AmygdalaGo-BOLT3D: A boundary learning transformer for tracing human amygdala

Bo Dong[1,2†], Quan Zhou[3,4†], Peng Gao[3,4†], Jintao Wei[1,2], Jiale Xiao[4], Wei Wang[4], Peipeng Liang[5], Danhua Lin[4], Hongjian He[1,6,7*], Xi-Nian Zuo[3,8,9,10*]

[1]Center for Brain Imaging Science and Technology, Zhejiang University, Hangzhou, 310027, Zhejiang, China.
[2]College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, 310027, China.
[3]State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, 100875, China.
[4]Faculty of Psychology, Beijing Normal University, Beijing, 100875, Beijing, China.
[5]School of Psychology, Capital Normal University, Beijing, 100190, China.
[6]School of Physics, Zhejiang University, Hangzhou, 310058, Zhejiang, China.
[7]State Key Laboratory of Brain-Machine Intelligence, Zhejiang University, Hangzhou, 310058, Zhejiang, China.
[8]Developmental Population Neuroscience Research Center, IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, 100875, China.
[9]National Basic Science Data Center, Beijing, 100190, China.
[10]Developmental Population Neuroscience Research Base for Education, Nanning Normal University, Nanning, 530001, China.

*Corresponding author(s). E-mail(s): hhezju@zju.edu.cn; xinian.zuo@bnu.edu.cn;
†These authors contributed equally to this work.

Abstract

Automated amygdala segmentation is one of the most common tasks in human neuroscience research. However, due to the small volume of the human amygdala, especially in developing brains, the precision and consistency of the segmentation results are often affected by individual differences and inconsistencies in data distribution. To address these challenges, we propose an algorithm for learning boundary contrast of 427 manually traced amygdalae in children and adolescents to generate a transformer, AmygdalaGo-BOLT3D, for automatic segmentation of human amygdala. This method focuses on the boundary to effectively address the issue of false positive recognition and inaccurate edges due to small amygdala volume. Firstly, AmygdalaGo-BOLT3D develops a basic architecture for an adaptive cooperation network with multiple granularities. Secondly, AmygdalaGo-BOLT3D builds the self-attention-based consistency module to address generalizability problems arising from individual differences and inconsistent data distributions. Third, AmygdalaGo-BOLT3D adapts the original sample-mask model for the amygdala scene, which consists of three parts, namely a lightweight volumetric feature encoder, a 3D cue encoder, and a volume mask decoder, to improve the generalized segmentation of the model. Finally, AmygdalaGo-BOLT3D implements a boundary contrastive learning framework that utilizes the interaction mechanism between a prior cue and the embedded magnetic resonance images to achieve effective integration between the two. Experimental results demonstrate that predictions of the overall structure and boundaries of the human amygdala exhibit highly improved precision and help maintain stability in multiple age groups and imaging centers. This verifies the stability and generalization of the algorithm designed for multiple tasks. AmygdalaGo-BOLT3D has been deployed for the community (GITHUB_LINK) to provide an open science foundation for its applications in population neuroscience.

Keywords: Amygdala, Transformer, Artificial Intelligence, Neuroscience, Development

Introduction

The amygdala is indispensable for mood regulation, cognitive learning, and memory consolidation, and has been closely associated with behavioral pathologies, particularly during critical stages of socio-emotional maturation such as childhood and adolescence (Sammallahti et al, 2023; Chai et al, 2023; Paré and Headley, 2023). Magnetic Resonance Imaging (MRI) of brain tissue structures enables a detailed examination and volumetric analysis of the amygdala, facilitating the construction of developmental trajectories and the investigation of neurodevelopmental processes underlying various

neuropsychiatric conditions (Song, 2023; Ben-Zion et al, 2023). Therefore, a volumetric analysis based on a precise segmentation of the amygdala is essential for these studies.

Amygdala segmentation has relied on histological atlases and manual segmentation techniques utilizing in vivo neuroimaging data. However, while serving as a golden standard, these methods predominantly involve manual labeling, making them labor intensive and prone to errors. Given the necessity to process large datasets, this task becomes particularly cumbersome and requires the development of automatic segmentation methods. A frequently used technique in automatic methods involves probabilistic estimates derived from Markov random fields (Fischl et al, 2002) to assign voxel labels in anatomical MRI images. This approach has been widely implemented in popular software such as FreeSurfer (Fischl et al, 2002), achieving notable segmentation performance (Trombini et al, 2023). Other research has used Bayesian methods to delineate brain anatomical boundaries (Patenaude et al, 2011), based on the signal intensity of MRI images and the structural configurations defined by established atlases, such as those used in FSL-FIRST and similar segmentation algorithms (Liu and Chou, 2023). In addition, multi-atlas label fusion segmentation techniques (Xie et al, 2023) integrate information from multiple atlases to reduce mislabeling errors resulting from inaccurate registrations and mitigate the inherent instability of MRI data. This methodology has proven particularly effective in scenarios involving high variability in Markov random fields covering individual disparities and pathological lesions, and has been incorporated into volBrain (Manjón and Coupé, 2016) and more recent algorithms (Xie et al, 2023; Liu et al, 2023b).

Despite the plethora of available automated brain image segmentation tools, including FreeSurfer and volBrain, it is noteworthy that these toolkits were not developed exclusively for amygdala segmentation. Recent investigations have revealed significant discrepancies in the quantitative estimates of amygdala volume when using existing software algorithms, which can affect the validity of amygdala-centric research

findings (Zhou et al, 2021). That is, inaccurate segmentation can introduce systematic biases, compromising the precision and reliability of volumetric measurements.

In summary, the accurate segmentation of the human amygdala poses several challenges. Firstly, the amygdala constitutes merely 0.05% of the entire brain, making it susceptible to false positive segmentations. Secondly, although there are distinct class differences between the amygdala and adjacent tissue structures, these differences are minimal, only about 12.6%, compared to surrounding tissues. The complexity of neighboring structures further complicates the segmentation task due to low contrast issues (Pruessner et al, 2000; Fischl et al, 2002; Manjón and Coupé, 2016). In this *Methods* paper, we propose an innovative algorithm to take advantage of manually traced large-scale amygdalae (Zhou et al, 2021) for the transformer model (Lee et al, 2022) that emphasizes boundary delineation to address the challenges mentioned above in automated amygdala segmentation.

Method

We delineate the amygdala segmentation into three critical phases: 1) robust global localization, 2) precise delineation and 3) generalizable segmentation. In examining the structure of the amygdala, the observer must first capture the entire cerebral image and subsequently estimate the approximate position of the amygdala. This estimate relies on prior knowledge of the anatomy of the limbic system, where the amygdala is located, and considers adjacent structures such as the hippocampus. Subsequent to this initial phase, the observer leverages a meticulous understanding of the amygdala's internal morphology, gray matter intensities, and other pertinent information to execute a detailed regional delineation. To enhance the segmentation process, a transformer-based algorithm is employed to discern boundary contrasts. This approach aims to improve both the robustness of global amygdala localization and the precision

of boundary delineation, thus improving inter-individual consistency, and includes the following key elements.

1. To achieve comprehensive detection of amygdala positioning in T1-weighted (T1w) images, we employed a transformer-based network named BOLT3D. The BOLT3D model constructs a multigranularity adaptive collaboration (MGAC) architecture, integrating a global coarse-grained perceptual module to elucidate the global semantics of the amygdala structure and enhance positional sensitivity. Given that intricate details are often neglected in global representations, a local fine-grained enhancement module was developed to augment the perception and retention of detailed information. To harmonize coarse-grained global and fine-grained local data, a multi-granular adaptive fusion module is introduced. This allows the model to concurrently consider the differences between the entire amygdala and adjacent structures, as well as the similarities within the amygdala. This design facilitates the recognition of global positioning and local disparity features, thereby improving the localization accuracy and robustness of amygdala structures. The overarching design objective is to accomplish a more precise and exhaustive detection of amygdala localization contours.
2. To address the challenge of dependency on manual segmentation boundaries and enhance the congruence between predicted and actual delineations, a novel learning strategy for boundary contrasts is introduced. This method enhances the discriminability of limbic structures from adjacent tissues by amplifying the characteristic differences between the amygdala and its surrounding tissues, while minimizing the intra-amygdala variations. Consequently, this refines the alignment between automated boundary segmentation and manual segmentation outcomes, thereby enabling a more precise demarcation of the amygdala boundaries.
3. To ensure the generalization efficacy of BOLT3D, a Segment Anything model (SAM) is devised for the refinement of coarse predictions. Specifically, the coarse

outputs of the MAGC module are sampled as point cues for the SAM model, utilizing the coarse segmentation as a pre-segmentation step for further refinement. Concurrently, the original SAM model is adapted to the amygdala context, comprising three essential components: a lightweight volumetric feature encoder, a 3D cue encoder, and a volumetric mask decoder, thereby enhancing the model's generalized segmentation performance.

Multi-granularity Adaptive Collaboration

Initially, we derived global and local semantic features from the original MRI image data employing a multigranularity semantic generation module as depicted in Figure 1(b). Specifically, for T1w image data I, the multigranularity semantic generation module exploits the local induction capability of convolution by employing two 3D convolutions to extract the primary local features F_l. This module facilitates the computation of the interrelationships among feature points, thereby enabling the extraction of global semantic information from regions that exhibit high response in the initial position features. Subsequently, we construct the preliminary global feature F_g. Given a feature map $F^{'}$ of dimensions $d \times H \times W$, $F^{'}$ is derived from the input image I via a single convolutional layer and subsequently projected onto a weight map and a mapping $F^{'}$ through two convolutional layers. The channel count on the weight map is hw ($h \leq H$, $w \leq W$), where h and w represent the dimensions of the predefined global semantic map. The weight map is then flattened for the Softmax computation, and the resulting weights are used to aggregate the semantic information within the $F^{'}$ mapping by matrix multiplication, resulting in F_g. For succinctness in the illustration, dimensions are presented in abbreviated form.

