UKBOB: One Billion MRI Labeled Masks for Generalizable 3D Medical Image Segmentation
Supplementary Material

Emmanuelle Bourigault Amir Jamaludin Abdullah Hamdi
Visual Geometry Group, University of Oxford
emmanuelle@robots.ox.ac.uk
amirj@robots.ox.ac.uk
abdullah.hamdi@eng.ox.ac.uk

A. Detailed Setup

A.1. Datasets

We conducted our experiments on four primary datasets: **UK Biobank.** A more comprehensive dataset of 51,761 full-body MRIs from more than 50,000 volunteers [16], capturing diverse physiological attributes across a broad demographic spectrum. UK Biobank MRIs are resampled to be isotropic and cropped to a consistent resolution (501 $\times$ 160 $\times$ 224).

BRATS. The largest public dataset of brain tumours consisting of 5,880 MRI scans from 1,470 brain diffuse glioma patients, and corresponding annotations of tumours [1, 2, 12]. All scans were skull-stripped and resampled to 1 mm isotropic resolution. All images have resolution 240 $\times$ 240 $\times$ 155. Tumours are annotated by expert clinicians for three classes: Whole Tumour (WT), Tumour Core (TC), and Enhanced Tumour Core (ET).

BTCV. BTCV (Beyond the Cranial Vault) abdomen dataset [5]. This dataset involves 30 training and 20 testing subjects and 13 labelled organs: spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal vein and splenic vein, pancreas, right adrenal gland and left adrenal gland. We combine the left and right adrenal gland into one. Scans are resampled to consistent resolution (224 $\times$ 224 $\times$85) and intensity scaled in the range [-175,250] Hounsfield Units (HU).

AMOS. AMOS Abdomen MRI [10] from the MICCAI AMOS Challenge, which consists of segmentation of abdominal organs from 100 MRI scans split equally into train and test sets. The organs include the liver, spleen, pancreas, kidneys, stomach, gallbladder, esophagus, aorta, inferior vena cava, adrenal glands, and duodenum. Scans are resampled to consistent resolution (256 $\times$ 256 $\times$ 125) and scans normalised for intensity channel wise in the range [0,1].

A.2. Evaluation Metrics

- **Dice Score** The Dice Score, or Dice Coefficient, is a statistical measure used to assess the similarity between two samples. It is widely utilized in medical image analysis due to its sensitivity to variations in object size. The Dice Score is calculated by doubling the area of overlap between the predicted and ground truth segmentations and dividing by the total area of both. The formula is:

$$\text{Dice} = \frac{2 \times \text{Area}(S_{\text{pred}} \cap S_{\text{gt}})}{\text{Area}(S_{\text{pred}}) + \text{Area}(S_{\text{gt}})}$$

  This metric ranges from 0 to 1, with a value of 1 indicating perfect agreement between the prediction and the ground truth. The Dice Score is particularly robust against variations in the size of the segmented objects, making it extremely useful in medical applications where such variability is common.
  Both IoU and Dice Score offer comprehensive insights into model accuracy, with the Dice Score being especially effective in scenarios involving significant variations in object size.
- **Hausdorff Distance** The Hausdorff Distance is a metric used to measure the extent of discrepancy between two sets of points, often applied to evaluate the accuracy of object boundaries in image segmentation tasks. It is particularly useful for quantifying the worst-case scenario of the distance between the predicted segmentation boundary and the ground truth boundary.
  The Hausdorff Distance calculates the greatest distance from a point in one set to the closest point in the other set. In image segmentation, this involves finding the largest distance from any point on the predicted boundary to the nearest point on the ground truth boundary, and vice versa. The mathematical definition is:

$$\text{HD} = \max\left\{\sup_{p\in P}\inf_{q\in Q} d(p,q), \sup_{q\in Q}\inf_{p\in P} d(p,q)\right\}$$

