Neuroverse3D: Developing In-Context Learning Universal Model for Neuroimaging in 3D

Jiesi Hu[*1,2], Chenfei Ye[*1], Yanwu Yang[3,4], Xutao Guo[1,2], Yang Shang[1], Pengcheng Shi[1], Hanyang Peng[†2], Ting Ma[†1, 2]

[1]Harbin Institute of Technology, Shenzhen, China [2]Peng Cheng Laboratory, Shenzhen, China

[3]University Hospital Tübingen, Tübingen, Germany [4]German Center for Mental Health (DZPG), partner site, Jena & Tübingen, Germany

22b352011@stu.hit.edu.cn, chenfei.ye@foxmail.com

philoso_phy0922@163.com, tma@hit.edu.cn

Abstract

In-context learning (ICL), a type of universal model, demonstrates exceptional generalization across a wide range of tasks without retraining by leveraging task-specific guidance from context, making it particularly effective for the intricate demands of neuroimaging. However, current ICL models, limited to 2D inputs and thus exhibiting suboptimal performance, struggle to extend to 3D inputs due to the high memory demands of ICL. In this regard, we introduce Neuroverse3D, an ICL model capable of performing multiple neuroimaging tasks in 3D (e.g., segmentation, denoising, inpainting). Neuroverse3D overcomes the large memory consumption associated with 3D inputs through adaptive parallel-sequential context processing and a U-shaped fusion strategy, allowing it to handle an unlimited number of context images. Additionally, we propose an optimized loss function to balance multi-task training and enhance focus on anatomical boundaries. Our study incorporates 43,674 3D multi-modal scans from 19 neuroimaging datasets and evaluates Neuroverse3D on 14 diverse tasks using held-out test sets. The results demonstrate that Neuroverse3D significantly outperforms existing ICL models and closely matches task-specific models, enabling flexible adaptation to medical center variations without retraining. The code and model weights are publicly available at https://github.com/jiesihu/Neuroverse3D.

1. Introduction

Computational neuroimage analysis has significantly advanced our understanding of the brain and non-invasive diagnostics, crucial for quantitative analysis and precision medicine. Recently, universal models trained on multi-domain datasets have gained attention for their ability to handle diverse tasks (e.g., segmentation, denoising, inpainting) and modalities (e.g., MRI-T1w, MRI-T2w, CT), demonstrating adaptability to domain shifts and even unseen tasks [12, 38, 43, 57]. In-context learning (ICL) has emerged as a promising paradigm [7, 10, 51], enabling models to adapt to domains without retraining by using image-label pairs as task-specific guidance. These informative contexts make ICL models particularly effective for segmenting tissues with intricate morphology and unifying diverse image generation tasks.

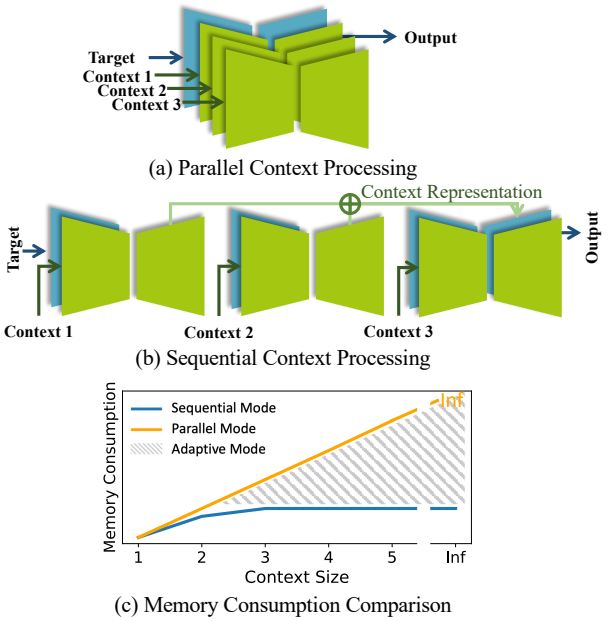
While ICL models show clinical potential in neuroimaging [12, 14, 54], they are limited by dimensionality. Existing models process 3D volumes as 2D slices, which results in the loss of inter-slice correlations and volumetric context. This lack of global spatial awareness hinders performance on neuroimaging tasks, such as hippocampus segmentation which requires 3D anatomical understanding [13, 20].

To address this limitations, it is crucial to develop an ICL model that captures 3D global information. However, creating a universal ICL model with 3D input presents significant challenges, as 3D images can be over 100 times larger than their 2D slices, leading to substantial memory requirements. This results in three critical issues: (1) Context size is significantly limited, hindering performance as larger contexts size demonstrably improve outputs [7, 12, 14, 52]. (2) Training models with more parameters becomes infeasible, restricting their capacity and potential performance. (3) Processing high-resolution 3D inputs becomes difficult.

To develop 3D universal models in neuroimaging, we introduce Neuroverse3D, an in-context learning model that takes 3D neuroimaging as input. To our knowledge, it is the first 3D ICL universal model for neuroimaging. To mitigate the significant memory demands, we propose Adaptive Parallel-Sequential Processing (APSP) and a U-shape fusion strategy to partition the entire context into multi-

[*]Equal contribution.

[†]Corresponding author: Hanyang Peng, Ting Ma.



(a) Parallel Context Processing

(b) Sequential Context Processing

(c) Memory Consumption Comparison

Figure 1. (a) Parallel processing of all contexts, which is applied by most existing ICL models. (b) Sequential processing of each context, which greatly reduces memory usage. (c) It shows the memory increment when increasing the context size. Our proposed model processes contexts in an adaptive mode, allowing for flexible parallel and sequential processing of contexts.

ple mini-contexts for processing, while ensuring the consistency of the model's gradient when training and output when inferencing. As shown in Figure 1, this approach can significantly reduce memory consumption, even accommodating unlimited context sizes. Additionally, to address task imbalance from increased dimensionality and enhance anatomical boundary focus, we propose an optimized loss function, further improving performance. As the first 3D ICL model in neuroimaging, we evaluate Neuroverse3D's potential across diverse tasks on held-out datasets. Our contributions are summarized as follows:

- We mitigate the high memory demands of 3D image contexts using a novel APSP approach with a U-shaped fusion strategy, enabling unlimited context processing for both training and inference. This significantly improves the feasibility of deploying 3D neuroimaging ICL models in clinical practice.
- We develop an optimized loss function to address class imbalance resulting from increased dimensionality and enhance focus on challenging segmentation regions and anatomical boundaries, further improving performance.
- We propose Neuroverse3D, the first 3D ICL universal model for neuroimaging, trained on extensive multi-center datasets. Neuroverse3D significantly outperforms other ICL models, achieving over 20 absolute Dice point gains across four tasks and matching task-specific performance, offering valuable insights for universal model development.

2. Related Work

2.1. Domain Challenges in Medical Imaging

Deep learning models in medical imaging frequently face domain shifts due to heterogeneous imaging data distributions, which reduce model performance on new, unseen domains. Solutions like domain adaptation [29, 53, 55, 59] require target-domain fine-tuning, limiting practical use due to the need for deep learning expertise. Domain generalization aims to produce models that generalize to new domains without fine-tuning [30, 32, 49, 58]. While it alleviates the effects of domain shift, generalization remains challenging because models lack knowledge about unseen domains during training. Different from domain adaptation and domain generalization methods, universal models have emerged as a promising solution by learning generalizable image representations from a large number of datasets.

2.2. Universal Models in Medical Imaging

Universal models in medical imaging can be classified into three categories based on their prompt types. The first category utilizes symbolic prompts, such as points or bounding boxes, as demonstrated by models like SAM [11, 34, 43]. The second category employs natural language prompts [38, 39]. The third category, also known as ICL models, uses image-label pairs as prompts [12, 14, 28, 54]. These universal models, trained on extensive datasets, exhibit exceptional generalization capabilities. Among these, image-label pairs provide rich contextual information, enabling ICL models to perform accurate segmentation on unseen tasks and even handle image generation tasks. However, the richness of these prompts in ICL models' inputs significantly increases memory consumption, a critical challenge that this study seeks to overcome.

2.3. In-Context Learning

Originally developed in natural language processing [10], ICL vision models has recently shown potential for creating universal models that can adapt to new tasks and domains by using image-label pairs as prompts to convey task-specific information. In natural image processing, ICL-based models such as LVM [7], Painter [51], and SegGPT [52] have demonstrated strong versatility across diverse tasks.

Recent studies indicate that ICL models achieve high accuracy and robust cross-domain generalization in neuroimaging, effectively addressing domain shifts across varied imaging modalities [12, 14, 28, 54]. Models like UniverSeg [12], Neuralizer [14], and One-Prompt [54] lever-

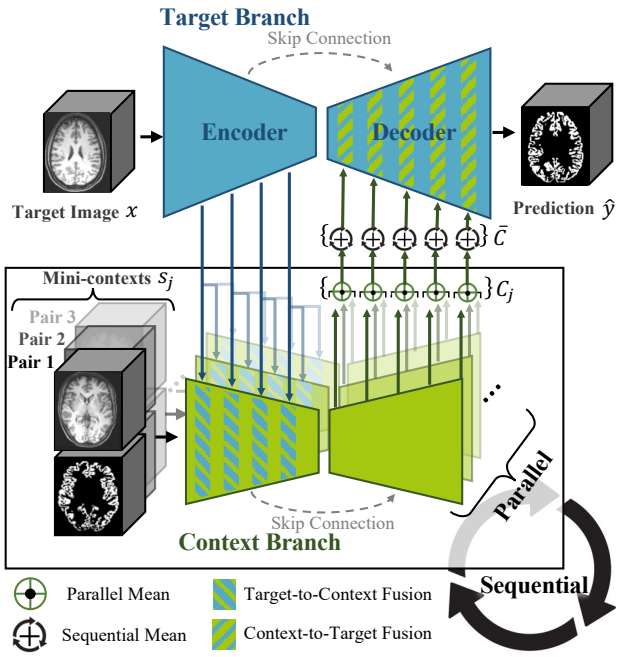


Figure 2. Illustration of our model architecture. The network consists of two branches for extracting representations from the target and context images, respectively. The Target-to-Context and Context-to-Target Fusion modules enable information exchange between the branches. Our model adaptively performs both parallel and sequential context processing.

age context sets to adapt to new domains and tasks without retraining, performing effectively in few-shot scenarios. However, current ICL models are unable to directly process 3D neuroimage data, limiting their ability to fully capture the volumetric information of 3D images.

3. Method

In-Context Learning (ICL) models for neuroimaging achieve universality by learning task-specific processing from context examples, typically image-label pairs. In the following sections, we detail our Adaptive Parallel-Sequential Processing (APSP) and U-shaped fusion strategy for efficient context processing, followed by our optimized loss function for further performance improvements.