The MGAC module is composed of three components: a global module for coarse-grained perception, a local module for fine-grained enhancement, and an adaptive module for multigranularity fusion, as depicted in Figure 1(c). These components

independently optimize global and local features and subsequently integrate them interactively to enhance the model's performance. Figure 1(b) illustrates the positioning of the adaptive MGAC module within the overall architecture, while Figure 1(c) details its internal structure.

Each MGAC module receives two inputs, specifically global coarse-grained semantic characteristics F_g and local fine-grained semantic features F_l. To mitigate the significant semantic distribution discrepancy between these two sets of characteristics, BatchNorm3D is used to normalize the inputs. Furthermore, the linear position projection from previous work is mapped according to the positional relationships of the elements, ensuring that the arrangement of tokens within the transformer remains unchanged while preserving their positional correlations (Dosovitskiy et al, 2021). However, this approach neglects local structural information, which is vital for image data. Although integrating positional encoding can enable the transformer to learn positional relations, such encoding from randomly initialized parameters requires extensive training data, which poses a challenge in the realm of medical image analysis. Consequently, we substitute traditional positional encoding with convolutional layers, thereby incorporating the inductive bias inherent to convolutions. This substitution facilitates the preservation of local structural information intrinsic to the image while avoiding the limitations associated with positional encoding.

The three-dimensional depth-wise convolutionary technique is employed to convolve the interrelationships among local neighborhoods within fine-grained semantic features, thus inducing visual bias. In contrast, the three-dimensional pointwise convolution serves to compute global coarse-grained semantic features. The resultant features, denoted F_g and F_l, are subsequently flattened into a one-dimensional sequence to facilitate Transformer processing. Upon this transformation, F_g is converted into a global query vector F_g^q and a global content vector F_g^v. Similarly, F_l,

which encapsulates detailed information, is transformed into a local index vector F_l^k and a local content vector F_l^v through analogous operations.

Regarding the global coarse-grained perception module and the local fine-grained enhancement module, their internal architectures remain consistent, as illustrated in Figure 1b. These modules primarily comprise two components: a multi-head self-attention mechanism and a feedforward neural network. Within the multi-head self-attention mechanism, Z_l replaces the input and is initially expanded to a one-dimensional sequence. A fully connected layer then serves as the transformation layer, converting the input into a query vector, an index vector, and a content vector. Subsequently, the multi-head attention matrix is computed by formula 1. The feedforward layer further refines the interrelationships among voxel features, while integration with the residual mechanism augments data flow. This approach proficiently preserves the original features and mitigates gradient vanishing, as detailed in formula 2.

$$Z_l' = MHSA\left(LN(Z_l)\right) + Z_l, \tag{1}$$

$$Z_{l'}^v = FFN\left(LN\left(Z_l'\right)\right) + Z_l', \tag{2}$$

In this context, $MHSA(\cdot)$ denotes the multi-head self-attention mechanism, $LN(\cdot)$ signifies layer normalization, and $FFN(\cdot)$ represents the feedforward neural network. The global coarse-grained perception module and the local fine-grained enhancement module employ the computational algorithms delineated in formulas 1 and 2, respectively, to transform the global content vector F_g^v and the local content vector F_l^v as inputs. Upon completion of the processing, the resulting updated global content vector $F_{g'}^v$ and the local content vector $F_{l'}^v$ are obtained.

Within the context of the adaptive multigranularity fusion module, as illustrated in Figure 1, the correlation matrix Att is derived using F_g^q as the query vector and F_l^k as the index vector. The correlation is computed by applying the formula 3, as follows:

$$Att_{i,j} = \frac{1}{\sqrt{d_k}}\left(e^{F_{g\,i}^{q} F_{l\,j}^{k\top}} / \sum_{j=1}^{n} e^{F_{g\,i}^{q}}\right), \tag{3}$$

$Att_{i,j}$ denotes the i-th voxel characteristic within the global query vector F_g^q and the corresponding i -th voxel characteristic in the local index vector F_l^k. The correlation score between the characteristics of the j th voxel is computed accordingly. To mitigate overly concentrated correlations in specific local regions and facilitate the model optimization process, a scaling factor of $1/\sqrt{d_k}$ is implemented, where d_k signifies the dimension of an individual attention head in terms of the number of channels within F_g^q.

Upon deriving the similarity matrix Att, it is necessary to remap it into the original value space, comprising both the global content vector space F_g^v and the local content vector space F_l^v. Specifically, given the similarity matrix Att, for the local fine-grained branch F_l^v, Att undergoes direct propagation into a Softmax function to produce a sparse feature distribution. This distribution is then subjected to matrix multiplication with F_l^v, followed by convolution operations to further refine the feature distribution. This operational sequence is encapsulated in Equation 4. In contrast, for the global coarse-grained branch F_g^v, the matrix Att requires transposition to determine the similarity score between the individual feature points and the query feature. Subsequently, matrix multiplication with F_g^v is performed, and the resultant product is processed through deep separable convolutions to enhance the smoothness of the feature distribution. In addition, a residual technique is employed to preserve the integrity of the information and mitigate issues of vanishing gradients. Ultimately, this module yields the outputs $F_g^{'}$ and $F_l^{'}$.

$$F_l^{'} = DW\left(softmax(Att)F_l^v\right) + F_{l'}^v, \tag{4}$$

$$F_g^{'} = DW\left(softmax(Att^{\top})F_g^v\right) + F_{g'}^v, \tag{5}$$

Boundary Contrast Learning

A vast majority of existing 3D segmentation algorithms (Çiçek et al, 2016; Chen et al, 2019; Jiang et al, 2022) typically neglect explicit segmentation of structural boundaries. However, certain algorithms strive to improve boundary delineation through attention mechanisms (Liu et al, 2023a; Cheng et al, 2019). Despite these efforts, there remains a paucity of comprehensive and targeted research focused on evaluating the influence of boundary regions on segmentation performance, resulting in suboptimal overall model performance. To address this gap, we propose to meticulously align boundary information from manual annotations with model segmentation outputs and introduce an innovative boundary contrast learning framework.

Previous contrastive learning methodologies have been predominantly applied at the image level. For example, positional (Tang et al, 2022) extends 3D images to 2D sequences and takes advantage of positional relationships within the 2D sequence to determine spatial relationships between positive and negative samples for contrastive learning. Mask Transfiner (Ke et al, 2022) adopts contrastive learning in image patches rather than at the image level, utilizing pseudolabels to derive patches with distinct semantic categories to form positive and negative sample pairs. However, these approaches are inadequate for high-resolution segmentation, since accurate amygdala segmentation requires proximity of same-type pixels throughout the 3D image, separation of different-type pixels, and clustering of pixels based on their labels. Therefore, in this study, the boundary pixels are used to construct positive and negative sample pairs to maximize the distinction between the boundary pixels and adjacent nonrelevant pixels, as illustrated in Figure 1(d). Specifically, we designate the local neighborhood around the boundary as a subset, creating the boundary contrast learning strategy from overlapping 3D patch areas distributed throughout the 3D space. For each subset, kernels adjacent to the boundary select a 3D patch space dimension. Manually

segmented amygdala data serve as a basis for constructing positive and negative samples. The InfoNCE information maximization algorithm (Radford et al, 2021) is then utilized to calculate the disparity between positive and negative samples in the latitude space on the hypersphere. Taking into account the anatomical dimensions of the amygdala, the size of the kernel is set to $5 \times 5 \times 5$, and each subset forms positive and negative samples within this specified area.

As illustrated in Figure 1, the formula 6 is specifically designed to focus on the boundary voxels by exclusively computing them. Initially, all boundary regions B_l in manual annotations are considered, and for each voxel x_i in B_l, the sampling scope for positive and negative voxels is confined to its local neighborhood N_i. Under this constraint, the positive sample $x_j \in N_i \wedge l_i = l_j$ for the boundary voxel x_i is identified, while the remaining neighboring voxels are treated as negative samples $x_j \in N_i \wedge l_i \neq l_j$. In this paper, j denotes the sample voxel within the neighborhood N_i corresponding to the voxel i, with l_i and l_j representing the categories of different voxels x_i and x_j, respectively. Consequently, the boundary voxel contrast learning algorithm enhances the feature discernment of structurally organized boundary voxels, thus establishing the target for contrast optimization within boundary voxels, which is crucial for enhancing the segmentation of boundary regions. Specifically, for voxels x_i within boundary regions B_l, learned representations are incentivized to be more congruent with voxel features of the same category while being less congruent with features of disparate categories. Similarly, there are:

$$L_{CBL} = \frac{-1}{B_l} \sum_{x_i \in B_l} log\Big(\frac{\sum_{x_j \in N_i \wedge l_j = l_i} exp(-d(f_i, f_j)/\tau)}{\sum_{x_k \in N_i} exp(-d(f_i, f_k)/\tau)}\Big), \tag{6}$$

Here, f_i denotes the representation of the characteristics of x_i, $d(\cdot, \cdot)$ denotes the distance function, and τ represents the temperature coefficient in contrastive learning. The numerator component captures the similarity between positive instances,

while the denominator reflects the similarity between positive and negative instances. Consequently, a higher intraclass similarity coupled with a lower interclass similarity results in a reduced loss. The temperature coefficient is a predefined hyperparameter that governs the model's sensitivity to negative samples. An increased temperature coefficient τ flattens the distribution of $d(f_i, f_j)$, leading to a uniform treatment of all negative samples by contrastive loss, which diminishes the model's learning efficacy. Conversely, setting the temperature coefficient too low directs the model's focus toward particularly challenging negative samples, which might actually be potential positive samples. This hampers the model's convergence and degrades its generalization capability. Hence, the determination of the temperature coefficient is crucial.