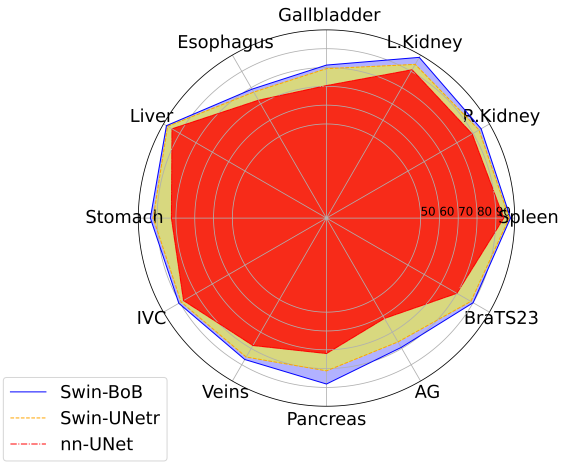

Figure I. **Per-Class Performance Comparison with Specialized Segmentation models.** We compare the Dice Score performance of our Swin-BOB model and baselines Swin-UNetr[7] and nn-UNet[9] on abdominal organ segmentation (BTCV) and brain tumour segmentation (BRATS).

where P and Q are the sets of boundary points of the predicted segmentation and the ground truth segmentation, respectively, and $d(p, q)$ represents the Euclidean distance between points p and q.

A.3. Segmentation Details

We perform a series of experiments to determine the best segmentation model on UKBOB using state-of-the-art multi-resolution CNN (UNet [15], SegResNet [13], nn-UNet [9]) and transformer-based networks (TransUNet [4], UNetr [8], Swin-UNetr [7]). We report segmentation performance in Table VII where Swin-UNetr outperforms baselines by a margin, followed by nn-UNet. We show visual examples of the 72 class labels in UKBOB in Figure IV and Figure V.

We also show detailed baseline comparison for BTCV and AMOS in Table I and Table II respectively. We provide radar plot in Figure I that summarizes the performance of our segmentation model Swin-BOB compared to baseline segmentation models on different classes from BTCV and BRATS23 class average.

We show visual comparison on BRATS (Figure III) of our model segmentation relative to ground-truth.

We also show the t-sne visualization of the features in Figure II illustrating the quality of the features.

Model	Mean Dice Score	Mean Hausdorff Distance
UNet[15]	0.782	8.374
SegResNet[13]	0.794	7.912
TransUNet[4]	0.838	6.258
UNetr[8]	0.856	4.317
Swin-Unetr[7]	0.869	3.801
nn-UNet[9]	0.802	6.782
AttentionUNet[14]	0.816	5.848

Table I. **Comparison of segmentation model performance on BTCV (n = 12 classes).**

Model	Mean Dice Score
TransBTS[19]	0.792
UNETR[8]	0.762
nnFormer[20]	0.790
SwinUNETR[7]	0.880
3D UX-Net[11]	0.900

Table II. **Comparison of Segmentation Models for AMOS Segmentation (n = 14 classes).**

Model	Mean Dice Score	Mean Hausdorff Distance
$\epsilon = 3$	0.891	7.126
$\epsilon = 2$	0.884	7.528
$\epsilon = 1$	0.792	8.247
$\epsilon = 4$	0.766	8.594
$\epsilon = 5$	0.745	8.972

Table III. **Effect of Filtration Threshold on Segmentation Performance on manual annotated set of abdomen organs (300) from UK Biobank.** The 11 abdomen organs and bones that have been manually annotated represent the overlap organs with BTCV[5] and UK Biobank[6].

Dataset	Mean Dice Score	Hausdorff Distance
AMOS	0.831	7.647
BTCV	0.837	5.138

Table IV. **Zero-shot performance on external datasets.**

B. Dataset Filtration Details

B.1. Threshold Selection

Full ablation experiments for threshold selection is available in Table III. Results on impact of filtration on BTCV and AMOS are reported in Table V. We therefore ensure high-quality labels by removing outliers adequately.

B.2. Zero-Shot Generalization

Our zero-shot evaluation on the AMOS and BTCV datasets highlights the robustness of filtered labels. Metrics are de-

Configuration	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	IVC	AG	Aorta
AMOS	0.9084	0.9311	0.9421	0.6516	0.6582	0.9581	0.8216	0.8740	0.5292	0.9062
AMOS + filtering	0.9102	0.9397	0.9508	0.6582	0.6673	0.9662	0.8315	0.8824	0.6209	0.9183
BTCV	0.883	0.884	0.932	0.795	0.790	0.946	0.885	0.871	0.784	0.799
BTCV + filtering	0.889	0.889	0.941	0.813	0.825	0.949	0.893	0.883	0.799	0.869

Table V. **Zero-shot 3D Segmentation Performance of Swin-BOB on AMOS external MRI data and CT (BTCV) for same organ classes**.