3.1. Model

As illustrated in Figure 2, the model contains two 3D U-Net branches: a target branch for extracting target image representations and a context branch for extracting context representations. These branches communicate through the U-shape fusion strategy at each stage. The context branch receives concatenated image-label pairs as input and uses APSP to compute the mean context representation, which is then passed to the target branch decoder for final prediction

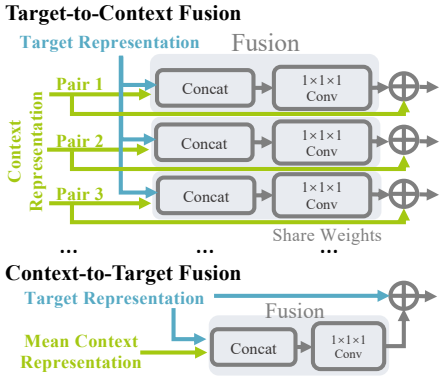


Figure 3. Illustration of the Target-to-Context Fusion and Context-to-Target Fusion modules.

generation.

Context Partition. Given a task-specific context $S = \{(x_i, y_i)\}_{i=1}^{L}$, where each (x_i, y_i) represents an image-label pair within a context set of size L, we perform context partitioning. This divides S into $n = \lceil \frac{L}{\ell} \rceil$ disjoint mini-contexts $\{s_j\}_{j=1}^{n}$, each containing ℓ image-label pairs, with ℓ determined by available memory resources. Larger values of ℓ decrease inference time but increase memory usage, and vice versa. If L is not divisible by ℓ, the final mini-context will contain fewer than ℓ pairs, ensuring that $\bigcup_{j=1}^{n} s_j = S$.

Adaptive Parallel-Sequential Processing. APSP accommodates arbitrary context partitions. Specifically, each mini-context is initially processed in parallel, and the resulting context representations are subsequently fused during sequential processing.

During parallel processing, all image-label pairs within a mini-context s_j are processed simultaneously through shared-weight 3D U-Nets in the context branch. The context representation for each mini-context s_j is computed as $C_j = g(s_j, f_{\text{enc}}(x))$, where x represents the target image, g denotes the context branch, and f_{enc} is the encoder of the target branch. Here, C_j is a set comprising the representations from each stage of the context decoder, defined as $C_j = \{\bar{c}^i_{\text{dec},j}\}_{i=1}^{\mathcal{D}}$, where $\mathcal{D}$ represents the number of decoder stages. $\bar{c}^i_{\text{dec},j}$ is computed via a parallel mean aggregation of image-label pairs inside s_j:

$$\bar{c}^i_{\text{dec},j} = \frac{1}{\ell} \sum_{k=1}^{\ell} c^i_{\text{dec},j,k}, \tag{1}$$

where $c^i_{\text{dec},j,k}$ denotes the output representation of the k-th image-label pair from the i-th decoder stage in mini-context s_j. This mean aggregation ensures the output is invariant to the order of image-label pairs within a mini-context.

The sequential computation procedure for $\bar{C}$ is outlined in Algorithm 1. The model processes each mini-context se-

Algorithm 1 Adaptive Parallel-Sequential Processing. Let x and S denote the target image and context set, respectively. $\bar{c}^i_{\text{dec},j}$ represents the context representation of the j-th mini-context at the i-th decoder stage. $\text{len}(s_j)$ indicates the number of image-label pairs in s_j.

$(x, y, S) \sim \mathcal{T}$ ▷ Sample data
$\{s_j\}_{j=1}^{n} \leftarrow S$ ▷ Split context set into mini-contexts
$\{\bar{c}^i_{\text{dec}}\}_{i=1}^{\mathcal{D}} \leftarrow \{0\}_{i=1}^{\mathcal{D}}$ ▷ Initialize representation
$w \leftarrow 0$ ▷ Initialize weight accumulator
for $j = 1, \ldots, n$ **do** ▷ Sequential processing
  $\{\bar{c}^i_{\text{dec},j}\}_{i=1}^{\mathcal{D}} = g(s_j, f_{\text{enc}}(x))$
  $\alpha \leftarrow \frac{w}{w+\text{len}(s_j)}$
  $\{\bar{c}^i_{\text{dec}}\}_{i=1}^{\mathcal{D}} \leftarrow \{\alpha \bar{c}^i_{\text{dec}} + (1-\alpha)\bar{c}^i_{\text{dec},j}\}_{i=1}^{\mathcal{D}}$ ▷ Sequential mean
  $w \leftarrow w + \text{len}(s_j)$ ▷ Update cumulative weight
  delete $\{\bar{c}^i_{\text{dec},j}\}_{i=1}^{\mathcal{D}}$ **if** $j \neq n$ ▷ Release memory
end for
$\bar{C} \leftarrow \{\bar{c}^i_{\text{dec}}\}_{i=1}^{\mathcal{D}}$
$\hat{y} \leftarrow f_{dec}(f_{\text{enc}}(x), \bar{C})$ ▷ Generate final prediction

quentially, updating $\bar{C}$ iteratively. The sequential mean operation ensures equal weighting of image-label pairs across the entire context set, making $\bar{C}$ invariant to the ordering of mini-contexts. After $\bar{C}$ is obtained, and the final prediction is computed as $\hat{y} = f_{\text{dec}}(f_{\text{enc}}(x), \bar{C})$, where f_{dec} denotes the decoder of the target branch. The combination of parallel and sequential mean operations in APSP guarantees consistent mean context representation computation for any context order or partition.

To save memory, intermediate features from processed mini-contexts are discarded after contributing to $\bar{C}$, with only the last mini-context retained for gradient computation during training. This reduces memory usage to the final mini-context rather than the entire context set. Moreover, during training, we randomly shuffle mini-contexts and scale the last mini-context's representation C_n by a factor of n during gradient computation. This ensures the expected value of the gradients computed using APSP is equivalent to the expectation when no mini-contexts are discarded, as demonstrated below:

$$\mathbb{E}_{\pi}\left[\nabla \mathcal{L}_{\text{APSP-scaled},\pi}\right] = \nabla \mathcal{L}_{\text{full}}, \tag{2}$$

where $\nabla \mathcal{L}_{\text{full}}$ and $\nabla \mathcal{L}_{\text{APSP-scaled},\pi}$ denote the gradients obtained from full-context processing and APSP, respectively. π represents the original index of the mini-context selected for gradient computation in APSP. A proof of this equivalence is provided in the supplemental Sec. A.1.

U-Shape Fusion Strategy. The U-shaped design makes feature propagation through two phases: first, from the target branch to the context branch at each encoder stage via Target-to-Context Fusion modules, and then from the context branch back to the target branch at each decoder stage through Context-to-Target Fusion modules, as illustrated in Fig. 2. This U-shaped representation flow is actually essential, as it allows the model to avoid storing features from previously processed mini-contexts, a limitation of alternative fusion strategies. Further explanation are provided in the supplemental Sec. A.2.

As illustrated in Fig. 3, the fusion components within the Target-to-Context Fusion and Context-to-Target Fusion modules are identical. The fusion operation is defined as

$$\text{Fusion}(c^i, t^i) = \text{Conv}(c^i \,||\, t^i), \tag{3}$$

where c^i and t^i denote the feature representations of the i-th stage from the context and target branches, respectively. Conv represents a convolutional operation, and $||$ signifies feature map concatenation.

Target-to-Context Fusion. This module is incorporated after each encoder stage of the context branch as follows:

$$\hat{c}^i = \mathcal{A}(\text{Fusion}(c^i, t^i) + c^i), \tag{4}$$

where $\hat{c}^i$ represents the fused context representation propagated to the next stage, and $\mathcal{A}$ denotes the activation function. Convolutional parameters are shared across all image-label pairs to enable parallel processing.

Context-to-Target Fusion. This module is integrated after each decoder stage of the target branch as follows:

$$\hat{t}^i = \mathcal{A}(\text{Fusion}(\bar{c}^i, t^i) + t^i), \tag{5}$$

where $\hat{t}^i$ denotes the fused target representation that is sent to the next decoder stage, and $\bar{c}^i$ represents the mean context representation.

3.2. Loss Function

The total loss function is defined as:

$$\mathcal{L}_{\text{total}} = \mathbb{E}_{\tau}\left[\mathbb{E}_{(x^\tau, y^\tau), S^\tau}\left[\lambda^\tau \mathcal{L}^\tau(\hat{y}^\tau, y^\tau)\right]\right],$$

where τ denotes the sampled tasks, and $\lambda^\tau \in \mathbb{R}^+$ represents the task-specific weighting coefficient. The procedure begins by sampling a task τ from all tasks, followed by sampling a target image x^τ, corresponding ground truth y^τ and a support set S^τ within that task. The task-specific loss function $\mathcal{L}^\tau$ is then applied to compute the loss between the $\hat{y}^\tau$ and y^τ.

For both segmentation and generation tasks, we employ an L_1-based loss function, with modifications to accommodate the characteristics of 3D neuroimages.

Specifically, 3D neuroimage segmentation, unlike its 2D counterpart, exhibits increased sparsity in structures like the hippocampus and amygdala due to the added dimension, resulting in greater class imbalance. To mitigate this, we propose a modified $\mathcal{L}_{\text{smooth}-L_1}$ loss:

$$\mathcal{L}^{\text{seg}}(\hat{y}, y) = \begin{cases} \frac{1}{3}|\hat{y} - y|^3, & \text{if } |\hat{y} - y| < 1, \\ |\hat{y} - y| - \frac{2}{3}, & \text{otherwise.} \end{cases} \quad (6)$$

This formulation, using a higher exponent than standard $\mathcal{L}_{\text{smooth}-L_1}$, prioritizes challenging anatomical regions over simpler ones, balancing performance across diverse tasks. We opted against Dice or Focal loss to avoid further complicating the balance between segmentation and generation tasks.

For generation tasks, we apply a standard $\mathcal{L}_{\text{smooth}-L_1}$ loss. Furthermore, to achieve better generative performance on brain images filled with intricate anatomical boundaries, we incorporate an additional $\mathcal{L}_{\text{smooth}-L_1}$ loss on the intensity difference of the image, inspired by [40]:

$$\mathcal{L}^{\text{gen}}(\hat{y}, y) = \frac{\mathcal{L}_{\text{smooth}-L_1}(\hat{y}, y)}{2} + \frac{\mathcal{L}_{\text{smooth}-L_1}(\Delta\hat{y}, \Delta y)}{2}, \quad (7)$$

where Δy denotes the intensity difference at the current voxel relative to its neighboring voxels.