Consistency Module with SAM

We use the SAM framework to develop a consistency strategy to mitigate the discrepancies that arise from the scanning of different subjects. This strategy comprises three integral components: the volume feature encoder, the 3D cue encoder, and the volume mask decoder. Specifically, the principal contributions can be delineated as follows. Firstly, we introduce an SAM model tailored for the amygdala segmentation task. Unlike existing approaches to improving SAM, the novelty of our model lies in its comprehensive consideration of factors such as the volume, shape, and relative position of the amygdala. It eschews the conventional slice-by-slice feature extraction in favor of a holistic total volume approach. This method establishes inter-slice associations, rendering it exceptionally suited for structures with intricate core regions. In contrast to traditional algorithms, this model provides a more comprehensive representation of anatomical structures, particularly for smaller entities. Secondly, the incorporation of a 3D prompt encoder ensures the preservation of prior spatial position information and the 3D spatial features provided by the user. In contrast to the 2D prompt encoder, it furnishes more precise relative position cues in a multi-dimensional space. To facilitate

the effective amalgamation of the 3D cue information with the embedding features derived from the volume feature encoder, we have devised a volume mask decoder. This decoder harnesses the interaction mechanism between a prior cue embedding and the MRI embedding, facilitating seamless integration. This fusion mechanism adeptly captures nuanced details within the feature space, offering a robust tool for the precise depiction of anatomical structures.

SAM constitutes an interactive segmentation methodology predicated on iterative operations, enhancing the model's representational capacity through the incorporation of positional cues and pre-segmentation data, thereby facilitating continuous optimization of segmentation accuracy via multiple interactive iterations. In particular, when existing SAM models are applied directly to medical imaging tasks, specifically the amygdala segmentation task, their accuracy is suboptimal. To address this issue, we have meticulously tailored the design of the 2D SAM model for the amygdala segmentation context and introduced an advanced SAM model, which comprises three integral encoders for the volume feature, the 3D cue and the volume mask.

The function of the Volume Feature Encoder is to transform the input T1w structural image into an MRI embedding vector. Compared to the 2D encoder in the traditional SAM, it allows simultaneous consideration of volume, shape, and relative position, thus offering a more comprehensive representation of anatomical structures and enhancing the processing of complex structures around the core area. To incorporate interactive cues and existing amygdala segmentation results, a 3D cue encoder is used to encode this prior information. The 3D cue encoder comprises two components: the click-cue branch and the pre-segmentation cue branch. The click-cue branch converts user-provided cue points into learnable location embedding features to more precisely delineate the location of the amygdala. The pre-segmentation prompt encoder uses the last click of the model during interactive segmentation as input or feeds the pre-segmentation result into the model, thereby prompting the model to

adjust and optimize the current segmentation result. To effectively integrate the cue encoder's information with the feature embedding produced by the feature encoder, and to retrieve and refine the information pertaining to the amygdala, we devise a volume mask detector to ensure the seamless alignment between cue feature information and the original feature encoder output.

As the pivotal component, the volume mask decoder is principally designed to exhaustively explore key characteristics within the amygdala, thereby enhancing the model's efficacy in intricate medical contexts. Its fundamental function is to amalgamate the MRI embedding vector resulting from the volume feature encoder with the user-provided prompt inputs to ascertain the correlation between them, culminating in precise predictive results. The 3D mask decoder emits MRI embedding and cue embedding through a multi-layered fusion 3D feature encoder, facilitating multi-level and multi-perspective information extraction. This comprehensive fusion mechanism enables the model to assimilate MRI images more profoundly and improves the discernment of medical structures, thus increasing predictive precision. In conclusion, the decoder produces precise predictions of specific structures or lesions within medical images, offering crucial support for clinicians to make accurate diagnoses and devise treatment strategies.

Loss Function

We implement an advanced deep-supervised strategy for the formulation of the loss function. Specifically, this involves the integration of supervisory signals in each intermediate module and the terminal module, thereby facilitating enhanced gradient propagation and optimizing the efficiency of network training. For each supervisory signal, we incorporate two distinct loss functions, namely, the binary cross-entropy (BCE) loss and the Dice loss. The overall loss function is constructed as follows.

$$L = \sum_{k=1}^{4} \left(L_{\text{BCE}}^{k} + L_{\text{Dice}}^{k} \right). \tag{7}$$

To address the issue of imbalance between positive (amygdala) and negative (other tissue) samples in the loss of BCE, an adapted version is employed. For simplicity, the superscript k is excluded and L_{extBCE} is delineated as follows.

$$L_{\text{BCE}} = -\frac{\sum_{i=1}^{H} \sum_{j=1}^{W} (1 + \alpha_{i,j}) \sum_{l=0}^{1} (g_{i,j} = l) logPr(p_{i,j} = l|\Psi)}{\sum_{i=1}^{H} \sum_{j=1}^{W} (\alpha_{i,j})}, \tag{8}$$

where $l \in (0, 1)$ denotes two categories of labels. The variables $p_{i,j}$ and $g_{i,j}$ represent the predicted and ground truth values, respectively, at the spatial coordinate (i, j) within a CT image characterized by a width of W and a height of H. The symbol Ψ encompasses all parameters of the model and $Pr(p_{i,j} = l)$ signifies the predicted probability. The term $\alpha_{i,j} \in [0, 1]$ stands for the pixel-specific weight. We define α as

$$\alpha_{i,j} = \left| \frac{\sum_{(m,n) \in A_{i,j}} g_{m,n}}{\sum_{(m,n) \in A_{i,j}} 1} - g_{i,j} \right|, \tag{9}$$

where $A_{i,j}$ denotes the region adjacent to the pixel (i, j), and $|\cdot|$ signifies the calculation of the absolute value. Moreover, the Dice loss is defined as

$$L_{\text{Dice}} = 1 - \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} (2 g_{i,j} p_{i,j})(1 + \alpha_{i,j})}{\sum_{i=1}^{H} \sum_{j=1}^{W} (g_{i,j} + p_{i,j} + g_{i,j} p_{i,j})(1 + \alpha_{i,j})}. \tag{10}$$

Implementation Details, Data and Metrics

Our algorithm was implemented utilizing the PyTorch framework and utilizes the full capabilities of the Tesla A10 GPU to optimize computational efficiency. AdamW optimizer (Xu et al, 2022), a common choice within the transformer architecture domain, was used for routine updates of network parameters. The learning rate was meticulously set to $1e - 4$ and the weight decay was equally adjusted to $1e - 4$ to achieve

an equilibrium between the stability of the training and the convergence speed of the model. To mitigate memory consumption, a batch size of 1 was designated, and the training process encompassed 100 iterations. This configuration ensures that the model undergoes sufficient training cycles to attain optimal performance and effectively discern feature representations. During the testing phase, the input image dimensions were reduced to $48 \times 64 \times 80$ to minimize computational demands. Subsequently, no post-processing optimization techniques were employed, thereby ensuring an objective assessment of model performance.

In this study, neuroimaging data of human brain scans from three cohorts, including the Chongqing cohort of the Chinese Color Nest Cohort (CKG-CCNC) (Fan et al, 2023; Liu et al, 2021; Yang et al, 2017), the Beijing Normal University Childhood Adversity Cohort (BNU-CAC) (Sun et al, 2022) and the Zhejiang University Multicenter Traveler Cohort (ZJU-MTC) (Tong et al, 2020), were used. CKG-CCNC is part of the development component of CCNP (devCCNP), which was collected at Southwest University in Chongqing, China. The devCCNP aims to map normative brain development trajectories within the Chinese population during school years. The devCCNP-CKG dataset included data from 201 participants aged 6-17 years, who were invited to participate in three consecutive waves of data collection at intervals of approximately 1.25 years. At these time points, T1-weighted magnetic resonance imaging (T1w) was performed and images were visually inspected to exclude those with significant head movement artifacts and structural abnormalities. After this initial quality control, the final sample included 427 scans from 198 participants (105 girls and 93 boys). The distribution of participants with one, two and three scans was 48, 71, and 79, respectively, with an average of 2.16 scans per participant. BNU-CAC aims to investigate the impact of childhood adversity on brain development. It consists of 116 subjects (60 girls and 56 boys) aged 8-13 years who have been scanned once per individual and passed quality control protocols. ZJU-MTC includes three adult

travelers (Tong et al, 2020) to evaluate image consistency between centers under identical parameters with data collection performed across 10 scanning centers, ensuring comprehensive inter-center comparability.

To rigorously assess the model's performance, standard evaluation metrics for medical image segmentation have been incorporated, primarily encompassing similarity metrics and distance-based metrics. These include the Dice coefficient (Milletari et al, 2016) and the Jaccard coefficient (JS), the average surface distance (ASD) and the average symmetric surface distance (ASSD) (Berhane et al, 2020), and the relative volume difference (RVD) (Wang et al, 2008). In addition, precision and recall metrics derived from the confusion matrix are also utilized (Borji et al, 2015). Details of these metrics are described below.

Dice similarity coefficient (Dice) is used to measure the similarity between the predicted and the ground truth (GT).

$$Dice = \frac{2|V_{\text{Seg}} \cap V_{\text{GT}}|}{|V_{\text{Seg}}| + |V_{\text{GT}}|}, \tag{11}$$

where V_{Seg} and V_{GT} represent the voxel sets of the segmentation and GT, respectively. $|\cdot|$ denotes the operation of cardinality computation, which provides the number of elements in a set.

Intersection over Union (IoU) is the area of intersection between the predicted segments and the labels divided by the area of union between the predicted segments and the labels.

$$IoU = \frac{|V_{\text{Seg}} \cap V_{\text{GT}}|}{|V_{\text{Seg}} \cup V_{\text{GT}}|}. \tag{12}$$

Average Surface Distance (ASSD) is a metric based on the surface distance defined as the average distance between the surface voxel $S(V_{\text{Seg}})$ of the segmentation

result V_{Seg} and the surface voxel $S(V_{\text{GT}})$ of GT V_{GT}.

$$ASSD = \frac{1}{|S(V_{\text{Seg}})| + |S(V_{\text{GT}})|} \left(\sum d\left(s_{V_{\text{Seg}}}, S(V_{\text{GT}})\right) + \sum d\left(s_{V_{\text{GT}}}, S(V_{\text{Seg}})\right) \right) \tag{13}$$

where $d(s, S(X))$ is defined as the minimum Euclidean distance from the voxel s to the surface voxel of the volume X.

