Ensembled Laterality-Aware Residual nnU-Net for Low-Field Pediatric Brain MRI Segmentation

Gianni Manuel Gagliardi*[1], Frederik Graewert*[1], Maximilian Palm*[1], Jonathan Deissler[2], and Constantin Ulrich[2]

[1] Heidelberg University, Heidelberg, Germany
gianni.gagliardi@stud.uni-heidelberg.de
frederik.graewert@stud.uni-heidelberg.de
maximilian.palm01@stud.uni-heidelberg.de

[2] Division of Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany
{jonathan.deissler,constantin.ulrich}@dkfz-heidelberg.de

*Equal contribution, Author order among the co-first authors may be adjusted for individual use

Abstract. Low-field magnetic resonance imaging (MRI) can broaden access to pediatric neuroimaging, but automatic segmentation remains difficult because contrast is weak, signal-to-noise ratio is limited, and several target structures are small. For LISA 2026 Task 2, we developed a reproducible nnU-Net v2 pipeline for multi-structure segmentation of pediatric brain T2 MRI. Following prior LISA work, we retained the official left/right labels and evaluated residual-encoder models without mirroring. A controlled three-way ablation showed that naive mirroring reduced mean performance, whereas anatomy-aware mirroring with explicit left/right label exchange recovered most of the loss and did not differ clearly in Dice from no mirroring after paired multiplicity-adjusted testing. No coarse global left/right swaps were detected. Five-fold cross-validation on 79 training cases otherwise showed only minor differences among a plain nnU-Net, standard ResEnc-M, and a cubic-patch ResEnc-M. We therefore evaluated a heterogeneous ensemble that adds stronger augmentation, supervised external pretraining, and self-supervised initialization. A hippocampus-only model, inspired by the specialist strategy used in LISA 2025, was combined with the generalist ensemble through explicit label-level merge rules. The deployable hippo-first merge produced the numerically highest mean and hippocampal Dice scores among the deployable specialist variants, whereas the generalist-only ensemble remained better on the distance and volume metrics. After multiplicity adjustment the ensemble improvement was statistically significant, whereas the specialist merge produced no detectable change. The observed differences are small and metric-dependent. We consequently present the method as an incremental, reproducible challenge pipeline and do not claim that one configuration is uniformly superior.

Keywords: Low-field MRI · Pediatric brain segmentation · nnU-Net · Residual encoder U-Net · Ensemble segmentation

1 Introduction

Structural brain MRI is central to studying early neurodevelopment and detecting deviations in anatomical maturation, yet high-field MRI systems remain expensive, infrastructure-intensive, and unevenly distributed globally. Low-field and ultra-low-field scanners, including portable 0.064 T systems, can reduce cost and logistical barriers, which may make them particularly relevant in low-resource clinical settings [10,11,1]. The LISA challenge targets this setting by focusing on quality assurance and segmentation of low-field pediatric brain MR images [1,2].

The segmentation task is technically challenging for several reasons. Ultra-low-field MR images have reduced signal-to-noise ratio and weaker tissue contrast compared with standard 1.5 T or 3 T acquisitions [10,11,1]. Pediatric brain anatomy changes rapidly during early development, making direct transfer of adult segmentation tools unreliable. The target structures include small bilateral regions such as the hippocampi and caudate nuclei, where small boundary errors strongly affect overlap and surface metrics. Finally, LISA-style labels may be affected by the difficulty of propagating or refining high-field-derived anatomical references in low-field image space, a limitation also discussed in prior LISA 2025 submissions [3,4].

In this work, we focus on a strong, reproducible challenge-style solution rather than proposing a new architecture. nnU-Net remains a competitive and carefully engineered framework for biomedical image segmentation [7], and recent validation studies have emphasized that strong baselines are essential before claiming methodological improvements [8]. Prior LISA 2025 work similarly showed that nnU-Net ResEnc-style baselines, combined with careful handling of data and annotations, can be competitive for low-field pediatric brain segmentation [3,4].