4. Experiments

4.1. Data

Datasets. To ensure robust cross-center generalization and diversity, we collected 19 datasets with multiple modalities and centers, comprising 43,674 scans. The dataset includes common neuroimaging modalities including T1, T2, FLAIR, MRA, DWI, ADC, PD, and CT. Data from 15 datasets, including 38,126 scans and 5,632 segmentation masks, were used for training and validation [1, 3–5, 16, 18, 23, 25, 26, 33, 37, 44, 46, 47, 50, 56], with a random 9:1 split. Data from the remaining 4 datasets, including 5,548 images and 1,096 segmentation masks, was used as held-out datasets [2, 35, 41, 45] to assess performance on unseen centers. The held-out dataset was divided into an 8:2 split for the meta context set (from which ICL models select context) and the test set. During training, we used both the original images and images aligned to the MNI 152 template space [17] to increase data diversity. We utilized FreeSurfer [15] to generate additional anatomical segmentation labels for 3 datasets [2, 18, 47].

Data Preprocessing. To mitigate variability in image size across neuroimaging datasets, we first resampled the voxel resolution to 1 mm^3. The brain images were then rescaled to fit within a $128 \times 128 \times 128$ 3D image. Intensity values were normalized to the [0, 1] range based on the 0.02 and 0.98 intensity percentiles. Segmentation masks were binarized by assigning 0 to the background and 1 to the foreground.

Neuroimaging Tasks. With these datasets, our model was trained on multiple segmentation tasks, including anatomical segmentation [8], tumor segmentation [46], vessel segmentation [56], and generation tasks, including bias field correction [19], inpainting [42], super-resolution [36], Gaussian noise removal, salt-and-pepper noise removal [36], 2D-to-3D reconstruction [31], modality transformation [48], and skull stripping [27]. For segmentation tasks, binary masks were produced by thresholding the model outputs at 0.5.

Image and Task Augmentation. During training, we applied data augmentation across two dimensions: image and task. In addition to standard image augmentations, we enhanced the contrast diversity of images using a randomly initialized convolutional network [49]. For task augmentation, we introduced random task overlapping and also followed the method introduced in [14]. Synthetic brain data with random modalities [9] was also added. Further details are provided in supplemental Sec. B.3.

Detailed information on datasets and tasks is provided in the supplemental Sec. B.

4.2. Compared Models

Neuroverse3D. The target and context branches are based on a 5-stage 3D U-Net architecture [13], with each stage comprising two residual blocks [21] constructed from 3×3 convolutional layers. The network initiates with 32 channels in the first stage, subsequently doubling the channel count at deeper stage. GELU activation functions [22] are employed throughout the network. Neuroverse3D is trained jointly across all tasks, whereas Neuroverse3D-unseen maintains an identical architecture but excludes the evaluated task during training. Consequently, we trained a group of Neuroverse3D-unseen models for unseen task evaluation.

Task-Specific Models. We compared Neuroverse3D against 3D task-specific models, which shared identical architecture and channel configurations but lacked the context branch. These task-specific models were trained on the held-out dataset, thereby mitigating domain shift concerns. This approach simulates the scenario where medical centers traditionally train models for specific scenario without employing ICL model. Identical data augmentation strategies were applied, excluding task augmentations that could potentially disadvantage task-specific models, such as the application of Sobel filters to segmentation masks. We evaluated performance under both few-shot learning scenarios, where models were trained with only context set data, and fully supervised learning scenarios, leveraging all available data from the meta-context set.

Other In-Context Learning Models. We compared our method with state-of-the-art ICL methods, including Painter [51], SegGPT [52], UniverSeg [12], and Neuralizer [14], all designed for 2D inputs. To accommodate the 3D context, we randomly sampled axial slices containing

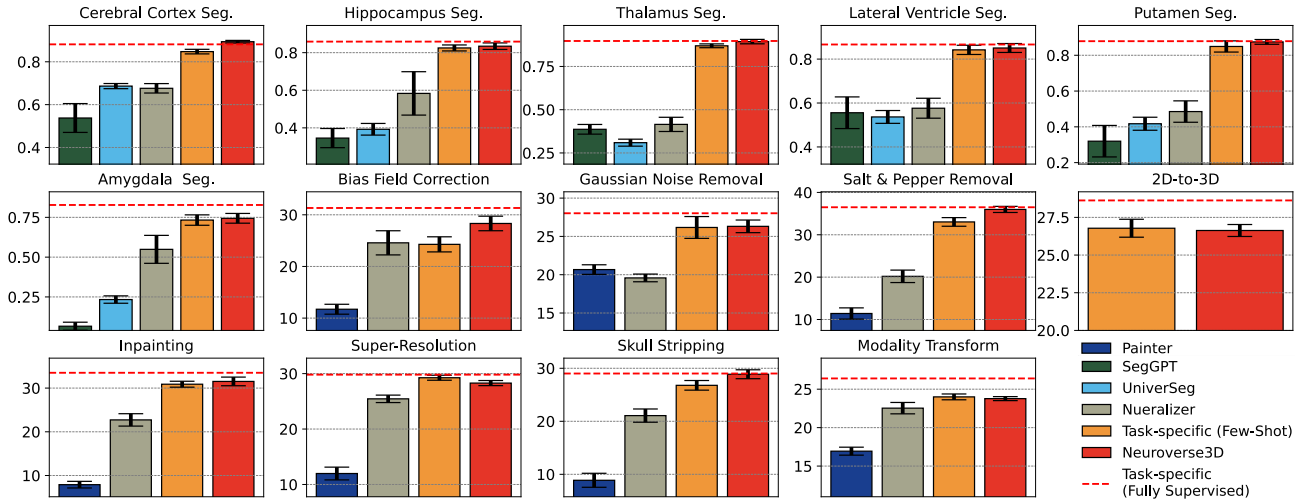


Figure 4. Performance comparison of Neuroverse3D with other models on held-out datasets, under a context size of 8. This includes ICL models trained on neuroimages (UniverSeg [12], Neuralizer [14]), models trained on natural images (Painter [51], SegGPT [52]), and task-specific models in few-shot and fully supervised settings. The Dice coefficient is used for segmentation tasks, and PSNR for generation tasks.

the segmentation target or brain to construct the 2D context set. The 2D context sizes were set to 1 for Painter [51], 8 for SegGPT [52], 32 for Neuralizer [14], and 64 for UniverSeg [12], corresponding to the reported optimal context sizes in their respective publications. The target 3D images were similarly decomposed into axial 2D slices and fed into these methods. The resulting 2D outputs were then reassembled into 3D images for metric evaluation. We downloaded the corresponding pretrained weights for these models to perform inference. UniverSeg [12] and SegGPT [52] are restricted to segmentation tasks, and all 2D models are incapable of performing 2D-to-3D reconstruction task.

Comprehensive details regarding the training and evaluation protocols are provided in supplemental Sec. C.

4.3. Model Comparison on Held-Out Datasets

As demonstrated in Fig. 4, we evaluated Neuroverse3D against SOTA ICL models and task-specific models. Comparisons with Painter were excluded for segmentation tasks due to its low Dice scores.

In segmentation tasks, Neuroverse3D significantly outperformed all other ICL models, with Dice score gains exceeding 20 percentage points for targets such as hippocampus, thalamus, lateral ventricle, and putamen. Furthermore, it surpassed the performance of few-shot models in all segmentation tasks, demonstrating that employing an ICL model in few-shot scenarios offers medical centers a more cost-effective and accurate solution. Moreover, the performance of our model closely approached that of fully supervised U-Net models. We attribute Neuroverse3D's substantial performance improvement primarily to its effective utilization of a 3D architecture, enabling it to not only better process 3D target images but also to capture the global information from the 3D context set, facilitating a more precise understanding of the segmentation targets.

In generation tasks, Neuroverse3D outperformed all other ICL models, surpassing Neuralizer and Painter across all tasks. Its performance closely matched few-shot models and nearly reached fully supervised levels in salt-and-pepper noise removal and skull stripping. This highlights the potential of 3D ICL models in generation tasks. Notably, our model's performance gain in some generation tasks was less significant. This may be attributed to certain tasks, like modality transform and super-resolution, where 2D context sufficiently conveys task information, reducing the advantage of a 3D architecture.

Fig. 5 illustrates qualitative predictions of ICL models, showcasing results for both axial and sagittal slices. As 2D ICL models generate predictions on axial slices, discontinuities are frequently observed on sagittal slices. In segmentation, compared to Neuroverse3D, other models demonstrate lower accuracy and frequent false positives/negatives in sagittal slices. In generation tasks, Neuralizer generates excessive background errors, while Painter, despite reasonable performance on Gaussian denoising, struggles with neuroimage-specific tasks like bias correction and skull stripping. Conversely, Neuroverse3D maintains consistent accuracy across all tasks.

These results demonstrate Neuroverse3D's accurate performance across diverse segmentation and generation tasks from different medical centers, attributed to its novel approach for 3D context processing and optimized loss func-

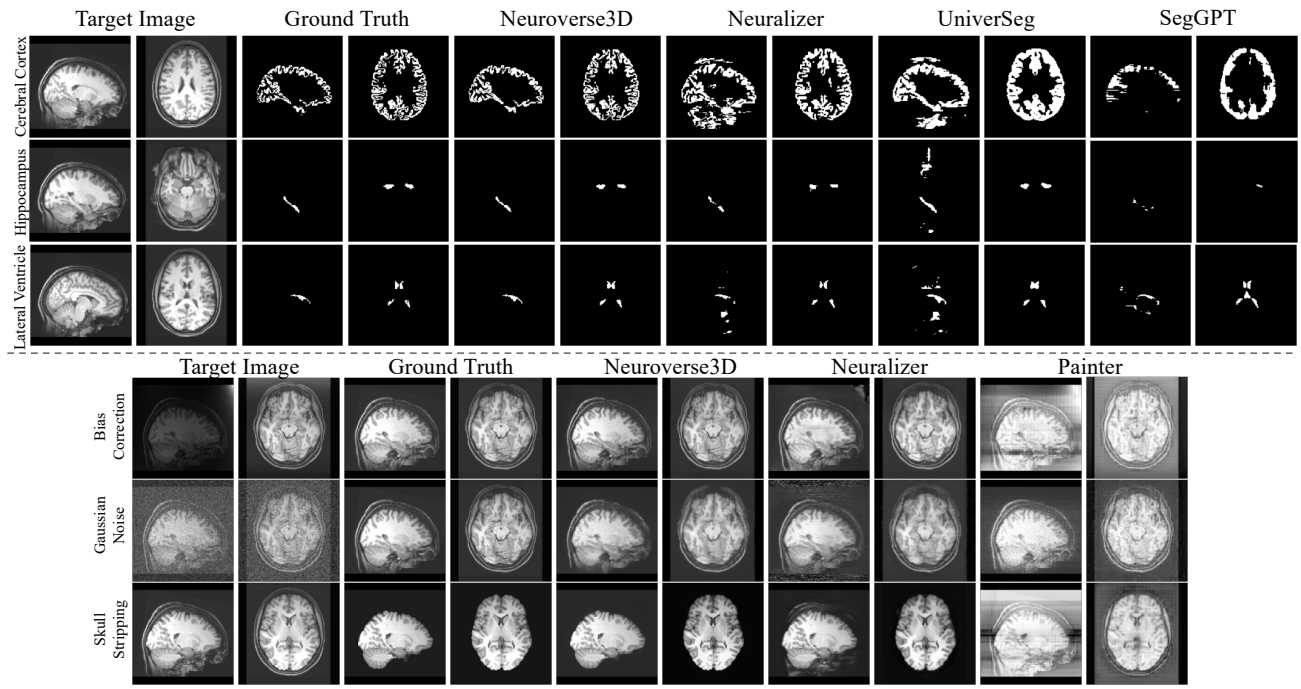


Figure 5. Qualitative results for four representative tasks are presented. To comprehensively illustrate the 3D image results, slices from both sagittal (first row) and axial (second row) views are displayed. Qualitative results for all tasks can be found in supplementary Sec. C.3

tion.