Relative Volume Difference (RVD) is used to measure the difference between predicted lung infections and GT. We take the absolute value of RVD and report the corresponding results in this work.

$$RVD = |\frac{|V_{\text{Seg}}| - |V_{\text{GT}}|}{|V_{\text{GT}}|}|. \tag{14}$$

Asymmetry index (AI) assesses the degree of lateralization of the left and right amygdala and is defined as

$$AI = \frac{V_{left} - V_{right}}{V_{left} + V_{right}} \times 100\%. \tag{15}$$

Interclass Correlation Coefficient (ICC) reflects the consistency of results across multiple centers and is defined as

$$ICC = \frac{(MSB - MSW)}{MSB + (k-1) * MSW}. \tag{16}$$

MSB denotes the mean square between groups, MSW denotes the mean square within groups, and k is the number of raters.

Results

Learning Capability Assessment

BOLT3D demonstrates unequivocal superiority within the test cohort, attaining exemplary performance in critical metrics including Dice, JS and HD95 relative to other prevalent algorithms. A comprehensive analysis of these individual metrics and the pronounced performance benefits conferred by BOLT3D on various parameters is presented in Table 1.

The Dice coefficient and the JS coefficient serve as imperative metrics for evaluating the extent of overlap in segmentation outcomes. BOLT3D achieves a Dice coefficient of 0.904, which surpasses alternative algorithms such as U-Net3D (0.851), U-Net3D++ (0.787), and ResU-Net3D (0.841). Analogously, the JS coefficient demonstrates a comparable trend. BOLT3D (0.825) clearly outperforms other algorithms, underscoring its superior segmentation precision. The HD95 and the ASSD quantify the spatial discrepancy between the segmentation outcomes and the reference annotations, particularly highlighting the interval between the segmented boundary of the amygdala and the true structural boundary. BOLT3D records a markedly reduced Hausdorff distance of 1.097 relative to other algorithms such as U-Net3D (1.565), U-Net3D++ (1.963), and ResU-Net3D (1.669). This shows that BOLT3D excels in preserving spatial congruence between segmented peripheries and actual structural demarcations. Precision and recall metrics are used to assess false positives. The precision metric quantifies the

Table 1: Amygdala index results of 347 scans in the CKG-CCNC cohort

Method	**Dice**	**JS**	**HD95**	**ASD**	**ASSD**	**Precision**	**Recall**	**RVD**
Freesurfer	0.776	0.636	2.530	1.093	0.954	0.732	0.832	0.142
U-Net3D	0.851	0.743	1.565	0.746	0.693	0.774	0.951	0.239
U-Net3D++	0.787	0.651	1.963	1.064	0.982	0.658	0.985	0.512
ResU-Net3D	0.841	0.728	1.669	0.806	0.736	0.754	0.957	0.279
BOLT3D	0.904	0.825	1.097	0.483	0.467	0.883	0.929	0.056

proportion of true positive instances among those designated as positive by the model, whereas the recall metric evaluates the proportion of true positive instances correctly identified by the model from all positive instances in the dataset. BOLT3D's precision and recall are 0.883 and 0.929, respectively, surpassing the performance of other algorithms. BOLT3D manifests an admirable balance between these metrics, resulting in an overall superior performance relative to other algorithms. Additionally, BOLT3D achieves a relative RVD of 0.056, significantly lower than that of other algorithms, thereby highlighting the algorithm's exceptional accuracy in volume prediction.

Figure 2 elucidates the comparative segmentation outcomes across multiple algorithms on a representative subject, illustrating the superior performance of the proposed algorithm with respect to volumetric and boundary concordance with manual segmentation. Concerning the overall segmentation, the amygdala delineated by Freesurfer deviates markedly from both the morphological and manual segmentation results. The presence of numerous spaces indicates an inability to accurately capture the structural integrity of the actual amygdala. A closer inspection of its peripheries reveals that the Freesurfer exhibits numerous irregularities when juxtaposed with the gold standard. The amygdala delineated by Unet exhibits morphological characteristics comparable to those of the gold standard. However, an assessment of the superposition effect and edge segmentation results reveals that the volumetric delineation by Unet is substantially overestimated. In contrast, our BOLT3D results exhibit a morphological resemblance closer to the gold standard. The superposition effect in BOLT3D is similarly congruent with the gold standard, demonstrating no significant volumetric overestimation or underestimation, which is corroborated by the edge segmentation results. The superiority of BOLT3D is primarily attributable to the integration of Transformer technology with a boundary contrast learning algorithm, thus improving its ability to capture intricate structural details in the image while preserving overall structural integrity.

Generalization Performance Assessment

To evaluate the generalization capabilities of the model, we initially validated it using data from previously uncounted centers, encompassing a cohort of 116 subjects (Table 2). The BOLT3D algorithm exhibited robust performance in similarity metrics, achieving a Dice coefficient of 0.845 and a JS of 0.733. This indicates that despite variations in the data distribution, the overall performance of the algorithm remained substantially unchanged. In terms of distance evaluation, BOLT3D demonstrated HD95 and ASD values of 1.594 and 0.654, respectively. Furthermore, the algorithm's ASSD of 0.666 exhibited less fluctuation compared to the Unet algorithm. BOLT3D achieved precision and recall scores of 0.933 and 0.775, respectively, underscoring its efficacy in correctly identifying true positives while minimizing false negatives. The equilibrium point of the precision-recall distribution further underscores BOLT3D's ability to maintain high performance across both precision and recall metrics. The RVD for BOLT3D was recorded at 0.168, reflecting its precision in predicting the amygdala volume with minimal relative volume deviation. These results suggest that the performance of the BOLT3D algorithm in these indicators is not significantly compromised despite the fact that distribution variations are encountered in new centers. Moreover, it significantly outperforms competing algorithms, thereby affirming the multicenter accuracy and consistency afforded by the integration of the consistency module.

Table 2: Amygdala index results of 116 subjects in the BNU-CAC cohort

Method	**Dice**	**JS**	**HD95**	**ASD**	**ASSD**	**Precision**	**Recall**	**RVD**
U-Net3D	0.705	0.545	2.860	1.445	1.286	0.551	0.980	0.789
U-Net3D++	0.773	0.631	3.551	1.423	1.129	0.665	0.927	0.404
ResU-Net3D	0.782	0.643	2.177	1.107	0.969	0.663	0.955	0.448
BOLT3D	**0.845**	**0.733**	**1.594**	**0.654**	**0.666**	**0.933**	**0.775**	**0.168**

Furthermore, this study presents the results for 116 unseen subjects in BNU-CAC, juxtaposed with 347 observed scans in CKG-CCNC (Figure 3). Given that the 116 BNU-CAC subjects represent distinct data distributions and that the model had not been previously trained on this data distribution, the overall performance index is lower compared to the 347 CKG-CCNC scans. This discrepancy arises because the model demonstrates diminished performance on unfamiliar data relative to familiar data, stemming from variations in feature distributions, inter-data noise, or other unidentified factors, resulting in performance deterioration. Despite a reduction in the overall index when applied to novel data, the model shows substantial performance improvements relative to the baseline comparison model. This indicates that our algorithm has commendable generalization capabilities.

We employ three travelers from ten distinct centers to determine the stability and generalizability of the model between multiple centers (Figure 4). It can be discerned that BOLT3D achieved superior segmentation accuracy for each of the three subjects in all ten centers, showing a notable improvement over U-Net 3D, with improvements ranging from 0.02 to 0.07. This underscores the robustness of BOLT3D in multicenter image segmentation tasks. Currently, BOLT3D clearly demonstrates its superiority over U-Net3D.

Contributions to Neuroscience

Segmentation Validity

For the assessment of volumetric measurements of the left amygdala in 347 scans, Pearson's correlation coefficients with manual segmentations were calculated for five algorithms. U-Net3D, U-Net3D++, ResU-Net3D, Freesurfer, and BOLT3D (Figure 5 illustrates). The correlation coefficients and corresponding P-values were determined for each algorithm. The findings indicate varying degrees of correlation with manual segmentation in all algorithms, although some exhibited substantial overestimation. In

particular, BOLT3D demonstrated exceptional consistency in volumetric estimation, closely aligning with manual segmentation and manifesting a substantial positive linear relationship. The correlation coefficient for BOLT3D reached a value of 0.74, exceeding the range of 0.46 to 0.61 observed with other algorithms. Despite the strong correlation results observed for algorithms such as Freesurfer and U-Net3D, BOLT3D provided a more definitive overall trend. Its elevated correlation coefficient in the analysis of the left amygdala underscores BOLT3D's superior performance in volumetric detection across different structures or tissues, achieving considerable concordance with manual segmentation. In addition, the narrow dispersion of the scatter points accentuates the exceptional consistency in individual measurements obtained through BOLT3D, further attesting to its remarkable stability.

Systematic Difference Analysis

We present a detailed analysis of systematic discrepancies in the quantification of amygdala volume, elucidating the consistency and comparability between automated segmentation methodologies and manual segmentation (Figure 6). Of the five algorithms evaluated for volume quantification, our method demonstrates superior consistency (Figure 6). In this representation, the upper and lower lines denote the 95% limits of agreement, while the central line signifies the mean difference. A qualitative examination of the data distribution around the mean reveals that our algorithm exhibits a moderate and concentrated spread within the volume measurement spectrum. Our BOLT3D algorithm registers an average volume discrepancy of -0.152mm^3. In contrast, U-Net3D, U-Net3D++, Freesurfer, and ResU-Net3D manifest average volume discrepancies of -1.433mm^3, -0.647mm^3, -1.874mm^3, and -0.775mm^3, respectively, indicative of significant overestimation. The inflated volumes reported by these algorithms are 9.427 times, 4.256 times and 5.098 times that of BOLT3D, underscoring the substantial advantage of BOLT3D in volumetric measurement precision. Minimal

systematic deviations were observed between BOLT3D and manually adjusted volumes, highlighting its precision and dependability in the quantification of amygdala volumes.