Morshuis, Hein, and Baumgartner already evaluated a ResEnc baseline without mirroring, a variant with symmetric labels and laterality restoration, and a separate U-Net trained only on the hippocampus [3]. Our corresponding configuration and specialist experiment are directly inspired by that study. In contrast to their symmetric-label branch, our models retain the official 11 foreground labels throughout training. Our incremental contribution is to combine a related specialist with a heterogeneous LISA 2026 generalist ensemble and compare explicit post-hoc merge rules.

We evaluate PlainConvUNet, standard residual-encoder medium nnU-Net (ResEnc-M), cubic 128^3 patch size ResEnc-M, a DA5 augmentation variant, and two pretraining/fine-tuning directions based on Baby Open Brains (BOB) and OpenMind pretrained MAEs [5,12]. We additionally evaluate the hippocampus specialist because the hippocampi are consistently the weakest labels. The experiments compare three matched mirroring policies, individual ensemble components, progressively larger ensembles, and alternative specialist merges.

2 Data and Task

2.1 Challenge task

LISA 2026 Task 2 requires automatic segmentation of low-field pediatric brain T2 MRI. According to the official challenge description, the target structures are the bilateral hippocampi, lateral ventricles, caudate nuclei, putamen and globus pallidus complex, and thalamus, together with the corpus callosum [1]. All experiments use the official segmentation masks, denoted `_seg.nii.gz`, as the training target.

2.2 Training data and preprocessing state

The local training set contains 79 complete image–label pairs. For each case, the input image is a combined isotropic low-field T2 volume, denoted `_ciso.nii.gz`, and the official training target is the corresponding segmentation mask, denoted `_seg.nii.gz`. A second mask type, `_LF_seg.nii.gz`, is available for the same cases as an additional low-field-refined reference. We keep `_LF_seg.nii.gz` separate from the main training target to avoid mixing label definitions. This follows annotation-quality concerns raised in prior LISA work, where the choice of high-field-derived versus low-field-refined labels can materially affect training behavior [4].

All 79 cases have matching `_ciso.nii.gz` and `_seg.nii.gz` files after combining the main Task 2 folder with the Task 2 Extra folder. All analyzed volumes have 1.0 mm isotropic voxel spacing. The two observed image shapes are $(197, 233, 189)$ and $(195, 233, 159)$. Images are stored as floating-point data, and the segmentation masks contain integer labels 0–11.

3 Methods

3.1 nnU-Net conversion and baseline

We use nnU-Net v2 as the base framework because it automatically configures preprocessing, normalization, patch size, network topology, training, validation, and inference for biomedical segmentation tasks [7]. The dataset was converted into a single-channel MRI dataset and normalized using z-score normalization. The PlainConvUNet baseline uses the standard nnU-Net v2 3D full-resolution setup with spacing $[1.0, 1.0, 1.0]$ mm, patch size $[112, 160, 128]$, and batch size 2. The model is trained with nnU-Net's default supervised multi-class segmentation objective, combining Dice and cross-entropy losses [9,7].

3.2 Residual-encoder and ensemble variants

Recent nnU-Net v2 presets include residual-encoder U-Net variants, motivated by stronger feature extraction in 3D medical segmentation [8]. We evaluate

residual-encoder medium plans (ResEnc-M) against the PlainConvUNet baseline. The standard ResEnc-M plan keeps the automatically planned patch size $[112, 160, 128]$, while the cubic ablation manually changes it to $[128, 128, 128]$. Both use the 3D full-resolution setting and batch size 2.

A design choice shared with the mirroring-free baseline of Morshuis et al. is to disable mirroring during residual-encoder training [3]. The task contains five bilateral label pairs: $(1, 2)$, $(3, 4)$, $(5, 6)$, $(7, 8)$, and $(9, 10)$. A left/right flip applied without exchanging these class identities places left-labelled anatomy on the opposite side of the augmented image. Morshuis et al. additionally investigated symmetric labels followed by laterality restoration. We retain the official labels and directly isolate mirroring in three otherwise matched standard-patch ResEnc-M variants: no mirroring, standard nnU-Net mirroring without class exchange, and anatomy-aware mirroring. The anatomy-aware transform retains nnU-Net's default mirroring policy but exchanges all five label pairs whenever a sampled flip includes the verified left/right axis; corpus-callosum label 11 remains unchanged. All variants use the same five folds, `checkpoint_best.pth`, and inference without test-time augmentation. The input NIfTI files use RAS orientation and the plans use identity `transpose_forward`; consequently, physical left/right corresponds to nnU-Net spatial axis 2. A preflight check verified this axis from all 79 masks and passed all 395 bilateral centroid-order checks.

