . The interpolation itself is performed by applying a sampling kernel to the input map  $U$  to retrieve the value at a particular pixel in the output map  $V$ . The sampling is identical for each channel, hence conserving spatial consistency.

### 2.4.3. Network blocks

**2.4.3.1. Competitive dense block (CDB) design.** In VINNA, a CDB is formed by repetitions of the basic composite function consisting of a probabilistic rectified linear unit (pReLU) activation function, a  $3 \times 3$  convolution, and a batch-normalization (BN). Feature competition within the block is achieved by using maxout (Goodfellow et al., 2013) instead of concatenations (Jégou et al., 2017) in the local skip connections. The maxout operation requires normalized inputs and is therefore always performed after the BN (see position of maxout in CDB design in Fig. 4).

**2.4.3.2. Input competitive dense block (IDB) design.** In contrast to the described CDB, the first two network blocks follow a different order of operation. Here, the raw inputs are normalized by first passing them through a BN-Conv-BN combination before adhering to the original composite function scheme (Conv-BN-pReLU) (see Fig. 4, IDB).

**2.4.3.3. Pre-IDB.** The first encoder block in VINNA called pre-IDB (see Fig. 4) transfers image intensity information from the native image to the latent space and encodes voxel size and subject-space-dependent information before the internal interpolation step. The composite function scheme is identical to the IDB, and the added prefix simply allows differentiation of the block placements.

**2.4.3.4. Post-CDB.** Akin to the pre-IDB, an additional CDB block in the decoder merges the non-interpolated feature information returned from the pre-IDB skip connection and the upsampled feature maps from the network-integrated 4-DOF transform modules. A concatenation operation combines both feature maps, before passing them to a standard CDB block (see Fig. 4, (post-)CDB). After the final  $1 \times 1$  convolution, a softmax operation returns the desired class probabilities.

## 2.5. Loss function

All networks are trained with a weighted composite loss function of logistic loss and Dice loss (Roy et al., 2017) combined with the high-resolution specific weighting from VINN (Henschel et al., 2022). In short, erosion and dilation of the cortex labels creates a binary mask of the outer cortex, small WM strands, and deep sulci. Wrong predictions in these areas result in a higher loss, hence guiding the network to focus on areas particularly affected by partial volume effect (PVE). The original publications' ablation experiments evaluated the impact of the different function elements: the logistic loss and Dice loss combination improves overall segmentation performance

(Roy et al., 2017), while the high-resolution weighting leads to higher segmentation accuracy on the cortical parcels (Henschel et al., 2022).

If we consider  $p_{l,i}(x)$  as the estimated probability of pixel  $i$  that belongs to class  $l,y$ , as the corresponding ground truth probability, and  $\omega_i$  as the associated weight given to the pixel  $i$  based the loss function can be formulated as

$$\mathcal{L} = - \underbrace{\sum_{l,i} \omega_i y_{l,i} \log p_{l,i}(x)}_{\text{Logistic loss}} - \underbrace{\sum_l \frac{2 \sum_i p_{l,i}(x) y_{l,i}}{\sum_i p_{l,i}(x) + \sum_i y_{l,i}}}_{\text{Soft Dice loss}} \quad (1)$$

with  $\omega_i = \omega_{\text{median freq.}} + \omega_{\text{gradient}} + \omega_{\text{GM}} + \omega_{\text{WM/Sulci}}$ .

Here,  $\omega_{\text{median freq.}}$  represents median frequency balancing addressing the class imbalance and  $\omega_{\text{gradient}}$  boundary refinement through a 2D gradient vector (Roy et al., 2017), while  $\omega_{\text{GM}}$  and  $\omega_{\text{WM/Sulci}}$  assign higher weights to PVE-affected areas (Henschel et al., 2022).

## 2.6. View aggregation

In order to account for the inherent 3D geometry of the brain, we adopt the 2.5D view aggregation scheme from (Henschel et al., 2020, 2022) for CNN\*, VINN, and VINNA. In short, we train one network instance per anatomical plane and calculate a weighted average of the resulting softmax probability maps. The weight of the sagittal predictions is reduced by one half compared to the other two views to account for the missing lateralization in the sagittal view. In this plane, the network predicts 23 instead of 88 classes.

## 2.7. Augmentations

### 2.7.1. External augmentation (exA)

The current state-of-the-art approach to introduce robustness to position changes into neural networks is extensive external augmentation (see Fig. 1B). Therefore, we contrast our proposed network-integrated 4-DOF transform module against this approach. We use random transforms with rotation parameters sampled from a uniform distribution of the predefined range  $-180^\circ$  to  $180^\circ$  and translation by 0 to 15 px to augment images during the training phase and interpolate linearly. For CNN\* and nnUNet, augmentation also includes sampling of scaling parameters from a uniform distribution of the predefined range 0.8 to 1.15. VINN's resolution-normalization module makes this step obsolete. Every minibatch hence consists of a potentially transformed MRI (using bi-linear interpolation) and a corresponding label map (using NN sampling). By exposing a network to a large variety of

possible image positions during training, orientation-robustness can be achieved. All external augmentation routines are implemented using torchIO (Pérez-García et al., 2021).

### 2.7.2. Image intensity augmentation

To allow generalization outside of the dHCP cohort, we apply a number of intensity or texture augmentations on the fly to the training batch, namely bias field changes, random gamma alterations, ghosting, spiking, blurring, and Gaussian noise. Each batch sampled from the original data is transformed by any of the operations above with a probability of 0.4. As before, all augmentations are implemented using torchIO.

## 2.8. Evaluation metrics

We use the Dice Similarity Coefficient (DSC) (Dice, 1945; Sørensen, 1948) and Average Surface Distance (ASD) to compare different network architectures and modifications against each other, and to estimate similarity of the predictions with a number of previously unseen scans. Both are standard metrics to evaluate segmentation performance. We establish improvements by statistical testing (Wilcoxon signed-rank test (Wilcoxon, 1945) after Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) for multiple testing) referred to as “corrected p” throughout the paper.

## 2.9. Training setup for all models

### 2.9.1. Training dataset

For training, we select 318 representative participants from the dHCP cohort. Resolutions are equally represented with 106 MRIs at 1.0 mm, 0.8 mm, and 0.5 mm, respectively. Empty slices are filtered from the volumes, leaving on average 137 single view planes per subject and a total training size of at least 20k images per network. The original nnUNet does not filter the volumes. The 3D version uses 3D patches instead of 2D slices. The parameters are automatically determined by nnUNet to guarantee an ideal set-up for the given segmentation task. Otherwise, we train all directly compared networks (CNN\*, VINN, VINNA, 2D nnUNet, and 3D nnUNet) under the same conditions.

### 2.9.2. Training parameters

We implement and train independent models to convergence (early stopping) for the coronal, axial, and sagittal planes with PyTorch (Paszke et al., 2017), using one

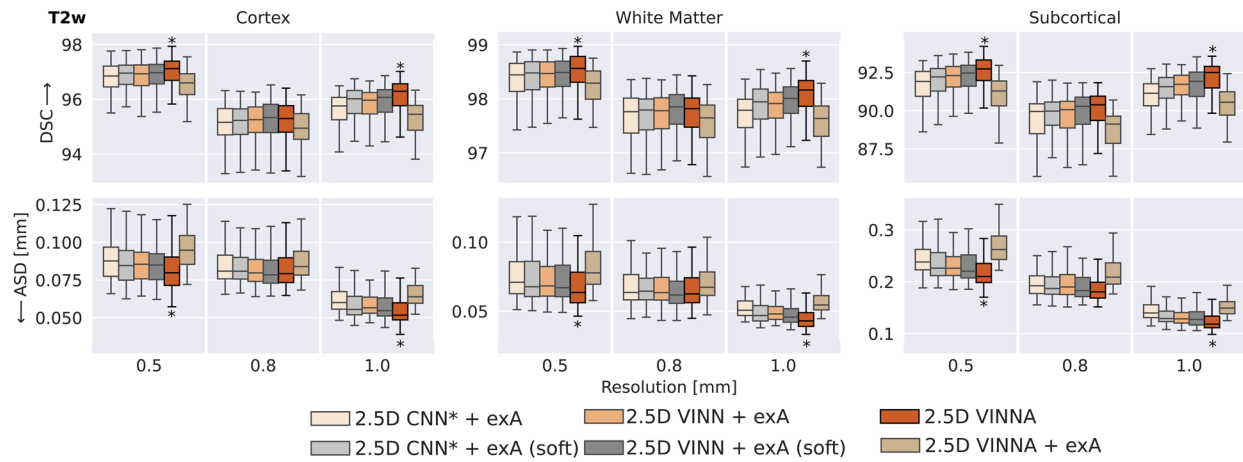
NVIDIA V100 GPU with 32GB RAM. During training, the modified Adam optimizer (Loshchilov & Hutter, 2019) is used with a learning rate of 0.001. Using a cosine annealing schedule (Loshchilov & Hutter, 2017) with warm restarts, the learning rate is adapted after initially 10 epochs. The epoch offset is subsequently increased by a factor of two. The momentum parameter is fixed at 0.95 to compensate for the relatively small mini batch size of 16 images for CNN\*, VINN, and VINNA. For nnUNet, the optimal batch-size is automatically determined (Isensee et al., 2021). For the given segmentation problem, the 2D nnUNet uses a batch size of 128 while the 3D version relies on a smaller batch size of 128. To ensure a fair comparison, all networks (CNN\*, VINN, VINNA, 2D nnUNet, and 3D nnUNet) have been trained under equal hardware and hyper-parameter settings otherwise. A comparing table is available in the Appendix (Section A.4).

## 3. RESULTS

We group the presentation of results into three blocks: 1. ablative architecture improvements to determine the best performing module for orientation and position transformation (Section 3.1), 2. performance analysis to comprehensively characterize the advantages of VINNA with respect to state-of-the-art traditional atlas- and deep-learning-based methods (Section 3.2), and 3. external validation on M-CRIB (Alexander et al., 2019) to assess generalizability and performance with respect to manual labels (Section 3.3). Following best practice in data-science, we utilize completely separate datasets during the evaluations: the validation set for Section 3.1 (Appendix Table A: Validation), and various test sets for Sections 3.2 and 3.3 (Appendix Table A: Testing). This avoids data leakage and ensures that training, method design decisions, and final testing do not influence each other, which could otherwise lead to overly optimistic results (overfitting).

### 3.1. External augmentation versus network-integrated 4-DOF transform module in VINNA

As high variances with respect to head orientations and spatial resolutions are common in newborns and are likely to be underrepresented in the limited available data cohorts, we first compare multiple approaches for extension of the training distribution for accurate (sub) millimeter newborn whole-brain segmentation. Traditionally, external data augmentation (exA), such as scaling, rotation, or translation, addresses this problem by interpolating both, the image and label maps, to a new, random position during training. Due to the discrete nature of the label maps, lossy NN interpolation cannot be avoided, unless the transformations are applied to the



**Fig. 5.** Comparison of approaches for external and internal spatial augmentation on T2w images: Our VINNA method—with the 4-DOF transform module—(fifth box in the group) outperforms both state-of-the-art approaches with external Augmentation (CNN\* + exA and VINN + exA, first and third box) in Dice Similarity Coefficient (DSC, upper row) and average surface distance (ASD, lower row). This performance advantage is significant (corrected  $p < 10^{-9}$ ) and consistent across resolutions and structures. Combining VINNA with external augmentation (VINNA + exA, last box) reduces performance. When avoiding nearest-neighbor label interpolation by applying exA on one-hot encoded logits (+exA (soft)) instead of the label maps, performance improves for both CNN\* and VINN (second box and fourth box, respectively). VINNA with the internal 4-DOF module outperforms the soft-loss methods on 1.0 and 0.5 mm by a significant margin (corrected  $p < 10^{-9}$ , indicated with\*). exA: external Augmentation.

one-hot-encoded logits (soft-loss). We therefore evaluate both, the traditional exA and soft-loss implementation (referred to as exA (soft)). Our new 4-DOF transform module in VINNA internally emulates possible head transformations and acts directly on the encoded feature maps. To evaluate effectiveness of the exA versus VINNA, we compare VINNA with parameter-identical CNN\*, VINN, and VINNA equipped with exA. Each subsequent improvement in segmentation performance is confirmed by statistical testing (corrected  $p < 0.05$ ).