4.4. Impact of Context Size

Figure 6 illustrates model performance across varying context sizes. Our model consistently outperforms other models across all context sizes in both segmentation and generation tasks. Besides, Neuroverse3D, along with other models, shows a significant performance increase with larger context sizes, emphasizing the critical role of context quantity for ICL models. The average metric improvement resulting from increased context size diminishes as the context size increases, reaching a plateau around 16. However, the standard deviation of predictions continues to decrease, leading to more stable predictions. APSP overcomes memory limitations, enabling Neuroverse3D to support larger or even unlimited context sizes under limited resources, thereby facilitating the practical use of Neuroverse3D. SegGPT, due to memory limitations, could only support a maximum context size of 16 in our experiments.

4.5. Performance on Unseen Tasks

In Fig. 6, Neuroverse3D-unseen shows the model's ability to generalize to tasks not encountered during training, which is of clinical interest.

For segmentation tasks, Neuroverse3D-unseen exhibited a slight performance reduction compared to Neuroverse3D, with approximately a 5-point decrease in the Dice coefficient after the context size was greater than or equal to 8. However, it still significantly outperformed other ICL models. For generation tasks, Neuroverse3D-unseen showed a more substantial performance decline, although it remained marginally superior to Neuralizer. We attribute the greater performance drop in generation tasks compared to segmentation tasks to the fact that, during training, Neuroverse3D-unseen encountered a wider variety of segmentation tasks, including segmentations of over thirty brain structures and random combinations of brain structures. This led to a learning process for segmentation tasks closer to meta-learning, enabling cross-task generalization. In contrast, the training set contained fewer generation tasks, leading the model to tend to memorize each generalization task. This suggests that incorporating a greater diversity of tasks into ICL models is a promising direction for further enhancing cross-task generalization capabilities.

4.6. Memory and Time Requirements

Fig. 7 illustrates the resource consumption under various settings, where ℓ denotes the size of the mini-context, and Inf represents the maximum memory usage for a given mini-context size ℓ. When $\ell = 1$, memory consumption is limited to 7.35GB, and the model operates in a purely sequential manner, substantially reducing memory demands and simplifying deployment. Increasing ℓ results in higher memory consumption but reduces computation time. Crucially, all settings yield consistent results, demonstrating the deployment flexibility of our model.

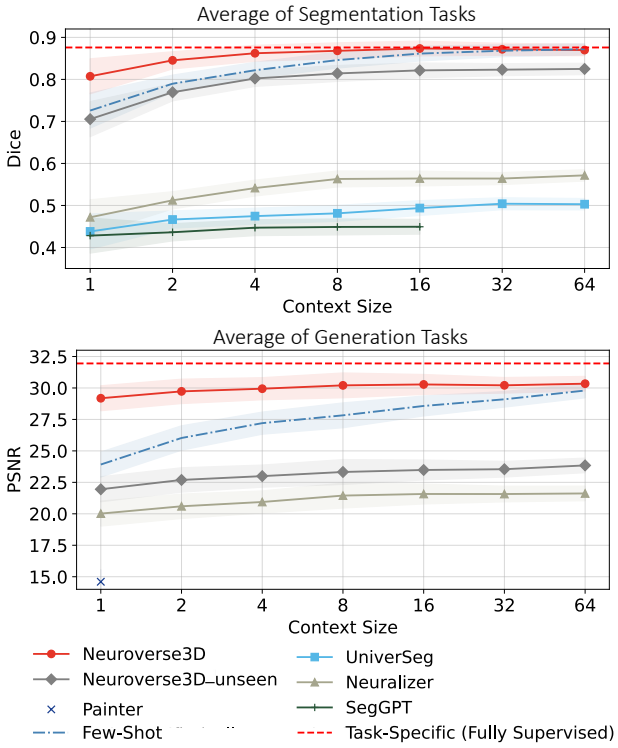


Figure 6. Performance across different 3D context sizes. A group of Neuroverse3D-unseen models was trained to assess performance on tasks not encountered during training. Due to computational constraints, segmentation task evaluations here only include cerebral cortex, hippocampus, thalamus, and lateral ventricle. Generation task evaluations include bias field correction, Gaussian noise removal, and salt-and-pepper noise removal.

Supplemental Tab. 7 presents a comparison of inference times across different models. Despite employing the most extensive context (8×128 2D slices), our model demonstrates superior inference speed compared to other ICL models. This efficiency stems from the fact that 2D models process images slice-by-slice, necessitating the recomputation of context for each slice, thereby resulting in redundant context feature computations and prolonged inference durations. This further underscore the importance of developing Neuroverse3D which inherently avoid this issue.

4.7. Ablation Study of Loss Functions

We evaluated the model's performance without the proposed modified $\mathcal{L}_{\text{smooth}-L_1}$ loss and gradient loss, as shown in Tab. 1.

Removing the modified $\mathcal{L}_{\text{smooth}-L_1}$ loss led to a decline in segmentation performance, particularly in small, complex regions such as the hippocampus (see supplemental Tab. 4 for details). This indicates that the modified $\mathcal{L}_{\text{smooth}-L_1}$ loss helps the model focus on more challenging regions, achieving a better balance in performance across various segmentation tasks.

Excluding the gradient loss resulted in a marked decrease in generation task performance, underscoring the importance of capturing edge information in brain images. Interestingly, even though the gradient loss was not applied to segmentation tasks, it improved segmentation performance. This improvement likely stems from the model's heightened sensitivity to boundaries.

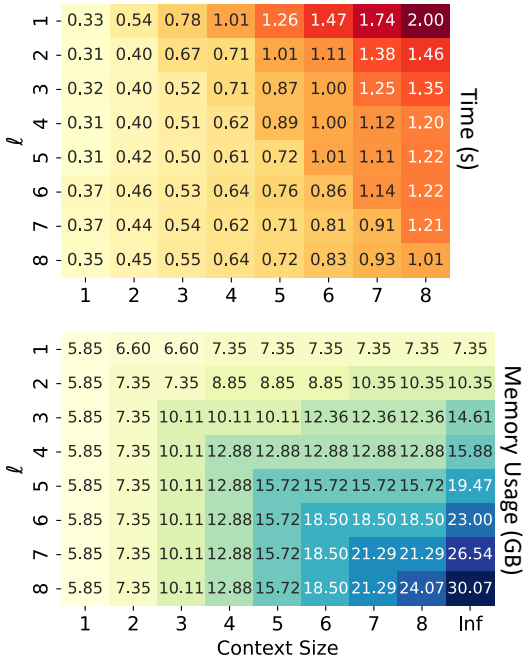


Figure 7. Time and memory consumption for different context sizes and mini-context sizes ℓ on an NVIDIA V100 GPU during inference.

Modified L1	Gradient Loss	Segmentation (Dice)	Generation (PSNR)
	✓	0.7977 ± 0.0313	28.19 ± 0.65
✓		0.8014 ± 0.0304	26.67 ± 0.69
✓	✓	$\mathbf{0.8173 \pm 0.0271}$	$\mathbf{28.38 \pm 0.76}$

Table 1. Ablation study of loss functions demonstrating the model performance across all segmentation and generation tasks, with results averaged over context sizes of 1, 2, 4, and 8.

5. Conclusion

This paper introduces Neuroverse3D, the first 3D In-Context Learning (ICL) universal model for neuroimaging, addressing the critical challenge of high memory consumption inherent in ICL. The proposed APSP and U-shaped fusion enable adaptive parallel-sequential context processing, supporting unlimited context. An optimized loss function

balances performance across diverse tasks and enhances anatomical focus. Evaluated on cross-center data, Neuroverse3D significantly outperforms other ICL models in all tasks, matching the performance of fully supervised models in segmentation. Trained on large datasets, Neuroverse3D demonstrates robust generalization and eliminates retraining, showing strong practical potential. By overcoming memory limitations for the ICL model, Neuroverse3D fundamentally paves the way for more diverse scenarios of ICL exploration such as processing other organs in 3D.

Acknowledgements

This work was supported in part by grants from the National Natural Science Foundation of P.R. China (62276081), Basic Foundation of Shenzhen Science and Technology National Key Program (CJGJZD20230724093959002), and The Major Key Project of PCL (PCL2023A09). Yanwu acknowledge the affiliation with the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

References

[1] Adhd-200. https://fcon_1000.projects.nitrc.org/indi/adhd200/. 5, 3

[2] Fcon-1000. https://fcon_1000.projects.nitrc.org/fcpClassic/FcpTable.html. 5, 3

[3] Ixi dataset. https://brain-development.org/ixi-dataset/. 5, 3

[4] Cerebral artery segmentation challenge (cas) 2023. https://codalab.lisn.upsaclay.fr/competitions/9804, 2023. 3

[5] Lindsay M Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega-Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific data*, 4 (1):1–26, 2017. 5, 3

[6] Arman Avesta, Sajid Hossain, MingDe Lin, Mariam Aboian, Harlan M Krumholz, and Sanjay Aneja. Comparing 3d, 2.5 d, and 2d approaches to brain image segmentation. *medRxiv*, pages 2022–11, 2022. 6

[7] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22861–22872, 2024. 1, 2

[8] Benjamin Billot, Douglas Greve, Koen Van Leemput, Bruce Fischl, Juan Eugenio Iglesias, and Adrian V Dalca. A learning strategy for contrast-agnostic mri segmentation. *arXiv preprint arXiv:2003.01995*, 2020. 5

[9] Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023. 5

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2

[11] Nhat-Tan Bui, Dinh-Hieu Hoang, Minh-Triet Tran, Gianfranco Doretto, Donald Adjeroh, Brijesh Patel, Arabinda Choudhary, and Ngan Le. Sam3d: Segment anything model in volumetric medical images. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2024. 2

[12] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21438–21451, 2023. 1, 2, 5, 6, 9