Brain Growth Curve Modeling

To provide insight into the effectiveness of these algorithms in forecasting the growth curve, and further assess alongside their alignment with manually annotated results, the growth curves of the volume change of the amygdala derived algorithmically with age were plotted. Specifically, as shown in Figure 7, the growth curve patterns were parallel in girls and boys for both manual and automatic tracing methods although boys showed larger volumes of their amygdala than girls across the entire age range. As a reference standard, the manual tracing method revealed that the human amygdala (both left and right) exhibited linear growth during childhood and adolescence in both boys and girls. BOLT3D produces a linear trend consistent with that of manual segmentation (Figure 7 and Table 3). Although the bilateral volumes of amygdala predicted by BOLT3D are slightly larger than those from manual annotations, the scatter plot shows a more confined aggregation among age cohorts. The FreeSurfer method illustrates inverted U-shaped curves, which diverge from the linear trend observed in manual segmentation (Figure 7). The scatter plot reveals an increase in variability between age cohorts, with bilateral amygdala volumes exceeding those derived from both manual annotation and BOLT3D. BOLT3D produced growth curves very similar to those established by the manual tracing method. This demonstrates significant advantages in fitting growth curves compared to other algorithms, attesting to BOLT3D's robustness and accuracy in accounting for gender- and age-related changes when estimating amygdala volume. Compared to traditional and other advanced algorithms, it excels at capturing subtle changes in growth curves, not only in terms of

Table 3: Statistics of brain amygdala volume growth and development curve indicators.

	Manual	Ours	Freesurfer	U-Net3D	U-Net3D++	ResU-Net3D
EDF	1.33	1.59	2.09	2.38	3.02	2.84
Volume (mm^3)	1.47	1.56	1.73	2.20	1.80	1.91

overall adaptation efficiency but also in accurately predicting curve details and particular features. These findings improve our understanding of the development of the amygdala over time with respect to sex differences and age-related factors.

Ablation Study

To evaluate the stability of the model, we analyzed the distribution of Dice and ASSD scores between individuals for both volume and boundary metrics. The consistency analysis depicted in Figure 8(a) illustrates that our algorithm demonstrates markedly lower discrepancies relative to U-Net3D and Freesurfer, thus confirming stability across a substantial cohort of subjects. The Dice coefficient of the BOLT3D algorithm almost reached a value of 0.9 (Figure 8.b), surpassing the average performance of both Freesurfer and U-Net3D. The proposed MGAC leverages the Transformer mechanism by amalgamating global, coarse-grained perceptual processes with local, fine-grained enhancement techniques. Such mechanisms enable the acquisition of both global and local features of the amygdala, while the adaptive multigranularity fusion mechanism extracts abundant semantic information from varied scales for optimization. Consequently, the MGAC algorithm shows remarkable stability and superior efficacy in comprehensive structural processing.

Evaluating the consistency of segmentation through multiple repeated scans of a representative subject at various centers is imperative to substantiate the effectiveness of the consistency module. A quantitative comparison between ten MR centers is

presented in Figure 8(c), which corroborates that the SAM module accurately identifies and maintains target consistency, producing stable results under diverse scanning conditions. The optimized segmentation algorithm exhibited exceptional performance in an experiment involving 116 subjects from the BNU-CAC cohort.

AmygdalaGo-BOLT3D model availability

The AI training platform supporting this work was provided by High-Flyer AI (Hangzhou High-Flyer AI Fundamental Research Co., Ltd.). Our dedicated web portal freely provides the optimal model parameters to be applied to any user-specified dataset in the context of open science. In addition to the pre-trained models, the website provides tutorials and codes (GITHUB_LINK).

Discussion

We propose a sophisticated transformer architecture explicitly tailored for amygdala segmentation. Our approach leverages the long-distance relational modeling capabilities inherent to the 3D Transformer, incorporating cross-attention mechanisms to enhance the extraction of salient regions across the entire image. This ensures comprehensive receptive field coverage across the full 3D MRI, transcending the limitations of local patch-based approaches. We address the intricacies posed by the low contrast and structural complexity distinguishing the amygdala from adjacent tissues. To align model-predicted boundaries with the ground truth more accurately, we introduce a boundary contrastive learning algorithm. This refinement substantially improves the delineation between the amygdala and its surrounding tissues and reduces intra-amygdala heterogeneity, culminating in more precise segmentation.

Our proposed model exhibits precise segmentation of the amygdala in a spectrum of data sets that cover a wide age range from childhood and adolescency to adulthood. It maintains remarkable consistent performance when deployed in multicenter

adult datasets. This underscores the model's exceptional generalization proficiency and robust performance across varied data sources and contexts. The model's ability to generalize effectively and maintain accuracy across heterogeneous datasets underscores its reliability and efficacy. Consequently, some controversy findings in the field of cognitive neuroscience of the human amygdala may benefit from the proposed BOLT3D transformer. An example of this situation is previous research on the relationship between the size of the social network and the volume of the amygdala (McGugin et al, 2024; Kanai et al, 2011; Bickart et al, 2011).

Given that the BOLT3D algorithm presented herein is primarily designed to segment low-contrast and diminutive target structures, it exhibits robust segmentation capabilities across diverse individuals and scanning conditions. There exists a substantial corpus of similar tasks within medical imaging that share considerable similarities. Consequently, the proposed BOLT3D algorithm is expected to demonstrate consistently stable and precise performance in additional segmentation tasks, such as facial nerve segmentation and hippocampal segmentation. This assertion corroborates the algorithm's stability and generalizability across multiple tasks, thereby establishing a strong foundation for medical image segmentation endeavors. Considering the primary design objective of the BOLT3D algorithm to effectively segment low-contrast and diminutive target structures, it manifests proficient segmentation capabilities across heterogeneous subjects and varied scanning environments. A substantial body of analogous tasks within the domain of medical imaging exhibit significant parallels to this endeavor. As such, the proposed BOLT3D algorithm is expected to consistently yield stable and precise segmentation results in ancillary tasks, including but not limited to facial nerve segmentation and hippocampal segmentation. This demonstrates the robustness and generalizability of the algorithm in multifarious segmentation tasks, thereby underpinning a robust foundation for advances in medical image segmentation.

Recent advances in open neuroimaging data are significantly augmenting our comprehension of the intricate structures and networks within the human brain (Lu et al, 2024). The aggregation of images from various large-scale data repositories has significantly improved statistical power, allowing the discernment of subtle variances within the human amygdala (Cotter et al, 2023; Hatzenbuehler et al, 2022). Nonetheless, there is a pressing need for new research methodologies to enhance the computational efficiency and practical applicability of these automatic extraction tools (Zhou et al, 2021; Ghanem et al, 2024). Global imaging collaborations have been instrumental in the evolution of population neuroscience (Paus, 2010; Falk et al, 2013), propelled by extensive magnetic resonance imaging initiatives (Bethlehem et al, 2022), and have been pivotal in driving innovations in analytic techniques and translational research (Zhou et al, 2023). AmygdalaGo-BOLT3D caters to the population neuroscience of the human amygdala through generative AI predicated on manually traced or labeled datasets.

Several potential limitations of the present study must be taken into account. First, the multicenter data quality of pediatric brain imaging could be suboptimal, which underscores the imperative for ongoing research and improved data collection in this domain. By addressing these deficiencies, the field of amygdala segmentation can be further advanced, thereby improving the precision and dependability of segmentation outcomes across diverse populations and data environments. Second, this study lacks a detailed subdivision of the amygdala, specifically finer-grained nuclei labeling. Future work will focus on providing more granular annotations and establishing a computational baseline to support nuanced analysis of the amygdala.

Conclusion

This investigation begins with the amygdala segmentation, engaging in a comprehensive analysis from various perspectives and advancing novel methodologies to address

the complex challenges inherent in segmentation. Primarily, we address the algorithmic predicament posed by the diminutive structure of the amygdala, and secondarily, we tackle issues such as diminished boundary contrast and interference from adjacent anatomical entities. In response to these challenges, this paper innovatively introduces an alternative boundary contrast learning strategy grounded in the Transformer architecture, aimed at improving the efficacy of the amygdala segmentation algorithm. To overcome the difficulties posed by individual variability and inconsistency in data distribution, we have incorporated a consistency module utilizing SAM, seeking to infuse additional prior knowledge or significantly refine existing outcomes via interactive algorithms. The imperative and relevance of model validation within cognitive neuroscience are underscored.

Acknowledgment

This work has been supported by the the STI 2030—Major Projects of the Brain Science and Brain-Inspired Intelligence Technology (2021ZD0200500), the Major Fund for International Collaboration of National Natural Science Foundation of China (81220108014), National Natural Science Foundation of China (82372036), Open Research Fund of the State Key Laboratory of Cognitive Neuroscience and Learning (CNLZD2001), Research of Basic Discipline for the 2.0 Base of Top-notch Students Training Program from The Ministry of Education of China (20211033), and the National Basic Science Data Center "Interdisciplinary Brain Database for In vivo Population Imaging" (ID-BRAIN).