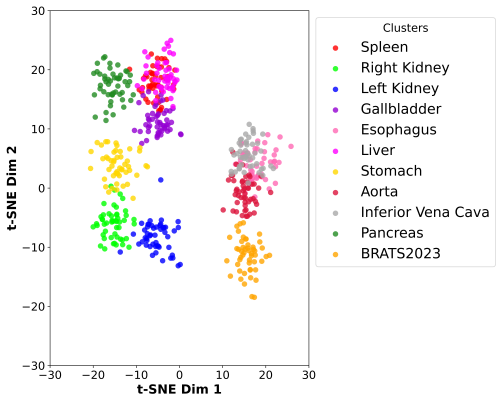

Figure II. **Distribution of Feature Embeddings on BTCV organs and BRATS23**. Each category is represented with a unique color. We reduce features embeddings to 2D for each class using t-sne [17]. The low dispersion of the clusters between each other indicates that the features of different classes probably share similar patterns and this explains the beneficial effect of large pre-training.

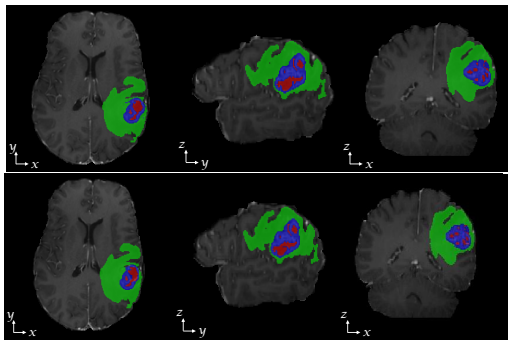

Figure III. **Qualitative Performance on BRATS.** We show the ground-truth *top* and output *bottom* of our pre-trained Swin-BOB model for 3D segmentation on the brain tumour BRATS dataset with 3 tumour class labels [1]. .

tailed in Table IV.

B.3. Residual Label Noise

While filtration reduces label noise, some false positives persist. To further improve the quality of the segmentation, we could incorporate human-in-the-loop approaches that turned efficient as shown in [3, 6].

Method	BTCV Dice	BTCV HD	AMOS Dice	AMOS HD	BRATS Dice	BRATS HD
no TTA	0.883	8.261	0.847	8.105	0.882	8.624
TENT [18]	0.887	7.940	0.852	7.861	0.886	8.042
ETTA (ours)	**0.892**	**7.381**	**0.864**	**7.191**	**0.894**	**7.130**

Table VI. **TTA baselines**. Comparison of Dice Score and Hausdorff Distance across three benchmarks using Swin-Bob model.

B.4. Filtering Out Patients Abnormalities

One concern of automatic filtration is that it might filter out some natural abnormalities or pathologies in the patients, mistaken as wrong labels. We visualize some of these filtered-out labels in Figure 9 (main paper) and show that indeed lack quality labels rather than the patients have obvious abnormalities. To quantify this behavior, we measure the 50-sample average LPIPS distance (the lower the more similar) between any two 3D mid-abdominal slices from full UKBOB (0.315), between filtered/filtered-out samples (0.329), between filtered/filtered samples (0.303), and between filtered-out/filtered-out samples (0.339). This shows that all the distances are almost identical, indicating mostly homogeneous organs in the dataset partitioning and hence the filtration is mostly about the quality of the labels rather than filtering out patients with abnormality.

C. Entropy Test-Time Adaptation (ETTA)

C.1. Algorithm Details

In this section, we detail the algorithmic process for our test-time adaptation (ETTA). It works by refining predictions minimizing entropy:

$$L_{\text{ent}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} p_{i,c}\log p_{i,c}.$$

During test time, only batch normalization parameters are fine-tuned while keeping other parameters fixed. The method is simple, and efficient computationally since it does not require retraining the full model. We show detailed step by step procedure in Algorithm 1.