In Figure 5, we compare the model performance of six approaches: CNN\* with exA and exA (soft), (Section 2.4, CNN\* + exA, left box; CNN\* + exA (soft), second box from the left), VINN with exA and exA (soft) (Section 2.4, VINN + exA, third box from the left; VINN + exA (soft), fourth box from the left), VINNA with the new 4-DOF transform module (VINNA, second box from the right), and finally VINNA with exA (VINNA + exA, right box). The analysis of the DSC (top) and ASD (bottom) is grouped for three groups of structures (cortex averages 32 labels, WM averages 32 labels, and subcortical structures average 20 labels) and three resolutions (from left to right 0.5 mm, 0.8 mm, and 1.0 mm). We present performance for T2w MRIs, but we found the same ranking for T1w MRIs.<sup>7</sup>

Looking at the T2w segmentation and focusing on the different resolutions, the differences between the approaches are largest on the subcortical structures. The

CNN\* with exA reaches an average DSC of 91.91, 89.95, and 91.14 and an ASD of 0.238 mm, 0.192 mm, and 0.193 mm for input data of 0.5 mm, 0.8 mm, and 1.0 mm, respectively. The slight reduction in performance for 0.8 mm resolution consistently occurs for all evaluated models and is probably caused by the necessary image resampling from the original resolution of 0.5 mm and subsequent reprocessing with the dHCP-pipeline (Section 2.2). Interpolation from 0.5 mm to 1.0 mm results in a well-aligned grid due to the even division by factor 2 (8 voxels get averaged into a single larger voxel). Resampling to 0.8 mm, on the other hand, requires an uneven interpolation grid with weighted averages and original voxels that contribute to multiple larger voxels. This more challenging setting could result in the slightly reduced segmentation performance. Optimization of the architecture design towards multi-resolution (VINN, Section 2.4) leads to significant improvement in the DSC and ASD across the cortical, WM, and subcortical structures (Fig. 5, VINN + exA). Particularly, the subcortical segmentations are improved by around 0.5%. Importantly, the internal 4-DOF transform module (VINNA), which avoids label interpolation all together, further reduces the error by one half and increases segmentation performance significantly compared to both CNN\* and VINN with exA. This effect is consistent across all structures and resolutions. Specifically, the ASD at the high-resolution benefits from the new module. Here, performance can be improved by 4.47% on the cortex, 5.19% on the WM, and 5.89% on the subcortical structures. For the lower

<sup>7</sup> To reduce redundancy, we relegate that analysis to the Appendix (Section A.1).

resolution, the improvement on the cortex and WM is slightly lower (around 2%) while the subcortical structures benefit from the 4-DOF transform module similarly to the 0.5 mm resolution experiments.

The soft-loss, implementing augmentations for the label maps via linear interpolation of the one-hot encoded feature maps, does improve performance compared to the nearest-neighbor based external augmentation for both CNN\* and VINN (see CNN\*/VINN + exA versus CNN\*/VINN + exA (soft)). The subcortical structures improve the most (0.27% DSC and 5.1% ASD for CNN\* + exA versus + exA (soft); 0.21% and 2.35% ASD for VINN + exA versus + exA (soft)), followed by the cortex (0.15% DSC and 3.7% ASD; 0.07% DSC and 1.8% ASD). The improved performance without nearest-neighbor interpolation further strengthens our recommendation to avoid this type of label interpolation wherever possible. The proposed VINNA architecture outperforms VINN + exA (soft) on the 0.5 and 1.0 mm resolution by 0.17% and 0.35% DSC and 5.27% and 6.23% ASD, respectively. The subcortical structures benefit the most with an average improvement by 0.34% DSC and 4.39% ASD, respectively. On the 0.8 mm resolution, no significant difference is detectable between the two approaches.

Overall, VINNA reaches the highest DSC and lowest ASD for the cortical structures (96.24, 0.079 mm), WM (98.18, 0.063 mm), and subcortical structures (91.87, 0.190 mm) across all resolutions. The addition of exA to the framework (VINNA + exA, right box in each plot) again reduces performance. The DSC drops by 0.4%, 0.28%, and 1.15% on the cortex, WM and subcortical structures on average, while the ASD worsens by 4.75%, 3.92%, and 4.84%. Overall, results with VINNA are significantly better on 0.5 mm and 1.0 mm compared to all ablations on the validation set (corrected  $p < 10^{-6}$ ).

### 3.2. Comparison to state-of-the-art neonate segmentation tools

To evaluate how VINNA compares to state-of-the-art neonate MRI segmentation tools, namely nnUNet (3D and 2D), iBEAT, and infantFS, we take a closer look at the DSC and ASD on the testing sets.

#### 3.2.1. Comparison of deep-learning networks

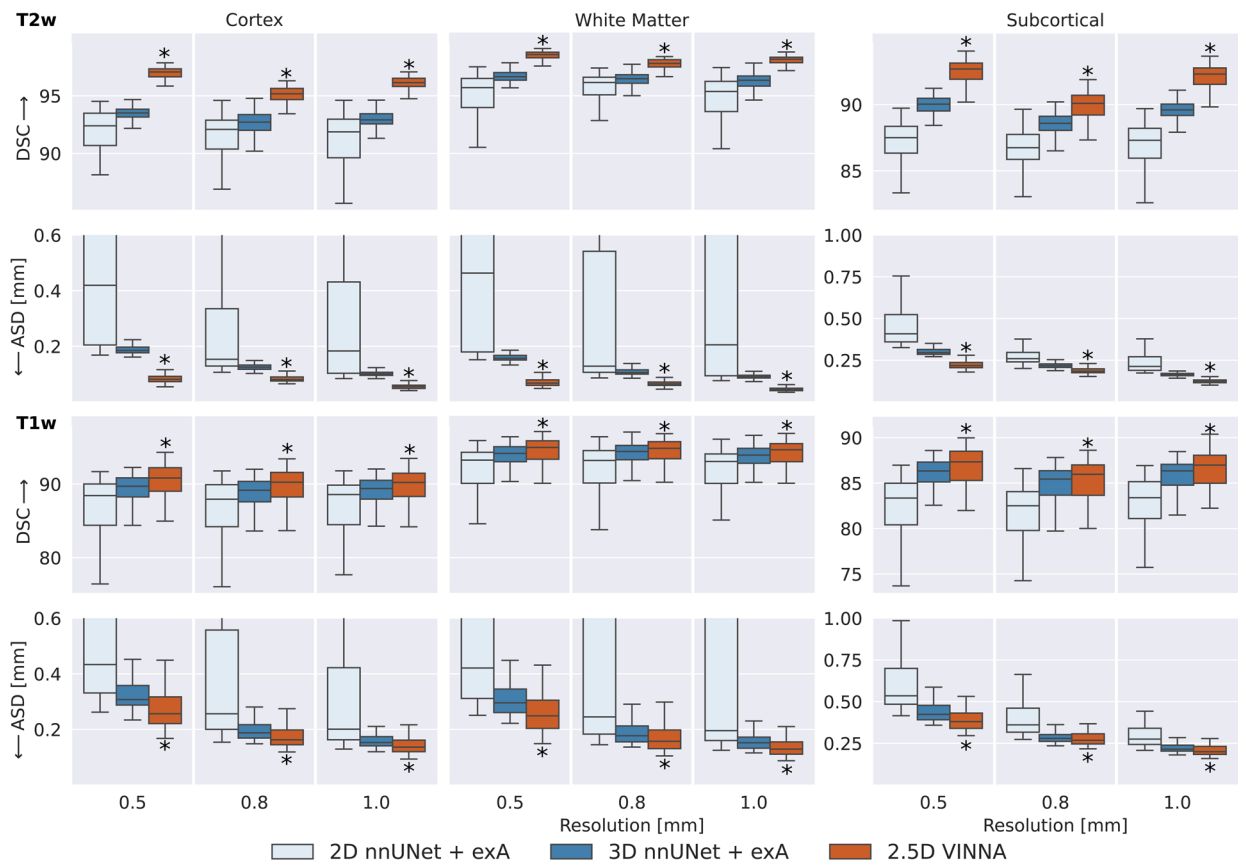
Figure 6 shows a detailed comparison of three different deep-learning based methods for neonate segmentation across modalities (T2w top, T1w bottom) and resolutions. All models are trained under equal parameter and dataset settings. Comparing performance between the two modalities shows that all models perform better on the T2w (top) than the T1w MRIs (bottom) across all struc-

tures. With VINNA (right box in each plot), the reduction is similar across all resolutions, with an average difference in DSC of 5.85, 3.47, and 5.16 on the cortex, WM, and subcortical structures. The ASD is, on average, improved by 0.09 mm when predicting on the T2w instead of T1w inputs. The nnUNet framework in 2D (left box) and 3D (second from left) has less improvement on the T2w images with an average difference between T1w and T2w of 3.75, 2.44, and 3.70 in DSC and 0.04 mm ASD on the aforementioned structures.

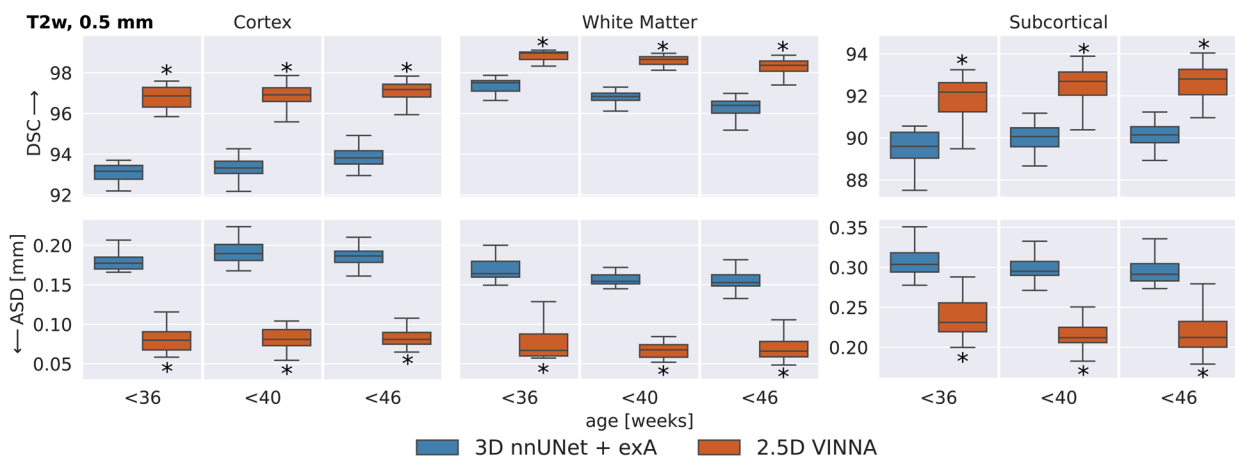
When comparing the four models, the 2D nnUNet + exA (left box) version performs worse than the 3D (second from left box), and 2.5D VINNA (right box) across all resolutions, structures, and modalities. Particularly notable are the large variations of 2D nnUNet + exA in prediction performance (large standard deviation) and large ASD (see Fig. 6), especially at the highest resolution (0.42 mm, for the cortex, 0.46 mm for WM and 0.41 mm for subcortical structures). This difference is less prominent in the DSC scores, but 2D nnUNet + exA also performs worst across all resolutions with respect to this metric (i.e., 92.39, 95.70, and 87.50 for a resolution of 0.5 mm). The 3D nnUNet + exA (second from left) improves accuracy by 24.5%, 12.3%, and 22.9% for ASD and 1.75%, 1.06%, and 1.64% for DSC across the three different resolutions. VINNA with its 4-DOF transform module (VINNA, right box) is again the best performing model, significantly outperforming all other networks. Compared to the 3D nnUNet + exA, ASD and DSC scores are significantly improved with the highest gain on the cortical structures (56%, 35%, and 45% ASD and 3.8%, 2.7%, and 3.5% DSC for 0.5 mm, 0.8 mm, and 1.0 mm, respectively).