[13] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 1, 5

[14] Steffen Czolbe and Adrian V Dalca. Neuralizer: General neuroimage analysis without re-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6217–6230, 2023. 1, 2, 5, 6, 9

[15] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012. 5, 3

[16] Adam E Flanders, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein, Felipe C Kitamura, Matthew P Lungren, et al. Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. *Radiology: Artificial Intelligence*, 2 (3):e190211, 2020. 5, 3

[17] Vladimir S Fonov, Alan C Evans, Robert C McKinstry, C Robert Almli, and DL Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102, 2009. 5

[18] Rani Gera, Maya Bar Or, Ido Tavor, Dana Roll, Jeffrey Cockburn, Segev Barak, Elizabeth Tricomi, John P O'Doherty, and Tom Schonberg. Characterizing habit learning in the human brain at the individual and group levels: A multi-modal mri study. *NeuroImage*, 272:120002, 2023. 5, 3

[19] Tal Goldfryd, Shiri Gordon, and Tammy Riklin Raviv. Deep semi-supervised bias field correction of mr images. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1836–1840. IEEE, 2021. 5

[20] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. *Medical Image Analysis*, 98:103324, 2024. 1

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[22] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[23] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1): 762, 2022. 5, 3

[24] Malte Hoffmann, Benjamin Billot, Douglas N Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V Dalca. Synthmorph: learning contrast-invariant registration without acquired images. *IEEE transactions on medical imaging*, 41 (3):543–558, 2021. 5, 6

[25] Avram J Holmes, Marisa O Hollinshead, Timothy M O'keefe, Victor I Petrov, Gabriele R Fariello, Lawrence L Wald, Bruce Fischl, Bruce R Rosen, Ross W Mair, Joshua L Roffman, et al. Brain genomics superstruct project initial data release with structural, functional, and behavioral measures. *Scientific data*, 2(1):1–16, 2015. 5, 3

[26] Andrew Hoopes, Malte Hoffmann, Douglas N Greve, Bruce Fischl, John Guttag, and Adrian V Dalca. Learning the effect of registration hyperparameters with hypermorph. *The journal of machine learning for biomedical imaging*, 1, 2022. 5

[27] Andrew Hoopes, Jocelyn S Mora, Adrian V Dalca, Bruce Fischl, and Malte Hoffmann. Synthstrip: skull-stripping for any brain image. *NeuroImage*, 260:119474, 2022. 5

[28] Jiesi Hu, Yang Shang, Yanwu Yang, Guo Xutao, Hanyang Peng, and Ting Ma. Icl-sam: Synergizing in-context learning model and sam in medical image segmentation. In *Medical Imaging with Deep Learning*, 2024. 2

[29] Jiesi Hu, Yanwu Yang, Xutao Guo, Jinghua Wang, and Ting Ma. A chebyshev confidence guided source-free domain adaptation framework for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2024. 2

[30] Shishuai Hu, Zehui Liao, Jianpeng Zhang, and Yong Xia. Domain and content adaptive convolution based multi-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(1):233–244, 2022. 2

[31] Juan Eugenio Iglesias, Benjamin Billot, Yaël Balbastre, Azadeh Tabari, John Conklin, R Gilberto González, Daniel C Alexander, Polina Golland, Brian L Edlow, Bruce Fischl, et al. 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, 2021. 5

[32] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pages 322–348. PMLR, 2020. 2

[33] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008. 5, 3

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[35] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11): 2556–2568, 2019. 5, 3

[36] Sonia Laguna, Riana Schleicher, Benjamin Billot, Pamela Schaefer, Brenna Mckaig, Joshua N Goldstein, Kevin N Sheth, Matthew S Rosen, W Taylor Kimberly, and Juan Eugenio Iglesias. Super-resolution of portable low-field mri in real scenarios: integration with denoising and domain adaptation. In *Medical Imaging with Deep Learning*, 2022. 5

[37] Sook-Lei Liew, Bethany P Lo, Miranda R Donnelly, Artemis Zavaliangos-Petropulu, Jessica N Jeong, Giuseppe Barisano, Alexandre Hutton, Julia P Simon, Julia M Juliano, Anisha Suri, et al. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific data*, 9(1):320, 2022. 5, 3

[38] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023. 1, 2

[39] Jie Liu, Yixiao Zhang, Kang Wang, Mehmet Can Yavuz, Xiaoxi Chen, Yixuan Yuan, Haoliang Li, Yang Yang, Alan Yuille, Yucheng Tang, et al. Universal and extensible language-vision models for organ segmentation and tumor detection from abdominal computed tomography. *Medical Image Analysis*, page 103226, 2024. 2

[40] Peirong Liu, Oula Puonti, Xiaoling Hu, Daniel C Alexander, and Juan Eugenio Iglesias. Brain-id: Learning robust feature representations for brain imaging. *arXiv preprint arXiv:2311.16914*, 2023. 5

[41] Siman Liu, Yin-Shan Wang, Qing Zhang, Quan Zhou, Li-Zhi Cao, Chao Jiang, Zhe Zhang, Ning Yang, Qi Dong, Xi-Nian Zuo, et al. Chinese color nest project: An accelerated longitudinal brain-mind cohort. *Developmental Cognitive Neuroscience*, 52:101020, 2021. 5, 3

[42] Xiaofeng Liu, Fangxu Xing, Chao Yang, C-C Jay Kuo, Georges El Fakhri, and Jonghye Woo. Symmetric-constrained irregular structure inpainting for brain mri registration with tumor pathology. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part I 6*, pages 80–91. Springer, 2021. 5

[43] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 1, 2

[44] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007. 5, 3

[45] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011. 5, 3

[46] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 5, 3

[47] Allison C Nugent, Adam G Thomas, Margaret Mahoney, Alison Gibbons, Jarrod T Smith, Antoinette J Charles, Jacob S Shaw, Jeffrey D Stout, Anna M Namyst, Arshitha Basavaraj, et al. The nimh intramural healthy volunteer dataset: A comprehensive meg, mri, and behavioral resource. *Scientific Data*, 9(1):518, 2022. 5, 3

[48] Alexander FI Osman and Nissren M Tamam. Deep learning-based convolutional neural network for intramodality brain mri synthesis. *Journal of Applied Clinical Medical Physics*, 23(4):e13530, 2022. 5

[49] Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2022. 2, 5, 6

[50] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3): e1001779, 2015. 5, 3

[51] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 1, 2, 5, 6, 9

[52] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 1, 2, 5, 6, 9

[53] Yan Wang, Jian Cheng, Yixin Chen, Shuai Shao, Lanyun Zhu, Zhenzhou Wu, Tao Liu, and Haogang Zhu. Fvp: Fourier visual prompting for source-free unsupervised domain adaptation of medical image segmentation. *IEEE Transactions on Medical Imaging*, 2023. 2

[54] Junde Wu and Min Xu. One-prompt to segment all medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11302–11312, 2024. 1, 2

[55] Chen Yang, Xiaoqing Guo, Zhen Chen, and Yixuan Yuan. Source free domain adaptation for medical image segmentation with fourier style mining. *Medical Image Analysis*, 79: 102457, 2022. 2

[56] Kaiyuan Yang, Fabio Musio, Yihui Ma, Norman Juchler, Johannes C Paetzold, Rami Al-Maskari, Luciano Höher, Hongwei Bran Li, Ibrahim Ethem Hamamci, Anjany Sekuboyina, et al. Benchmarking the cow with the topcow challenge: Topology-aware anatomical segmentation of the circle of willis for cta and mra. *ArXiv*, 2023. 5, 3

[57] Yanwu Yang, Chenfei Ye, Guinan Su, Ziyao Zhang, Zhikai Chang, Hairui Chen, Piu Chan, Yue Yu, and Ting Ma. Brainmass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. *IEEE Transactions on Medical Imaging*, 2024. 1

[58] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J Wood, Holger Roth, Andriy Myronenko, Daguang Xu, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging*, 39(7):2531–2540, 2020. 2

[59] Xuzhe Zhang, Yuhao Wu, Elsa Angelini, Ang Li, Jia Guo, Jerod M Rasmussen, Thomas G O'Connor, Pathik D Wadhwa, Andrea Parolin Jackowski, Hai Li, et al. Mapseg: Unified unsupervised domain adaptation for heterogeneous medical image segmentation based on 3d masked autoencoding and pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5851–5862, 2024. 2

Neuroverse3D: Developing 3D In-Context Learning Universal Model in Neuroimaging

Supplementary Material

A. Method

A.1. Gradient Equivalence

Here, we prove that the expected value of the gradient using Adaptive Parallel-Sequential Processing (APSP) is the same as the full-context gradient when no mini-contexts are discarded. The full-context gradient, exemplified by the i-th stage output $\bar{c}^i_{\text{dec},j}$, with respect to parameters θ is defined as:

$$\nabla_\theta \mathcal{L}^i_{\text{full}} = \nabla_\theta \sum_{j=1}^{n} \frac{\text{len}(s_j)}{L} \bar{c}^i_{\text{dec},j}, \tag{8}$$

where $L = \sum_{j=1}^{n} \text{len}(s_j)$ ensures equal weighting of image-label pairs across mini-contexts. For brevity, we focus on the i-th stage, with the understanding that the analysis applies analogously to other stages.

In APSP, mini-contexts are randomly shuffled, and we select the last mini-context to compute the gradient. Numerically, this is equivalent to randomly selecting a mini-context with index π from $\{1, 2, ..., n\}$ with a uniform distribution, denoted as $\pi \sim \mathcal{U}\{1, ..., n\}$. The APSP gradient using only the π-th mini-context is:

$$\nabla_\theta \mathcal{L}^i_{\text{APSP},\pi} = \nabla_\theta \frac{\text{len}(s_\pi)}{L} \bar{c}^i_{\text{dec},\pi}. \tag{9}$$

To ensure the expected APSP gradient matches the full-context gradient, during gradient computation we scale $\bar{c}^i_{\text{dec},\pi}$ by a factor of n, making $\nabla_\theta \mathcal{L}^i_{\text{APSP-scaled},\pi} = \nabla_\theta \frac{\text{len}(s_\pi)}{L} n \bar{c}^i_{\text{dec},\pi}$ the scaled gradient for the π-th mini-context. It is important to note that the scaling factor of n is exclusively applied during the gradient computation phase and is omitted during the forward computation, thus having no impact on the forward computation of the model. The expected APSP gradient with scaling is:

$$\begin{aligned}
\mathbb{E}_\pi \left[\nabla_\theta \mathcal{L}^i_{\text{APSP-scaled},\pi} \right] &= \mathbb{E}_\pi \left[\nabla_\theta \frac{\text{len}(s_\pi)}{L} n \bar{c}^i_{\text{dec},\pi} \right] \\
&= \sum_{j=1}^{n} P(\pi = j) \left[n \nabla_\theta \frac{\text{len}(s_j)}{L} \bar{c}^i_{\text{dec},j} \right] \\
&= \sum_{j=1}^{n} \frac{1}{n} \left[n \nabla_\theta \frac{\text{len}(s_j)}{L} \bar{c}^i_{\text{dec},j} \right] \\
&= \sum_{j=1}^{n} \nabla_\theta \frac{\text{len}(s_j)}{L} \bar{c}^i_{\text{dec},j} \\
&= \nabla_\theta \sum_{j=1}^{n} \frac{\text{len}(s_j)}{L} \bar{c}^i_{\text{dec},j} \\
&= \nabla_\theta \mathcal{L}^i_{\text{full}}.
\end{aligned} \tag{10}$$

Equation (10) shows that the expected APSP gradient equals the full-context gradient $\nabla_\theta \mathcal{L}^i_{\text{full}}$. This analysis, exemplified for the i-th stage, extends to all network stages and parameters.