Beyond the three base configurations, we evaluate complementary components for ensembling. DA5 denotes a residual-encoder trainer with stronger spatial/intensity augmentation and no mirroring. BOB denotes supervised external pretraining/fine-tuning using Baby Open Brains, a dataset of manually curated infant brain segmentations from 71 imaging visits across 51 participants [5]. SSL denotes self-supervised initialization and fine-tuning using the nnssl framework, which provides 3D medical vision self-supervised learning code and MAE-pretrained checkpoints from the OpenMind dataset [12].

3.3 Five-fold validation and final inference

All reported validation results are based on five-fold cross validation. Each case is evaluated only by the fold model for which that case was held out during training. Metrics are averaged over all 11 foreground labels for each case and then summarized across cases using mean $\pm$ sample standard deviation. We report Dice similarity coefficient (DSC), Hausdorff distance (HD), 95th percentile Hausdorff distance (HD95), average symmetric surface distance (ASSD), and absolute relative volume error (RVE). DSC, HD, ASSD, and RVE are official LISA segmentation metrics [1]; HD95 is additionally reported as a robust surface-distance metric.

For the controlled mirroring ablation, the two planned paired DSC contrasts compare no mirroring with standard mirroring and with anatomy-aware mirroring. We use two-sided paired Wilcoxon signed-rank tests and control the family-wise error rate across these contrasts with Holm correction. Effect estimates are mean within-case DSC differences, with paired bias-corrected and accelerated

(BCa) bootstrap 95% confidence intervals based on 100,000 bootstrapping iterations. A separate laterality audit compares direct and left/right-swapped correspondence across the five bilateral label pairs and flags a case when swapped correspondence exceeds direct correspondence. The same paired framework is applied to the ensemble and specialist comparisons, with Holm correction across the four planned DSC contrasts and, separately, across the secondary metrics within each contrast.

The generalist ensemble contains five components: standard ResEnc-M, cubic 128^3 patch size ResEnc-M, DA5 ResEnc-M, BOB-fine-tuned cubic ResEnc-M, and SSL-initialized ResEnc-M fine-tuning. During OOF evaluation, each component predicts a case only with the fold for which that case was held out. The code verifies that all generalists use the same 12-class map, exports their softmax arrays, and combines them with an equal-weight arithmetic mean using `nnUNetv2_ensemble`; the final generalist label at voxel v is

$$\hat{y}(v) = \arg\max_{c} \frac{1}{M} \sum_{m=1}^{M} p_{m,c}(v), \tag{1}$$

where $M = 5$. Because nnU-Net resamples the network logits back to the original image geometry before exporting the class posteriors, all components emit arrays on the same voxel grid, which is what makes it possible to average components whose plans differ in patch size and target spacing. No component, class, or voxel weighting is applied and no temperature or logit calibration is used; ties in the arg max are resolved toward the smallest label index, which makes the operation deterministic. For challenge inference, each component first averages the probabilities from its five fold checkpoints; the five component-level probability arrays are then fused in the same way. Mirroring-based test-time augmentation is disabled in both workflows. Each validation case is therefore the fusion of five networks, none of which saw that case during training, whereas the submission fuses all $5 \times 5 = 25$ generalist fold models; the reported out-of-fold figures thus describe a strictly smaller ensemble than the submitted one.