To evaluate if this trend is consistent across age groups, the 0.5 mm MRI images are split into three approximately equal-sized groups based on the participants' age-at-scan information (32–36, 36–40, and 40–46 weeks). Figure 6 shows DSC (top) and ASD (bottom) calculated for T2w (Fig. 7) for the 3D nnUNet + exA (left box) and VINNA with its 4-DOF transform module (right box) in each of the categories.

Consistent with the previous section, 3D nnUNet + exA reaches the weakest ASD and DSC across all age groups. On average, the VINNA with its 4-DOF transform module (right box) improves performance compared to nnUNet by 2.8%, 2.9%, and 2.8% DSC and 41.5%, 43.6%, and 43.1% ASD from the youngest (32–36) to the oldest (< 46) age group and reaches a DSC of 96.86, 96.913, and 97.17 for the cortical structures, 98.96, 98.652, and 98.362 for the WM structures, and 92.18, 92.69, and 92.80 for the subcortical structures across all age groups (youngest to oldest) on the T2w MRIs (Fig. 7, top). Here, VINNA also reaches the lowest ASD



**Fig. 6.** SOTA Segmentation performance: Our VINNA with the 4-DOF transform module (last box in group) outperforms the three state-of-the-art deep-learning approaches, 2D nnUNet + exA and 3D (first and second box) in Dice Similarity Coefficient (DSC, upper row) and average surface distance (ASD, lower row). This performance advantage is significant (corrected  $p < 10^{-10}$ , indicated with \*) and consistent across three resolutions (0.5 mm, 0.8 mm, and 1.0 mm), two modalities (T2w, top and T1w, bottom), and three structure groups (cortex, WM, and subcortical structures).



**Fig. 7.** SOTA Segmentation performance across age groups: VINNA equipped with the 4-DOF transform module (right box) consistently outperforms state-of-the-art deep-learning approaches 3D nnUNet with external augmentation (+exA, left box) across four age groups. Improvement in dice similarity coefficient (DSC, top) and average surface distance (ASD, bottom) on T2w at 0.5 mm is significant (corrected  $p < 10^{-4}$ , indicated with \*) for < 36, < 40, and < 46 week-old newborns.

(0.080 mm, 0.081 mm, and 0.081 mm for the cortical structures, 0.067 mm, 0.068 mm, and 0.066 mm for the WM structures, and 0.231 mm, 0.212 mm, and 0.212 mm for the subcortical structures). The results are significantly better compared to 3D nnUNet + exA (corrected  $p < 10^{-9}$ ). As seen by the increasing DSC and decreasing ASD, the younger age groups (< 32 – 36) have proved to be more challenging to segment. For VINNA, the performance decreases most significantly for the subcortical structures (DSC by 1.82% and ASD by 20.09%) and least on the cortex (0.35% DSC and 4.65% ASD). This trend is consistent for the other two models. Assessment of qualitative differences between the 3D nnUNet + exA and VINNA on a representative participant at 40 weeks of age (Fig. 10) shows slight over-segmentation of the cortex and loss of small WM strands with the 3D nnUNet + exA (third row, second column, arrows). Overall, the segmentation with VINNA (fourth row) appears less smoothed and closer to the ground truth (second row).

### 3.2.2. Comparison to iBEAT

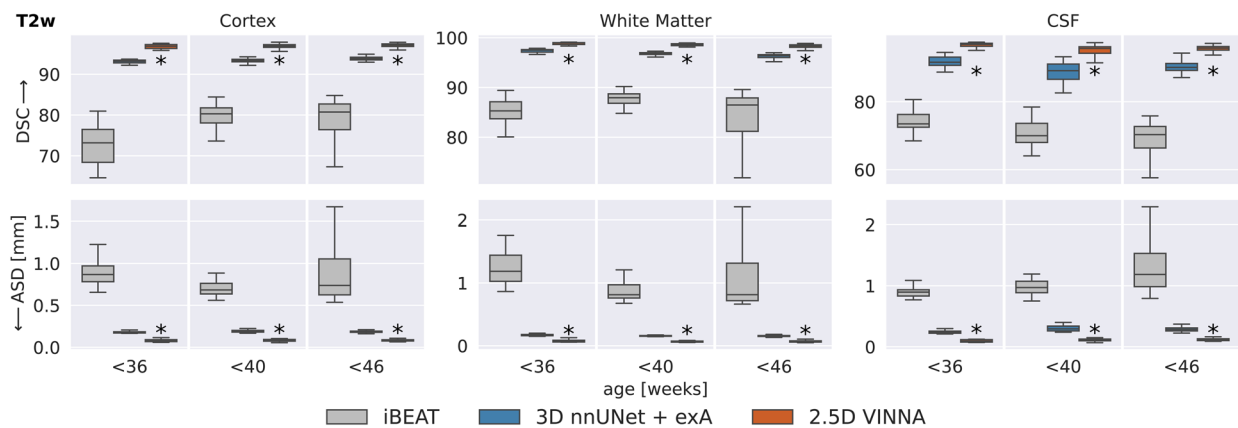
In Figure 8, the deep-learning models are compared to the docker version of iBEAT on the T2w images at 0.5 mm. iBEAT is officially designed for ages 0–6 years and the docker version we used for processing returns three labels (WM, GM, CSF). The definition of these labels is different from that of the dHCP-atlas (see Section 2.3.3), so we map the ground truth as well as the predictions from the deep-learning networks (nnUNet3D and VINNA) to be able to compare segmentation similarity. As described in Section 2.3, retraining the CNN part of iBEAT under same data and label definitions is not possible, as neither the source code nor the original training data is

available online. Note, even though we do not need to interpolate, cross-protocol comparisons include atlas differences and may introduce additional errors due to the mapping. While results should be interpreted with the caveat that iBEAT uses a different atlas and training dataset than 3D nnUNet + exA and VINNA, the label harmonization allows an as-fair-as-possible comparison with this state-of-the-art method.

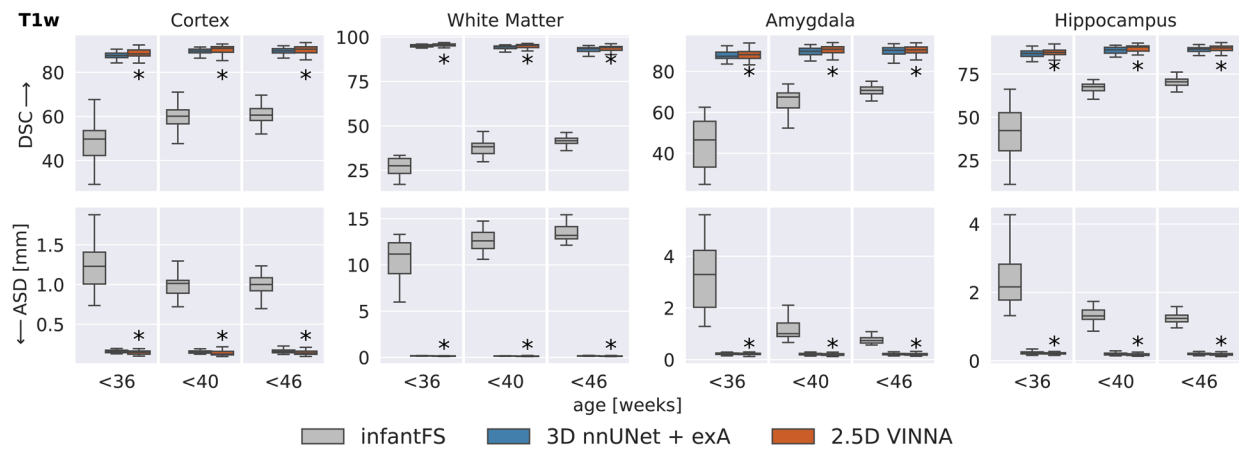
With respect to the mapped dHCP-reference segmentation, DSC (top) and ASD (bottom) are lower for iBEAT (left box in each plot) compared to the deep-learning methods (3D nnUNet + exA, middle box, and VINNA, right box in each plot). Performance of iBEAT on the GM, WM, and CSF improves with age. GM and WM are closest to the reference at 36–40 weeks (DSC 80.29/87.94 and ASD 0.68/0.813 mm), while CSF peaks at 32–26 weeks (73.49 and 0.896 mm). The 3D nnUNet + exA and VINNA show a similar trend, but performance is more consistent. As mentioned before, the differences are not necessarily due to wrong predictions made by iBEAT. Looking at the qualitative comparison in Figure 10, differences appear small, with iBEAT (third row) missing a few WM strands (arrow) and slightly over-segmenting the cortex compared to the mapped ground truth (second row).

### 3.2.3. Comparison to infantFS

Figure 9 shows performance comparison between infantFS (left box) and the deep-learning methods, 3D nnUNet + exA (middle box) and VINNA (left box), but in contrast to previous evaluations on T1w images at 1.0 mm, the operating resolution and modality of infantFS. Note that the infantFS labels are also different from those of dHCP and the ground truth labels must be mapped



**Fig. 8.** Deep-learning networks versus iBEAT: Similarity to the dHCP reference is higher with VINNA (third box) and 3D nnUNet + exA (second box) than iBEAT (first box) with respect to dice similarity coefficient (DSC, top) and average surface distance (ASD, bottom) on T2w MRIs at 0.5 mm across three age groups. Segmentation results with VINNA are significantly closer to dHCP (corrected  $p < 10^{-6}$ , indicated with \*) for CSF, GM, and WM. Note that iBEAT's structure definition is not identical to the dHCP-ALBERTs atlas and analysis is based on harmonized, merged labels.



**Fig. 9.** Deep-learning networks versus infantFS: VINNA with the 4-DOF transform module (third box) and 3D nnUNet + exA (second box) are closer to the dHCP reference than infantFS (first box) on T1w at 1.0 mm, the supported modality and resolution for infantFS. Dice Similarity Coefficient (DSC, top) and average surface distance (ASD, bottom) significantly improve with VINNA (corrected  $p < 10^{-6}$ , indicated with \*) on the cortex, WM, hippocampus, and amygdala across all age groups. Note that definition of subcortical structures differs in infantFS and predictions are harmonized to allow comparison to the deep-learning models.

(see Section 2.3.3 for details). As for iBEAT, the cross-protocol comparison can put the traditional method at an unfair disadvantage and results should be considered with this caveat.

Overall, the infantFS predictions differ strongly from the mapped dHCP ground truth, specifically for the younger age ranges. The highest similarity is reached for the subcortical structures (amygdala and hippocampus) at 40–46 weeks (DSC of 70.72/70.56 and ASD of 0.7361.23 mm, respectively). The cortex and WM reach a maximum DSC of 60.60/41.62 and ASD of 1.013.20 mm. Qualitative comparison (see Fig. 10, third row, left panel) shows difficulties with the correct location of the GM and WM border on the dHCP T1w MRI. Larger portions of the cortex are under-segmented, and strands of WM are lost compared to the ground truth. The deep-learning methods reach a DSC above 80 and an ASD below 0.5 mm for all structures and age groups. The method closest to the dHCP reference is again VINNA with the 4-DOF transform module (left box), followed by 3D nnUNet + exA (middle box). For a detailed comparison between nnUNet and VINNA on T1w, see Figure 7.

