arXiv:2605.20405v1 [eess.IV] 19 May 2026

DISENTANGLING SAMPLING FROM TRAINING BUDGET IN CLASS-IMBALANCED CT BODY COMPOSITION SEGMENTATION

Iason Skylitsis
Department of Biomedical Engineering & Physics
Amsterdam University Medical Center
Amsterdam, The Netherlands
Informatics Institute, Faculty of Science
University of Amsterdam
Amsterdam, The Netherlands
i.skylitsis@amsterdamumc.nl

Dimitrios Karkalousos
Department of Biomedical Engineering & Physics
Amsterdam University Medical Center
Amsterdam, The Netherlands
Informatics Institute, Faculty of Science
University of Amsterdam
Amsterdam, The Netherlands
d.karkalousos@amsterdamumc.nl

Ivana Išgum
Department of Radiology
Mayo Clinic
Rochester, United States
Department of Biomedical Engineering & Physics
Amsterdam University Medical Center
Amsterdam, The Netherlands
Department of Radiology
Amsterdam University Medical Center
Amsterdam, The Netherlands
isgum.ivana@mayo.edu

ABSTRACT

Class imbalance is a fundamental challenge in medical image segmentation, where frequent classes typically dominate training at the expense of rare classes. Loss-based approaches mitigate imbalance by reweighting the per-pixel loss within the batch, while sampling strategies control which images enter the batch. Yet neither explicitly controls which classes appear within the batch, leaving rare-class exposure only partially rebalanced. In this work, we adopt episodic sampling from few-shot learning to promote class-balanced batch construction in a fully supervised setting. We decouple episodic sampling from its conventional metric-learning context and evaluate it in body composition segmentation in CT. We compare episodic sampling against random and weighted sampling on nine muscle and adipose tissues, derived from 210 scans of the public SAROS dataset. Training is performed under full- and low-data regimes, with additional comparisons under matched training iteration budgets. Under full-data training, all three strategies performed comparably (mean Dice 0.882 for episodic, 0.878 for random and weighted). Under low-data training, episodic sampling outperformed random and weighted (0.787 vs. 0.758 and 0.762), driven by a 12-fold difference in training iterations. Under matched training budgets, random and weighted overfit earlier, while episodic improved for approximately three times more iterations before plateauing. Our findings identify the training iteration budget as under-recognized confound in sampling strategies, motivating iteration-aware evaluation protocols for small datasets. Furthermore, the residual advantage of episodic sampling is consistent with an implicit regularization effect of class-balanced batches, offering a low-cost, model-agnostic strategy for class-imbalanced medical image segmentation. Code is available at `https://github.com/iasonsky/episodic-sampling`.

Keywords: Class Imbalance · Sampling Strategies · Training Budget · Medical Image Segmentation · Body Composition · Computed Tomography

1 Introduction

Standard supervised learning typically samples training instances uniformly, implicitly assuming a balanced data distribution. In dense prediction tasks, such as semantic segmentation of medical images, this assumption rarely holds. Classes like background and large anatomical structures comprise orders of magnitude more pixels than small tissues or lesions. In addition, since segmentation models are trained by computing a loss over every pixel in each image, the gradient updates that drive learning are dominated by frequent classes, which contribute the majority of the per-pixel loss terms. As a result, rare classes receive proportionally fewer gradient updates, leading to models biased toward frequent classes, overfitting, and reduced segmentation accuracy under class imbalance.

Class imbalance in medical image segmentation is typically mitigated at the loss level. For example, weighted cross-entropy assigns higher penalties to underrepresented classes. Dice loss [Sudre et al., 2017] is inherently robust to class frequency disparities by optimizing for region overlap. Focal loss [Lin et al., 2018] down-weights easy examples to focus training on hard ones. In addition, compound losses, such as cross-entropy combined with Dice, have been shown to handle class imbalance more robustly than single losses [Ma et al., 2021], establishing them as a standard practice [Isensee et al., 2020]. Despite their effectiveness on the gradient signal, loss-based approaches do not alter the training distribution itself.