References

Ben-Zion Z, Korem N, Fine N, et al (2023) Structural neuroimaging of hippocampus and amygdala subregions in post-traumatic stress disorder (ptsd): A scoping review. Biological Psychiatry Global Open Science

Berhane H, Scott M, Elbaz M, et al (2020) Fully automated 3d aortic segmentation of 4d flow mri for hemodynamic analysis using deep learning. Magnetic resonance in medicine 84(4):2204–2218

Bethlehem RA, Seidlitz J, White SR, et al (2022) Brain charts for the human lifespan. Nature 604(7906):525–533

Bickart K, Wright C, Dautoff R, et al (2011) Amygdala volume and social network size in humans. Nature Neuroscience 14:163–164

Borji A, Cheng MM, Jiang H, et al (2015) Salient object detection: A benchmark. IEEE transactions on image processing 24(12):5706–5722

Chai Y, Gehrman P, Yu M, et al (2023) Enhanced amygdala–cingulate connectivity associates with better mood in both healthy and depressive individuals after sleep deprivation. Proceedings of the National Academy of Sciences 120(26):e2214505120

Chen W, Liu B, Peng S, et al (2019) S3d-unet: separable 3d u-net for brain tumor segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4, Springer, pp 358–368

Cheng J, Liu J, Liu L, et al (2019) Multi-level glioma segmentation using 3d u-net combined attention mechanism with atrous convolution. In: 2019 ieee international

conference on bioinformatics and biomedicine (BIBM), IEEE, pp 1031–1036

Çiçek Ö, Abdulkadir A, Lienkamp SS, et al (2016) 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 424–432

Cotter D, Campbell C, Sukumaran K, et al (2023) Effects of ambient fine particulates, nitrogen dioxide, and ozone on maturation of functional brain networks across early adolescence. Environment international 117:108001

Dosovitskiy A, Beyer L, Kolesnikov A, et al (2021) An image is worth 16x16 words: Transformers for image recognition at scale. ICLR

Falk EB, Hyde LW, Mitchell C, et al (2013) What is a representative brain? Neuroscience meets population science. Proceedings of the National Academy of Sciences 110(44):17615–17622

Fan XR, Wang YS, Chang D, et al (2023) A longitudinal resource for population neuroscience of school-age children and adolescents in china. Scientific data 10(1):545

Fischl B, Salat DH, Busa E, et al (2002) Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33(3):341–355

Ghanem K, K S, A S, et al (2024) Longitudinal microstructural changes in 18 amygdala nuclei resonate with cortical circuits and phenomics. Communications Biology 7:477

Hatzenbuehler M, Weissman D, McKetta S, et al (2022) Smaller hippocampal volume among black and latinx youth living in high-stigma contexts. Journal of the American Academy of Child and Adolescent Psychiatry 61:809–819

Jiang Y, Zhang Y, Lin X, et al (2022) Swinbts: A method for 3d multimodal brain tumor segmentation using swin transformer. Brain sciences 12(6):797

Kanai R, Bahrami B, Roylance R, et al (2011) Online social network size is reflected in human brain structure. Proceedings of The Royal Society B: Biological Sciences 279:1327–1334

Ke L, Danelljan M, Li X, et al (2022) Mask transfiner for high-quality instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4412–4421

Lee Y, Kim J, Willette J, et al (2022) Mpvit: Multi-path vision transformer for dense prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

Liu CY, Chou JS (2023) Bayesian-optimized deep learning model to segment deterioration patterns underneath bridge decks photographed by unmanned aerial vehicle. Automation in Construction 146:104666

Liu H, Huo G, Li Q, et al (2023a) Multiscale lightweight 3d segmentation algorithm with attention mechanism: Brain tumor image segmentation. Expert Systems with Applications 214:119166

Liu H, Nie D, Yang J, et al (2023b) A new multi-atlas based deep learning segmentation framework with differentiable atlas feature warping. IEEE Journal of Biomedical and Health Informatics

Liu S, Wang YS, Zhang Q, et al (2021) Chinese color nest project: An accelerated longitudinal brain-mind cohort. Developmental Cognitive Neuroscience 52:101020

Lu B, Chen X, Castellanos FX, et al (2024) The power of many brains: Catalyzing neuropsychiatric discovery through open neuroimaging data and large-scale collaboration. Science Bulletin 69:1536–1555

Manjón JV, Coupé P (2016) volbrain: an online mri brain volumetry system. Frontiers in neuroinformatics 10:30

McGugin R, Roche A, Ma J, et al (2024) Challenges in replication: Does amygdala gray matter volume relate to social network size? Cognitive Affective Behavior Neuroscience 24:707–719

Milletari F, Navab N, Ahmadi SA (2016) V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision (3DV), Ieee, pp 565–571

Paré D, Headley DB (2023) The amygdala mediates the facilitating influence of emotions on memory through multiple interacting mechanisms. Neurobiology of Stress p 100529

Patenaude B, Smith SM, Kennedy DN, et al (2011) A bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage 56(3):907–922

Paus T (2010) Population neuroscience: why and how. Human brain mapping 31(6):891–903

Pruessner JC, Li LM, Serles W, et al (2000) Volumetry of hippocampus and amygdala with high-resolution mri and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cerebral cortex 10(4):433–442

Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning,

PMLR, pp 8748–8763

Sammallahti S, Serdarevic F, Tiemeier H (2023) Excessive crying, behavior problems, and amygdala volume: A study from infancy to adolescence. Journal of the American Academy of Child & Adolescent Psychiatry 62(6):675–683

Song J (2023) Amygdala activity and amygdala-hippocampus connectivity: Metabolic diseases, dementia, and neuropsychiatric issues. Biomedicine & Pharmacotherapy 162:114647

Sun J, Jiang Y, Wang X, et al (2022) Cortisol reactivity as a mediator of peer victimization on child internalizing and externalizing problems: The role of gender differences. Research on Child and Adolescent Psychopathology 50(3):283–294

Tang L, Zhan Y, Chen Z, et al (2022) Contrastive boundary learning for point cloud segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8489–8499

Tong Q, He H, Gong T, et al (2020) Multicenter dataset of multi-shell diffusion mri in healthy traveling adults with identical settings. Scientific Data 7(1):157

Trombini M, Solarna D, Moser G, et al (2023) A goal-driven unsupervised image segmentation method combining graph-based processing and markov random fields. Pattern Recognition 134:109082

Wang Y, van Klaveren RJ, van der Zaag-Loonen HJ, et al (2008) Effect of nodule characteristics on variability of semiautomated volume measurements in pulmonary nodules detected in a lung cancer screening program. Radiology 248(2):625–631

Xie L, Wisse LE, Wang J, et al (2023) Deep label fusion: A generalizable hybrid multi-atlas and deep convolutional neural network for medical image segmentation.

Medical Image Analysis 83:102683

Xu Y, Wei H, Lin M, et al (2022) Transformers in computational visual media: A survey. Computational Visual Media 8:33–62

Yang N, He Y, Zhang Z, et al (2017) Chinese color nest project: Growing up in china. Chinese Science Bulletin 62:3008–3022. URL https://api.semanticscholar.org/CorpusID:103639957

Zhou Q, Liu S, Jiang C, et al (2021) Charting the human amygdala development across childhood and adolescence: Manual and automatic segmentation. Developmental Cognitive Neuroscience 52:101028

Zhou Z, Chen L, Milham M, et al (2023) Six cornerstones for translational brain charts. Science Bulletin 68(8):795–799

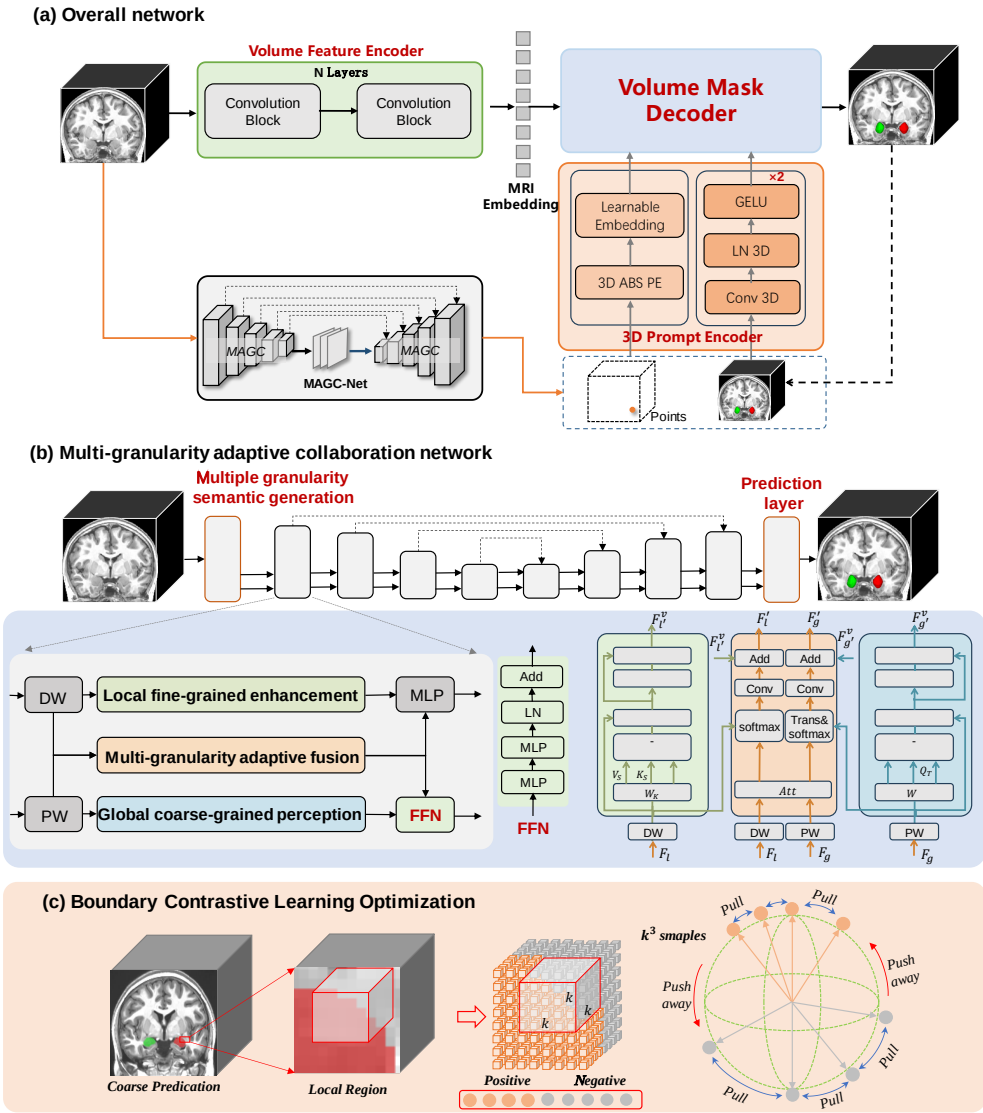

Fig. 1: A schematic diagram of the AmygdalaGo-BOLT3D's architecture. (a) The model's overall structure consists of two principal components: the MAGC-Net responsible for initial segmentation, and the SAM, tasked with the refinement of the segmentation including three essential components: a lightweight Volumetric Feature Encoder, a 3D Prompt Encoder, and a Volumetric Mask Decoder. The MAGC network produces preliminary segmentation outcomes and strategically sampling voxels, which are then fed into the SAM's prompt encoder to guide the refinement process. (b) Elaborates on the MAGC network's key module, the MAGC block, designed to address the challenges posed by the amygdala's minute size, ensuring the extraction of accurate semantic features. Moreover, (c) outlines the edge contrast learning optimization strategy, incorporated to further augment edge delineation and enhance the overall segmentation precision.

