

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

Emmanuelle Bourigault

Amir Jamaludin Visual Geometry Group, University of Oxford

Abdullah Hamdi

emmanuelle@robots.ox.ac.uk

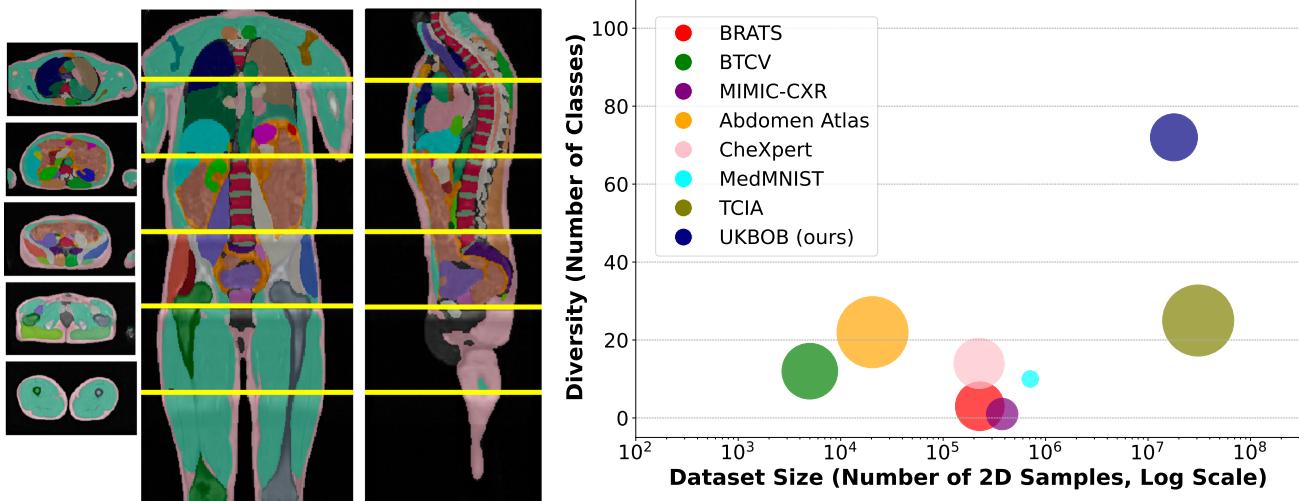


Figure 1. UKBOB Size and Diversity. Our proposed UK Biobank Organs and Bones (UKBOB) is the largest labeled medical imaging dataset for segmentation, comprising body organs of 51,761 MRI 3D samples (17.9 M 2D images) and a total of more than 1.37 billion 2D masks of 72 organs. *Left:* we show label examples from UKBOB from axial, coronal, and sagittal views. *Right:* We show a plot of the size (number of 2D images) and diversity (number of classes) of our UKBOB compared to other medical images datasets. The size of the bubbles indicates 2D image resolution. This new scale in dataset size and diversity should unlock a new wave of applications and methods in the computer vision and medical imaging communities.

Abstract

In medical imaging, the primary challenge is collecting large-scale labeled data due to privacy concerns, logistics, and high labeling costs. In this work, we present the UK Biobank Organs and Bones (UKBOB), the largest labeled dataset of body organs of 51,761 MRI 3D samples (17.9 M 2D images) and a total of more than 1.37 billion 2D segmentation masks of 72 organs based on the UK Biobank MRI dataset. We utilize automatic labeling, introduce an automated label cleaning pipeline with organ-specific filters, and manually annotate a subset of 300 MRIs with 11 abdominal classes to validate the quality (UKBOB-manual). This approach allows for scaling up the dataset collection while maintaining confidence in the labels. We further confirm the validity of the labels by the zero-shot generalization of trained models on the filtered UKBOB to other small labeled datasets from a similar domain (e.g. abdominal MRI). To fur-

ther elevate the effect of the noisy labels, we propose a novel Entropy Test-time Adaptation (ETTA) to refine the segmentation output. We use UKBOB to train a foundation model (Swin-BOB) for 3D medical image segmentation based on Swin-UNetr, achieving state-of-the-art results in several benchmarks in 3D medical imaging, including BRATS brain MRI tumour challenge (+0.4% improvement), and BTCV abdominal CT scan benchmark (+1.3% improvement). The pre-trained models and the code are available at <https://emmanuelleb985.github.io/ukbob>, while filtered labels will be made available with the UK Biobank.

1. Introduction

The advent of large-scale labeled datasets such as ImageNet [62] and LAION [63] has been a cornerstone in the remarkable progress of computer vision, enabling the devel-

opment of powerful foundation models [57, 60] that excel across various tasks. These models benefit immensely from the abundance of labeled data, which allows them to learn rich and generalizable representations. In stark contrast, the medical imaging domain grapples with a significant scarcity of large-scale labeled datasets due to stringent privacy regulations, complex logistics, and the high costs associated with expert annotations [61, 26]. This limitation hampers the development of robust and generalizable models for critical tasks such as 3D medical image segmentation, which is essential for disease diagnosis, treatment planning, and patient monitoring.

Previous medical image datasets either lack diversity or are too small for generalization [15, 2, 25]. Recognizing the pressing need for extensive and diverse medical imaging datasets, we introduce the **UK Biobank Organs and Bones (UKBOB)**, the largest labeled segmentation medical imaging dataset to date. Based on the UK Biobank MRI dataset [66], UKBOB comprises 51,761 3D MRI scans and over *one billion* 2D segmentation masks covering 72 organs. This dataset not only surpasses existing medical imaging datasets in scale but also in anatomical diversity, providing an unprecedented resource for training robust and generalizable models (see Figure 1). Table 1 highlights the differences in scope and other aspects between the different datasets. To collect the labels of UKBOB, we leverage automatic labeling based on the newly released TotalVibe Segmentator [16]. However, the automatic labeling of such a vast dataset introduces challenges related to label noise and quality assurance. To tackle this, we propose a novel mechanism for filtering organ labels based on a statistical Specialized Organ Labels Filter (SOLF). We also collect manual labels from 300 MRIs for 11 abdominal organs acting as validation (UKBOB-manual). We further account for noisy labels and dynamically refine the segmentation based on the model’s confidence using a novel Entropy Test-Time Adaptation (ETTA). These approaches ensure high-quality labels and enhances the model’s robustness. We validate the validity of the labels by demonstrating zero-shot generalization of the trained models on the filtered UKBOB dataset to other datasets from similar domains, such as the AMOS abdomen MRI dataset [30] and the BTCV abdomen CT dataset [15].

Leveraging the extensive UKBOB dataset, we train *Swin-BOB*, a foundation model for 3D medical image segmentation based on Swin-UNetr [21]. Our model achieves state-of-the-art performance on several benchmarks in 3D medical imaging, including the BRATS brain tumor MRI challenge [2] and the BTCV abdominal CT scan benchmark [15]. Our contributions can be summarized as follows:

Contributions: (i) We introduce UK Biobank Organs and Bones (UKBOB), the largest labeled dataset of organs, consisting of 51,761 MRI 3D samples and a total of 1.37 bil-

lion 2D segmentation masks of 72 organs based on the UK Biobank MRI dataset. (ii) We leverage automatic mechanisms for cleaning and filtering the labels based on body statistics and specialized organ filter, allowing for high-quality scale-up of the labels. The collected labels are validated by a subset of 300 manually annotated labels of 11 abdominal organs. (iii) To further elevate the effect of the noisy labels, we propose a novel Entropy Test-time Adaptation (ETTA) to refine the segmentation outputs. (iv) We train *Swin-BOB*, a foundation model for 3D medical image segmentation based on Swin-UNetr network [21], achieving state-of-the-art results on standard benchmarks in 3D medical imaging.