Complementary to gradient-level mitigation, class imbalance can also be addressed at the input level by shaping batch composition through the sampling process. For example, standard weighted sampling assigns higher selection probabilities to images containing rare classes. More sophisticated approaches include oversampling and undersampling [He and Garcia, 2009], class-aware and repeat-factor sampling [Gupta et al., 2019, Yaman et al., 2023], patch- and volume-level sampling weighted by class presence [Kamnitsas et al., 2017], and per-image imbalance-ratio weighting [Roshan et al., 2024a]. Despite this diversity, such methods control which images enter the batch without controlling class composition within it. Rare-class voxels therefore remain embedded in dominant-class context, so the per-voxel gradient signal remains only partially rebalanced.

Sampling has also been used for variance reduction, refining optimization by reducing gradient noise during batch construction. Zhao and Zhang [2014] showed that stratified mini-batch sampling tightens convergence bounds relative to uniform sampling, requiring fewer iterations to reach a given error level. Subsequent work formalized this concept into importance sampling, a complementary variance-reduction tool for deep learning. Katharopoulos and Fleuret [2018] and You et al. [2023] applied importance sampling to prioritize the most representative pixels within semantically similar groups. However, later work showed minimal effect of importance sampling on the asymptotic decision boundary of overparameterized networks [Byrd and Lipton, 2019], often underperforming fine-tuned baselines [Shwartz-Ziv et al., 2023] or even compromising representation quality [Kang et al., 2020, Zhou et al., 2020].

Across input-level rebalancing and stratified or importance sampling, existing methods adjust how often individual images are drawn into a batch, whether to compensate for class imbalance or to reduce gradient variance. The class composition within each batch, however, is not explicitly controlled. A notable exception comes from few-shot prototypical learning [Snell et al., 2017], where training mini-batches (episodes) are sampled from a controlled subset of classes, with each episode containing a support and query set. Episodic sampling has shown promising results on imbalanced medical image segmentation [Ouyang et al., 2020, Guo et al., 2025, Roshan et al., 2024b, Tian et al., 2024], yet its mechanism is typically entangled with metric-based learning objectives. The episodic batch-construction logic, however, is independent of metric learning and model-agnostic, suggesting plug-and-play applications in fully supervised learning.

Nevertheless, adapting episodic sampling in supervised training, raises a methodological challenge that has received limited attention in medical image segmentation. Sampling-strategy comparisons typically specify training schedules in epochs, including learning rate milestones, early stopping patience, and maximum training duration, implicitly coupling the effective training iterations budget to dataset size. When samplers with different numbers of iterations per epoch are compared under such schedules, this coupling introduces a confound. Previous work in classification has shown that the apparent gains of specialized imbalanced sampling schemes can shrink substantially when iteration budgets are matched [Li et al., 2020, Arazo et al., 2021], or when compared against fine-tuned baselines [Shwartz-Ziv et al., 2023]. The state-of-the-art nnU-Net framework [Isensee et al., 2020] sidesteps this by setting a budget of $250{,}000$ training iterations regardless of dataset size, yet typically sampling schemes specify schedules in epochs.

In this work, we decouple the episodic batch construction from metric-based learning and apply episodic sampling in standard supervised training. We compare episodic sampling against standard random and weighted sampling under two training-data regimes: a full-data setting using all annotated volumes, and a low-data setting retaining 10% via patient-level subsampling, which sharpens class underexposure. To isolate the contribution of the sampling mechanism from the training budget, we further evaluate the strategies under matched iteration budgets and examine their interaction with epoch-based scheduling. We focus on multi-class body composition segmentation in Computed Tomography

(CT), an inherently class-imbalanced task where large adipose and muscle compartments coexist with small, spatially localized structures, yielding class frequencies that differ by several orders of magnitude within each scan. Beyond the methodological fit, our work targets fine-grained segmentation of multiple muscle structures, a setting under-explored by existing body composition analysis pipelines, which typically operate on a single 2D slice or collapse the problem to coarse tissue labels [Blankemeier et al., 2023, Hofmann et al., 2025].