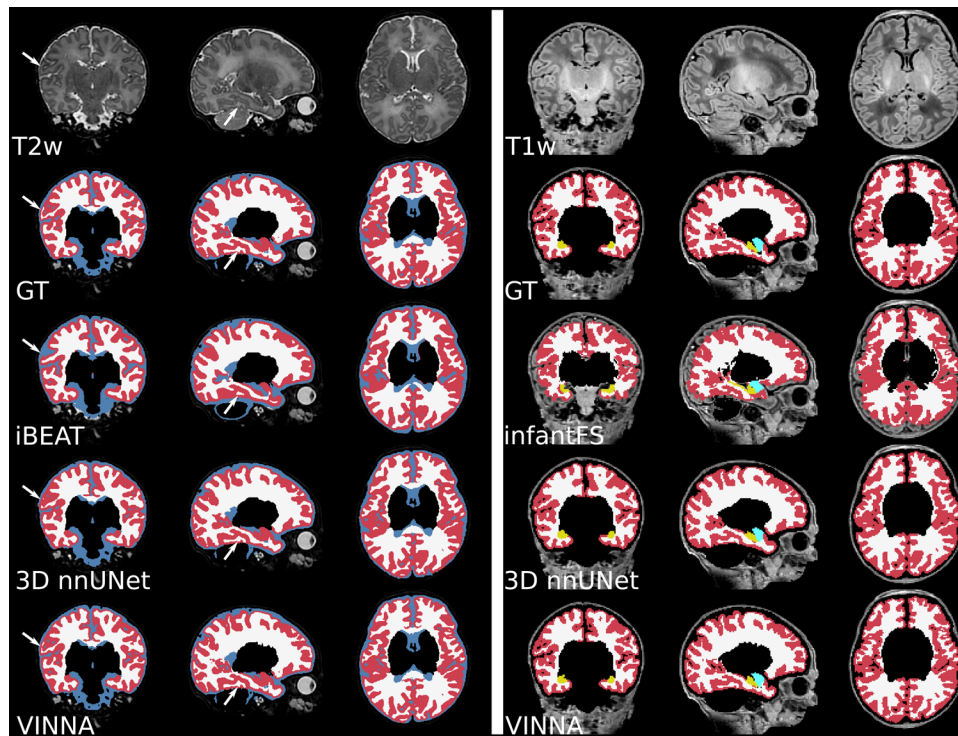
### 3.3. External validation on manual labels (M-CRIB)

To assess generalizability to a different dataset in our target age range (24–44 weeks) and to provide results with respect to a manual reference, we compare the segmentations produced by VINNA, 3D nnUNet + exA, and the dHCP-minimal-processing-pipeline to the 0.62 mm high-resolution T2w scans forming the M-CRIB atlas (Alexander et al., 2019). This dataset contains T2w MRIs from 10 participants and accompanying label maps based on the

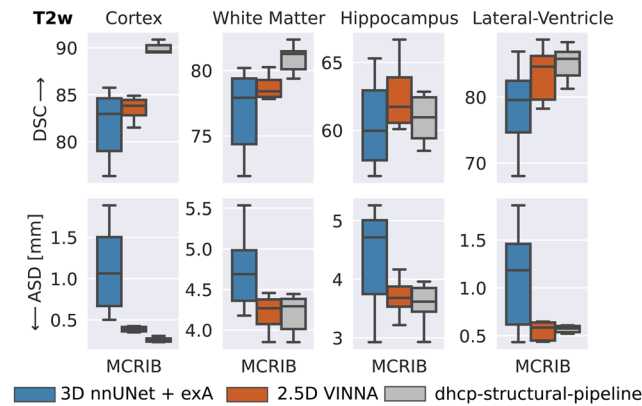
Desikian-Killiany-Tourville (DKT) atlas (Klein & Tourville, 2012). Note that the labels are not identical to the dHCP-ALBERTs atlas. Hence, we combine the cortical parcels to one label (cortex) for the segmentation comparison and mask all but three structures (WM, hippocampus, and lateral ventricles). As iBEAT does not differentiate between subcortical structures and GM or WM, nor CSF and ventricles, mapping of both, the ground truth and prediction, would be different compared to nnUNet, VINNA, and the dHCP-minimal-processing-pipeline. A fair comparison is therefore only possible between the latter methods, and iBEAT is thus not included in the following section.

In Figure 11, the DSC (top) and ASD (bottom) are compared over four structures (from left to right: cortex, WM, hippocampus, and lateral-ventricles) for the deep-learning methods (3D nnUNet + exA and VINNA) as well as the dHCP-minimal-processing-pipeline. Predictions for these methods are all based on the same label definition (dHCP-ALBERTs atlas (Gousias et al., 2012; Makropoulos et al., 2014)). As for the dHCP test set, VINNA outperforms the 3D versions of nnUNet across all four structures with a DSC of 83.83 and an ASD of 0.387 mm on the cortex, 78.41 and 4.268 mm on the WM, 61.73 and 3.682 mm on the hippocampus, and 84.62 and 0.585 mm on the lateral ventricles. Compared to nnUNet, the performance improves on average by 3.36% for the DSC and 19.40% for the ASD. Furthermore, 3D nnUNet + exA incorrectly flips left-right labels for five participants (P02-P04, P08-P09). We restore the lateralization in the presented results as DSC and ASD would have otherwise been close to zero for half of the participants. On the hippocampus, ventricles, and WM performance of VINNA is





**Fig. 10.** Qualitative T1w and T2w MRI segmentations on a representative scan at 41 weeks. VINNA with the 4-DOF transform module (last row) captures structural details lost in other methods. Comparison of the mapped ground truth (top) and segmentations from iBEAT (third row), 3D nnUNet + exA (fourth row) on a representative participant’s T2w MRI (left). The right side shows the T1w-scan from the same participant at 1.0 mm with ground truth (top), infantFS (third row), and the deep-learning methods.



**Fig. 11.** External validation of segmentation performance on the M-CRIB dataset. VINNA (second box) outperforms nnUNet3D (first box) on all four structures and the dhCP-structural-pipeline (third box) on the hippocampus. Dice Similarity Coefficient (DSC) and average surface distance (ASD) for cortex, WM, hippocampus, and lateral-ventricles are calculated with respect to harmonized manual labels on 10 subjects from M-CRIB. Note that atlas definition differs between ground truth and predictions.

similar to the dhcp-minimal-processing-pipeline, which is the best performing method. Specifically, predictions of the cortex are closer to the M-CRIB manual labels with the dhcp-pipeline (DSC of 89.60 and ASD of 0.257 mm),

which is likely due to the reliance on surface models, while the trained deep-learning models seem to over-segment the cortex (indicated by the similarity of the ASD, but larger differences in the DSC). Note that VINNA performs slightly better on the hippocampus with an increase in DSC by 1.25% and ASD by 0.58%. Due to the low number of participants, significance tests are not applicable for these experiments.

#### 4. DISCUSSION

In this paper, we present VINNA—a resolution-independent network for native-resolution neonatal brain MRI segmentation. With VINNA and our novel network-integrated 4-DOF transform module, we address two main difficulties associated with neonate segmentation: resolution non-uniformity across data cohorts and the extended range of head positions in infants.

In contrast to adults, newborn head positioning in the scanner varies significantly due to imaging during sleep, smaller head sizes, and relevant necessary modifications to scanner equipment, such as the padding of head coils. Additionally, while scans are commonly recorded at high resolutions, no uniform standard exists across cohorts. The availability of newborn datasets is also scarcer than

that of adult subjects, and the existing collections to date are unlikely to represent that wide diversity in resolutions and head positions.

The current state-of-the-art to address spatial variability such as head positions is data augmentation, which applies randomly sampled scale, rotation, and translation transformations in the native imaging space (i.e., externally to both intensity and label map). In VINNA, we introduce the 4-DOF transform module that can apply such transformations internally as part of the network. As the parameters to the transformation are inputs to the network, they can be randomized during training, similarly to external data augmentation methods. Moving the augmentation operation into the network, so it acts upon feature maps instead of inputs, marks a novel shift for data augmentation strategies. While we only implemented an augmentation of 4 DOFs here, the concept may be generalized to 9 DOFs for 3D or even to warp fields as well as to other tasks such as classification, regression, etc.

We demonstrate that the new network-integrated 4-DOF transform with internal augmentation outperforms state-of-the-art external augmentation approaches in CNN\*, VINN, and nnUNet (Isensee et al., 2021) on the dHCP cohort. Across three different resolutions and two modalities, our VINNA achieves the highest DSC (95.33 on average), as well as the lowest ASD (0.102 mm on average). Metric evaluation combined with qualitative inspection indicates that the internal 4-DOF transform module in VINNA better retains high-level details across all resolutions and age groups. It should be noted that VINNA performs augmentation of the feature maps in the first-scale transition. The first (pre-IDB) and last (post-CDB) blocks, therefore, act independent of any interpolation and promote feature detection at an unaltered scale. While the ablative results in Figure 5 indicate better performance than with external augmentation, generalization beyond the scales encountered during training is not assured in these two blocks. Due to their simplicity, the initial, low-level features are, however, often empirically resolution-independent by nature (e.g., Line-Detector).

To better explain the factors and mechanisms driving the performance improvements of VINNA, we review the observation from FastSurferVINN (Henschel et al., 2022) that motivated the extension presented here: in one-to-one comparisons, external augmentation reduces the segmentation performance on sub-millimeter MRIs. While—at first sight—the addition of data augmentation reducing performance seems contradictory, the one-to-one comparison of VINNA and VINNA + exA (see Fig. 5, the only difference is added external augmentation) robustly confirms the observation and extends it from just scaling to rigid transforms.

The positive effect of data augmentation is usually associated with an expansion of the input dataset through equivariant operations. Implementing operations for image and label pairs that are truly equivariant can be difficult. We believe that the loss of information due to image interpolation (lossy interpolation of the label map and image) is larger than previously believed. Internal augmentation, for the first time, offers an alternative approach with interpolation of continuous values in multiple feature maps, reducing the information loss.

Furthermore, the 4-DOF transform module together with the internal augmentation regularizes the latent space of the network, because it imposes an additional constraint: spatial consistency of the feature maps. Compared to equivalent CNN architectures, VINNA (and VINN) also benefit from a reduced capacity requirement to capture a large range of resolutions.

Our comparison to nnUNet highlights that 2D approaches lack contextual information, and fail to provide reliable predictions for whole-brain segmentation. The compromise between mid-range and long-range context in the 2.5D VINNA recovers structural information better and achieves higher segmentation performance across all age groups and structures—even compared to 3D methods. As full-view 3D networks are currently not applicable for high-resolution MRI segmentation due to memory requirements, nnUNet and other 3D networks rely on patch-based processing. In this case, the increased 3D context comes at the cost of limited long-range information and features a smaller field of view, potentially explaining the observed reduction in accuracy compared to 2.5D networks. This finding is in line with previous investigations which found limited performance differences between 2.5D and 3D approaches, even after extensive optimization of the 3D network architectures (Roy et al., 2022).

On the dHCP cohort, VINNA and its 4-DOF transform module also emulates the segmented structures better than traditional state-of-the-art infant pipelines, namely infantFS (Zöllei et al., 2020) and iBEAT (Wang et al., 2023). Notably, infantFS relies on traditional atlas-based segmentation while iBEAT uses a combination of deep-learning and traditional tools with a number of CNNs trained on defined target age ranges. While the re-trained networks (nnUNet + exA and VINNA) reach better results with respect to DSC and ASD, it should be noted that both, infantFS and iBEAT, differ significantly with respect to the returned number and definition of segmented regions. The necessary mapping between the segmentations is bound to introduce a bias, which can not be easily assessed. Additionally, both pipelines cater to a slightly different, larger age range (0–2 years for infantFS and 0–6 years for iBEAT). Consequently, predictions from

both methods of the cortex and WM improve for participants closer to the officially supported age range (>40 weeks). Qualitative assessment also shows good performance for the older newborns in iBEAT. infantFS unfortunately fails to correctly capture the cortex and WM on the majority of participants.