2. Related Work

3D Segmentation in Medical Imaging. Advancements in deep learning have significantly influenced 3D data processing, leading to various approaches such as point-based methods [54, 55], voxel-based methods [46, 11], and view-based methods [65, 19, 20, 45, 44, 18, 23]. In medical imaging, the U-Net architecture [61] revolutionized image segmentation with its symmetric encoder-decoder structure and skip connections, becoming widely used for tasks such as organ segmentation and tumor detection [53, 52, 36]. For 3D volumetric data like MRI or CT scans, 3D U-Net variants have extended this architecture by replacing 2D operations with 3D counterparts, enhancing performance in volumetric segmentation tasks. Recent developments in label-free segmentation utilize self-supervised learning and multimodal foundation models [81, 24, 17, 51, 35, 37, 12, 42, 79, 67, 8, 44] to segment 3D scenes without explicit labels. However, all of these models require large-scale medical datasets to generalize well [32, 25, 58, 78]. While datasets like MIMIC-CXR [32] and CheXpert [25] focus on chest imaging with extensive collections of X-ray images and associated clinical labels, they lack detailed segmentation masks necessary for advanced anatomical analysis. Datasets such as AbdomenAtlas-8K and Abdomen Atlas 1.1 [56, 41] provide valuable multi-organ CT scans with organ-level annotations but are limited to specific regions or modalities. The UK Biobank Imaging Study [66], one of the largest clinical trials, has collected extensive MRI data; however, prior works have not fully leveraged its potential for comprehensive organ segmentation. Our proposed *UK Biobank Organs and Bones (UKBOB)* dataset leverages this resource, presenting the largest labeled collection of MRI scans with detailed segmentation masks for 72 organs. By introducing a novel filtering mechanism based on normalized body statistics, we ensure high-quality labels while scaling up dataset collection, enabling the training of foundational models for 3D medical image segmentation with significant improvements over existing benchmarks.

Full-Body MRI Analysis. Most automatic MRI methods

Attribute	Medical Imaging Segmentation Datasets					
	BRATS [2]	BTCV[15]	MIMIC-CXR [31]	Abd.Atlas [40]	Total Segmentator [72]	UKBOB (ours)
Number of Classes	3	12	1	25	104	72
Number of 3D Samples	1,470	50	N/A	20,460	1,204	51,761
Total Number of 2D Images	227,850	5,000	377,110	673,000	400,000	17,902,080
Number of 2D Label Masks	581,715	425,000	N/A	16,825,000	5,800,000	1,378,913,040
Number of Patients	1,470	50	227,835	N/A	1,204	50,000
Meta Information	N/A	N/A	Text Reports	N/A	N/A	Bone Density + Fat %
Scope	Brain	Abdomen	Chest	Abdomen	Full-Body	Full-Body
Modality	MRI	CT	X-rays	CT	CT	MRI
Specialty	Tumour	Organs	COVID	Organs	Organs/Bones	Organs/Bones
2D Image Resolution (axial)	240 × 240	314 × 214	2500 × 3056	280 × 280	512 × 512	224 × 174

Table 1. **Comparison of Different Medical Imaging Segmentation Datasets.** We compare our proposed UKBOB to other well-known medical image segmentation datasets in terms of scope, size, and modality.

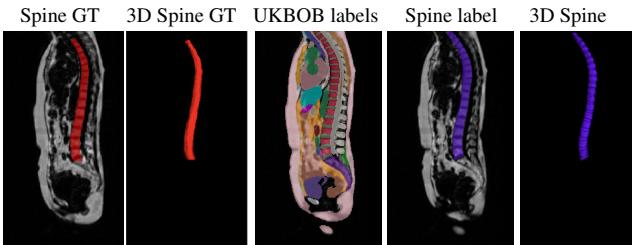


Figure 2. **Accuracy of UKBOB Labels.** An example of segmentation labels in UKBOB is shown in the sagittal view. The labels include “spine” (in purple) which we can compare to previously collected hand labels of the spine [5] (in red). We note that the newly collected labels match the manual labels in the spine with a total Dice score of 81.1% on a set of 250 manually annotated test samples, indicating accurate labels.

have focused on segmenting individual organs or tumors [9, 14, 73, 59], with limited research on whole-body scans. Studies that consider full-body imaging often emphasize the spine [28, 27, 74, 75, 6, 4], which is crucial for applications like scoliosis detection. Recently, [16, 1] released a full torso TotalVibeSegmentator first trained on a subset of NAKO (85 subjects) and UK Biobank (16 subjects) with a nnUNet[26] network. Their network, while useful, does not provide rich enough information to show improved performance on medical image segmentation tasks. Our work builds upon these efforts by using the TotalVibeSegmentator network to collect labels for the 51,761 samples of UK Biobank , filtering MRI labels and verifying their segmentation quality. This enables the training of a general MRI foundation model (Swin-BOB) that can generalize to various tasks and modalities.

Domain Adaptation in Medical Imaging. While U-Net-like networks and their variants perform well in supervised medical image segmentation, significant performance degradation occurs when the test data differs from training data due to variations in protocols, scanners, or modalities [43]. Test-time adaptation (TTA) addresses this by fine-tuning model parameters at test time using only test data without ground-truth [33]. Methods like TENT [69] minimize prediction entropy at test time to improve robustness and

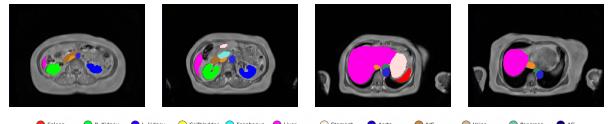


Figure 3. **UKBOB-Manual.** We collect manual labels for 300 samples of UKBOB for 11 abdominal organs totaling 3,000 images. UKBOB-manual acts as manual validation for the large UKBOB. Examples of axial slices are shown here.

segmentation performance. Augmentation-based Test-Time Adaptation is proposed [80] to improve on the domain gap issue. However, these approaches rely on well-calibrated models and may be sensitive to augmentation procedures. Recent work [13] integrates different predictions using various target domain statistics to enhance performance. Our work tries to address the issue of domain gap when the training domain has noisy labels making the adaptation even more challenging. Our model utilizes the confidence in predictions to adapt based on the entropy map on the test samples. This increases the model’s robustness across a wider domain gap.

3. Methodology

3.1. UKBOB Dataset Labels Collection

UKBiobank is a comprehensive dataset of 51,761 full-body MRIs from more than 50,000 volunteers[66], capturing diverse physiological attributes across a broad demographic spectrum. This dataset is unlabeled, which limits the potential applications for medical image understanding. We construct the UKBOB dataset by leveraging the UK Biobank MRI Study [66], which consists of 51,761? neck-to-knee 3D MRI scans. Each scan includes four sequences: fat-only, water-only, in-phase, and out-of-phase images. To obtain segmentation labels for $C = 72$ organs, we employ the TotalVibeSegmentator [16], an automatic segmentation tool trained on a subset of UK Biobank data. This approach allows us to generate over 1.37 billion 2D segmentation masks. Automatic labeling at this scale is crucial due to the impracticality of manual annotation. While it is not feasible to confirm the quality of 17.9M annotated images manually,

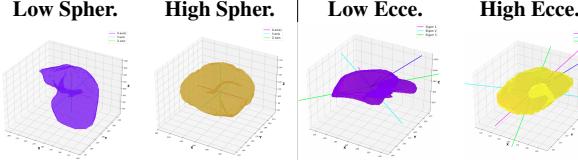


Figure 4. Specialized Organ Label Filter (SOLF). SOLF integrates sphericity, eccentricity, and normalized volume to statistically filter out inaccurate organ labels. From left to right, the panels display examples of low sphericity (0.21), high sphericity (0.95), low eccentricity (0.14), and high eccentricity (0.87).

we describe next a robust quality control mechanism on the collected labels to insure accurate labels.

3.2. Organ Labels Quality Control

Manual Labels for Verification. We design a mechanism to validate these collected labels by humans. To do so we collect manual labels from 3000 2D image from 300 MRI samples for 10 abdominal organs (UKBOB-manual). These manual labels (see examples in Figure 3) act as a validation for the large UKBOB dataset. On these labels, the UKBOB automatic labels obtain an average Dice Score of 0.891 (see Table 3). Furthermore, we verify the spine labels of UKBOB using previously collected manual labels of 200 3D spine labels [5]. We show an example in Figure 2 and we see how the new collected labels match the manual labels in the spine with a total Dice score of 0.811, indicating accurate labels. We discuss in Section 5.1 another mechanism for verifying the labels by zero-shot generalization of trained models to other similar datasets that has manual labels.