Following the hippocampus-only experiment of Morshuis et al. [3], we train an additional standard-patch ResEnc-M specialist on the same 79 cases and fold splits. Labels 3–11 are mapped to background, yielding the classes background, left hippocampus, and right hippocampus. The specialist uses 1 mm isotropic spacing, patch size $[112, 160, 128]$, batch size 2, the default Dice–cross-entropy objective, and no mirroring. Its three-class probabilities cannot be directly averaged with the generalists' 12-class probabilities, so it is incorporated after argmax without label remapping because its hippocampus IDs 1 and 2 match the generalist label map.

We evaluate two deployable label-level merges. For *hippo-first*, the specialist labels are written to an empty label map and protected; every voxel that remains background is then filled from the generalist prediction. The resulting hippocampus is the union of the specialist and generalist hippocampus extents. For *hippo-overwrite*, the generalist hippocampus labels are first removed and then replaced by the specialist labels, so the final hippocampus extent equals

the specialist prediction. Writing $S(v)$ for the specialist label and $H = \{1, 2\}$ for the two hippocampal labels, both rules are voxel-wise, deterministic, and parameter-free:

$$y_{\mathrm{hf}}(v) = \begin{cases} S(v) & \text{if } S(v) \in H, \\ \hat{y}(v) & \text{otherwise,} \end{cases} \qquad y_{\mathrm{ow}}(v) = \begin{cases} S(v) & \text{if } S(v) \in H, \\ 0 & \text{if } S(v) = 0 \text{ and } \hat{y}(v) \in H, \\ \hat{y}(v) & \text{otherwise.} \end{cases} \tag{2}$$

No morphological operation, connected-component filtering, or other postprocessing is applied at any point in the pipeline, and if a specialist prediction is missing for a case the generalist prediction is passed through unchanged so that no case is dropped from the evaluation. A validation-only signed-RVE oracle additionally chooses between these rules using $(V_{pred} - V_{gt})/(V_{gt} + 10^{-5})$: a negative value selects hippo-first and a non-negative value selects hippo-overwrite. Because this per-case decision requires the reference mask, it cannot be used for hidden validation or test cases. The repository also uses the cohort-level mean signed RVE only to recommend one fixed deployable strategy; the final candidate uses this fixed hippo-first rule.

Three properties follow directly from eq. (2). Here, $\{S \in H\}$ denotes the set of voxels v for which $S(v) \in H$.

The hippo-first rule forms the union of specialist and generalist hippocampal predictions, whereas hippo-overwrite replaces the generalist hippocampus entirely. The two rules therefore differ only where the specialist predicts background and the generalist predicts hippocampus; non-hippocampal labels can only lose voxels through the merge. Since the aggregate metrics average over all eleven labels, we additionally report hippocampal DSC.

3.4 Reproducibility and implementation details

The pipeline was structured so that the same dataset conversion, preprocessing, trainer class, plans file, and checkpoint choice can be reused for validation and test inference. The raw nnU-Net dataset contains 79 training images in `imagesTr` and 79 labels in `labelsTr`. We verified the dataset with nnU-Net's integrity check before training. The final OOF-ensemble and packaged inference workflows use `checkpoint_best.pth` consistently for all included folds. Large data artifacts, preprocessed arrays, checkpoints, and predictions were kept outside version control; only scripts, configuration files, evaluation code, and tabulated results should be versioned. All supervised nnU-Net models were trained on bwUniCluster 3.0 using one NVIDIA H100 GPU per fold, 8 CPU cores, and 64 GB RAM. A single PlainConvUNet fold required approximately 30 hours in our setup, while a ResEnc-M fold required approximately 32 hours depending on the configuration. The final ensemble consists of independently trained fold models from the selected components.

Table 1. Controlled mirroring ablation on the same 79 OOF cases and folds using `checkpoint_best.pth` and no inference TTA. Values are mean (sample SD). Paired contrasts are reported in the text.