The original iBEAT v2.0 paper (Wang et al., 2023) also evaluates performance on the dHCP data and reports a DSC of 0.9 for the WM and 0.85 for the GM. Our results are in concordance with this assessment for the >40 week-old participants (DSC of 0.83 on the GM and 0.88 on the WM). The authors do not provide information on their label harmonization, therefore we cannot infer their reference standard. In contrast to the docker v2.0 version, the cloud version of iBEAT (not available for local data processing) does provide cortical parcellations. Extracting the cortex from the GM label (a combination of both cortical and subcortical GM in the docker version) allows direct comparison to the dHCP solution after merging its cortical structures, possibly explaining performance differences. In summary, iBEAT seems to work well on the supported age range while infantFS is less precise on the dHCP population. Other features included in the pipelines, such as surface generations, are an advantage compared to the proposed VINNA and can help to refine the segmentation of convoluted structures such as the cortex (Fischl, 2012). For our target domain in this paper, however, the VINNA architecture appears to emulate the investigated tissue classes more precisely.

Due to the limited extrapolation capabilities of neural networks, generalizability beyond the training set distribution is, however, uncertain. While the 4-DOF transform module in VINNA serves as a diversification of the training distribution with respect to spatial orientations and image resolution, the base cohort is still only a representation of the dHCP population, that is, all scans encountered during training represent newborns between 24–44 weeks post-menstrual age from a control cohort acquired on the same 3 T Phillips scanner. Therefore, dedicated experimental validation is required to confirm the models' effectiveness under differing conditions. As for all automated methods, manual quality checks of the predictions are recommended. While VINNA does perform well on M-CRIB, which covers the same age range as the dHCP, generalization to other cohorts is not necessarily guaranteed. Specifically, the T1w image intensities in dHCP appear significantly different from other cohorts which might also explain why infantFS performs poorly on the testing set. The T2w MRIs in dHCP are, on average, of better quality and the dhcp-minimal-processing-pipeline builds the ground truth segmentations based on it (Makropoulos et al., 2018). Additionally, in the early weeks of life, tissue contrast is higher in T2w recordings

as the brain is not fully matured and myelination is still ongoing (Dubois et al., 2021; Miller et al., 2012). Structural details and specifically tissue boundaries might be missing, blurred, or ambiguous in T1w MRI. Hence, the imaging data may lack sufficient information to allow correct delineation of the (sub-)cortical structures. This may also explain why the deep-learning networks (i.e., nnUNet, CNN\*, VINN), and VINNA are not able to emulate the ground truth on the T1w MRI as closely as on the T2w images. Including a SynthSeg-like intensity augmentation can potentially aid generalization across a wider age range. The method has previously been used in adults to generate contrast-agnostic networks that are able to segment both T1- and T2-weighted images (Billot et al., 2020; Iglesias et al., 2021). Due to the strong contrast changes in the early developmental years, implementing such a generative model may be an interesting direction for future work. However, as mentioned in the introduction, adaptation to the newborn cohort (Shang et al., 2022) showed that the synthetic images still differ considerably from real data leading to performance reduction compared to age-specific models trained on real images. Successful adaptation requires a strategy to close this performance gap.

Better accessibility of newborn datasets would allow diversification of the training sets and subsequently a better representation of the newborn MRI distribution with respect to both, T1w and T2w modalities. It has been shown that an increase in the training corpus alone is extremely effective to boost performance (Henschel et al., 2022; Sun et al., 2017). Age-specific models, such as our VINNA or iBEAT's CNNs, are another way to reduce variations and therefore, segregate the problem (i.e., less variations within one age group). However, the limited data availability and non-uniform segmentation labels still pose significant barriers. Models for more specific segmentations than just CSF, WM, and GM are currently not trainable in a supervised fashion due to missing ground truth. In addition, definition of these three structures alone already varies across different atlases and tools, which makes fair method comparisons challenging. Neither manual labels nor automated segmentation tools exist for a unified segmentation definition across different resolutions, modalities, and age ranges. The M-CRIB atlas (Alexander et al., 2019), an infant-specific version of the Desikan-Killiany-Tourville-Atlas (Klein & Tourville, 2012) that is commonly used in adults, provides a first step towards this goal. A consistent structure definition across different stages of life is especially important in the context of longitudinal studies, as segmentation with age-dependent models can induce biases and reduce anatomical consistency (Reuter et al., 2012). How to solve this conundrum is an open question for the

future. Several NIH- and internationally-funded initiatives have recently been dedicated to acquire data from newborns (Makropoulos et al., 2018), infants, and pediatric age ranges (Howell et al., 2019; Jernigan et al., 2016; Volkow et al., 2021) as well as adolescence (Karcher & Barch, 2021). Due to the easy integration of varying data resolutions and accommodation for head position variations between infants, toddlers, and adults, our VINNA architecture might prove to be useful in this area once the data and label availability problem is resolved.

Overall, with VINNA, we provide a fast and accurate method for high-resolution subcortical structure segmentation, cortical and WM parcellation of neonatal T1w and T2w MRI which generalizes well across the dHCP cohort. The presented 4-DOF transform module is also easy to integrate into other network architectures and might prove useful in different areas dealing with strong orientation variations. The application to neonates will be made available under VINNA4neonates as an open source package.<sup>8</sup> Adaptation of the infantFS surface pipeline to the VINNA4neonates predictions, similar to the approach taken in FastSurfer (Henschel et al., 2020) for adults, is an exciting direction for future work.

## DATA AND CODE AVAILABILITY

All MRI datasets used within this article are publicly available, and the open source repositories are cited within the article (Section 2.1). The source code of VINNA4neonates will be made publicly available on Github (<https://github.com/deep-mi/VINNA4neonates>) upon acceptance.

## AUTHOR CONTRIBUTIONS

Leonie Henschel: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing—original draft, Writing—review & editing, and Visualization. David Kügler: Conceptualization, Methodology, Software, Writing—review & editing, and Supervision. Martin Reuter: Conceptualization, Methodology, Resources, Writing—review & editing, Supervision, Project administration, and Funding acquisition. Lilla Zöllei: Conceptualization, Methodology, Resources, Writing—review & editing, Supervision, Project administration, and Funding acquisition.

## FUNDING

This work was supported by DZNE institutional funds, by the Federal Ministry of Education and Research of Germany (031L0206, 01GQ1801), by NIH (R01 LM012719,

R01 AG064027, R56 MH121426, P41 EB030006, and 5R01 HD109436-01), NICHD, and an NIH Shared Instrument Grant (S10RR023043). Funding was further received from the Chan-Zuckerberg Initiatives Essential Open Source Software for Science RFA (EOSS5 2022-252594).

## DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## ACKNOWLEDGEMENTS

Data were provided by the Melbourne Children's Regional Infant Brain Atlas and the Developing Human Connectome Project, KCL-Imperial-Oxford Consortium funded by the European Research Council under the European Union Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement no. [319456]. We are grateful to the families who generously supported this trial.

## REFERENCES

- Ahmad, S., Wu, Z., Li, G., Wang, L., Lin, W., Yap, P.-T., & Shen, D. (2019). Surface-constrained volumetric registration for the early developing brain. *Medical Image Analysis*, 58, 101540. <https://doi.org/10.1016/j.media.2019.101540>
- Ajayi-Obe, M., Saeed, N., Cowan, F., Rutherford, M., & Edwards, A. (2000). Reduced development of cerebral cortex in extremely preterm infants. *The Lancet*, 356(9236), 1162–1163. [https://doi.org/10.1016/s0140-6736\(00\)02761-6](https://doi.org/10.1016/s0140-6736(00)02761-6)
- Alexander, B., Loh, W. Y., Matthews, L. G., Murray, A. L., Adamson, C., Beare, R., Chen, J., Kelly, C. E., Anderson, P. J., Doyle, L. W., Spittle, A. J., Cheong, J. L. Y., Seal, M. L., & Thompson, D. K. (2019). Desikan-killiany-tourville atlas compatible version of M-CRIB neonatal parcellated whole brain atlas: The M-CRIB 2.0. *Frontiers in Neuroscience*, 13, 34. <https://doi.org/10.3389/fnins.2019.00034>
- Bastiani, M., Andersson, J. L., Cordero-Grande, L., Murgasova, M., Hutter, J., Price, A. N., Makropoulos, A., Fitzgibbon, S. P., Hughes, E., Rueckert, D., Victor, S., Rutherford, M., Edwards, A. D., Smith, S. M., Tournier, J.-D., Hajnal, J. V., Jbabdi, S., & Sotiropoulos, S. N. (2019). Automated processing pipeline for neonatal diffusion MRI in the developing human connectome project. *NeuroImage*, 185, 750–763. <https://doi.org/10.1016/j.neuroimage.2018.05.064>
- Beare, R. J., Chen, J., Kelly, C. E., Alexopoulos, D., Smyser, C. D., Rogers, C. E., Loh, W. Y., Matthews, L. G., Cheong, J. L. Y., Spittle, A. J., Anderson, P. J., Doyle, L. W., Inder, T. E., Seal, M. L., & Thompson, D. K. (2016). Neonatal brain tissue classification with morphological adaptation and unified segmentation. *Frontiers in Neuroinformatics*, 10, 12. <https://doi.org/10.3389/fninf.2016.00012>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to

<sup>8</sup> on our github pages.

- multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Billot, B., Robinson, E., Dalca, A. V., & Iglesias, J. E. (2020). Partial volume segmentation of brain MRI scans of any resolution and contrast. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020* (pp. 177–187). Springer International Publishing. [https://doi.org/10.1007/978-3-030-59728-3\\_18](https://doi.org/10.1007/978-3-030-59728-3_18)
- Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., & de Bruijne, M. (2019). Semi-supervised medical image segmentation via learning consistency under transformations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer International Publishing. [https://doi.org/10.1007/978-3-030-32226-7\\_90](https://doi.org/10.1007/978-3-030-32226-7_90)
- Bozek, J., Makropoulos, A., Schuh, A., Fitzgibbon, S., Wright, R., Glasser, M. F., Coalson, T. S., O’Muircheartaigh, J., Hutter, J., Price, A. N., Cordero-Grande, L., Teixeira, R. P. A., Hughes, E., Tusor, N., Baruteau, K. P., Rutherford, M. A., Edwards, A. D., Hajnal, J. V., ... Robinson, E. C. (2018). Construction of a neonatal cortical surface atlas using multimodal surface matching in the developing human connectome project. *NeuroImage*, 179, 11–29. <https://doi.org/10.1016/j.neuroimage.2018.06.018>
- Bui, T. D., Wang, L., Chen, J., Lin, W., Li, G., & Shen, D. (2019). Multi-task learning for neonatal brain segmentation using 3D Dense-Unet with dense attention guided by geodesic distance. In Q. Wang, F. Milletari, H. V. Nguyen, S. Albarqouni, M. J. Cardoso, N. Rieke, Z. Xu, K. Kamnitsas, V. Patel, B. Roysam, S. Jiang, K. Zhou, K. Luu, & N. Le (Eds), *Domain adaptation and representation transfer and medical image learning with less labels and imperfect data* (Vol. 11795, pp. 243–251). Springer International Publishing. [https://doi.org/10.1007/978-3-030-33391-1\\_28](https://doi.org/10.1007/978-3-030-33391-1_28)
- Chen, H., Dou, Q., Yu, L., Qin, J., & Heng, P.-A. (2018). Voxresnet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage*, 170, 446–455. <https://doi.org/10.1016/j.neuroimage.2017.04.041>
- Cherel, M., Budin, F., Prastawa, M., Gerig, G., Lee, K., Buss, C., Lyall, A., Zaldarriaga Consing, K., & Styner, M. (2015). Automatic tissue segmentation of neonate brain MR Images with subject-specific atlases. *Proceedings of SPIE the International Society for Optical Engineering*, 941311. <https://doi.org/10.1117/12.2082209>
- Copeland, A., Silver, E., Korja, R., Lehtola, S. J., Merisaari, H., Saukko, E., Sinisalo, S., Saunavaara, J., Lähdesmäki, T., Parkkola, R., Nolvi, S., Karlsson, L., Karlsson, H., & Tuulari, J. J. (2021). Infant and child MRI: A review of scanning procedures. *Frontiers in Neuroscience*, 15, 666020. <https://doi.org/10.3389/fnins.2021.666020>
- Cordero-Grande, L., Hughes, E. J., Hutter, J., Price, A. N., & Hajnal, J. V. (2018). Three-dimensional motion corrected sensitivity encoding reconstruction for multi-shot multi-slice MRI: Application to neonatal brain imaging: Aligned Multi-Shot Multi-Slice MRI. *Magnetic Resonance in Medicine*, 79(3), 1365–1376. <https://doi.org/10.1002/mrm.26796>
- Coupé, P., Mansencal, B., Clément, M., Giraud, R., Denis de Senneville, B., Ta, V.-T., Lepetit, V., & Manjon, J. V. (2020). Assemblynet: A large ensemble of CNNs for 3D whole brain MRI segmentation. *NeuroImage*, 219, 117026. <https://doi.org/10.1016/j.neuroimage.2020.117026>
- Dai, Y., Shi, F., Wang, L., Wu, G., & Shen, D. (2013). iBEAT: A toolbox for infant brain magnetic resonance image processing. *Neuroinformatics*, 11(2), 211–225. <https://doi.org/10.1007/s12021-012-9164-z>
- de Macedo Rodrigues, K., Ben-Avi, E., Sliva, D. D., Choe, M.-s., Drottar, M., Wang, R., Fischl, B., Grant, P. E., & Zöllei, L. (2015). A freesurfer-compliant consistent manual segmentation of infant brains spanning the 0–2 year age range. *Frontiers in Human Neuroscience*, 9, 21. <https://doi.org/10.3389/fnhum.2015.00021>
- Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3), 968–980. <https://doi.org/10.1016/j.neuroimage.2006.01.021>
- Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297–302. <https://doi.org/10.2307/1932409>
- Ding, W., Abdel-Basset, M., Hawash, H., & Pedrycz, W. (2022). Multimodal infant brain segmentation by fuzzy-informed deep learning. *IEEE Transactions on Fuzzy Systems*, 30(4), 1088–1101. <https://doi.org/10.1109/tfuzz.2021.3052461>
- Dolz, J., Desrosiers, C., Wang, L., Yuan, J., Shen, D., & Ben Ayed, I. (2020). Deep CNN ensembles and suggestive annotations for infant brain MRI segmentation. *Computerized Medical Imaging and Graphics*, 79, 101660. <https://doi.org/10.1016/j.compmedimag.2019.101660>
- Dubois, J., Alison, M., Counsell, S. J., Hertz-Pannier, L., Hüppi, P. S., & Benders, M. J. (2021). MRI of the neonatal brain: A review of methodological challenges and neuroscientific advances. *Journal of Magnetic Resonance Imaging*, 53(5), 1318–1343. <https://doi.org/10.1002/jmri.27192>
- Dubois, J., Dehaene-Lambertz, G., Kulikova, S., Poupon, C., Hüppi, P., & Hertz-Pannier, L. (2014). The early development of brain white matter: A review of imaging studies in fetuses, newborns and infants. *Neuroscience*, 276, 48–71. <https://doi.org/10.1016/j.neuroscience.2013.12.044>
- Dubois, J., Lefèvre, J., Angleys, H., Leroy, F., Fischer, C., Lebenberg, J., Dehaene-Lambertz, G., Borradori-Tolsa, C., Lazeyras, F., Hertz-Pannier, L., Mangin, J.-F., Hüppi, P. S., & Germanaud, D. (2019). The dynamics of cortical folding waves and prematurity-related deviations revealed by spatial and spectral analysis of gyrification. *NeuroImage*, 185, 934–946. <https://doi.org/10.1016/j.neuroimage.2018.03.005>
- Evans, A. C. (2006). The NIH MRI study of normal brain development. *NeuroImage*, 30(1), 184–202. <https://doi.org/10.1016/j.neuroimage.2005.09.068>
- Fischl, B. (2012). Freesurfer. *NeuroImage*, 62(2), 774–781. <https://doi.org/10.1016/j.neuroimage.2012.01.021>
- Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3), 341–355. [https://doi.org/10.1016/s0896-6273\(02\)00569-x](https://doi.org/10.1016/s0896-6273(02)00569-x)
- Friston, K., Ashburner, J., Kiebel, S., Nichols, T., & Penny, W. (2007). *Statistical parametric mapping: The analysis of functional brain images*. Academic Press. <https://doi.org/10.1016/b978-012372560-8/50024-3>
- Garcia, K. E., Robinson, E. C., Alexopoulos, D., Dierker, D. L., Glasser, M. F., Coalson, T. S., Ortinau, C. M., Rueckert, D., Taber, L. A., Essen, D. C. V., Rogers, C. E., Smyser, C. D., & Bayly, P. V. (2018). Dynamic patterns of cortical expansion during folding of the preterm human

- brain. *Proceedings of the National Academy of Sciences of the United States of America*, 115(12), 3156–3161. <https://doi.org/10.1073/pnas.1715451115>
- Gilmore, J. H., Shi, F., Woolson, S. L., Knickmeyer, R. C., Short, S. J., Lin, W., Zhu, H., Hamer, R. M., Styner, M., & Shen, D. (2012). Longitudinal development of cortical and subcortical gray matter from birth to 2 years. *Cerebral Cortex*, 22(11), 2478–2485. <https://doi.org/10.1093/cercor/bhr327>
- Goodfellow, I., Warde-Farley, D., Mirza, M., Courville, A., & Bengio, Y. (2013). Maxout networks. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning* (Vol. 28, pp. 1319–1327). Atlanta, Georgia, USA: PMLR. <https://proceedings.mlr.press/v28/goodfellow13.html>
- Gousias, I. S., Edwards, A. D., Rutherford, M. A., Counsell, S. J., Hajnal, J. V., Rueckert, D., & Hammers, A. (2012). Magnetic resonance imaging of the newborn brain: Manual segmentation of labelled atlases in term-born and preterm infants. *NeuroImage*, 62(3), 1499–1509. <https://doi.org/10.1016/j.neuroimage.2012.05.083>
- Gui, L., Lisowski, R., Faundez, T., Hüppi, P. S., Lazeyras, F., & Kocher, M. (2012). Morphology-driven automatic segmentation of MR images of the neonatal brain. *Medical Image Analysis*, 16(8), 1565–1579. <https://doi.org/10.1016/j.media.2012.07.006>
- Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., & Reuter, M. (2020). FastSurfer—A fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219, 117012. <https://doi.org/10.1016/j.neuroimage.2020.117012>
- Henschel, L., Kügler, D., & Reuter, M. (2022). Fastsurfervinn: Building resolution-independence into deep learning segmentation methods—A solution for highres brain MRI. *NeuroImage*, 251, 118933. <https://doi.org/10.1016/j.neuroimage.2022.118933>
- Howell, B. R., Styner, M. A., Gao, W., Yap, P.-T., Wang, L., Baluyot, K., Yacoub, E., Chen, G., Potts, T., Salzwedel, A., Li, G., Gilmore, J. H., Piven, J., Smith, J. K., Shen, D., Ugurbil, K., Zhu, H., Lin, W., & Ellison, J. T. (2019). The UNC/UMN baby connectome project (BCP): An overview of the study design and protocol development. *NeuroImage*, 185, 891–905. <https://doi.org/10.1016/j.neuroimage.2018.03.049>
- Hughes, E. J., Winchman, T., Padormo, F., Teixeira, R., Wurie, J., Sharma, M., Fox, M., Hutter, J., Cordero-Grande, L., Price, A. N., Allsop, J., Bueno-Conde, J., Tusor, N., Arichi, T., Edwards, A. D., Rutherford, M. A., Counsell, S. J., & Hajnal, J. V. (2017). A dedicated neonatal brain imaging system: A dedicated neonatal brain imaging system. *Magnetic Resonance in Medicine*, 78(2), 794–804. <https://doi.org/10.1002/mrm.26462>
- Huo, Y., Xu, Z., Xiong, Y., Aboud, K., Parvathaneni, P., Bao, S., Bermudez, C., Resnick, S. M., Cutting, L. E., & Landman, B. A. (2019). 3D whole brain segmentation using spatially localized atlas network tiles. *NeuroImage*, 194, 105–119. <https://doi.org/10.1016/j.neuroimage.2019.03.041>
- Iglesias, J. E., Billot, B., Balbastre, Y., Tabari, A., Conklin, J., Gilberto González, R., Alexander, D. C., Golland, P., Edlow, B. L., & Fischl, B. (2021). Joint super-resolution and synthesis of 1 mm isotropic mp-rage volumes from clinical MRI exams with scans of different orientation, resolution and contrast. *NeuroImage*, 237, 118206. <https://doi.org/10.1016/j.neuroimage.2021.118206>
- Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203–211. <https://doi.org/10.1038/s41592-020-01008-z>
- Ito, R., Nakae, K., Hata, J., Okano, H., & Ishii, S. (2019). Semi-supervised deep learning of brain tissue segmentation. *Neural Networks*, 116, 25–34. <https://doi.org/10.1016/j.neunet.2019.03.014>
- Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. (2015). Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 28). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2015/hash/33ceb07bf4eeb3da587e268d663aba1a-Abstract.html>
- Jégou, S., Drozdal, M., Vazquez, D., Romero, A., & Bengio, Y. (2017). The one hundred layers tiramisú: Fully convolutional densenets for semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1175–1183). IEEE. <https://doi.org/10.1109/cvprw.2017.156>
- Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, 62(2), 782–790. <https://doi.org/10.1016/j.neuroimage.2011.09.015>
- Jernigan, T. L., Brown, T. T., Hagler, D. J., Akshoomoff, N., Bartsch, H., Newman, E., Thompson, W. K., Bloss, C. S., Murray, S. S., Schork, N., Kennedy, D. N., Kuperman, J. M., McCabe, C., Chung, Y., Libiger, O., Maddox, M., Casey, B., Chang, L., Ernst, T. M., ... Dale, A. M. (2016). The pediatric imaging, neurocognition, and genetics (PING) data repository. *NeuroImage*, 124, 1149–1154. <https://doi.org/10.1016/j.neuroimage.2015.04.057>
- Karcher, N. R., & Barch, D. M. (2021). The ABCD study: Understanding the development of risk for mental and physical health outcomes. *Neuropsychopharmacology*, 46(1), 131–142. <https://doi.org/10.1038/s41386-020-0736-6>
- Klein, A., & Tourville, J. (2012). 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in Neuroscience*, 6, 171. <https://doi.org/10.3389/fnins.2012.00171>
- Kuklisova-Murgasova, M., Aljabar, P., Srinivasan, L., Counsell, S. J., Doria, V., Serag, A., Gousias, I. S., Boardman, J. P., Rutherford, M. A., Edwards, A. D., Hajnal, J. V., & Rueckert, D. (2011). A dynamic 4D probabilistic atlas of the developing brain. *NeuroImage*, 54(4), 2750–2763. <https://doi.org/10.1016/j.neuroimage.2010.10.019>
- Kuklisova-Murgasova, M., Quaghebeur, G., Rutherford, M. A., Hajnal, J. V., & Schnabel, J. A. (2012). Reconstruction of fetal brain MRI with intensity matching and complete outlier removal. *Medical Image Analysis*, 16(8), 1550–1564. <https://doi.org/10.1016/j.media.2012.07.004>
- Kumar, S., Conjeti, S., Roy, A. G., Wachinger, C., & Navab, N. (2018). Infinet: Fully convolutional networks for infant brain MRI segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (pp. 145–148). IEEE. <https://doi.org/10.1109/isbi.2018.8363542>
- Lebenberg, J., Labit, M., Auzias, G., Mohlberg, H., Fischer, C., Rivière, D., Duchesnay, E., Kabdebon, C., Leroy, F., Labra, N., Poupon, F., Dickscheid, T., Hertz-Pannier, L., Poupon, C., Dehaene-Lambertz, G., Hüppi, P., Amunts, K., Dubois, J., & Mangin, J.-F. (2018). A framework based on sulcal constraints to align preterm, infant and adult human brain images acquired in vivo and post mortem. *Brain Structure and Function*, 223(9), 4153–4168. <https://doi.org/10.1007/s00429-018-1735-9>
- Liu, P., Parkinson, C., Jiang, D., Ouyang, M., De Vis, J. B., Northington, F. J., Tekes, A., Huang, H., Huisman, T. A., & Golden, W. C. (2019). Characterization of MRI techniques to assess neonatal brain oxygenation and blood flow. *NMR in Biomedicine*, 32(7), e4103. <https://doi.org/10.1002/nbm.4103>

- Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. *International Conference on Learning Representations*. <https://openreview.net/forum?id=Skq89Scxx>
- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations*. <https://openreview.net/forum?id=Bkg6RiCqY7>
- Makropoulos, A., Aljabar, P., Wright, R., Hüning, B., Merchant, N., Arichi, T., Tusor, N., Hajnal, J. V., Edwards, A. D., Counsell, S. J., & Rueckert, D. (2016). Regional growth and atlas of the developing human brain. *NeuroImage*, 125, 456–478. <https://doi.org/10.1016/j.neuroimage.2015.10.047>
- Makropoulos, A., Gousias, I. S., Ledig, C., Aljabar, P., Serag, A., Hajnal, J. V., Edwards, A. D., Counsell, S. J., & Rueckert, D. (2014). Automatic whole brain MRI segmentation of the developing neonatal brain. *IEEE Transactions on Medical Imaging*, 33(9), 1818–1831. <https://doi.org/10.1109/tmi.2014.2322280>
- Makropoulos, A., Robinson, E. C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., Counsell, S. J., Steinweg, J., Vecchiato, K., Passerat-Palmbach, J., Lenz, G., Mortari, F., Tenev, T., Duff, E. P., Bastiani, M., Cordero-Grande, L., Hughes, E., Tusor, N., Tournier, J.-D., ... Rueckert, D. (2018). The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. *NeuroImage*, 173, 88–112. <https://doi.org/10.1016/j.neuroimage.2018.01.054>
- McClure, P., Rho, N., Lee, J. A., Kaczmarzyk, J. R., Zheng, C. Y., Ghosh, S. S., Nielson, D. M., Thomas, A. G., Bandettini, P., & Pereira, F. (2019). Knowing what you know in brain segmentation using bayesian deep neural networks. *Frontiers in Neuroinformatics*, 13, 67. <https://doi.org/10.3389/fninf.2019.00067>
- Mehta, R., Majumdar, A., & Sivaswamy, J. (2017). BrainSegNet: A convolutional neural network architecture for automated segmentation of human brain structures. *Journal of Medical Imaging*, 4(2), 1–11. <https://doi.org/10.1117/1.jmi.4.2.024003>
- Miller, D. J., Duka, T., Stimpson, C. D., Schapiro, S. J., Baze, W. B., McArthur, M. J., Fobbs, A. J., Sousa, A. M. M., Šestan, N., Wildman, D. E., Lipovich, L., Kuzawa, C. W., Hof, P. R., & Sherwood, C. C. (2012). Prolonged myelination in human neocortical evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 109(41), 16480–16485. <https://doi.org/10.1073/pnas.1117943109>
- Moeskops, P., Benders, M. J., Chitã, S. M., Kersbergen, K. J., Groenendaal, F., De Vries, L. S., Viergever, M. A., & Išgum, I. (2015). Automatic segmentation of MR brain images of preterm infants using supervised classification. *NeuroImage*, 118, 628–641. <https://doi.org/10.1016/j.neuroimage.2015.06.007>
- Nie, D., Wang, L., Gao, Y., & Shen, D. (2016). Fully convolutional networks for multi-modality iso-intense infant brain image segmentation. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* (pp. 1342–1345). IEEE. <https://doi.org/10.1109/isbi.2016.7493515>
- Pasban, S., Mohamadzadeh, S., Zeraatkar-Moghaddam, J., & Shafiei, A. K. (2020). Infant brain segmentation based on a combination of VGG-16 and U-Net deep neural networks. *IET Image Processing*, 14(17), 4756–4765. <https://doi.org/10.1049/iet-ipr.2020.0469>
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch. *NIPS Workshop Autodiff*. <https://openreview.net/forum?id=BJJsrmfCZ>
- Pérez-García, F., Sparks, R., & Ourselin, S. (2021). TorchIO: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, 208, 106236. <https://doi.org/10.1016/j.cmpb.2021.106236>
- Prastawa, M., Gilmore, J. H., Lin, W., & Gerig, G. (2005). Automatic segmentation of MR images of the developing newborn brain. *Medical Image Analysis*, 9(5), 457–466. <https://doi.org/10.1016/j.media.2005.05.007>
- Prieto, J. C., Styner, M. A., Pham, K., Yang, X., & Niethammer, M. (2019). Multiseg pipeline: Automatic tissue segmentation of brain MR images with subject-specific atlases. In B. Gimi & A. Krol (Eds.), *Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging* (p. 18). SPIE. <https://doi.org/10.1117/12.2513237>
- Qamar, S., Jin, H., Zheng, R., Ahmad, P., & Usama, M. (2020). A variant form of 3D-UNet for infant brain segmentation. *Future Generation Computer Systems*, 108, 613–623. <https://doi.org/10.1016/j.future.2019.11.021>
- Rajchl, M., Lee, M. C. H., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M. A., Hajnal, J. V., Kainz, B., & Rueckert, D. (2017). Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE Transactions on Medical Imaging*, 36(2), 674–683. <https://doi.org/10.1109/tmi.2016.2621185>
- Reuter, M., Schmansky, N. J., Rosas, H. D., & Fischl, B. (2012). Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4), 1402–1418. <https://doi.org/10.1016/j.neuroimage.2012.02.084>
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer Assisted Intervention—MICCAI 2015*, 9351, 234–241. [https://doi.org/10.1007/978-3-319-24574-4\\_28](https://doi.org/10.1007/978-3-319-24574-4_28)
- Roy, A. G., Conjeti, S., Navab, N., Wachinger, C., & Alzheimer's Disease Neuroimaging Initiative. (2019). Quicknat: A fully convolutional network for quick and accurate segmentation of neuroanatomy. *NeuroImage*, 186, 713–727. <https://doi.org/10.1016/j.neuroimage.2018.11.042>
- Roy, A. G., Conjeti, S., Sheet, D., Katouzian, A., Navab, N., & Wachinger, C. (2017). Error corrective boosting for learning fully convolutional networks with limited data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 231–239). Springer. [https://doi.org/10.1007/978-3-319-66179-7\\_27](https://doi.org/10.1007/978-3-319-66179-7_27)
- Roy, S., Kügler, D., & Reuter, M. (2022). Are 2.5D approaches superior to 3D deep networks in whole brain segmentation? In E. Konukoglu, B. Menze, A. Venkataraman, C. Baumgartner, Q. Dou, & S. Albarqouni (Eds.), *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning* (Vol. 172, pp. 988–1004). PMLR. <https://proceedings.mlr.press/v172/roy22a.html>
- Serag, A., Aljabar, P., Ball, G., Counsell, S. J., Boardman, J. P., Rutherford, M. A., Edwards, A. D., Hajnal, J. V., & Rueckert, D. (2012). Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *NeuroImage*, 59(3), 2255–2265. <https://doi.org/10.1016/j.neuroimage.2011.09.062>
- Shang, Z., Turja, M. A., Feczko, E., Houghton, A., Rueter, A., Moore, L. A., Snider, K., Hendrickson, T., Reiners, P., Stoyell, S., Kardan, O., Rosenberg, M., Elison, J. T., Fair, D. A., & Styner, M. A. (2022). Learning strategies for contrast-agnostic segmentation via SynthSeg for infant MRI data. In E. Konukoglu, B. Menze, A. Venkataraman, C. Baumgartner, Q. Dou, & S. Albarqouni (Eds.), *Proceedings*