Specialized Organ Label Filter (SOLF). While automatic labeling enables creating large datasets, it introduces the possibility of noisy or erroneous labels. To mitigate this, we propose a filtration mechanism that removes outliers from segmentation. A question arises on how to distinguish segmentation failures from common patient abnormalities, *e.g.* enlarged liver. It's important to note that human organs follow typical geometric properties that arise from the body's need to optimize function while minimizing energy expenditure and structural stress. They reflect the underlying biological "blueprint" that has been honed by evolution [64], inspired by the evolutionary regularity of human organs [64], we propose the Specialized Organ Label Filter (SOLF), using three features *jointly*: *normalized volume*, *eccentricity*, and *sphericity* (illustrated in Figure 4).

For each organ class c , the normalized volume for some 3D sample is computed as $v_c = \frac{V_c}{V_{\text{body}}}$, where V_c is the voxel count for the organ $c \in \{1, 2, \dots, C\}$ and V_{body} is the total body voxel count. We define acceptable bounds for each feature by excluding the extreme ϵ percentiles. For example, the bounds for volume are set as

$$v_c^{\min} = P_{\epsilon/2}(\{v_c\}_{n=1}^N), \quad v_c^{\max} = P_{100-\epsilon/2}(\{v_c\}_{n=1}^N) \quad (1)$$

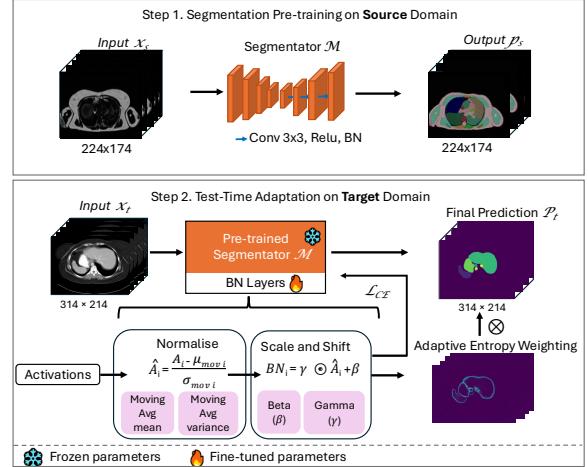


Figure 5. Entropy Test-Time Adaptation for Image Segmentation. We use a test-time entropy map to refine the batch norm layer of the network for robust segmentation output. This module is agnostic to the architecture of the deep neural network. Therefore, It can be used with any segmentation network to increase consistency and robustness, especially when trained with noisy labels.

, where $P_p(\cdot)$ denotes the p -th percentile function. Sphericity is defined as $\Phi_c = \pi^{1/3}(6V_c)^{2/3}/A_c$, with V_c computed from voxel counts and A_c as the surface area measured by counting the exposed voxel faces of the organ. Finally, eccentricity is defined as $E_c = \sqrt{1 - \lambda_{\min}/\lambda_{\max}}$, where λ_{\min} and λ_{\max} are the smallest and largest eigenvalues of the covariance matrix of organ c voxel coordinates. A sample is flagged as *inaccurate* if at least two of the three features (normalized volume, eccentricity, and sphericity) fall outside their respective acceptable ranges. Setting ϵ for SOLF effectively discards samples with anomalous organ characteristics while retaining valid labels. A single patient with abnormal organs is extremely unlikely to have more than a single independent aspect of deviation *at the same time*, hence indicating inaccurate labels.

We filter collected organ labels using the *patient's full-body statistics*, a novel approach compared to previous methods that rely on flat label statistics (IQR) [10, 38] rather than patient meta-information and organ-specific features .

3.3. Entropy Test-Time Adaptation (ETTA)

A common practice in medical imaging to addresses the labels distribution-shift is to employ Test-Time Adaptation (TTA) by fine-tuning model parameters at test time using only test data without ground-truth [33]. To mitigate the impact of any residual label noise at UKBOB, we introduce Entropy Test-Time Adaptation (ETTA). ETTA refines the model's predictions on test samples by fine-tuning the batch normalization parameters using the test data itself, guided by minimizing the prediction entropy. Given a test sample x , we first obtain the network's initial prediction $p = f_\theta(x)$, where

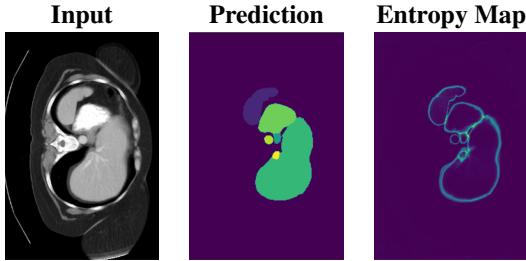


Figure 6. Entropy Map Visualization. We show from left to right an example of input, prediction and entropy map used in the Entropy Test-Time Adaptation on BTCV dataset [15] that can be leveraged to refine the output. Brighter regions indicate higher entropy.

$\mathbf{p} \in [0, 1]^{N \times C}$ is the softmax probability over C classes at each of the N voxels in the sample. Here, f_θ represents the segmentation network parameterized by θ . We define the entropy loss \mathcal{L}_{ent} over the predicted probabilities:

$$\mathcal{L}_{\text{ent}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C p_{i,c} \log p_{i,c} \quad (2)$$

where $p_{i,c}$ is the probability of class c at voxel i .

We update only the batch normalization parameters θ_{BN} while keeping the other network parameters θ_{fixed} frozen (see pipeline in Figure 5). The adaptation process involves minimizing the entropy loss with respect to θ_{BN} :

$$\theta_{\text{BN}}^* = \arg \min_{\theta_{\text{BN}}} \mathcal{L}_{\text{ent}}(f_{\theta_{\text{fixed}}, \theta_{\text{BN}}}(\mathbf{x})) \quad (3)$$

This process adapts the model to the test sample by encouraging confident (low-entropy) predictions, thereby refining the segmentation output. The adaptation is efficient as it involves updating a small subset of parameters and can be performed online during inference. ETTA leverages the entropy of the network’s predictions (Figure 6) to guide the adaptation, improving robustness to domain shifts and label noise.

4. Experiments

4.1. Evaluation Datasets and Metrics

Evaluation Datasets. We evaluate our model on multi-modal publicly available datasets i.e AMOS [30] of 13 abdominal organs from 100 MRI scans split equally into train and test sets, BTCV (Beyond the Cranial Vault) abdomen CT dataset [15] of 30 training and 20 testing subjects and 13 labelled organs, and BRATS [2, 47, 3], the largest publicly available dataset for brain tumors 5,880 MRI scans and corresponding annotations.

Evaluation Metrics. We evaluate our model using the Dice Score and the Hausdorff Distance Metric, which are widely used in medical image segmentation [43, 34]. The *Dice Score* measures the overlap between predicted and ground

Model	Dice Score	Hausdorff Distance
ResUNet++[29]	0.876	9.431
MedFormer [71]	0.881	8.822
nnUNet[26]	0.915	6.442
UNetr[22]	0.902	7.968
Swin-UNetr[21]	0.918	5.984

Table 2. UKBOB 3D Segmentation Benchmark. We show results on test mean Dice Score (%) and mean Hausdorff Distance ($n = 72$ classes) of our proposed benchmark on UKBOB. Note how Swin-UNetr [21] achieves the best results, resulting in our Swin-BOB foundation model.

truth masks, while the *Hausdorff Distance* assesses the boundary discrepancy, providing a comprehensive evaluation of segmentation performance.

4.2. Baselines

To evaluate our model, we compared it with a variety of established baselines across the BTCV, BRATS, and UK-BOB benchmarks. For the BTCV dataset, baseline models included UNet [61], SegResNet [49], TransUNet [7], UNetr [22], Swin-UNetr[21], nn-UNet [26], and AttentionUNet [50]. We also evaluate the base TotalVibe Segmentator (TVS) [16] in zero-shot and finetuning settings. These models represent widely adopted architectures in medical image segmentation, offering a range of network designs from classic CNN-based approaches to transformer-based architectures. For the BTCV dataset, diffusion-based segmentation methods were also considered, including MedSegDiff[76] and MedSegDiff-V2[77]. Together, these baselines provide a comprehensive foundation for evaluating our model’s performance in 3D medical segmentation. On the BRATS2023 benchmark, additional baselines incorporated V-Net [48], ResUNet++ [29], and nnFormer [82], along with UNETR [22] and Swin-UNetr [21]. These models were selected to encompass a spectrum of segmentation methods specifically suited to brain tumor segmentation tasks.