2 Methods

We investigate three sampling strategies for class-imbalanced body composition segmentation: random, weighted, and episodic. To isolate the effect of the sampling strategy, the network architecture, loss function, and optimization settings are held constant across all experiments. The following sections detail the dataset and the construction of reference annotations (Sec. 2.1), the sampling strategies (Sec. 2.2), the network architecture, training protocol, and evaluation metrics (Sec. 2.3), and the experimental setup (Sec. 2.4).

2.1 Data

We used 210 CT scans from the publicly available Sparsely Annotated Region and Organ Segmentation (SAROS) dataset [Koitka et al., 2024]. SAROS comprises 900 CT scans curated from 28 collections within The Cancer Imaging Archive (TCIA), of which only 210 were freely available without additional licensing requirements. Our experiments were therefore restricted to this subset, the characteristics of which are summarized in Table 1. Ethics approval for the use of these data was granted by the Medical Ethical Committee (METC) of Amsterdam UMC.

SAROS provides annotations for thirteen semantic body regions and six body-part labels, including subcutaneous adipose tissue (SAT) and skeletal muscle (SM). Reference annotations were created sparsely, by annotating every fifth axial slice. However, as noted in the dataset description, in the reference annotations the skin was merged into SAT and SM was segmented as a single contiguous structure rather than separated into individual muscles, with fascias and intermuscular adipose tissue (IMAT) incorporated into the muscle label. Additionally, after manual inspection we identified residual SM overestimation, including expansion into physiologically implausible regions (e.g., underneath the neural spine and along the vertebral surfaces), and remaining inclusion of fascia, scar tissue, or IMAT.

Table 1: Overview of the 210 publicly available scans from the SAROS dataset [Koitka et al., 2024].

Characteristic	**Value**
Demographics	
Sex	
Female / Male / Not reported	124 (59.0%) / 64 (30.5%) / 22 (10.5%)
Age	
Mean $\pm$ SD (yrs) / Range / Not reported	58.4 ± 17 / 2 mo – 90 yr / 42
Imaging type	
Contrast-Enhanced CT Single-Phase / Multi-Phase	95 (45.2%) / 60 (28.6%)
PET-CT	50 (23.8%)
Non-contrast CT	5 (2.4%)
Scanner vendor	
Siemens	114 (54.3%)
GE Medical Systems	49 (23.3%)
Not reported	47 (22.4%)
Longitudinal temporal event	
Surgery	86 (41.0%)
Diagnosis	41 (19.5%)
Baseline	30 (14.3%)
Not reported	53 (25.2%)

To address these issues, we refined and expanded the existing labels with additional muscle and adipose tissue segmentations obtained from the Body-and-Organ Analysis (BOA) tool [Haubold et al., 2024]. Nine tissue classes were defined: erector spinae muscle (ESM), intermuscular adipose tissue (IMAT), pectoral muscle (PEM), psoas major (PSM), quadratus lumborum (QLM), rectus abdominis (RAM), subcutaneous adipose tissue (SAT), skeletal muscle (SM), and visceral adipose tissue (VAT). SM was defined as the residual region after exclusion of the five

muscle subgroups. Hounsfield Unit (HU) thresholds were applied to constrain the masks to physiologically plausible attenuation ranges: $[-29, +150]$ HU for muscle tissues, $[-190, -30]$ HU for subcutaneous and intermuscular adipose tissue, and $[-150, -50]$ HU for visceral adipose tissue. VAT was further refined to prevent overlap with organs or bones by subtracting an organ-and-bone mask obtained with BOA, and isolated clusters of fewer than five voxels were removed via 3D connected-component analysis. IMAT was defined as the thresholded region within all muscle masks, with overlap with SAT and VAT explicitly subtracted. The refined labels were restricted to the anatomical trunk and extremities using the provided body-part annotations.