Training policy	DSC ↑	HD ↓	HD95 ↓
No mirroring	0.8026 (0.0652)	3.9307 (2.1826)	1.7994 (0.5530)
Standard mirroring	0.7926 (0.0725)	4.5419 (3.3944)	2.2866 (2.3879)
LR-swap mirroring	0.8008 (0.0640)	3.8731 (2.1948)	1.8372 (0.5515)

Training policy	ASSD ↓	RVE ↓
No mirroring	0.5245 (0.2087)	0.1212 (0.0555)
Standard mirroring	0.6139 (0.4396)	0.1409 (0.0652)
LR-swap mirroring	0.5331 (0.2071)	0.1297 (0.0613)

4 Experiments

We conduct the controlled mirroring ablation on 79 identical OOF cases with the standard-patch ResEnc-M plan, compare Plain nnU-Net against the standard and cubic ResEnc-M variants, and extend the comparison to DA5, BOB and SSL fine-tuning and to progressively larger ensembles. The hippocampus-specialist ablation compares no specialist, hippo-first, hippo-overwrite, and the validation-only signed-RVE oracle of Sect. 3.3; the oracle is included to study the two merge behaviors, not as a candidate for hidden-data inference. A further ablation removes the five worst-performing training cases, but is treated only as a diagnostic because its exclusion rule is derived from validation behavior and can introduce selection bias.

5 Results

5.1 Controlled mirroring ablation

Table 1 reports the matched three-way ablation. Standard mirroring without class exchange produced the weakest mean result on every metric. Relative to this variant, no mirroring increased mean DSC by 0.0099 (paired BCa 95% CI [0.0048, 0.0195]; Holm-adjusted $p = 0.0150$). Anatomy-aware mirroring recovered most of this loss. Its mean DSC was 0.0017 below no mirroring, with a confidence interval spanning zero (no mirroring minus anatomy-aware mirroring: 95% CI [−0.0012, 0.0047]; Holm-adjusted $p = 0.1829$). Thus, the experiment supports avoiding naive mirroring, but does not establish a clear DSC difference between no mirroring and anatomically correct label-swapping mirroring.

The orientation preflight passed all 395 reference-mask centroid-order checks (79 cases times five bilateral pairs). In the prediction audit, swapped left/right correspondence never exceeded direct correspondence for any of the 79 cases in any variant. These checks found no coarse global left/right swap, although they cannot exclude local boundary or class-assignment errors.

Table 2. Overall OOF validation results, reported as mean (sample SD) over cases. R, C, D, B, and S denote standard ResEnc-M, cubic ResEnc-M, DA5, BOB, and SSL, respectively. DSC uses four decimals; other metrics use three to retain the LLNCS body font without scaling.

Model	DSC ↑	HD ↓	HD95 ↓	ASSD ↓	RVE ↓
Plain	.8000 (.0651)	3.851 (2.193)	1.834 (.540)	.533 (.209)	.129 (.055)
R	.8026 (.0652)	3.931 (2.183)	1.799 (.553)	.525 (.209)	.121 (.055)
C	.8038 (.0659)	3.865 (2.120)	1.798 (.553)	.523 (.213)	.122 (.055)
B	.8022 (.0668)	3.878 (2.139)	1.803 (.557)	.526 (.216)	.121 (.055)
D	.8019 (.0644)	3.827 (2.091)	1.817 (.542)	.528 (.208)	.124 (.057)
S	.8019 (.0651)	3.958 (2.158)	1.809 (.539)	.528 (.206)	.125 (.058)
B+C+Plain	.8056 (.0657)	3.809 (2.126)	1.772 (.547)	.516 (.211)	**.120 (.054)**
R+C	.8051 (.0656)	3.804 (2.144)	1.779 (.546)	.517 (.211)	.121 (.054)
R+C+D	.8059 (.0650)	3.796 (2.133)	1.776 (.537)	.516 (.209)	.122 (.055)
R+C+D+B	.8064 (.0655)	3.785 (2.133)	1.769 (.536)	.514 (.210)	.121 (.054)
All generalists	.8066 (.0657)	**3.774 (2.128)**	**1.768 (.535)**	**.513 (.211)**	.121 (.055)
+ hippo-first	**.8071 (.0647)**	3.861 (2.195)	1.779 (.551)	.514 (.210)	.124 (.059)

5.2 Overall ensemble comparison

Table 2 summarizes the internal model comparison. All base models are close: cubic ResEnc-M improves DSC over Plain nnU-Net by only 0.0038, much smaller than the case-level standard deviation of about 0.065. The generalist ensemble yields the clearest numerical improvement over the individual models. Adding the specialist with hippo-first raises mean DSC by only 0.0005; it is therefore not presented as a uniformly better model.