4.3. Implementation Details

Pre-training. For pre-training on UKBiobank, we split the dataset into 80-10-10 for training, validation, and testing. The input is cropped to $96 \times 96 \times 96$ voxels from the 3D MRI. In training, for data augmentation, scans are intensity scaled, with random flipping along the 3 axes, random foreground cropping, random rotation 90 degrees with probability 10%, and random intensity shift with an offset of 0.1. In validation, scans are intensity-scaled. We use a batch size of 8, AdamW optimizer, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and cosine learning rate scheduler with a warm restart every 200 epochs. We start with the initial learning rate of 10^{-4} and decay of 10^{-5} . We train the model on 2 A6000 GPU for 3,000 epochs. We use binary Cross-Entropy and Dice Similarity Coefficient (DSC) as our loss function. The best model achieving the highest

Configuration	No Filter	IQR filter	SOLF Filter
Spine labels [5]	0.811	0.849	0.867
UKBOB-manual	0.873	0.877	0.891

Table 3. Precision of the collected UKBOB compared to Manual labels. We show the Dice Score of the collected UKBOB labels on subsets of manual labels on UKBOB for 200 Spines [5] and on 300 manual abdominal labels we collect (UKBOB-manual). Even without any filtration, UKBOB labels are precise, improving in precision with our designed SOLF statistical filter.

overall Dice score on the 72 classes is saved at validation.

Fine-tuning. The pre-trained Swin-BOB is used on BTCV, BRATS, and AMOS, keeping the same configuration as in training while reducing the warm-up scheduler to 50 epochs for a total of 500 epochs.

5. Results

5.1. Validating UKBOB Labels

Manual Verification. In Table 3 we validate the quality of the UKBOB labels and the organ quality control against manual spine annotation from [5] and our manually annotated 11 abdomen UKBOB organs (UKBOB-manual). Even without any filtration, UKBOB labels are precise, achieving Dice score of 0.811 and 0.873 on manual spines and UKBOB-manual respectively. We Also show that our SOLF filtering approach (when $\epsilon = 2$) increases Dice score by 0.056 compared to no filtering on the spine labels and by 0.018 on the labelled abdomen organs. We also show standard inter-quartile range filtering (IQR) [30] for comparison.

Zero-Shot Evaluations. We also rely on the zero-shot segmentation performance of a model trained solely on those collected labels to be evaluated on a similar domain, namely the AMOS Abdomen MRI dataset [30] and on MRI on BTCV dataset[15] in Table 5. AMOS shares 12 class labels with UKBOB while BTCV shares 11 class labels. In Table 5, we show 10 organs overlap between BTCV and AMOS where we omitted small organ i.e duodenum as it has also been omitted in baseline papers. We combined the left and adrenal gland into one class named AG. We train Swin-UNetr [21] from scratch on different filtration schemes of the UKBOB and run the evaluation on the test sets given by BTCV and AMOS (on the shared class labels) and report the mean Dice score and mean Haussdorff distance. We adjust the preprocessing (normalization to [0,1], and resizing) to ensure compatibility with the model’s pretrained dataset. These results highlight the importance of the filtration we followed ensuring better quality labels.

5.2. Swin-BOB: A Foundation Model for 3D Medical Image Segmentation

We train Swin-UNetr [21] on our filtered UKBOB for a foundation model (Swin-BOB) for 3D segmentation. We evaluate test performance on multiple downstream tasks

Model	Dice Score	Hausdorff Distance
UNet[61]	0.544	39.090
V-Net[48]	0.842	10.891
ResUNet++[29]	0.784	22.249
AttentionUNet[50]	0.798	20.048
nnFormer[82]	0.812	10.070
UNETR[22]	0.871	9.924
SegResNet[49]	0.890	8.650
Total Vibe Seg. [16]	0.830	8.973
Swin-UNetr[21]	0.886	9.016
Swin-BOB (ours)	0.894	8.650

Table 4. BRATS 3D Segmentation Benchmark. The proposed Swin-BOB model, pre-trained on UK Biobank organs and fine-tuned on BRATS2023 [2] archives state-of-the-art results on test mean Dice Score (%) and mean Hausdorff Distance ($n = 3$ classes). Baseline results are reported from the Swin-UNetr paper [21].

including BRATS brain MRI benchmark [2] in Table 4 and BTCV Abdomen CT [15] in Table 7 (examples in Figure 7). In both benchmarks, our Swin-BOB achieves state-of-the-art with up to 0.02 Dice score improvement and reduction of 2.4 in Mean Hausdorff Distance. We also establish a UKBOB benchmark with reported Dice score and Mean Hausdorff Distance of different networks in Table 2 to aid research in this direction.

5.3. Entropy Test Time Adaptation Results

In Table 6 we show the benefit of endowing different fine-tuned models with our proposed ETTA and show improvement on 3 different datasets’ test performance for 3D segmentation using 3 different networks (including Swin-BOB). In all the 3 networks, we compare the ETTA against augmentation-based test-time adaptation baseline [80]. This highlights the importance of ETTA in tackling the issue of domain shift in medical imaging especially when the training includes noisy labels, as in the case of the Swin-BOB model.

5.4. Analysis and Insights

Filtration Ablation Study. We study the effect of the filtration threshold ϵ in the SOLF filter (Eq (1) of the three features) on the zero-shot generalization of the models trained on the filtered subsets of UKBOB. For $\epsilon = 0, 1, 2, 3, 4$, and 5, the performance Dice score on BTCV is 0.792, 0.884, 0.892, 0.766, and 0.745, respectively. We also ablate the features used in the SOLF filter. In Table 5, when only the normalized volume is used in the SOLF filter(no Sphericity or Eccentricity), the quality of the filtration degrades considerably, highlighting the importance of each aspect of the SOLF filter to clean the labels.

Filtering Out Patients Abnormalities. One concern of automatic filtration in Section 3.2 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 and show that indeed lack quality labels

Dataset	Config.	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	IVC	AG	Aorta	Mean
AMOS	TVS	0.823	0.697	0.814	0.782	0.786	0.802	0.759	0.738	0.879	0.881	0.796
	no filter	0.908	0.931	0.942	0.657	0.658	0.958	0.822	0.874	0.529	0.906	0.818
	+ vol. filter	0.910	0.940	0.951	0.658	0.667	0.966	0.832	0.882	0.621	0.918	0.832
	+ SOLF filter	0.919	0.943	0.962	0.664	0.672	0.969	0.838	0.882	0.631	0.924	0.840
BTCV	TVs	0.848	0.721	0.801	0.785	0.797	0.795	0.737	0.715	0.861	0.862	0.792
	no filter	0.883	0.884	0.932	0.795	0.790	0.946	0.885	0.871	0.784	0.799	0.856
	+ vol. filter	0.881	0.870	0.939	0.812	0.801	0.925	0.873	0.858	0.781	0.852	0.875
	+ SOLF filter	0.891	0.890	0.949	0.823	0.881	0.897	0.899	0.891	0.824	0.871	0.882

Table 5. **Detailed Zero-shot 3D Segmentation Performance.** We show Zero-shot Test Dice Score of Swin-BOB on AMOS external MRI data and CT (BTCV) for same organ classes. We show 10 organs that overlap between BTCV and AMOS where we combined left and right adrenal gland into one class named AG while inferior vena cava is briefed as IVC. TotalVibe Segmentator model (TVs) [16] results are shown for reference, while the Swin-BOB model is trained on complete UKBOB without filtration, with normalized volume statistical filter, and with full SOLF filter respectively. Note the significant benefit of filtering the collected UKBOB labels.