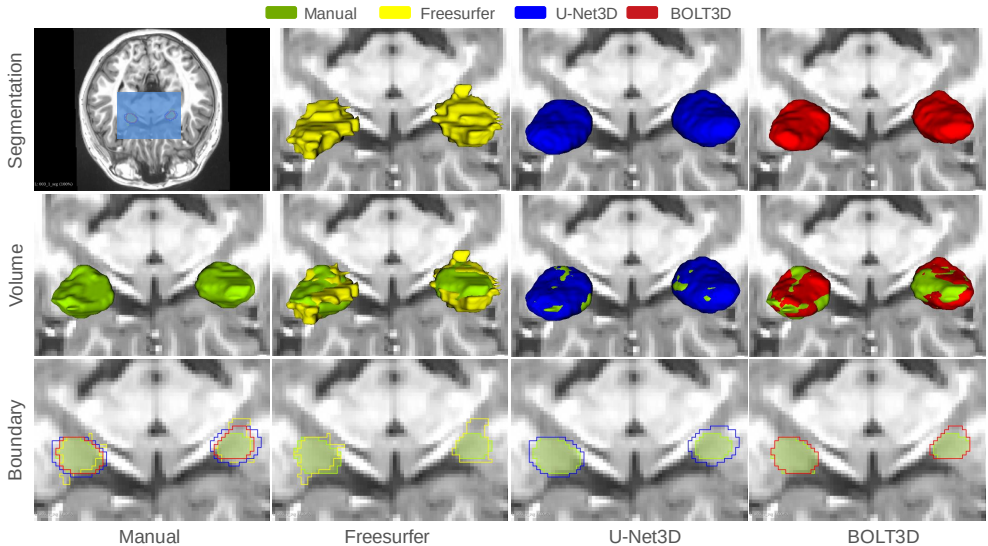

Fig. 2: A comparative visual examination of the amygdala in a representative subject, illustrating the preservation of structural integrity (middle row) and precision in boundary delineation (bottom row). The manual segmentation (left column) is established as the reference standard for accuracy. Our proposed BOLT3D method (right column) demonstrates superior performance in both volumetry and boundary definition, outperforming Freesurfer (second column) and U-Nest3D (third column).

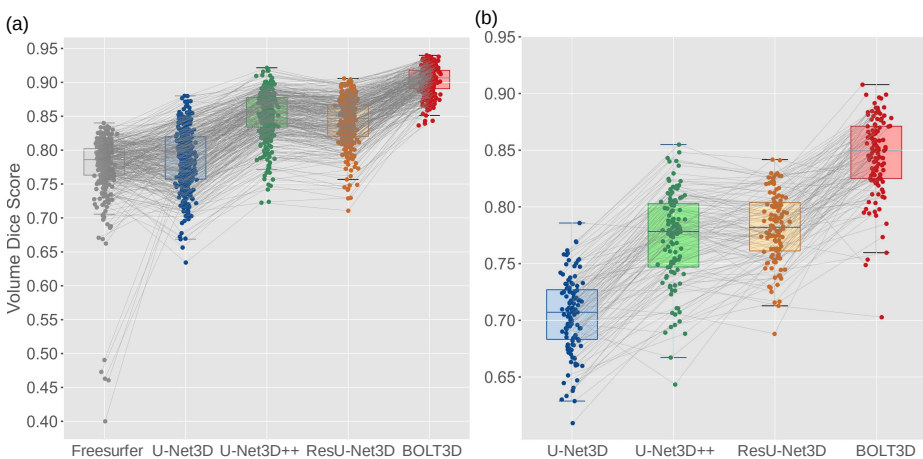

Fig. 3: Volume dice analysis in CKG-CCNC (a) and BNU-CAC (b) for the overall amygdala structure. As the CKG-CCNC was used during training, its overall performance is slightly higher than the unseen dataset of BNU-CAC for all deep-learning network methods, but the proposed BOLT3D achieves a dice score above 0.8 for both cases, even better than FreeSurfer.

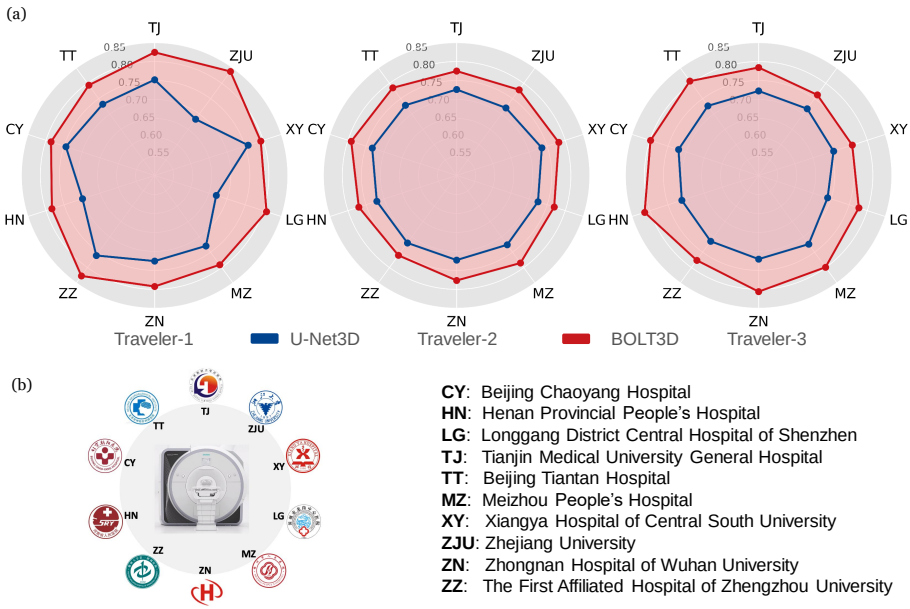

Fig. 4: Comparison of amygdala segmentation outcomes in itinerant volunteers from ZJU-MTC. (a) The proposed BOLT3D methodology (depicted in red) demonstrates superior performance relative to U-Net3D (depicted in blue) consistently across all centers for each participant, thereby underscoring its robust cross-center generalization efficacy. (b) All centers are equipped with identical model 3T scanners, geographically dispersed across China, with their respective names and abbreviations provided on the right.

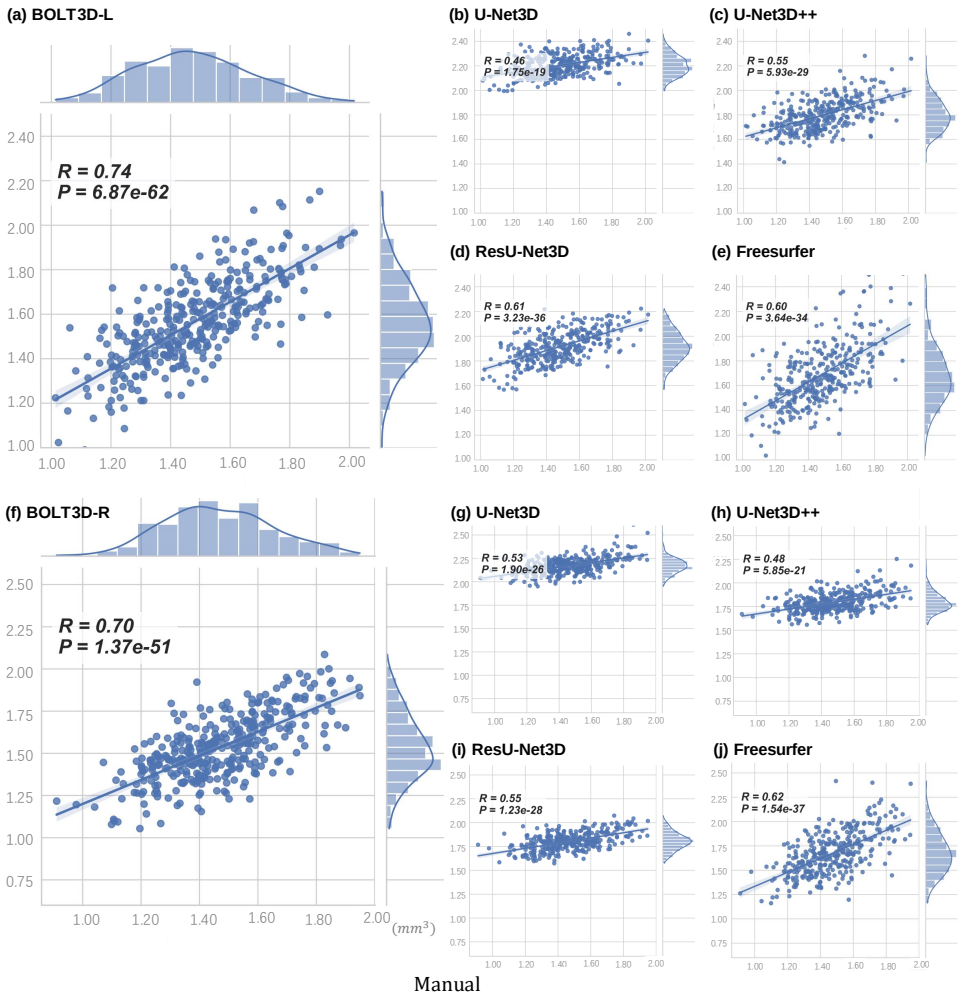

Fig. 5: Correlation analysis of amygdala volume as determined by segmentation algorithms in comparison with manual reference standards. The data set includes 347 scans, with the results of manual segmentation depicted on the horizontal axis. Among the evaluated methods, the proposed BOLT3D algorithm demonstrates the highest correlation with manual references for the left (a) and right (f) parts of the amygdala, signifying greater consistency in volumetric measurements. In contrast, the correlation coefficients for alternative segmentation tools, including U-Net3D (b, g), ResU-Net3D (c, h), U-Net3D++ (d, i) and Freesurfer (e, j), fall within a range of 0.46 to 0.60. The volume distributions derived from each algorithm are represented on the respective vertical axes.