Thus, by employing the gradient of only the last mini-context and scaling by n, APSP preserves the full-context gradient expectation while enabling memory-efficient computation.

A.2. U-Shape Fusion Strategy Explanation

In Figure 8, (a) illustrates our proposed U-Shape fusion strategy, while (b) shows a fusion strategy with alternating feature transmission, similar to those presented in [12] and [14] for comparison in the following analysis. The red semi-transparent arrows indicate the order of computation for feature maps in the network.

In case (b), there is a problematic trade-off between memory usage and computational cost. For example, when computing layer 2 of the target branch, it is necessary to calculate the feature maps for all image-label pairs in the context branch at layer 2, which are then passed to the target branch. After this, two options arise:

1. Retaining the context branch layer 2 features for subsequent computations requires storing feature maps for all image-label pairs across all mini-contexts during sequential processing, which significantly increases memory usage.
2. Discarding the context branch layer 2 features saves memory but requires recomputing them for subsequent layers, which substantially increases computational cost.

Both options severely hinder efficient sequential processing. In contrast, strategy (a) avoids these issues. For each image-label pair, all required context representations are obtained in a single pass, allowing computed feature maps to be discarded during sequential processing, thereby ensuring computational efficiency.

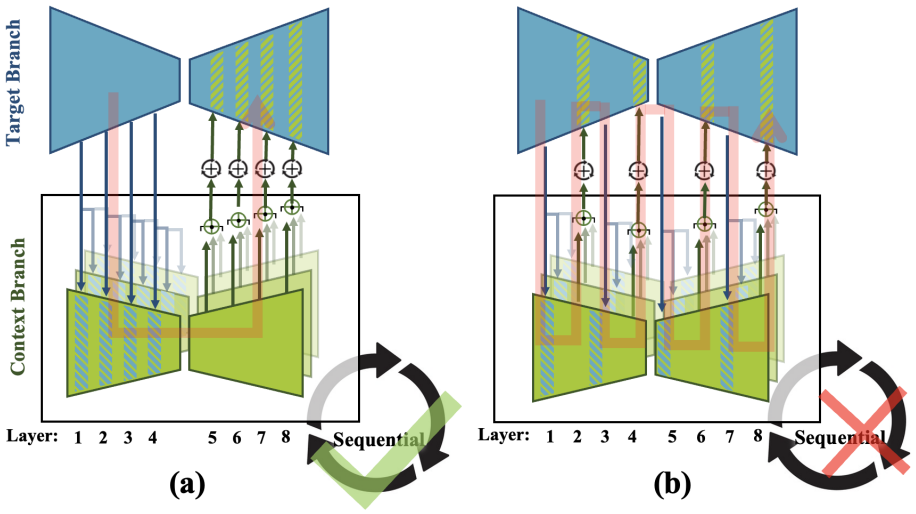


Figure 8. Illustration of different fusion strategies. (a) The proposed U-Shape fusion strategy. (b) The alternating fusion strategy. The red semi-transparent arrows indicate the computation order of the feature maps in the network.

B. Data

B.1. Domain Shift Between Training and Test Data

We trained a discriminator to distinguish between images from the training set and those from the held-out set, achieving accuracy, precision, recall, and F1 scores of 0.934, 0.932, 0.918, and 0.925, respectively. Additionally, we trained a nnUNet model for brain anatomical segmentation using the training set. These results serve as proxy evidence for the presence of domain shift.

B.2. Data Sampling

During training, tasks are selected based on a predefined sampling rate, followed by the random selection of a dataset within the chosen task with equal probability. Given a context size L, we randomly sample $L + 1$ image-label pairs from the training set of the selected dataset. One of these samples is alternately designated as the target image and ground truth, while the remaining L samples serve as the context set for the model. This process generates $L + 1$ unique target image and corresponding context set, significantly reducing the time consumption associated with data I/O.

Since the model is designed for binary classification, for multi-class segmentation datasets, we iteratively select each class as the foreground and the remaining classes as the background during training.

In generation tasks, we simulate various scenarios. For bias field correction, 3D bias fields are generated using Legendre polynomials with random coefficients. Gaussian noise removal involves simulating noise with a mean of 0 and a standard deviation randomly selected within the range 0.15 to 0.25. For salt-and-pepper noise removal, noise is applied with random probabilities equal to 0.04, where salt noise (value of 1) and pepper noise (value of 0) are added separately. For the inpainting task, binary masks are created using random 3D Perlin noise to occlude specific regions of the input image. The 2D-to-3D

Type for use	Dataset	Task	# Scans	# Masks	Modality
Training and Validation Set	TopCow[56]	Seg., Gen.	90	90	MRA
	CAS2023[4]	Seg., Gen.	100	100	MRA
	ISLES2022[23]	Gen., Mod.	750	0	DWI, ADC, FLAIR
	ATLAS[37]	Seg., Gen.	655	655	T1w
	IXI[3]	Gen., Mod.	2268	0	T1, T2, MRA, PD
	ICH Unlabeled[16]	Gen.	2000	0	CT
	ADHD[1]	Gen.	950	0	T1
	ADNI[33]	Gen., Mod.	9923	0	T1
	CMI[5]	Gen.	5146	0	T1
	GSP[25]	Gen.	2616	0	T1
	HAB[18]	Seg., Gen.	460	460	T1
	NIMH[47]	Seg., Gen.	248	248	T1
	OASIS[44]	Gen.	3916	828	T1
	UKBiobank[50]	Seg., Gen.	4000	2000	T1, T2
	BraTS[46]	Seg., Gen., Mod.	5004	1251	FLAIR, T1, T1CE, T2
Held-out Set	WMH[35]	Mod.	120	0	T1, FLAIR
	CCNP[41]	Gen.	1580	0	T1
	FCON1000[2]	Seg., Gen.	1096	1096	T1
	PPMI[45]	Gen.	2752	0	T1
	Total	Seg., Gen., Mod.	43674	6728	T1, T2, FLAIR, MRA, DWI, ADC, PD, CT

Table 2. Summary of Datasets. Seg., Gen., and Mod. represent segmentation, generation (excluding modality transformation), and modality transformation tasks, respectively.

task requires the model to reconstruct a complete 3D brain volume from only three central brain slices, with this task restricted to images in MNI space. In the super-resolution task, images are downsampled by a factor of 2. For skull stripping, input images include the skull, while ground truth images, with the skull removed, are generated using FreeSurfer [15]. Lastly, the modality transformation task uses registered pairs of different imaging modalities as input and output.

Figure 9 illustrates the target image, ground truth, and context set (with two image-label pairs shown as examples) for all tasks. The model takes the target image and context set as input, using the context set to infer the required task.

Task	Sampled Rate	Weight
Segmentation	2	50
Bias Remove	1	1
Gaussian Noise Remove	1	1
Salt & Pepper Noise Remove	1	1
2D to 3D Generation	1	0.5
Inpainting	1	1
Super-Resolution	1	1
Skull Stripping	1	1
Modality Transform	1	1

Table 3. Sampling rate and weight assigned to each task when training.

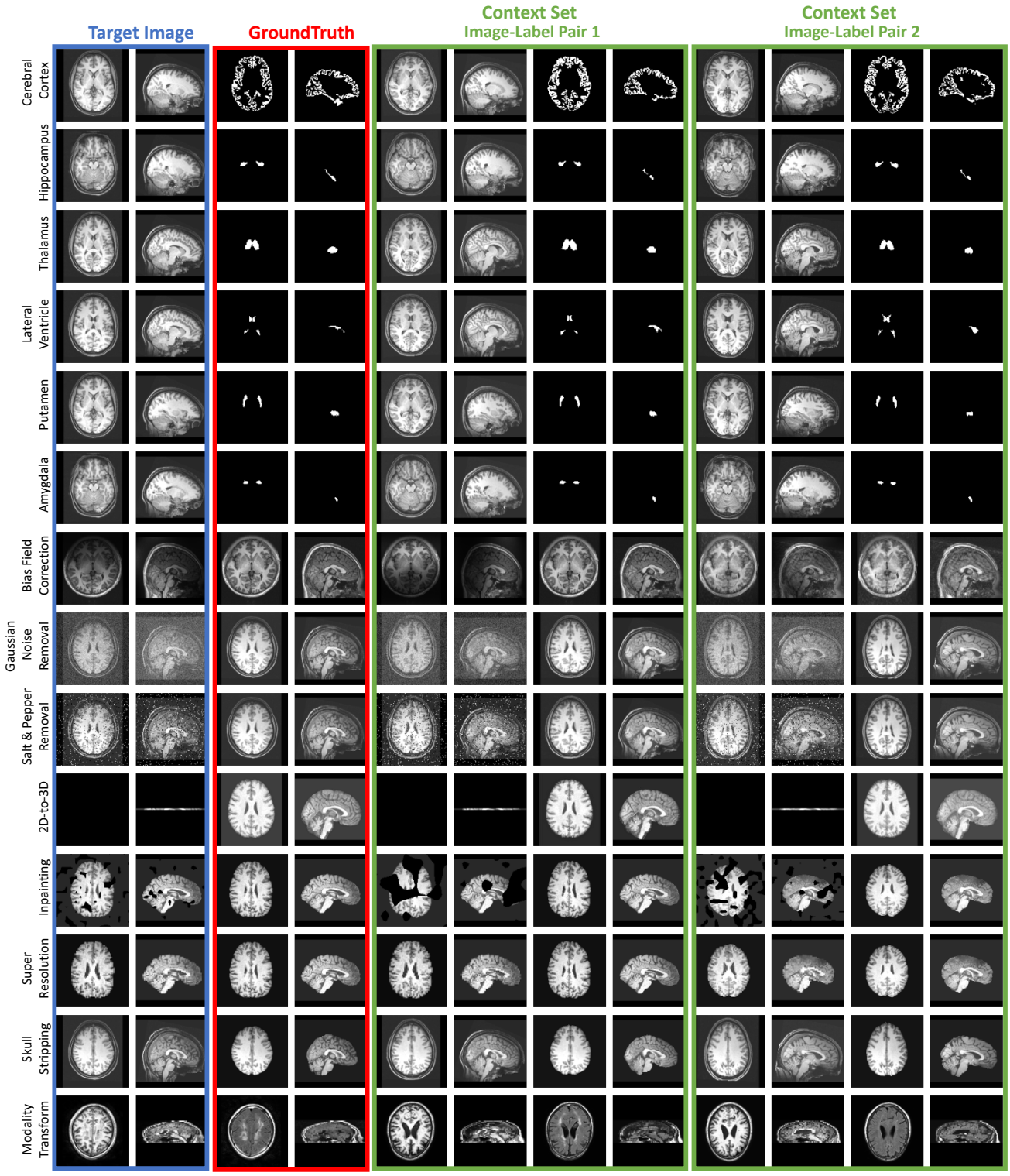


Figure 9. Visualization of the target image, ground truth, and image-label pairs in the context set.