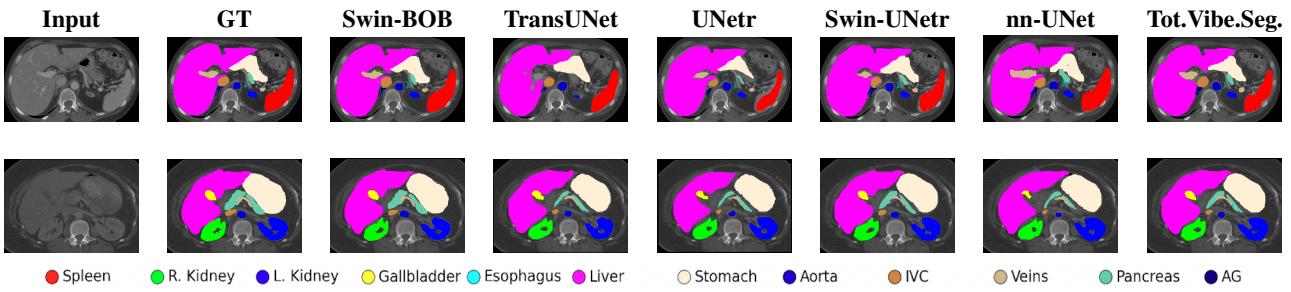


Figure 7. **Qualitative Results on BTCV.** We show comparisons of 3D segmentation on the abdomen BTCV dataset with 12 organ labels [15]. Note the significant improvement of our Swin-BOB especially on *Stomach* in the first row and *Pancreas* in the second row

Configuration	BTCV		AMOS		BRATS	
	Mean Dice	Mean HD	Mean Dice	Mean HD	Mean Dice	Mean HD
nn-UNet [26]	0.804	12.141	0.795	9.623	0.812	9.787
nn-UNet + TTA [80]	0.811	10.901	0.830	8.465	0.832	8.327
nn-UNet + ETTA (ours)	0.831	8.652	0.826	7.683	0.848	7.874
Swin-UNetr [21]	0.872	8.517	0.822	8.390	0.885	8.929
Swin-UNetr + TTA [80]	0.870	8.280	0.839	7.726	0.880	8.654
Swin-UNetr + ETTA (ours)	0.886	7.221	0.858	5.812	0.894	7.463
Swin-BOB (ours)	0.883	8.261	0.847	8.105	0.882	8.624
Swin-BOB + TTA [80]	0.883	7.901	0.857	5.651	0.887	7.712
Swin-BOB + ETTA(ours)	0.892	7.381	0.864	7.191	0.894	7.130

Table 6. **Effect of Entropy Test-Time Adaptation (ETTA).** We demonstrate that the proposed ETTA enhances the performance of fine-tuned models on the BTCV [15], BRATS [2], and AMOS [30] datasets. The best results are achieved by fine-tuning our baseline Swin-BOB—pre-trained on the UKBOB dataset. Our ETTA consistently improves performance across various networks and downstream tasks, outperforming the standard TTA baseline [80].

rather than the patients have obvious abnormalities. The combination of normalized volume, sphericity and eccentricity makes the filtration mostly about the quality of the labels rather than filtering out patients with abnormality. Figure 8 shows the distribution of organs normalized volumes in UKBOB before and after filtration.

Scaling-Law of 3D Medical Segmentation. We study the impact of the scale of training data of UKBOB on the downstream 3D segmentation performance, to justify the large scale UKBOB. We independently trained the Swin-UNetr network on subsets of UKBOB (*i.e.* 10%, 20%, 40%, 60%,

and 80% and show in Figure 10 the Average test Dice Score for both BTCV [15] and BRATS [2]. We also show the t-sne visualization of the features in Figure 11 illustrating the quality of the features.

6. Conclusions and Future Works

In this work, we introduced the UKBOB dataset, the largest labeled medical imaging dataset to date, comprising 51,761 3D MRI scans and over 1.37 billion 2D segmentation masks covering 72 organs. Our models trained on UKBOB demonstrate strong zero-shot generalization to other medical

Model	Spleen	R.Kid.	L.Kid.	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Veins	Panc.	AG	Avg.
TransUNet [7]	0.952	0.927	0.929	0.662	0.757	0.969	0.889	0.920	0.833	0.791	0.775	0.637	0.838
UNet [22]	0.968	0.924	0.941	0.750	0.766	0.971	0.913	0.890	0.847	0.788	0.767	0.741	0.856
Swin-UNet [21]	0.971	0.936	0.943	0.794	0.773	0.975	0.921	0.892	0.853	0.812	0.794	0.765	0.869
nnUNet [26]	0.942	0.894	0.910	0.704	0.723	0.948	0.824	0.877	0.782	0.720	0.680	0.616	0.802
Total Vibe Seg. [16]	0.948	0.914	0.917	0.736	0.741	0.954	0.859	0.881	0.794	0.752	0.718	0.699	0.826
MedSegDiff [76]	0.973	0.930	0.955	0.812	0.815	0.973	0.924	0.907	0.868	0.825	0.788	0.779	0.879
Swin-BOB (ours)	0.979	0.951	0.967	0.815	0.792	0.984	0.937	0.909	0.870	0.882	0.832	0.796	0.892
MedSegDiff-V2 (ens.)[77]	0.978	0.941	0.963	0.848	0.818	0.985	0.940	0.928	0.869	0.823	0.831	0.817	0.895
Swin-BOB (ens.)	0.981	0.958	0.971	0.817	0.796	0.988	0.942	0.912	0.874	0.886	0.836	0.799	0.897

Table 7. **3D Segmentation Performance on the BTCV Benchmark.** We evaluate our approach on the 3D segmentation task of the BTCV dataset using the Dice score. For a fair comparison, we also report 10-fold ensembling results (denoted as *ens.*) as presented in MedSegDiff-V2 [77].

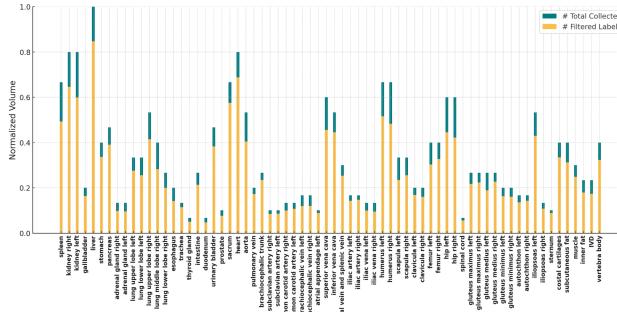


Figure 8. **UKBOB Distribution of Labels with our Filtration.** We show the distribution of mean normalized volumes of 72 labels before and after SOLF filtration. More examples and classes details are available in *Appendix*.

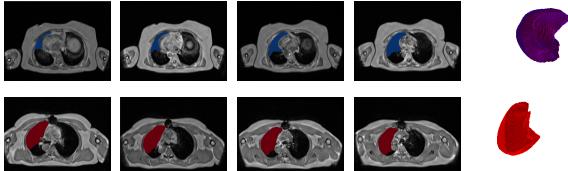


Figure 9. **Filtration of Inaccurate Labels.** *top* are manual labels of upper left lung overlaid on scans for several slice indices. *bottom* are filtered out labels (red) of the upper left lung overlaid on scans with corresponding slice indices. The filtered out lung is incomplete and erroneous.

imaging datasets and achieve state-of-the-art performance on several benchmarks in 3D medical image segmentation.

Limitations and Future Works. While UKBOB significantly expands the availability of large-scale labeled data for medical imaging, it is limited to neck-to-knee MRI scans and may not encompass the full diversity of imaging modalities and anatomical regions. Despite our filtration process, the automatic labeling may still introduce residual label noise that could impact model training. Future work includes extending the dataset to cover additional imaging modalities such as CT scans and incorporating more anatomical regions. Additionally, exploring advanced adaptation techniques and

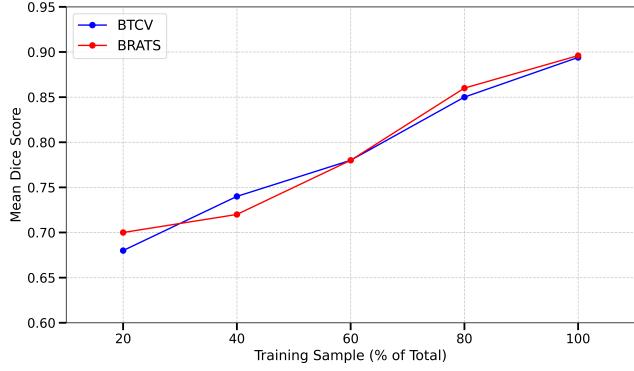


Figure 10. **Effect of UKBOB Pre-Training Dataset Size on Downstream Segmentation Performance.** We observe a consistent increase in test Dice Score for both BRATS [2] and BTCV [15] when doubling the size of pre-training Swin-BOB on UKBOB, acting as a foundation model.