All scans and segmentation maps were standardized to the Right-Anterior-Superior (RAS) anatomical coordinate system. Scans were then cropped along the longitudinal axis to the levels relevant for body composition analysis. Specifically, between the highest detected thoracic vertebra (up to T1) and the lowest detected lumbar vertebra (up to L4), with per-scan boundaries identified from a pre-existing whole-body segmentation map. This yielded a total of 10,920 slices across the 210 scans. The resulting slice-wise prevalence of each tissue class is shown in Fig. 1. Fig. 2 shows three representative examples comparing the original scans, reference annotations, and our refined reference annotations across varying vertebral levels.

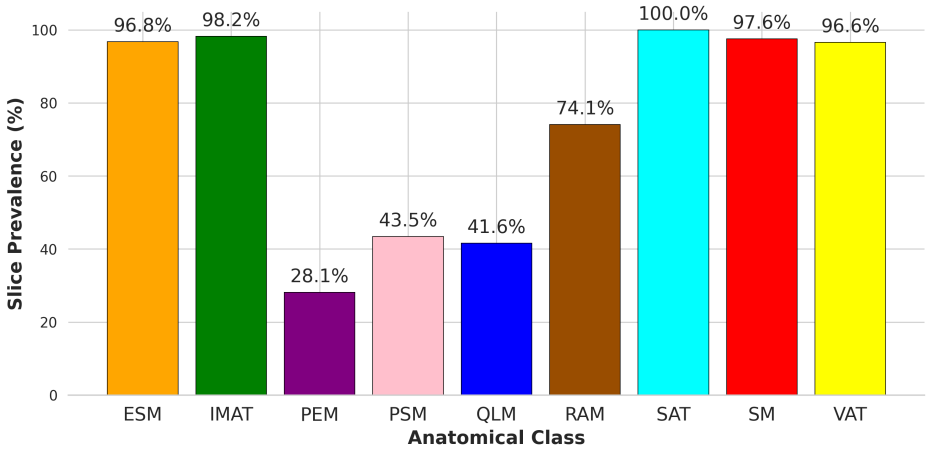


Figure 1: Slice-wise prevalence of the nine tissue classes. The y-axis indicates the percentage of slices in which each class is present. ESM: erector spinae muscle (orange); IMAT: inter-/intramuscular adipose tissue (green); PEM: pectoral muscle (purple); PSM: psoas muscle (pink); QLM: quadratus lumborum muscle (blue); RAM: rectus abdominis muscle (brown); SAT: subcutaneous adipose tissue (cyan); SM: skeletal muscle (red); VAT: visceral adipose tissue (yellow).

2.2 Sampling Strategies

We evaluated three sampling strategies spanning from class-agnostic to class-structured sampling. Specifically, random sampling drew slices uniformly from the training pool and served as the unweighted control. Weighted sampling operated at the slice level, biasing draws toward slices containing rare classes, but without constraining the within-batch composition. Episodic sampling promoted structured class composition within each mini-batch by sampling random subsets of foreground classes across consecutive iterations. The three strategies differed solely in how slices are drawn, with training performed fully supervised and the network architecture, loss function, and optimization settings held constant across all conditions.

Random Sampling. Slices are drawn uniformly from the training set, such that each mini-batch of size B is an i.i.d. sample from the full training pool.

Weighted Sampling. Each slice i is assigned a sampling probability, proportional to the inverse frequency of the rarest present foreground class:

$$p_i \propto \frac{1}{f_{c_i^*}}, \quad c_i^* = \arg\min_{c \in \mathcal{C}_i} f_c, \tag{1}$$

where $\mathcal{C}_i$ is the set of classes present in slice i and f_c the frequency of class c across the training set.