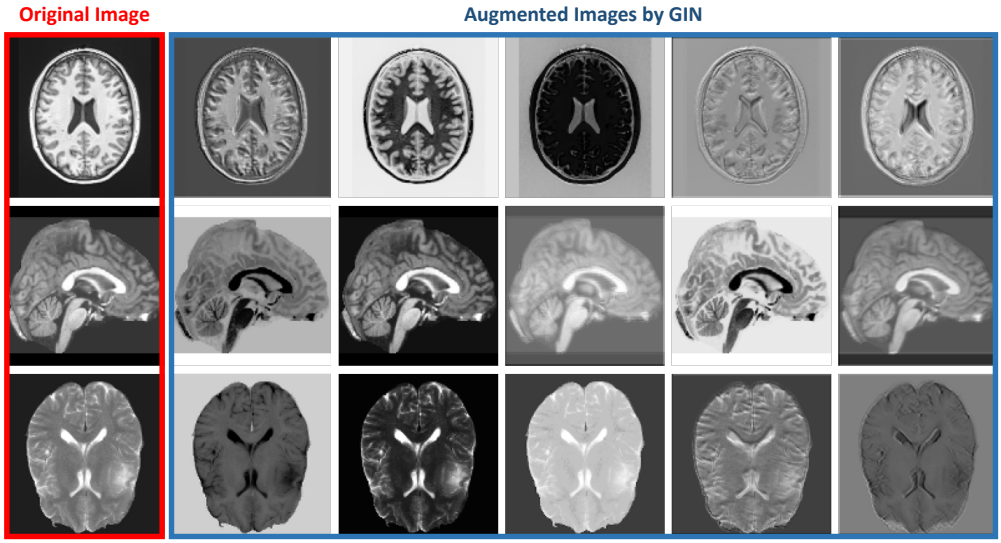


Figure 10. Visualization of the images augmented by GIN [49].

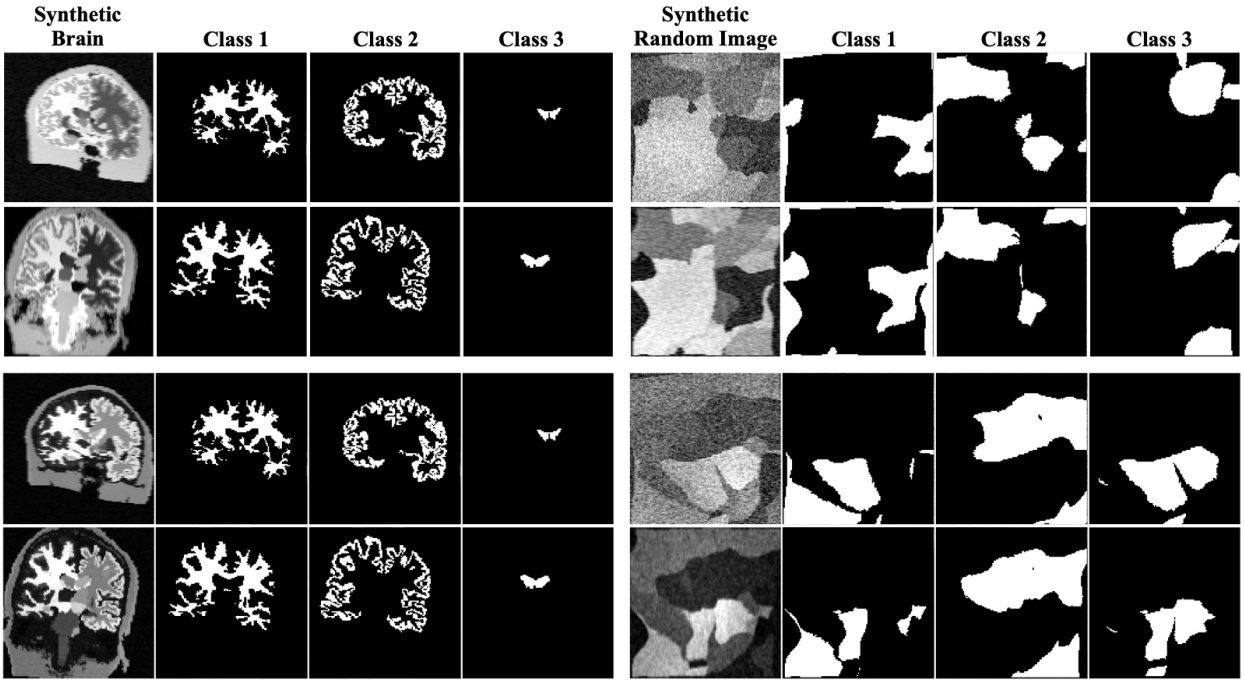


Figure 11. Visualization of the synthetic images [12, 24] for the segmentation task and corresponding segmentation masks.

B.3. Image and Task Augmentation

Image and task augmentation are performed after obtaining data sampling. For image augmentation, we applied the following transformations: random affine transformations (p=0.05), elastic deformations (p=0.05), flips (p=0.05), and rotations (p=0.05). To extend the diversity of input images, we introduced random intensity shifts (p=0.2), intensity scaling (p=0.2), Gaussian noise (p=0.1), intensity inversion (p=0.05), and contrast enhancement using a random convolutional network called GIN [49] (p=0.05). Figure 10 showcases examples of applying GIN to process images, resulting in the generation of random and diverse modalities.

For task augmentation, we employed several methods as follows:

- **Random Task Overlapping**: To enhance the model's ability to perform multiple tasks in a single run, we randomly overlay tasks during training: bias field correction, Gaussian noise removal, salt and pepper noise removal, inpainting, and super-resolution, each with a probability of 0.05. These overlays are independent.
- **Random Modalities** (p=0.05): In modality transformation tasks, we also applied GIN [49] to the target and ground truth images. The same parameters were used for both context and target images.
- **Random Foreground** (p=0.5): In multi-class segmentation datasets, multiple classes were randomly grouped into a single foreground class, with remaining classes assigned as background. This approach increases the variety of segmentation tasks. Specifically, we randomly sampled k classes in the dataset (capped at 10) to form the foreground.
- **Sobel Filter** (p=0.05): In segmentation tasks, the Sobel filter was applied to the labels, with the model tasked to predict the filtered output.
- **Mask Inversion** (p=0.05): In segmentation tasks, the foreground and background masks were swapped to introduce variability.
- **Random Dilation** (p=0.05): The segmentation masks of the target and context are dilated by 1 voxel.
- **Random Erosion** (p=0.05): The segmentation masks of the target and context are eroded by 1 voxel.

Each image and task augmentation was applied independently according to its probability, and these augmentations are mutually compatible.

Synthetic Data. To enhance generalization, we incorporated synthetic data, following [12, 24], as shown in Figure 11. This added 100 segmentation datasets with varying contrasts, each containing 100 3D image samples.

C. Experiments

C.1. Training

Neuroverse3D was trained on eight NVIDIA V100 GPUs, with a batch size of 1 per GPU, using the ADAM optimizer. Training ran for 120K steps with an initial learning rate of 10^{-4}. Validation loss was evaluated every 1.2K steps, and the learning rate was halved if no improvement was observed over 20 evaluations. To improve training efficiency, the context size and mini-context size were fixed at 3 for the first 100K steps, requiring the model to process only one mini-context. Then, the context size was uniformly selected between 1 and 8 for the final 20K steps. Total training took approximately 8 days, with the model achieving the lowest validation loss selected. All task-specific models were trained on a single GPU for 36K steps, due to the much smaller training set and faster convergence compared to the multi-task ICL model.

C.2. Evaluation

Performance was assessed using the Dice coefficient for segmentation tasks and the Peak Signal-to-Noise Ratio (PSNR) for generation tasks on held-out datasets. Each task and setting were evaluated 10 times with randomly selected context sets to compute the average and standard deviation, ensuring robust performance measurements.

C.3. Qualitative Results

Figure 12 present the comparisons between our model and other ICL models. In segmentation tasks, other ICL models often produce false positive results on background slices, while our Neuroverse3D effectively captures global information, eliminating this issue. For generation tasks, Neuroverse3D demonstrates improved slice-to-slice consistency compared to 2D ICL models, underscoring the critical importance of leveraging a 3D model.

Moreover, Painter, an ICL model trained on natural images, struggles to handle these specialized medical imaging tasks despite being exposed to extensive data and having a large number of parameters. This highlights that for ICL models, merely providing context might be insufficient for adaptation to the medical domain without incorporating domain-specific knowledge.

Figure 13 presents the comparison results between our model and task-specific models. Overall, in 3D scenarios, with a limited number of samples and well designed data augmentation, few-shot task-specific models can achieve impressive results, which aligns with findings from previous studies [6]. For 3D segmentation tasks, the performance of Neuroverse3D is on par with both few-shot and fully supervised models.