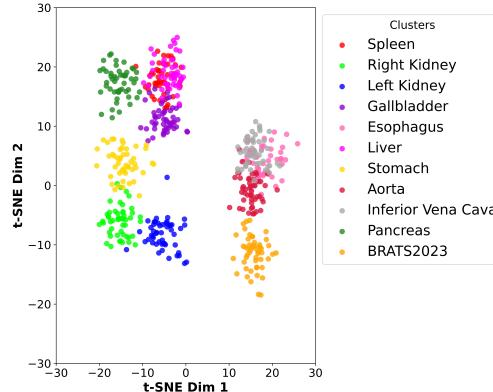


Figure 11. **Distribution of Feature Embeddings on BTCV organs and BRATS23.** Each category is represented with a unique color. We reduce features embeddings to 2D for each class using t-sne [68]. The low dispersion of the clusters between each other indicates that the features of different classes probably share similar patterns and this explains the beneficial effect of large pre-training.

integrating clinical metadata could enhance model robustness and applicability across diverse clinical settings.

Acknowledgments. This work was supported by the Centre

for Doctoral Training in Sustainable Approaches to Biomedical Science: Responsible and Reproducible Research (SABS: R3), University of Oxford (EP/S024093/1), and by the EPSRC Programme Grant Visual AI (EP/T025872/1). We are also grateful for the support from the Novartis-BDI Collaboration for AI in Medicine. Part of the support is also coming from KAUST Ibn Rushd Postdoc Fellowship program.

References

- [1] Tugba Akinci D'Antonoli, Lucas K. Berger, Ashraya K. Indrakanti, Nathan Vishwanathan, Jakob Weiss, Matthias Jung, Zeynep Berkarda, Alexander Rau, Marco Reisert, Thomas Küstner, Alexandra Walter, Elmar M. Merkle, Daniel T. Boll, Hanns-Christian Breit, Andrew Phillip Nicoli, Martin Segeroth, Joshy Cyriac, Shan Yang, and Jakob Wasserthal. Totalsegmentator mri: Robust sequence-independent segmentation of multiple anatomic structures in mri. *Radiology*, 314(2):e241613, 2025. PMID: 39964271. [3](#)
- [2] 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, Chi-haru 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. [2](#), [3](#), [5](#), [6](#), [7](#), [8](#), [13](#), [14](#)
- [3] 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. [5](#), [13](#)
- [4] Emmanuelle Bourigault, Abdullah Hamdi, and Amir Jamaludin. X-diffusion: Generating detailed 3d mri volumes from a single image using cross-sectional diffusion models, 2024. [3](#)
- [5] 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](#), [6](#), [14](#)
- [6] Emmanuelle Bourigault, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Scoliosis measurement on DXA scans using a combined deep learning and spinal geometry approach. In *Medical Imaging with Deep Learning*, 2022. [3](#)
- [7] 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. [5](#), [8](#), [13](#), [14](#)
- [8] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. [2](#)
- [9] Yuhua Chen, Dan Ruan, Jiayu Xiao, Lixia Wang, Bin Sun, Rola Saouaf, Wensha Yang, Debiao Li, and Zhaoyang Fan. Fully automated multi-organ segmentation in abdominal magnetic resonance imaging with deep neural networks. *Medical physics*, 2019. [3](#)
- [10] Junlong Cheng, Bin Fu, Jin Ye, Guoan Wang, Tianbin Li, Haoyu Wang, Ruoyu Li, He Yao, Junren Chen, JingWen Li, et al. Interactive medical image segmentation: A benchmark dataset and baseline. *arXiv preprint arXiv:2411.12814*, 2024. [4](#)
- [11] Christopher Choy, Jun Young Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. [2](#)
- [12] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. [2](#)
- [13] Haoyu Dong, Nicholas Konz, Hanxue Gu, and Maciej A. Mazurowski. Medical image segmentation with intent: Integrated entropy weighting for single image test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5046–5055, June 2024. [3](#)
- [14] Simon J. Doran, John H. Hipwell, Rachel Denholm, Björn Eiben, Marta, Busana, David John Hawkes, Martin O. Leach, and Isabel dos Santos Silva. Breast mri segmentation for density estimation breast mri segmentation for density estimation: Do different methods. 2017. [3](#)
- [15] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction, 2020. [2](#), [3](#), [5](#), [6](#), [7](#), [8](#), [13](#), [14](#)
- [16] 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](#), [5](#), [6](#), [7](#), [8](#), [14](#)
- [17] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2022. [2](#)
- [18] Abdullah Hamdi, Bernard Ghanem, and Matthias Nießner. Sparf: Large-scale learning of 3d sparse radiance fields from few input images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2930–2940, 2023. [2](#)
- [19] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtm: Multi-view transformation network for 3d shape recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. [2](#)

- ence on Computer Vision (ICCV), pages 1–11, October 2021. 2
- [20] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Voint cloud: Multi-view point cloud representation for 3d understanding. In *The Eleventh International Conference on Learning Representations*, 2023. 2
- [21] 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, 5, 6, 7, 8, 13, 14
- [22] 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. 5, 6, 8, 13, 14
- [23] Jan Held, Anthony Cioppa, Silvio Giancola, Abdullah Hamdi, Bernard Ghanem, and Marc Van Droogenbroeck. Vars: Video assistant referee system for automated soccer decision making from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5085–5096, 2023. 2
- [24] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson W. H. Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [25] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019. 2
- [26] 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, 3, 5, 7, 8, 13, 14
- [27] Amir Jamaludin, Jeremy Fairbank, Ian Harding, Timor Kadir, Tim J. Peters, Andrew Zisserman, and Emma M. Clark. Identifying scoliosis in population-based cohorts: Automation of a validated method based on total body dual energy x-ray absorptiometry scans. 2020. 3
- [28] Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Self-supervised learning for spinal mrис. *ArXiv*, abs/1708.00367, 2017. 3
- [29] Debeh Jha, Pia Helen Smedsrød, M. Riegler, Dag Johansen, Thomas de Lange, P. Halvorsen, and Håvard Dagenborg Johansen. Resunet++: An advanced architecture for medical image segmentation. *2019 IEEE International Symposium on Multimedia (ISM)*, pages 225–2255, 2019. 5, 6
- [30] Yuanfeng Ji, Haotian Bai, Chongjian GE, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, 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. 2, 5, 6, 7, 13
- [31] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6, 2019. 3
- [32] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6, 2019. 2
- [33] Neerav Karani, Ertunc Erdil, Krishna Chaitanya, and Endre Konukoglu. Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68:101907, 2021. 3, 4
- [34] Davood Karimi and Septimiu E Salcudean. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on medical imaging*, 39(2):499–513, 2019. 5
- [35] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [36] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 2
- [37] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2
- [38] Zeki Kuş and Musa Aydin. Medsegbench: A comprehensive benchmark for medical image segmentation in diverse data modalities. *Scientific Data*, 11(1):1283, 2024. 4
- [39] 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. 14
- [40] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro R. A. S. Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, Yutong Tang, Yining Cao, Haoqi Han, Zheyuan Zhang, Jiawei Liu, Tiezheng Zhang, Yujiu Ma, Jincheng Wang, Guang Zhang, Alan Yuille, and Zongwei Zhou. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking, 2024. 3
- [41] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro R. A. S. Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, Yutong Tang, Yining Cao, Haoqi Han, Zheyuan Zhang, Jiawei Liu, Tiezheng Zhang, Yujiu Ma, Jincheng Wang, Guang Zhang, Alan Yuille, and Zongwei Zhou. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking, 2024. 2
- [42] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3d an-