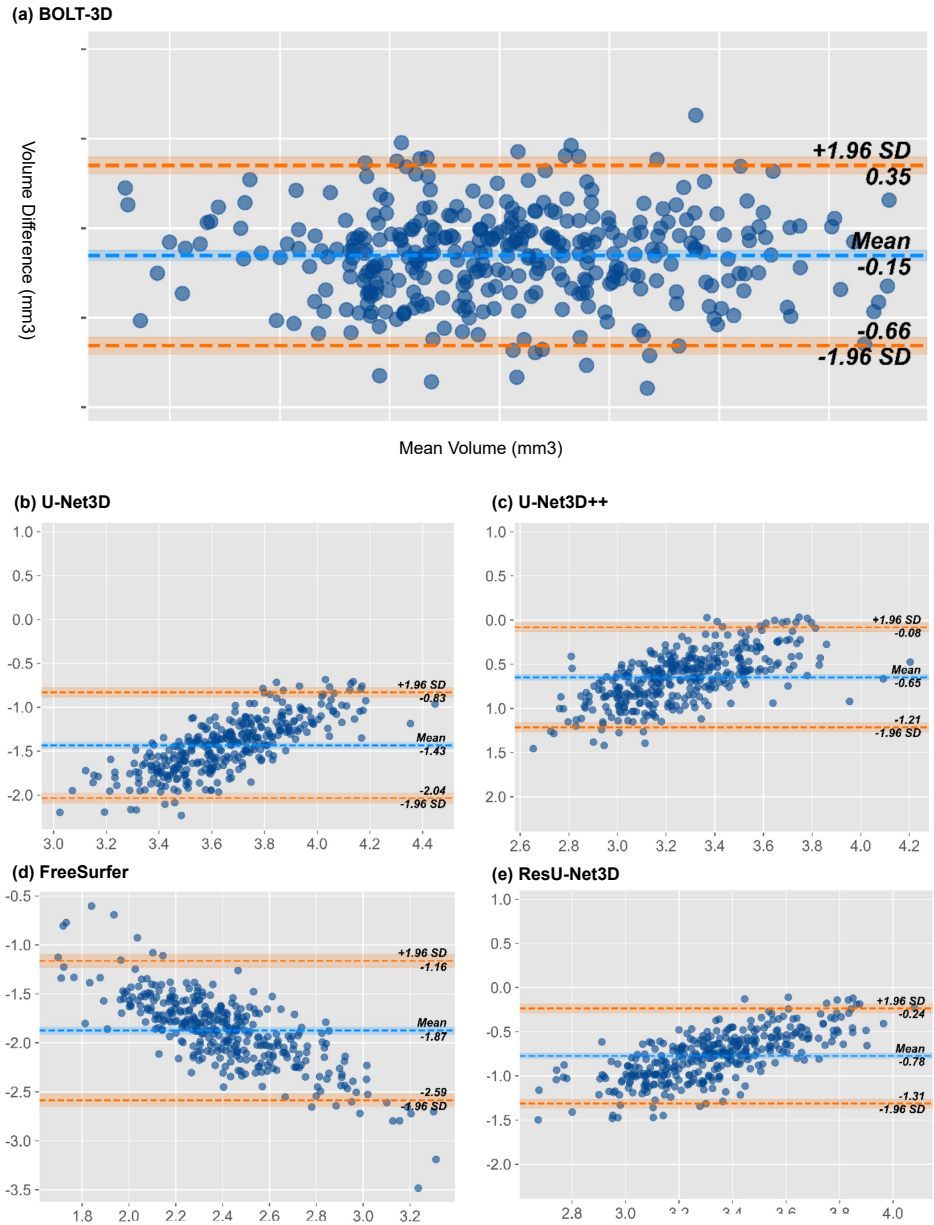

Fig. 6: Bland-Altman plots are employed to assess the concordance in volume size estimation between various segmentation algorithms and the manual reference standard across 347 scans for the left portion of the amygdala. The central dashed blue lines denote the mean differences, while the dashed yellow lines represent the limits of agreement at ± 1.96 standard deviations from the mean. Each blue dot represents the data for an individual subject.

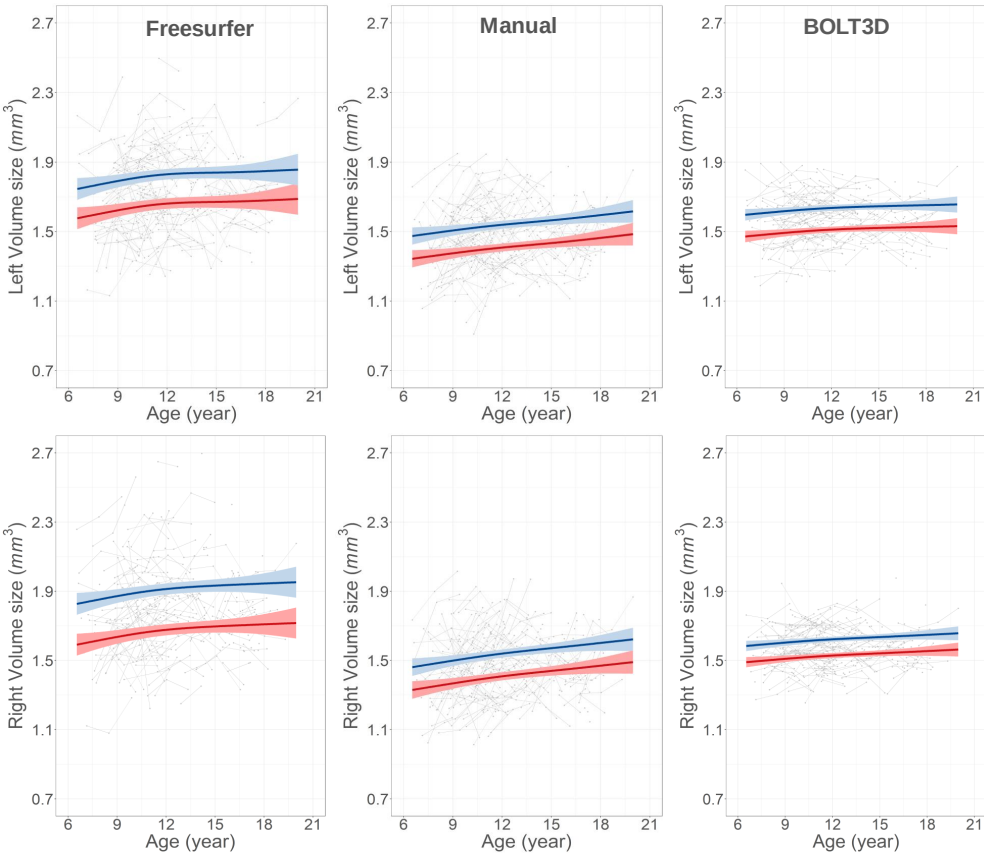

Fig. 7: The developmental trajectory of the amygdala, as delineated by our model and comparative methodologies, is presented. The FreeSurfer illustration (left) demonstrates an inverted U-shaped curve, indicative of a non-linear growth pattern. Conversely, the growth curves predicted by our BOLT3D model (right) exhibit a striking concordance with the linear trajectory observed in the manual tracing method (middle). Blue denotes trajectories for males, whereas red represents trajectories for females. The trajectories are enveloped by shaded 95% confidence intervals. Individual scatter points represent segmentation outcomes from distinct cases, with the lines connecting these points indicating results from multiple scans of the same participant.

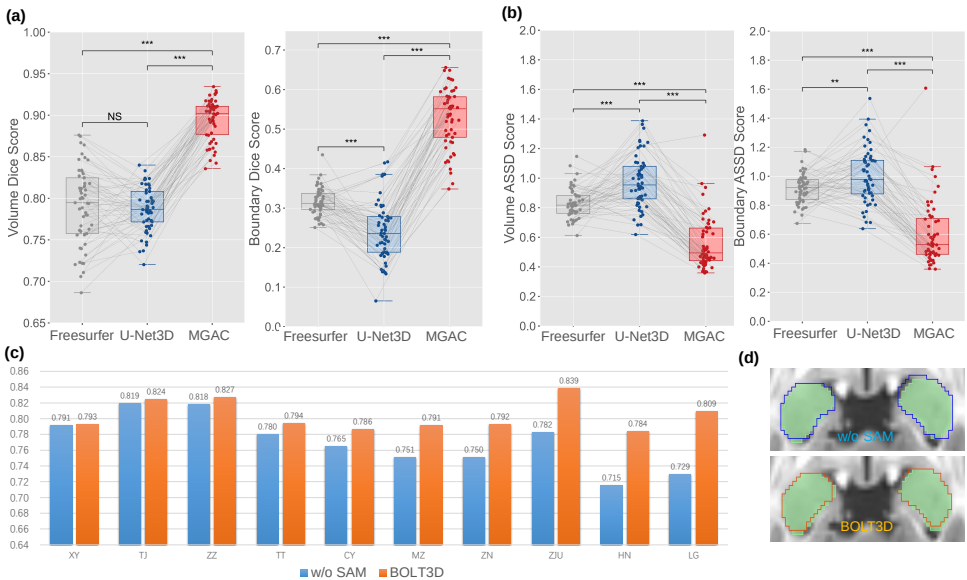

Fig. 8: Ablation experiments. The comparative analysis of consistency manifests that the transformer-based algorithm (MGAC) within BOLT3D demonstrates a significant diminution in discrepancies when contrasted with U-Net3D and Freesurfer, in terms of volumetric measurement and boundary delineation, as evidenced by their corresponding (a) Dice and (b) ASSD metrics. This illustrates the superior precision of our algorithm in accurately tracing structural margins and effectively reducing the risk of volumetric overestimation. (c) The integration of the SAM module, in conjunction with the MGAC module, markedly enhances the robustness of BOLT3D in the ZJU-MTC study, as quantitatively assessed through Dice scores. (d) A representative instance both with and without the SAM module, to elucidate the impact of these enhancements on segmentation quality.