Episodic Sampling. Each mini-batch is constructed as an episode. In each episode, N_C foreground classes are sampled, and for each class, N_S support slices and N_Q query slices are drawn. Support slices are sampled from the

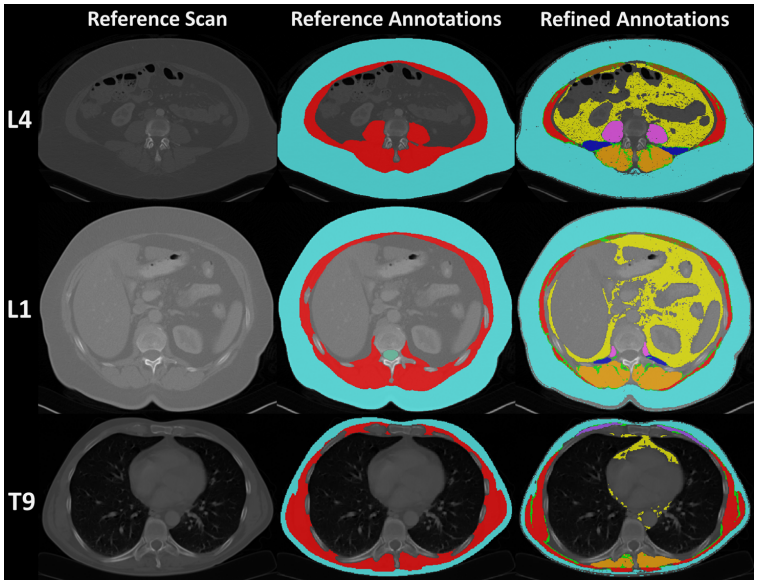


Figure 2: Example axial slices at three vertebral levels (L4–first, L1–second, and T9–third row), showing the reference SAROS CT scan, the reference annotations, and our refined nine-class reference annotations (left, center, and right column).

pool of slices containing the target class, while query slices are drawn from the same class-restricted pool. Since classes are sampled uniformly rather than in proportion to their frequency, rare and frequent classes appear as episode targets with equal probability, yielding approximately balanced class exposure over the course of training. The model can then be trained on either the support or the query slices, using the full multi-class labels in both cases.

2.3 Network Architecture & Training Protocol

Inputs were preprocessed by windowing HU (width 400, level 40) followed by linear normalization to $[-1, 1]$. For all experiments, we used a baseline 2D U-Net adopted from the nnU-Net implementation [Isensee et al., 2020]. The encoder consisted of six levels with convolutional pooling, beginning at a base feature width of 32 channels and doubling at each subsequent level up to a maximum of 480. Each level contained two convolutional blocks comprising 3×3 convolutions, instance normalization, and leaky ReLU activation (negative slope 10^{-2}). The decoder mirrored the encoder, using convolutional upsampling with skip connections and dropout ($p = 0.1$). The output channels corresponded to the nine tissue classes and background.

We used the AdamW optimizer [Loshchilov and Hutter, 2019] with an initial learning rate of 10^{-4} and weight decay of 10^{-2}. The learning rate was reduced by a factor of 0.1 at epochs 30 and 45 using a MultiStepLR scheduler. For random and weighted sampling, the batch size was set to 16. For episodic sampling, training comprised 500 episodes per epoch, with each episode sampling $N_C = 2$ foreground classes and drawing $N_S = 3$ support and $N_Q = 3$ query slices per class. Models were trained for a maximum of 200 epochs with early stopping triggered by mean foreground validation Dice (patience of 20 epochs). The loss function combined cross-entropy and Dice loss with equal weighting.

Segmentation performance was evaluated using two complementary metrics: the Dice similarity coefficient for quantifying area overlap [Dice, 1945], and the 95th-percentile Hausdorff Distance (HD95) for quantifying boundary accuracy [Taha and Hanbury, 2015]. Metrics were computed per class across all foreground classes.

All experiments were implemented in PyTorch and performed on NVIDIA V100 GPUs with 32 GB VRAM. The code is available at `https://github.com/iasonsky/episodic-sampling`.