- of *The 5th International Conference on Medical Imaging with Deep Learning*, (Vol. 172, pp. 1075–1084). PMLR. <https://proceedings.mlr.press/v172/shang22a.html>
- Shen, D. D., Bao, S. L., Wang, Y., Chen, Y. C., Zhang, Y. C., Li, X. C., Ding, Y. C., & Jia, Z. Z. (2023). An automatic and accurate deep learning-based neuroimaging pipeline for the neonatal brain. *Pediatric Radiology*, 53(8), 1685–1697. <https://doi.org/10.1007/s00247-023-05620-x>
- Shi, F., Yap, P.-T., Fan, Y., Gilmore, J. H., Lin, W., & Shen, D. (2010). Construction of multi-region-multi-reference atlases for neonatal brain MRI segmentation. *NeuroImage*, 51(2), 684–693. <https://doi.org/10.1016/j.neuroimage.2010.02.025>
- Shi, F., Yap, P.-T., Wu, G., Jia, H., Gilmore, J. H., Lin, W., & Shen, D. (2011). Infant brain atlases from neonates to 1- and 2-year-olds. *PLoS One*, 6(4), 1–11. <https://doi.org/10.1371/journal.pone.0018746>
- Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 843–852). IEEE. <https://doi.org/10.1109/iccv.2017.97>
- Sun, L., Ma, W., Ding, X., Huang, Y., Liang, D., & Paisley, J. (2019). A 3D spatially-weighted network for segmentation of brain tissue from MRI. *IEEE Transactions on Medical Imaging*, 39(4), 898–909. <https://doi.org/10.1109/tmi.2019.2937271>
- Sun, Y., Gao, K., Wu, Z., Li, G., Zong, X., Lei, Z., Wei, Y., Ma, J., Yang, X., Feng, X., Zhao, L., Le Phan, T., Shin, J., Zhong, T., Zhang, Y., Yu, L., Li, C., Basnet, R., Ahmad, M. O., ... Wang, L. (2021). Multi-site infant brain segmentation algorithms: The iSeg-2019 challenge. *IEEE Transactions on Medical Imaging*, 40(5), 1363–1376. <https://doi.org/10.1109/tmi.2021.3055428>
- Sørensen, T. J. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter/Kongelige Danske Videnskabernes Selskab*, 5, 1–34. [https://www.royalacademy.dk/Publications/High/295\\_S%C3%B8rensen,%20Thorvald.pdf](https://www.royalacademy.dk/Publications/High/295_S%C3%B8rensen,%20Thorvald.pdf)
- Volkow, N. D., Gordon, J. A., & Freund, M. P. (2021). The healthy brain and child development study—Shedding light on opioid exposure, COVID-19, and health disparities. *JAMA Psychiatry*, 78(5), 471–472. <https://doi.org/10.1001/jamapsychiatry.2020.3803>
- Wachinger, C., Reuter, M., & Klein, T. (2018). Deepnat: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage*, 170, 434–445. <https://doi.org/10.1016/j.neuroimage.2017.02.035>
- Wang, L., Nie, D., Li, G., Puybareau, E., Dolz, J., Zhang, Q., Wang, F., Xia, J., Wu, Z., Chen, J.-W., Thung, K.-H., Bui, T. D., Shin, J., Zeng, G., Zheng, G., Fonov, V. S., Doyle, A., Xu, Y., Moeskops, P., ... Shen, D. (2019). Benchmark on automatic six-month-old infant brain segmentation algorithms: The iSeg-2017 challenge. *IEEE Transactions on Medical Imaging*, 38(9), 2219–2230. <https://doi.org/10.1109/tmi.2019.2901712>
- Wang, L., Shi, F., Gao, Y., Li, G., Lin, W., & Shen, D. (2015). Isointense infant brain segmentation by stacked kernel canonical correlation analysis. In G. Wu, P. Coupé, Y. Zhan, B. Munsell, & D. Rueckert (Eds.), *Patch-based techniques in medical imaging* (pp. 28–36). Springer International Publishing. [https://doi.org/10.1007/978-3-319-28194-0\\_4](https://doi.org/10.1007/978-3-319-28194-0_4)
- Wang, L., Wu, Z., Chen, L., Sun, Y., Lin, W., & Li, G. (2023). iBEAT V2.0: A multisite-applicable, deep learning-based pipeline for infant cerebral cortical surface reconstruction. *Nature Protocols*, 18(5), 1488–1509. <https://doi.org/10.1038/s41596-023-00806-x>
- Wang, Y., Haghpanah, F. S., Zhang, X., Santamaria, K., Da Costa Aguiar Alves, G. K., Bruno, E., Aw, N., Maddocks, A., Duarte, C. S., Monk, C., Laine, A., Posner, J., & program collaborators for Environmental influences on Child Health Outcomes (2022). ID-Seg: An infant deep learning-based segmentation framework to improve limbic structure estimates. *Brain Informatics*, 9(1), 12. <https://doi.org/10.1186/s40708-022-00161-9>
- Weisenfeld, N. I., & Warfield, S. K. (2009). Automatic segmentation of newborn brain MRI. *NeuroImage*, 47(2), 564–572. <https://doi.org/10.1016/j.neuroimage.2009.04.068>
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83. <https://doi.org/10.2307/3001968>
- Zeng, G., & Zheng, G. (2018). Multi-stream 3D FCN with multi-scale deep supervision for multi-modality isointense infant brain MR image segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (pp. 136–140). IEEE. <https://doi.org/10.1109/isbi.2018.8363540>
- Zeng, Z., Zhao, T., Sun, L., Zhang, Y., Xia, M., Liao, X., Zhang, J., Shen, D., Wang, L., & He, Y. (2023). 3D-MASNet: Mixed-scale asymmetric convolutional segmentation network for 6-month-old infant brain MR images. *Human Brain Mapping*, 44(4), 1779–1792. <https://doi.org/10.1002/hbm.26174>
- Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., & Shen, D. (2015). Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*, 108, 214–224. <https://doi.org/10.1016/j.neuroimage.2014.12.061>
- Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1), 45–57. <https://doi.org/10.1109/42.906424>
- Zöllei, L., Iglesias, J. E., Ou, Y., Grant, P. E., & Fischl, B. (2020). Infant freesurfer: An automated segmentation and surface extraction pipeline for t1-weighted neuroimaging data of infants 0–2 years. *NeuroImage*, 218, 116946. <https://doi.org/10.1016/j.neuroimage.2020.116946>

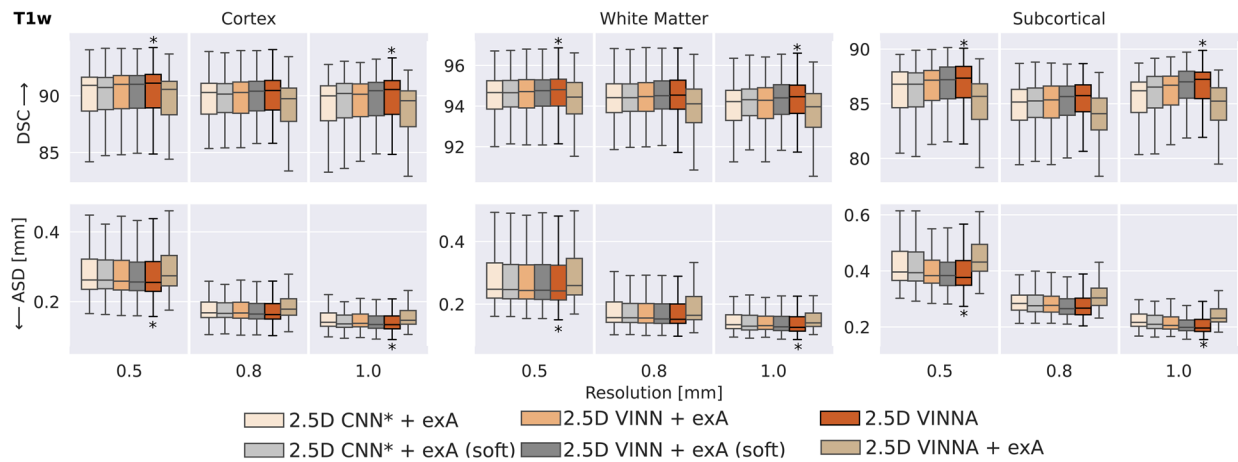
## APPENDIX

### A.1. Ablation on T1w

Appendix Figure A1 shows DSC (top) and ASD (bottom) for the ablative evaluation external augmentation (exA), optionally with linear interpolation of the one-hot encoded feature maps (exA (soft)) instead of hard label resampling, versus the internal 4-DOF transform modul (VINNA) on

the T1w. CNN\* + exA (first box) and + exA (soft) (second box) reach the weakest results, followed by VINN + exA (third box) and VINN + exA (soft) (fourth box). Finally VINNA (fifth box) reaches the best result. The performance improves the most on the subcortical structures (2.44% ASD and 0.4% DSC when switching from CNN\* to VINN, 5.91% ASD and 0.85% ASD when switching to VINNA). Performance improvements with the 4-DOF transform





**Appendix Fig. A1.** Comparison of approaches for external and internal spatial augmentation on T1w images: Our VINNA method—with the 4-DOF transform module—(fifth box in group) outperforms both state-of-the-art approaches with external Augmentation (CNN\* + exA, CNN\* + exA (soft), VINN + exA, and VINN + exA (soft); first to fourth box) in Dice Similarity Coefficient (DSC, upper row) and average surface distance (ASD, lower row). This performance advantage is significant on 0.5 mm and 1.0 mm (corrected  $p < 0.05$ , indicated with \*) and consistent across structures. Combining VINNA with external augmentation (VINNA + exA, last box) reduces performance. exA: external Augmentation. soft: linear interpolation of one-hot encoded feature maps.

module in VINNA, is significant for 0.5 mm and 1.0 mm (corrected  $p < 0.05$ ) and results in an average final DSC of 90.95, 90.03 and 90.36 and ASD of 0.295 mm, 0.2 mm, 0.2 mm for the 0.5 mm, 0.8 mm, and 1.0 mm resolution, respectively. Re-addition of exA to VINNA (sixth box) reduces performance to a similar level as CNN\* + exA.

Overall, all models perform better on T2w (Fig. 5) rather than T1w MRIs (Appendix Fig. A1) across all structures. When segmenting a T2w MRI of the same individual, performance improves for all networks, on average, by 5.7, 3.63 and 5.09 DSC points for the cortex, WM, and subcortical structures, respectively. The ASD difference between T1w and T2w applications nearly doubles for the highest resolution (0.5 mm) with an average of 0.18 mm compared to both lower resolutions (ASD of 0.9 mm for 1.0 and 0.8 mm) for the three structure averages. The DSC point improvement for the T2w images is similar across all resolutions. The trend is consistent for both networks (CNN\* and VINN) and augmentation schemes (external versus internal).

## A.2. Dataset summary

**Appendix Table A.** Summary of datasets used for training, validation, and testing.

Usage	Dataset	#P	Age [weeks]	Res [mm]
Training	dHCP	318	26-44	0.5, 0.8, 1.0
Validation	dHCP	90	27-41	0.5, 0.8, 1.0
Testing	dHCP	170	30-44	0.5, 0.8, 1.0
	M-CRIB	10	40-43	0.63

Table lists the dataset, number of participants (#P), range of post-menstrual age at scan in weeks and isotropic resolution in millimeter (Res [mm]) used in the paper.

## A.3. Labels

**Appendix Table B.** Subcortical segmentations of the dHCP-ALBERTs atlas and matching identifier (ID) equaling the number returned by VINNA for the structure.

ID		
Left	Right	Subcortical structure
0	0	Background
1	2	Hippocampus
3	4	Amygdala
17	18	Cerebellum
19	19	Brainstem
41	40	Caudate_nucleus
43	42	Thalamus_high_intensity_part_in_T2
45	44	Subthalamic_nucleus
47	46	Lentiform_Nucleus
48	48	Corpus_Callosum
49	50	Lateral-Ventricle
83	83	CSF
84	84	Extra_cranial_background
88	85	IntraCranialBackground
87	86	Thalamus_low_intensity_part_in_T2

Left and Right indicate the respective hemisphere. Four structures are not lateralized (same ID for Left and Right).

## A.4. Model parameters

**Appendix Table C.** Segmentation parcels of the dHCP-ALBERTs atlas for gray matter (GM) and white matter (WM).

GM ID		WM ID		Structure
Left	Right	Left	Right	
5	6	52	51	Anterior_temporal_lobe_medial_part
7	8	54	53	Anterior_temporal_lobe_lateral_part
9	10	56	55	Gyri parahippocampalis_et_ambiens_anterior_part
11	12	58	57	Superior_temporal_gyrus_middle_part
13	14	60	59	Medial_and_inferior_temporal_gyri_anterior_part
15	16	62	61	Lateral_occipitotemporal_gyrus_gyrus_fusiformis_anterior_part
21	20	64	63	Insula
23	22	66	65	Occipital_lobe
25	24	68	67	Gyri parahippocampalis_et_ambiens_posterior_part
27	26	70	69	Lateral_occipitotemporal_gyrus_gyrus_fusiformis_posterior_part
29	28	72	71	Medial_and_inferior_temporal_gyri_posterior_part
31	30	74	73	Superior_temporal_gyrus_posterior_part
33	32	76	75	Cingulate_gyrus_anterior_part
35	34	78	77	Cingulate_gyrus_posterior_part
37	36	80	79	Frontal_lobe
39	38	82	81	Parietal_lobe

Left and Right indicate the lateralization. The identifier (ID) equals the number returned by VINNA for the structure.

**Appendix Table D.** Summary of training parameters for CNN\*, VINN, VINNA, nnUNet 2D and nnUNet 3D.

Network	Dim	Time [h]	# Param [ $10^6$ ]	# Nets	Patch size	# Filter	# Layers	Kernel size	Batch size
CNN*	2.5D	75	1.52	3	256 x 256	64	5	3 x 3	16
VINN	2.5D	122	1.52	3	256 x 256	64	5	3 x 3	16
VINNA	2.5D	123	1.52	3	256 x 256	64	5	3 x 3	16
nnUNet	2D	88	18.75	1	160 x 128	32-512	5	3 x 3	125
nnUNet	3D	104	30.65	1	128 x 128 x 112	32-320	5	3 x 3 x 3	2

The different augmentations (external or internal) do not affect the parameters. The table lists the network name (Network), dimension (Dim), training time in hours (Time [h]), number of trainable parameters (# Param), image patch size (Patch Size), number of filters per convolution (# Filter), number of layers (# Layers), convolution kernel size (Kernel Size), and the batch size (Batch Size).