- notation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [43] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 3, 5
- [44] Jinjie Mai, Abdullah Hamdi, Silvio Giancola, Chen Zhao, and Bernard Ghanem. Egoloc: Revisiting 3d object localization from egocentric videos with visual queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 45–57, 2023. 2
- [45] Jinjie Mai, Wenxuan Zhu, Sara Rojas, Jesus Zarzar, Abdullah Hamdi, Guocheng Qian, Bing Li, Silvio Giancola, and Bernard Ghanem. Tracknerf: Bundle adjusting nerf from sparse and noisy views via feature tracks, 2024. 2
- [46] Daniel Maturana and Sebastian Scherer. VoxNet: A 3D Convolutional Neural Network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928, 2015-09. 2
- [47] 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. 5, 13
- [48] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016. 5, 6
- [49] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *BrainLes@MICCAI*, 2018. 5, 6, 13, 14
- [50] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Matthias 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. 5, 6, 14
- [51] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [52] Maximilian Achim Pfeffer and Sai Ho Ling. Evolving optimised convolutional neural networks for lung cancer classification. *Signals*, 3(2):284–295, 2022. 2
- [53] Adnan Qayyum, Syed Muhammad Anwar, Muhammad Majid, M. Awais, and Majdi R. Alnowami. Medical image analysis using convolutional neural networks: A review. *Journal of Medical Systems*, 42:1–13, 2017. 2
- [54] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017. 2
- [55] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems (NIPS)*, pages 5099–5108, 2017. 2
- [56] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Jie Liu, Yucheng Tang, Alan Yuille, and Zongwei Zhou. Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks, 2023. 2
- [57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2
- [58] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Mura: Large dataset for abnormality detection in musculoskeletal radiographs, 2018. 2
- [59] Ramin Ranjbarzadeh, Abbas Bagherian Kasgari, Saeid Jafarzadeh Ghoushchi, Shokofeh Anari, Maryam Naseri, and Malika Bendechache. Brain tumor segmentation based on deep learning and an attention mechanism using mri multi-modalities brain images. *Scientific Reports*, 11, 2021. 3
- [60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2
- [61] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. 2015. 2, 5, 6, 13, 14
- [62] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. 1
- [63] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmareczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1
- [64] Karthik Shetty, Annette Birkhold, Srikrishna Jaganathan, Norbert Strobel, Bernhard Egger, Markus Kowarschik, and Andreas Maier. Boss: Bones, organs and skin shape model. *Computers in Biology and Medicine*, 165:107383, 2023. 4
- [65] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 2
- [66] 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. 2, 3, 13
- [67] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Open-mask3d: Open-vocabulary 3d instance segmentation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

- [68] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8
- [69] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization, 2021. 3
- [70] Wenxuan Wang, Chen Chen, Meng Ding, Jiangyun Li, Hong Yu, and Sen Zha. Transbts: Multimodal brain tumor segmentation using transformer, 2021. 14
- [71] Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. *Advances in Neural Information Processing Systems*, 2024. 5
- [72] Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heße, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), Sept. 2023. 3
- [73] Rhydian Windsor and Amir Jamaludin. The ladder algorithm: Finding repetitive structures in medical images by induction. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1729–1733, 2020. 3
- [74] Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. A convolutional approach to vertebrae detection and labelling in whole spine mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020. 3
- [75] Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Self-supervised multi-modal alignment for whole body medical imaging. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2021. 3
- [76] Junde Wu, Huihui Fang, Yu Zhang, Yehui Yang, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *International Conference on Medical Imaging with Deep Learning*, 2022. 5, 8
- [77] Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: diffusion-based medical image segmentation with transformer. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI’24/IAAI’24/EAAI’24. AAAI Press, 2024. 5, 8
- [78] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1), Jan. 2023. 2
- [79] Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [80] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie A. Harmon, Baris I Turkbey, Bradford J. Wood, Holger R. Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging*, 39:2531–2540, 2020. 3, 6, 7
- [81] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [82] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation, 2022. 5, 6, 14

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[66], 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[2, 47, 3]. 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[15]. 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 [30] 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:

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

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[61], SegResNet[49], nn-UNet[26]) and transformer-based networks (TransUNet[7], UNetr[22], Swin-UNetr[21]). We report segmentation performance in Table VI 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 III and Figure IV.

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 II) of our model segmentation relative to ground-truth.

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.

Model	Mean Dice Score	Mean Hausdorff Distance
UNet[61]	0.782	8.374
SegResNet[49]	0.794	7.912
TransUNet[7]	0.838	6.258
UNetr[22]	0.856	4.317
Swin-Unetr[21]	0.869	3.801
nn-UNet[26]	0.802	6.782
AttentionUNet[50]	0.816	5.848

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

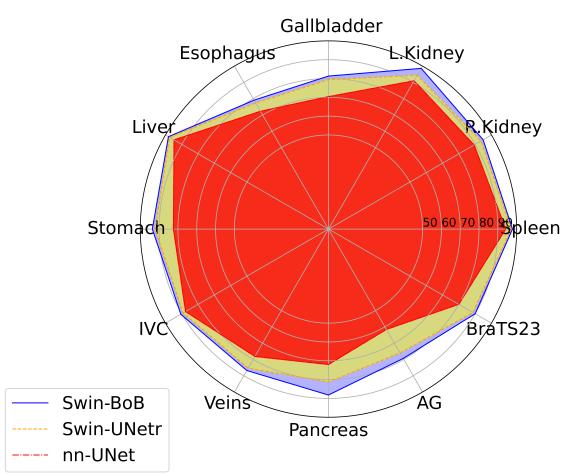


Figure I. **Per-Class Performance Comparison with Specialized Segmentation models.** We compare the Dice Score performance of our Swin-BOB model and baselines Swin-UNetr[21] and nn-UNet[26] on abdominal organ segmentation (BTCV) and brain tumour segmentation (BRATS).

Model	Mean Dice Score
TransBTS[70]	0.792
UNETR[22]	0.762
nnFormer[82]	0.790
SwinUNETR[21]	0.880
3D UX-Net[39]	0.900

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

B.2. Zero-Shot Generalization

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

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[15] and UK Biobank[16].

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

Table IV. Zero-shot performance on external datasets.

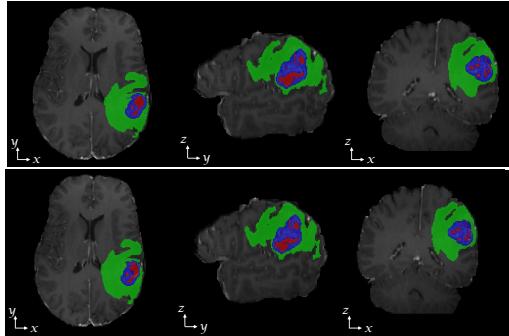


Figure II. **Qualitative Performance on BRATS.** We show the ground-truth *top* and output *bottom* of our pre-trained Swin-BOB model for 3D segmentation on the brain tumour BRATS dataset with 3 tumour class labels [2].

we could incorporate human-in-the-loop approaches that turned efficient as shown in [16, 5].

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

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.

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.