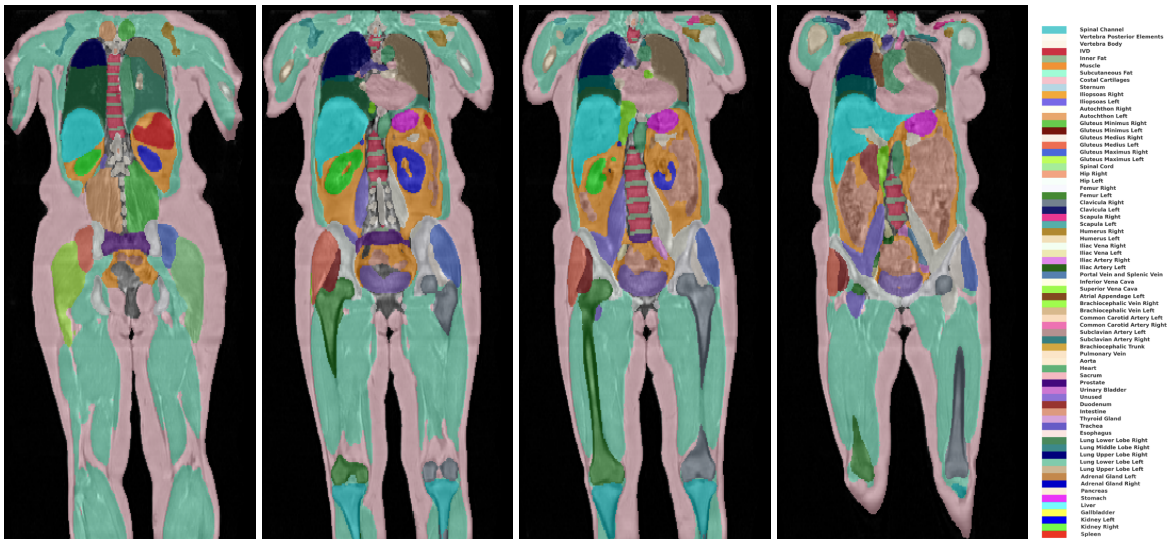

Figure IV. **Visualisation of UKBOB Segmentation Coronal Plane.** We show an example of 3D MRI from UKBOB for on coronal plane.

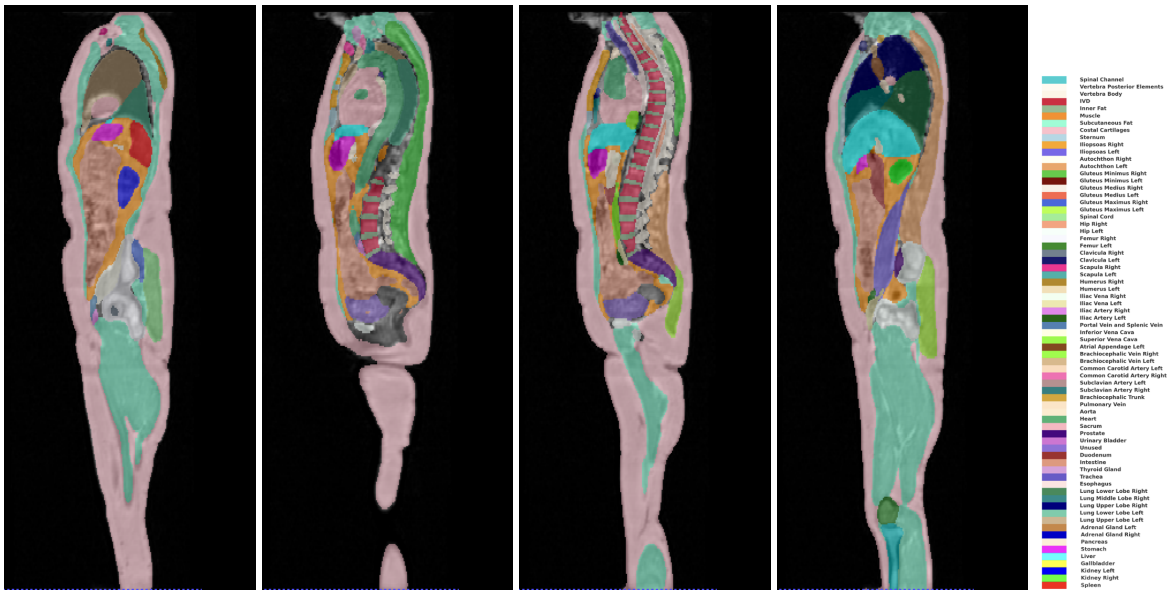

Figure V. **Visualisation of UKBOB Segmentation Sagittal Plane.** We show an example of 3D MRI from UKBOB for on sagittal plane.

D. Dataset Access and Code for Reproducibility

The dataset and pre-trained Swin-BOB models will be made available publicly via UK Biobank.