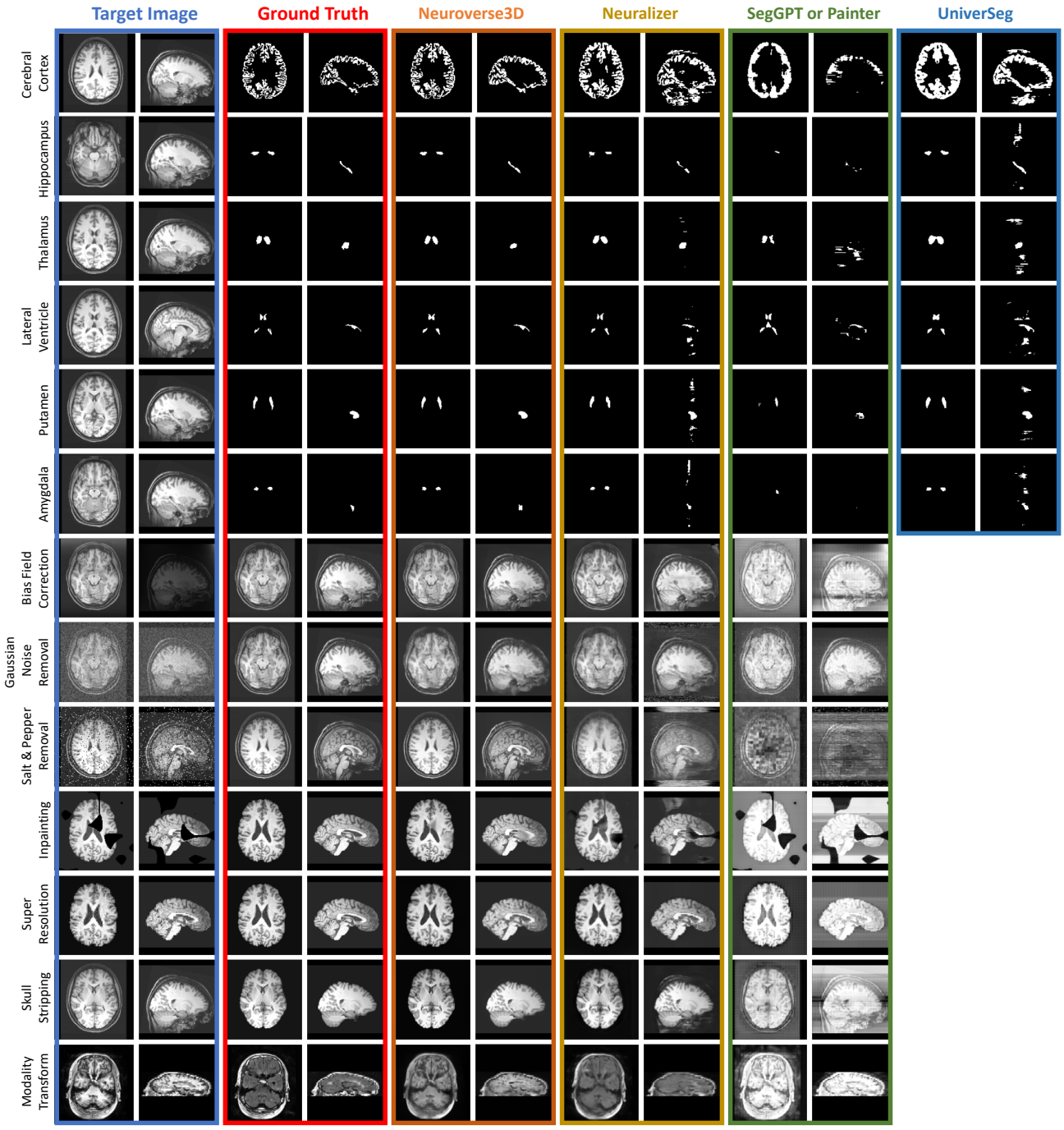


Figure 12. Qualitative results comparing of ICL models. In the SegGPT or Painter column, the results for segmentation tasks are from SegGPT, and the results for generation tasks are from Painter.

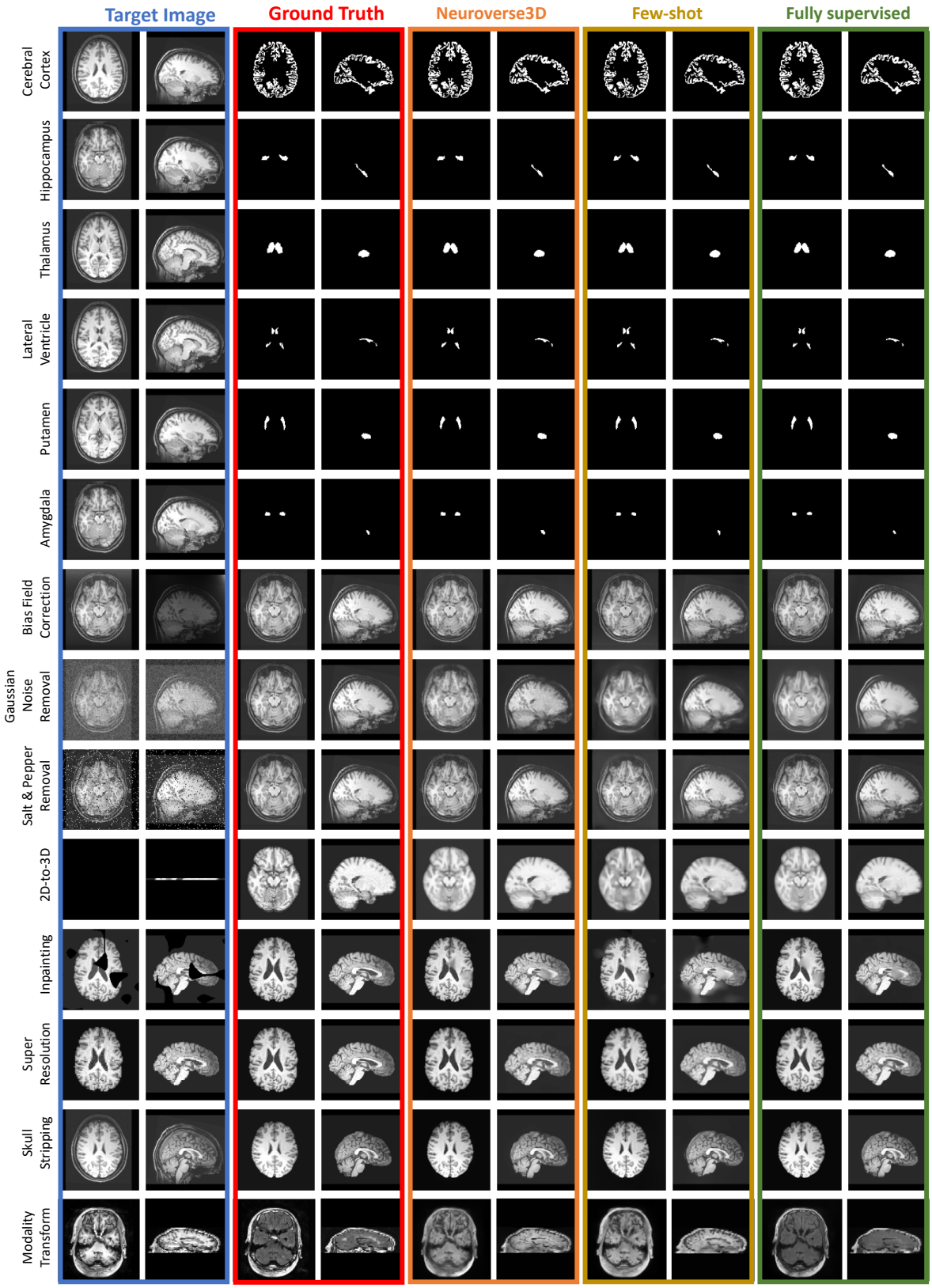


Figure 13. Qualitative comparison with task-specific models.

Smodified L1	Gradient Loss	Cerebral Cortex	Hippocampus	Thalamus	Lateral Ventricle	Putamen	Amygdala	Average
	✓	0.8847	0.7618	0.8586	0.8150	0.8220	0.6439	0.7977
✓		0.8724	0.7791	0.8648	0.8307	0.8167	0.6446	0.8014
✓	✓	0.8898	0.7917	0.8726	0.8290	0.8398	0.6812	0.8173

Table 4. Detailed ablation study of segmentation tasks.

Smodified L1	Gradient Loss	Bias Field Correction	Gaussian Noise Removal	Salt-and-Pepper Noise Removal	2D-to-3D	Inpainting	Super Resolution	Skull Stripping	Modality Transformation	Average
	✓	27.63	25.57	35.85	25.50	31.49	27.71	27.99	23.79	28.19
✓		24.16	24.93	30.82	25.55	29.19	27.48	27.44	23.81	26.67
✓	✓	27.68	25.89	35.72	26.08	31.21	28.08	28.61	23.74	28.38

Table 5. Detailed ablation study of generation tasks.

C.4. Inference Cost and Model Size

Table 6 summarizes the FLOPs and parameter counts for the models during inference. The task-specific model is implemented as a 5-stage U-Net with channel sizes of (32, 64, 128, 256, 512). Similarly, both the target and context branches of Neuroverse3D utilize a 5-stage U-Net with the same channel configuration.

Model	Parameters (M)	Inference TFLOPs
Task-specific	35.02	1.71
Neuroverse3D ($L = 1$)	70.85	4.35
Neuroverse3D ($L = 8$)	70.85	16.8

Table 6. Model parameters and inference FLOPs.

	Inference Time (s)	Context (pair)	Parameters (M)
Neuroverse3D	1.01	8 3D	70.85
Neuralizer [14]	4.96	32 2D	1.27
UniverSeg [12]	8.36	64 2D	1.18
Painter [51]	31.35	1 2D	307.72
SegGPT [52]	184.89	8 2D	307.72

Table 7. Inference time for a single 3D image (128 2D slices) and the corresponding model settings on a V100 GPU. For other comparison methods, we adopt the optimal context settings reported in their respective papers, consistent with the settings in Fig. 4.

C.5. Different Fusion Stage Configurations

Fusion Stage(s)	5	5–4	5–3	5–2	5–1 (Ours)
Segmentation (Dice)	0.8019	0.8118	0.8274	0.8347	**0.8446**
PET Segmentation (Dice)	0.2870	0.3162	0.3714	0.4945	**0.5314**
Generation (PSNR)	25.82	26.36	27.41	27.94	**28.87**

Table 8. Performance under different fusion stage configurations. Stage 5 corresponds to the deepest stage.

Tab. 8 summarizes the results of the ablation study on fusion stages. Denser connections improved the model's performance, especially on PET segmentation—an unseen modality.

Context Modality	Bias Removal	Gaussian	Salt & Pepper	Inpainting	Super-res.	2D to 3D
T1	**28.32**	**26.31**	**35.98**	**31.52**	**28.29**	**26.63**
T2	24.93	18.88	34.86	29.18	25.34	14.15
CT	24.95	23.32	33.97	28.67	24.62	12.27

Table 9. Generation performance (PSNR) under different context modalities, with T1 as the target modality.

C.6. Impact of Different Context Modalities

Tab. 9 presents the impact of different context modalities. For all tasks, performance is optimal when the context modality matches the target modality. The performance drop in salt-and-pepper noise removal is relatively small, potentially because this task is less dependent on contextual information. In contrast, Gaussian noise removal and 2D-to-3D transformation show more pronounced performance degradation, likely due to their greater reliance on context to convey task-relevant knowledge. These findings underscore the importance of modality alignment for the effective performance of ICL models.