DA5, BOB, and SSL do not justify separate claims as standalone improvements, but the heterogeneous generalist ensemble is numerically stronger than the individual components. It has better HD, HD95, ASSD, and RVE than hippo-first. We retain hippo-first as the deployable specialist candidate only because it has the highest numerical mean DSC and hippocampal DSC among deployable variants. This is an overlap-oriented preference, not dominance across metrics. Paired case-level testing supports the ensemble effect but not the specialist merge. Relative to the plain baseline, the generalist ensemble increased mean DSC by 0.0066 (paired BCa 95% CI [0.0034, 0.0098]; Holm-adjusted $p < 0.0001$) and improved every reported metric (adjusted $p \leq 0.0014$); relative to the cubic ResEnc-M variant, the strongest single model, it increased mean DSC by 0.0029 (95% CI [0.0016, 0.0042]; adjusted $p < 0.0001$), with HD, HD95, and ASSD also better (adjusted $p \leq 0.01$) and RVE unchanged (adjusted $p = 0.43$). Adding the specialist changed nothing detectable: hippo-first differed from the generalist ensemble by 0.0005 mean DSC (95% CI [−0.0002, 0.0017]; adjusted $p = 0.284$) and 0.0027 hippocampal DSC (95% CI [−0.0013, 0.0091]; adjusted $p = 0.278$), while hippo-overwrite was significantly worse on both (adjusted $p = 0.0057$ and $p = 0.0063$). The ensemble gains are therefore small yet systematic, whereas hippo-first is retained on numerical grounds alone.

Table 3. Hippocampus-specialist merge ablation, reported as mean (sample SD) over 79 cases. DSC_{hippo} averages the left and right hippocampus labels. The signed-RVE oracle uses the reference mask and is validation-only.

Variant	DSC ↑	HD ↓	HD95 ↓
No specialist	0.8066 (0.0657)	**3.7744 (2.1280)**	1.7683 (0.5347)
Hippo-first	0.8071 (0.0647)	3.8612 (2.1946)	1.7788 (0.5508)
Hippo-overwrite	0.8049 (0.0651)	3.8865 (2.1830)	1.7929 (0.5485)
Signed-RVE oracle	**0.8072 (0.0646)**	3.8580 (2.1946)	**1.7138 (0.5080)**

Variant	ASSD ↓	RVE ↓	DSC_{hippo} ↑
No specialist	**0.5126 (0.2105)**	**0.1211 (0.0546)**	0.6762 (0.1704)
Hippo-first	0.5143 (0.2099)	0.1242 (0.0587)	0.6789 (0.1631)
Hippo-overwrite	0.5169 (0.2108)	0.1218 (0.0518)	0.6666 (0.1634)
Signed-RVE oracle	0.5615 (0.2170)	0.1235 (0.0566)	**0.6794 (0.1630)**

5.3 Hippocampus-specialist merge ablation

Table 3 compares the generalist ensemble against the specialist merges. Hippo-first is numerically higher than no specialist by 0.0005 DSC and 0.0027 hippocampal DSC, but is worse on HD, HD95, ASSD, and RVE. The signed-RVE oracle is marginally higher on DSC and hippocampal DSC and lower on HD95, but it substantially worsens ASSD and requires a reference mask for each decision. It is therefore not a deployable submission method.

5.4 Per-structure final ensemble performance

Table 4 reports per-label metrics for the selected hippo-first final ensemble. Hippocampi remain the weakest labels by DSC and RVE even after specialist merging, while thalamus, lentiform nuclei, and caudate nuclei are more stable, consistent with their larger size and higher contrast.