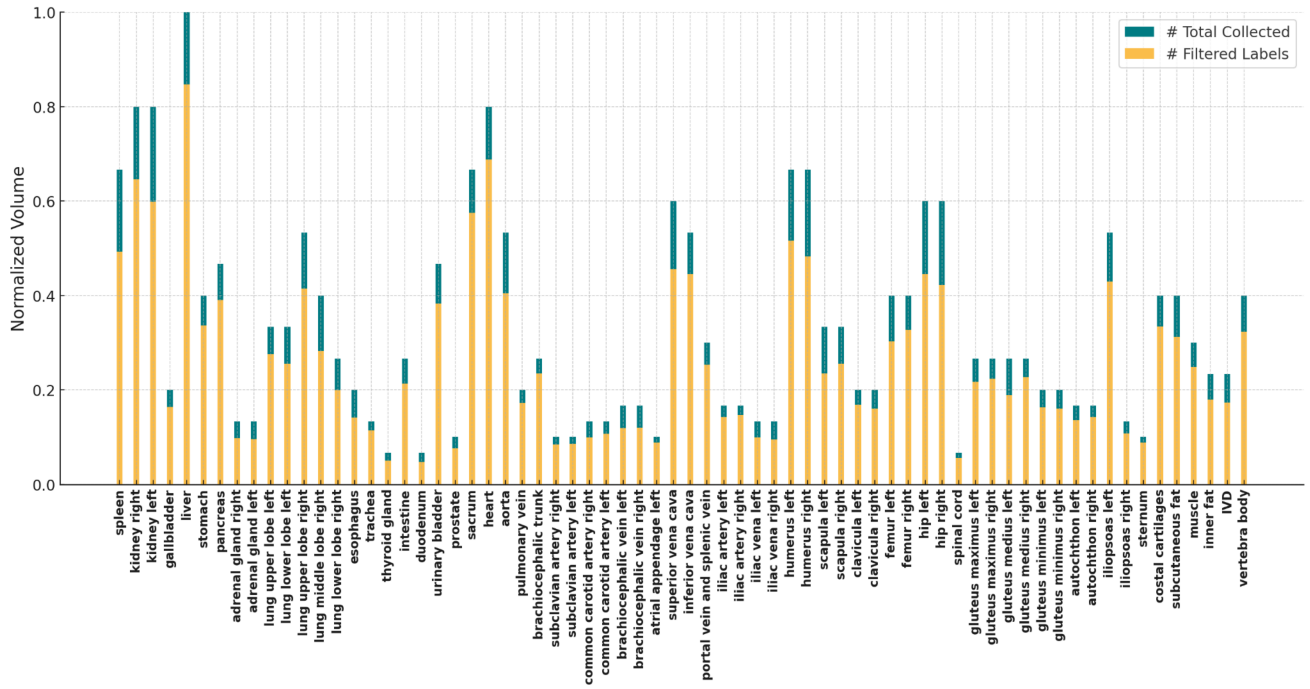

Figure VI. **UKBOB Distribution of Labels with our Filtration.** We show the distribution mean normalised volumes of 72 labels before and after filtration.