D. Dataset Access and Code for Reproducibility

The dataset and pre-trained Swin-BOB models will be made available publicly via UK Biobank. Documentation for reproducing experiments is provided in supplementary materials. Code is available at https://anonymous.4open.science/r/UK_BOB-ACF1/

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_{epoch} (optional)

Ensure: Adapted model M'

- 1: **Function** FREEZEEXCEPTBN(M):
- 2: **For each parameter** p **in** M :
- 3: **If** p does not belong to a BatchNorm layer:
- 4: $p.requires_grad \leftarrow \text{False}$
- 5: **End For**
- 6: **End Function**
- 7: **Function** UPDABNSTATISTICS(M, D_{test}):
- 8: Set M to training mode: $M.train()$
- 9: FREEZEEXCEPTBN(M)
- 10: **For each batch** $x \in D_{test}$:
- 11: Compute predictions: $M(x)$
- 12: **End For**
- 13: **End Function**
- 14: **Function** FINETUNEBN($M, D_{test}, \mathcal{L}, O, N_{epoch}$):
- 15: Set M to training mode: $M.train()$
- 16: FREEZEEXCEPTBN(M)
- 17: **For** $epoch \leftarrow 1$ **to** N_{epoch} :
- 18: **For each batch** $(x, y_{true}) \in D_{test}$:
- 19: Compute predictions: $y \leftarrow M(x)$
- 20: Compute loss: $\ell \leftarrow \mathcal{L}(y, y_{true})$
- 21: Zero gradients: $O.zero_grad()$
- 22: Backpropagate: $\ell.backward()$
- 23: Update parameters: $O.step()$
- 24: **End For**
- 25: **End For**
- 26: **End Function**
- 27: **Define Function** INFER(M, D_{test}):
- 28: Set M to evaluation mode: $M.eval()$
- 29: **For each batch** $x \in D_{test}$:
- 30: Compute predictions: $y \leftarrow M(x)$
- 31: **End For**
- 32: **End Function**

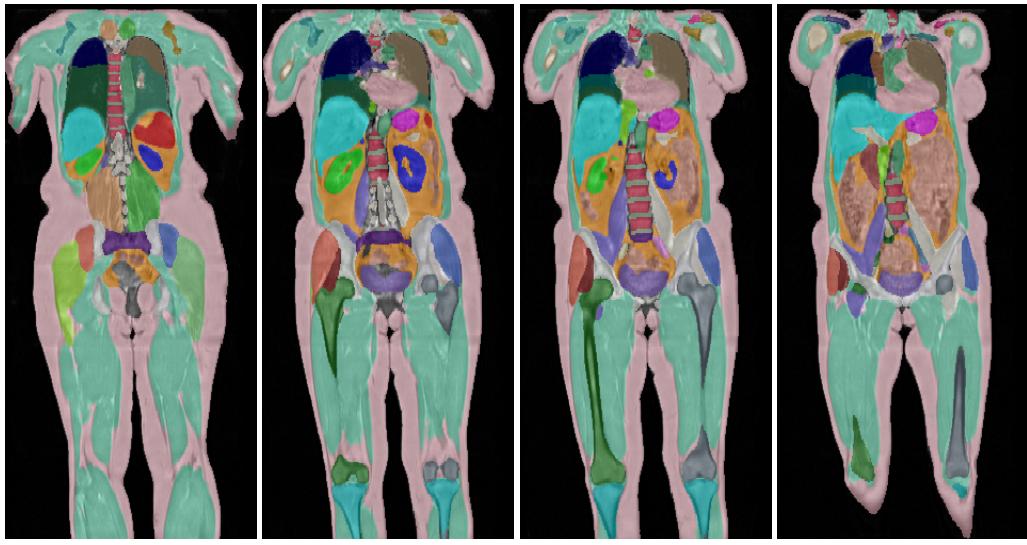


Figure III. Visualisation of UKBOB Segmentation Coronal Plane. We show an example of 3D MRI from UKBOB for on coronal plane.

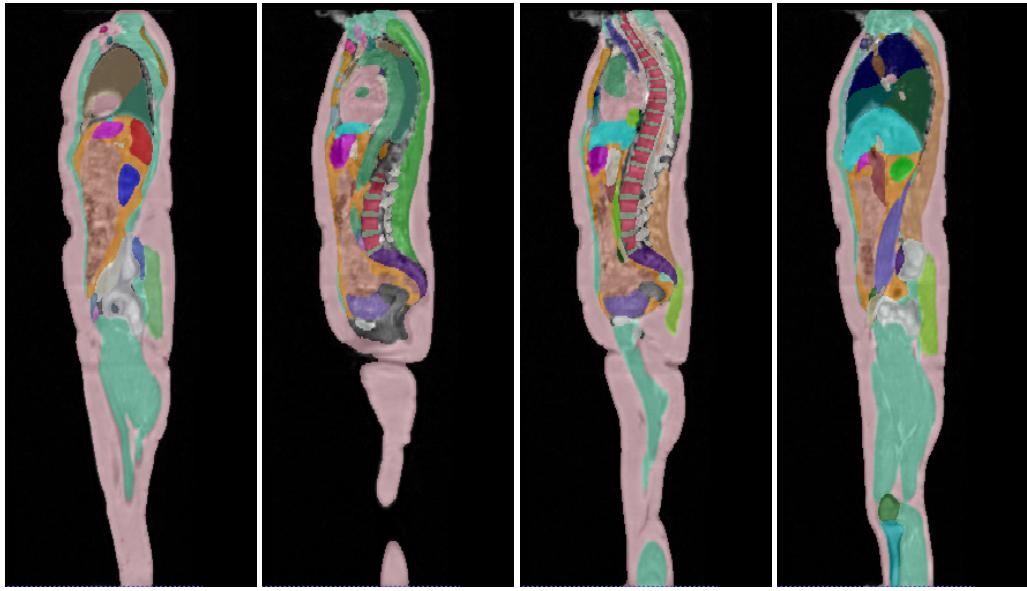


Figure IV. Visualisation of UKBOB Segmentation Sagittal Plane. We show an example of 3D MRI from UKBOB for on sagittal plane.

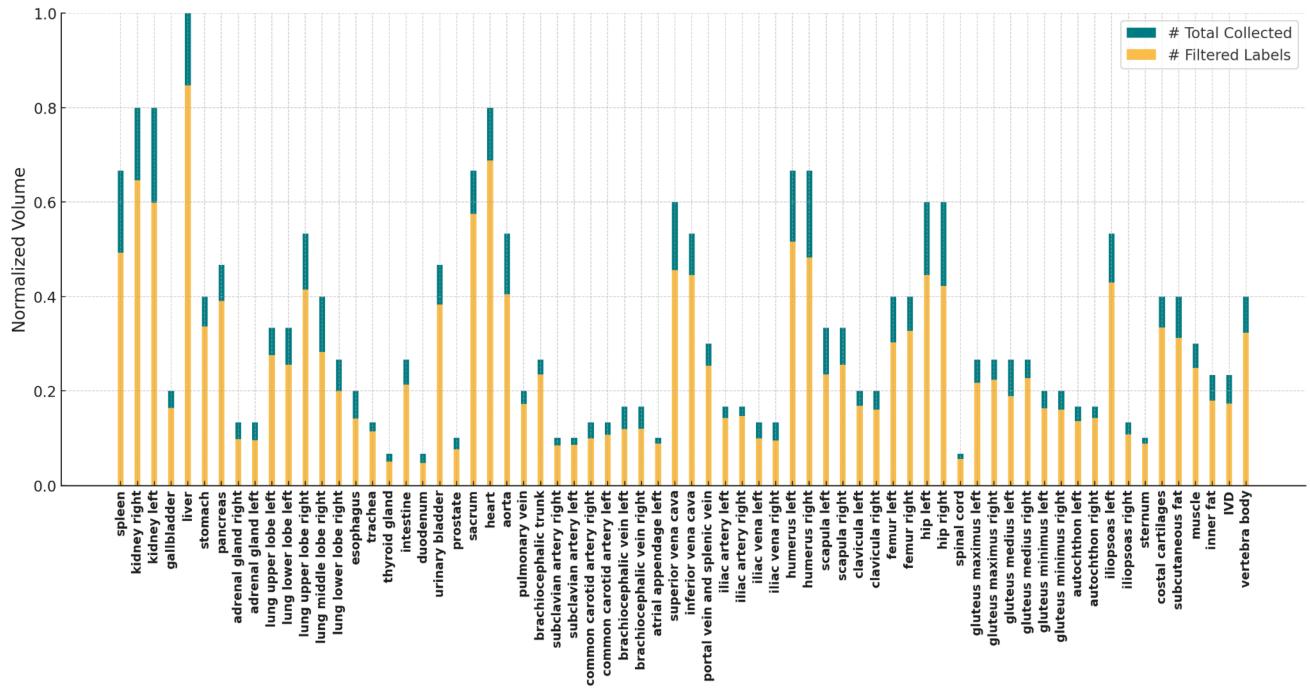


Figure V. UKBOB Distribution of Labels with our Filtration. We show the distribution mean normalised volumes of 72 labels before and after filtration.

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
clavicular 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 VI. 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.