5.5 Failure analysis

The lowest mean Dice cases are consistent across configurations. In previous full-data single-model analyses, LISA_0021, LISA_0001, LISA_0037, LISA_0005, and LISA_0033 repeatedly appeared among the hardest cases, and LISA_0021 was the hardest case for all evaluated full-data models. This repeated pattern suggests that errors are concentrated in specific cases and structures rather than arising only from random training noise. Prior LISA work has shown that annotation drift and label curation can strongly influence training and evaluation [4].

Table 4. Per-structure metrics of hippo-first, reported as mean (sample SD) over 79 cases.

Structure	DSC ↑	HD ↓	HD95 ↓	ASSD ↓	RVE ↓
Hippocampus L	.672 (.178)	4.694 (6.735)	2.259 (1.290)	.720 (.556)	.219 (.272)
Hippocampus R	.686 (.166)	3.603 (1.619)	2.243 (1.299)	.690 (.477)	.239 (.301)
Ventricle L	.802 (.089)	4.306 (2.017)	1.811 (1.025)	.414 (.214)	.111 (.087)
Ventricle R	.787 (.085)	4.666 (2.326)	1.917 (1.078)	.441 (.211)	.114 (.082)
Caudate L	.833 (.078)	3.202 (1.494)	1.527 (.790)	.437 (.230)	.115 (.091)
Caudate R	.830 (.066)	3.167 (1.239)	1.560 (.731)	.453 (.214)	.097 (.087)
Lentiform L	.858 (.054)	4.682 (15.215)	1.800 (.725)	.575 (.275)	.104 (.083)
Lentiform R	.856 (.049)	2.987 (.811)	1.831 (.698)	.568 (.240)	.119 (.097)
Thalamus L	.898 (.043)	3.056 (4.162)	1.576 (.648)	.485 (.233)	.095 (.086)
Thalamus R	.900 (.043)	3.378 (5.931)	1.566 (.654)	.476 (.239)	.081 (.070)
Corpus callosum	.757 (.108)	4.732 (10.297)	1.477 (.666)	.398 (.177)	.072 (.059)

6 Discussion

The main finding is that model-family differences are small, while complementary ensembling provides the most useful measured improvement. Plain nnU-Net already reaches 0.8000 mean DSC, and both residual-encoder variants improve the aggregate only slightly. This agrees with the broader nnU-Net literature: carefully validated baselines often close much of the performance gap, and small changes should not be overinterpreted without rigorous validation [7,8]. For this reason, the paper frames the method as a robust challenge pipeline rather than a novel segmentation architecture.

Laterality remains an important design constraint rather than a novel contribution. Morshuis et al. previously examined both a mirroring-free ResEnc baseline and symmetric labels with laterality restoration, and their work also motivated our hippocampus specialist [3]. Our controlled ablation clarifies the corresponding design choice while retaining the official labels. Under the matched setup, naive standard mirroring was inferior to no mirroring, whereas exchanging class identities for a verified left/right flip recovered most of the mean DSC loss. The paired analysis found no clear DSC difference between anatomy-aware mirroring and no mirroring. Moreover, the audit detected no coarse global left/right swaps in any variant. The lower performance of naive mirroring may therefore reflect inconsistent augmented supervision or more local errors rather than frequent whole-case label reversal. We use no mirroring as the simpler final policy, not because it was shown to be superior to anatomically correct mirroring.

BOB and SSL are treated as complementary rather than standalone improvements. BOB supplies curated infant segmentations, but differs from LISA in age, field strength, and label coverage [5]. OpenMind pretraining likewise introduces a modality and acquisition-domain gap [12]. Neither fine-tuned model is superior alone; their role is to diversify the generalist ensemble.

Limitations. Hippo-first is numerically higher on mean and hippocampal DSC, but the generalist ensemble remains better on HD, HD95, ASSD, and RVE; the reference-dependent signed-RVE oracle is not deployable. The 0.0005 DSC specialist difference is not statistically detectable (Holm-adjusted $p = 0.284$), so hippo-first is retained on numerical grounds only. The mirroring ablation is limited to one ResEnc-M configuration, internal OOF data, and two DSC contrasts, and its swap audit cannot exclude local errors. No hidden test result is reported. The filtered 74-case ablation is excluded because its rule was derived from validation performance, and `_LF_seg.nii.gz` masks remain unused.