Model	ResUNet	UNetr	nnUNet	Swin-UNetr	MedFormer
spleen	0.91	0.92	0.94	0.94	0.93
kidney right	0.87	0.89	0.91	0.92	0.90
kidney left	0.88	0.90	0.92	0.93	0.91
gallbladder	0.82	0.84	0.85	0.85	0.84
liver	0.94	0.96	0.97	0.96	0.96
stomach	0.88	0.89	0.90	0.91	0.89
pancreas	0.85	0.87	0.89	0.90	0.88
adrenal gland right	0.81	0.83	0.84	0.86	0.84
adrenal gland left	0.81	0.83	0.84	0.86	0.83
lung upper lobe left	0.93	0.94	0.96	0.96	0.95
lung lower lobe left	0.94	0.95	0.96	0.96	0.94
lung upper lobe right	0.94	0.95	0.96	0.96	0.94
lung middle lobe right	0.93	0.94	0.96	0.96	0.95
lung lower lobe right	0.93	0.95	0.96	0.96	0,95
esophagus	0.86	0.88	0.9	0.91	0.89
trachea	0.87	0.89	0.92	0.92	0.91
thyroid gland	0.74	0.75	0.76	0.77	0.75
intestine	0.87	0.91	0.93	0.92	0.91
duodenum	0.81	0.84	0.86	0.87	0.85
urinary bladder	0.89	0.93	0.95	0.96	0.94
prostate	0.91	0.92	0.94	0.94	0.94
sacrum	0.91	0.92	0.96	0.95	0.04
heart	0.92	0.96	0.97	0.97	0.96
aorta	0.91	0.93	0.95	0.94	0.93
pulmonary vein	0.87	0.89	0.91	0.92	0.91
brachiocephalic trunk	0.83	0.86	0.88	0.89	0.88
subclavian artery right	0.81	0.85	0.86	0.88	0.86
subclavian artery left	0.81	0.85	0.86	0.88	0.86
common carotid artery right	0.81	0.83	0.84	0.86	0.86
common carotid artery left	0.81	0.83	0.85	0.87	0.85
brachiocephalic vein left	0.82	0.85	0.88	0.89	0.85
brachiocephalic vein right	0.83	0.85	0.87	0.88	0.85
atrial appendage left	0.79	0.82	0.84	0.84	0.83
superior vena cava	0.89	0.91	0.93	0.93	0.91
inferior vena cava	0.89	0.90	0.92	0.92	0.90
portal vein and splenic vein	0.76	0.79	0.82	0.82	0.80
iliac artery left	0.83	0.85	0.87	0.87	0.85
iliac artery right	0.82	0.84	0.86	0.86	0.84
iliac vena left	0.85	0.88	0.91	0.91	0.89
iliac vena right	0.85	0.88	0.90	0.90	0.88
humerus left	0.90	0.93	0.94	0.93	0.93
humerus right	0.90	0.93	0.94	0.94	0.93
scapula left	0.86	0.89	0.91	0.91	0.89
scapula right	0.88	0.89	0.91	0.91	0.89
clavicula left	0.86	0.88	0.90	0.90	0.88
clavicula right	0.86	0.88	0.90	0.90	0.88
femur left	0.91	0.94	0.97	0.96	0.95
femur right	0.90	0.93	0.96	0.95	0.93
hip left	0.92	0.95	0.97	0.98	0.96
hip right	0.91	0.94	0.96	0.97	0.95
spinal cord	0.85	0.87	0.88	0.90	0.88
gluteus maximus left	0.93	0.95	0.98	0.98	0.95
gluteus maximus right	0.93	0.95	0.98	0.98	0.95
gluteus medius left	0.94	0.97	0.98	0.98	0.97
gluteus medius right	0.94	0.97	0.97	0.97	0.97
gluteus minimus left	0.94	0.97	0.94	0.95	0.97
gluteus minimus right	0.94	0.97	0.94	0.95	0.97
autochthon left	0.94	0.96	0.97	0.97	0.96
autochthon right	0.94	0.96	0.97	0.97	0.96
iliopsoas left	0.92	0.94	0.96	0.96	0.95
iliopsoas right	0.91	0.93	0.96	0.96	0.95
sternum	0.86	0.88	0.92	0.92	0.89
costal cartilages	0.85	0.87	0.90	0.91	0.88
subcutaneous fat	0.89	0.92	0.95	0.96	0.93
muscle	0.91	0.93	0.96	0.97	0.94
inner fat	0.86	0.88	0.90	0.91	0.89
IVD	0.86	0.88	0.90	0.91	0.88
vertebra body	0.89	0.92	0.94	0.94	0.93
vertebra posterior elements	0.82	0.84	0.86	0.88	0.85
spinal channel	0.87	0.89	0.91	0.91	0.89
bone other	0.82	0.84	0.86	0.87	0.84

Table VII. **3D Segmentation Performance on UK Biobank dataset**. We compare our UKBOB on 3D medical segmentation task on the UK Biobank test set (n=10,353) with 5-fold cross validation compared to other methods using the average Dice score and average Haussdorff Distance (HD) per class as metric. Standard deviations are shown next to the mean Dice Score and HD values.

Algorithm 1 Algorithm for Test-Time Adaptation Using Batch Normalization

Require: Pre-trained segmentation model M, test dataset D_{test}, loss function $\mathcal{L}$ (optional), optimizer O (optional), epochs N_{epochs} (optional)
Ensure: Adapted model M'

Function FREEZEEXCEPTBN(M):
  For each parameter p **in** M:
    If p does not belong to a BatchNorm layer:
      $p.requires\_grad \leftarrow$ **False**
  End For
End Function
Function UPDATEBNSTATISTICS(M, D_{test}):
  Set M to training mode: $M.train()$
  FREEZEEXCEPTBN(M)
  For each batch $x \in D_{test}$:
    Compute predictions: $M(x)$
  End For
End Function
Function FINETUNEBN($M, D_{test}, \mathcal{L}, O, N_{epochs}$):
  Set M to training mode: $M.train()$
  FREEZEEXCEPTBN(M)
  For $epoch \leftarrow 1$ **to** N_{epochs}:
    For each batch $(x, y_{true}) \in D_{test}$:
      Compute predictions: $y \leftarrow M(x)$
      Compute loss: $\ell \leftarrow \mathcal{L}(y, y_{true})$
      Zero gradients: $O.zero\_grad()$
      Backpropagate: $\ell.backward()$
      Update parameters: $O.step()$
    End For
  End For
End Function
Define Function INFER(M, D_{test}):
  Set M to evaluation mode: $M.eval()$
  For each batch $x \in D_{test}$:
    Compute predictions: $y \leftarrow M(x)$
  End For
End Function

References

[1] Ujjwal Baid, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe Campos Kitamura, Sarthak Pati, Luciano M. Prevedello, Jeffrey D. Rudie, Chiharu Sako, Russell T. Shinohara, Timothy Bergquist, Rong Chai, James A. Eddy, Julia Elliott, Walter Caswell Reade, Thomas Schaffter, Thomas Yu, Jiaxin Zheng, BraTS Annotators, Christos Davatzikos, John T Mongan, Christopher Paul Hess, Soonmee Cha, Javier E. Villanueva-Meyer, John B. Freymann, Justin S. Kirby, Benedikt Wiestler, Priscila Sacilotto Crivellaro, Rivka R.Colen, Aikaterini Kotrotsou, Daniel Marcus, Mikhail Milchenko, Arash Nazeri, Hassan M. Fathallah-Shaykh, Roland Wiest, András Jakab, Marc-André Weber, Abhishek Mahajan, Bjoern H Menze, Adam E. Flanders, and Spyridon Bakas. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *ArXiv*, abs/2107.02314, 2021. 1, 3

[2] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data*, 4, Sept. 2017. 1

[3] Emmanuelle Bourigault, Amir Jamaludin, Emma Clark, Jeremy Fairbank, Timor Kadir, and Andrew Zisserman. 3d shape analysis of scoliosis. In *MICCAI Workshop on Shape in Medical Imaging*. Springer, 2023. 3

[4] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation, 2021. 2

[5] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction, 2020. 1, 2

[6] Robert Graf, Paul-Sören Platzek, Evamaria Olga Riedel, Constanze Ramschütz, Sophie Starck, Hendrik Kristian Möller, Matan Atad, Henry Völzke, Robin Bülow, Carsten Oliver Schmidt, et al. Totalvibesegmentator: Full body mri segmentation for the nako and uk biobank. *arXiv preprint arXiv:2406.00125*, 2024. 2, 3

[7] Ali Hatamizadeh, V. Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *ArXiv*, abs/2201.01266, 2022. 2

[8] Ali Hatamizadeh, Dong Yang, Holger R. Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1748–1758, 2021. 2

[9] Fabian Isensee, Paul F Jaeger, Simon A A Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203—211, February 2021. 2

[10] Yuanfeng Ji, Haotian Bai, Chongjian GE, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, and Ping Luo. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36722–36732. Curran Associates, Inc., 2022. 1

[11] Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A. Landman. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation, 2023. 2

[12] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *I E E E Transactions on Medical Imaging*, 34(10):1993 – 2024, 2015. 1

[13] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *BrainLes@MICCAI*, 2018. 2

[14] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018. 2

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. 2015. 2

[16] Cathie L. M. Sudlow, John E. Gallacher, Naomi E. Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin J Landray, Bette C Liu, Paul M. Matthews, Giok Ong, Jill P. Pell, Alan J Silman, Alan Young, Tim Sprosen, Tim C Peakman, and Rory Collins. 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, 2015. 1

[17] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 3

[18] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization, 2021. 3

[19] Wenxuan Wang, Chen Chen, Meng Ding, Jiangyun Li, Hong Yu, and Sen Zha. Transbts: Multimodal brain tumor segmentation using transformer, 2021. 2

[20] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation, 2022. 2