Future work. The mirroring ablation should be replicated across ensemble components and hidden data. Further work should inspect weak hippocampal cases and compare `_seg.nii.gz` with `_LF_seg.nii.gz`; any data-exclusion rule must be justified independently of validation performance.

7 Conclusion

We developed a reproducible residual nnU-Net ensemble for LISA 2026 Task 2. Naive mirroring reduced mean performance, while anatomy-aware label exchange recovered most of the loss and did not differ clearly in DSC from no mirroring. The generalist ensemble was numerically strongest on most metrics; hippo-first traded slightly higher overlap for worse distance and volume metrics. These prior-work-inspired choices are pragmatic, reproducible trade-offs rather than novel or uniformly superior methods.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Acknowledgments The authors acknowledge support by the state of Baden-Württemberg through bwHPC (`https://wiki.bwhpc.de/e/BwUniCluster3.0/Acknowledgement`).

References

1. Lepore, N., Linguraru, M.G., et al.: Low field pediatric brain magnetic resonance Image Segmentation and quality Assurance (LISA 2026) challenge documentation. Synapse project syn72118611, `https://www.synapse.org/Synapse:syn72118611/wiki/` (2026)
2. Lepore, N., Linguraru, M.G. (eds.): Low Field Pediatric Brain Magnetic Resonance Image Segmentation and Quality Assurance. Second MICCAI Challenge, LISA 2025, Held in Conjunction with MICCAI 2025, Daejeon, South Korea, September 27, 2025, Proceedings. Lecture Notes in Computer Science, vol. 16411. Springer, Cham (2026). `https://doi.org/10.1007/978-3-032-14417-1`

3. Morshuis, J.N., Hein, M., Baumgartner, C.F.: Segmenting brain regions in low field pediatric brain MR images using (symmetric) nnU-Net ResEnc. In: Lepore, N., Linguraru, M.G. (eds.) LISA 2025. LNCS, vol. 16411, pp. 41–49. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-14417-1_4
4. Zalevskyi, V., Bulut, D.-B., Sanchez, T., Bach Cuadra, M.: Segmenting infant brains across magnetic fields: domain randomization and annotation curation in ultra-low field MRI. In: Lepore, N., Linguraru, M.G. (eds.) LISA 2025. LNCS, vol. 16411, pp. 50–62. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-14417-1_5
5. Feczko, E., Stoyell, S.M., Moore, L.A., et al.: Baby Open Brains: an open-source dataset of infant brain segmentations. Scientific Data 12(1), 1423 (2025). https://doi.org/10.1038/s41597-025-05404-y; PubMed: https://pubmed.ncbi.nlm.nih.gov/39464007/
6. MIC-DKFZ: nnssl: An OpenMind for 3D medical vision self-supervised learning. GitHub repository, https://github.com/MIC-DKFZ/nnssl (2026)
7. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
8. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K.H., Jäger, P.F.: nnU-Net revisited: a call for rigorous validation in 3D medical image segmentation. In: MICCAI 2024. LNCS, vol. 15009, pp. 488–498. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72114-4_47
9. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
10. Marques, J.P., Simonis, F.F.J., Webb, A.G.: Low-field MRI: an MR physics perspective. Journal of Magnetic Resonance Imaging 49(6), 1528–1542 (2019). https://doi.org/10.1002/jmri.26637
11. Sarracanie, M., Salameh, N.: Low-field MRI: how low can we go? A fresh view on an old debate. Frontiers in Physics 8, 172 (2020). https://doi.org/10.3389/fphy.2020.00172
12. Wald, T., Ulrich, C., Suprijadi, J., Ziegler, S., Nohel, M., Peretzke, R., Köhler, G., Maier-Hein, K. H.: An OpenMind for 3D Medical Vision Self-supervised Learning. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 23839–23879 (2025).