2.4 Experiments

Data were split into 85% for development and 15% for testing. The development set comprised 144 scans for training and 36 for validation, with five-fold cross-validation applied at the patient level. The test set comprised 30 held-out scans. To assess whether episodic sampling yields greater benefit under data scarcity and more severe class imbalance, experiments were conducted under two data regimes: (i) a full-data regime (100%), using all 144 training and 36

validation scans, and (ii) a low-data regime, retaining 10% of training and validation scans via random subsampling at the patient level.

As detailed below, in the full-data regime, all sampling strategies required a comparable number of iterations per epoch. As a result, epoch-based scheduling decisions, including learning rate milestones and early stopping patience, corresponded to similar iteration budgets across strategies. In the low-data regime, a $12\times$ disparity arose between random/weighted and episodic sampling. Learning rate milestones at epoch 30 corresponded to 1,290 iterations under random and weighted sampling versus 15,000 under episodic, and early stopping patience of 20 epochs corresponded to 860 versus 10,000 iterations, respectively.

- Random/weighted, full-data regime: $\lceil 8{,}369/16 \rceil = 523$ iterations per epoch.
- Random/weighted, low-data regime: $\lceil 684/16 \rceil = 43$ iterations per epoch.
- Episodic, both regimes: 500 iterations per epoch (fixed by the number of episodes).

Therefore, for a fair comparison in the low-data regime, we disentangled the sampling mechanism from the training budget and systematically evaluated performance under equivalent iterations.

2.4.1 Fixed iterations with constant learning rate.

We evaluated the per-iteration effectiveness of each sampling strategy by equalizing the training iterations and removing all epoch-based scheduling decisions. To that end, we trained all three samplers for exactly 3,000 iterations with a constant learning rate and without early stopping.

2.4.2 Iteration-calibrated schedule.

We tested whether random and weighted sampling can match episodic performance under the same effective training budget. To that end, we rescaled the random and weighted schedules from epoch-based to iteration-equivalent specifications, using episodic's 500 iterations per epoch as a reference. In episodic sampling, milestones at epochs 30 and 45 correspond to 15,000 and 22,500 iterations, the patience of 20 epochs to 10,000 iterations, and the maximum of 200 epochs to 100,000 iterations. For random and weighted sampling this corresponded to milestones at epochs 349 and 523, patience of 233 epochs, and a maximum of 2,500 epochs.

3 Results

Table 2 compares the performance of episodic, random, and weighted sampling under both the full-data (100%) and low-data (10%) regimes. In the full-data regime, the choice of sampling strategy had minimal impact. Episodic achieved a mean Dice of 0.882, compared to 0.878 for both random and weighted, with a corresponding advantage in HD95 (6.77 mm vs. 7.98 mm and 7.80 mm). This modest effect was consistent with the near-matched iteration budgets across strategies in this regime (523 vs. 500 iterations per epoch). In the low-data regime, the advantage of episodic became pronounced, with a mean Dice of 0.787, compared to 0.758 for random and 0.762 for weighted sampling. Performance improved on eight of the nine foreground classes, with the largest gains observed on the least prevalent classes (IMAT, QLM, PEM, PSM; Fig. 1). In addition, episodic achieved the lowest HD95 on seven of nine classes, while random sampling achieved the best average HD95 (15.89 mm vs. 16.05 mm for episodic). However, as detailed in Sec. 2.4 and shown in Fig. 3, episodic sampling ran $12\times$ more training iterations per epoch in the low-data regime. Across both regimes, random and weighted sampling performed comparably, with neither consistently outperforming the other. Fig. 4 shows representative segmentations for each sampling strategy under both training regimes.

For episodic sampling, we ran an ablation study comparing using the supports and queries as training inputs. As shown in the Appendix Table A, query- or support-based supervision yielded near-identical performance in both data regimes, with a slight advantage on queries (mean Dice of 0.882 [queries] vs. 0.881 [supports] at 100%, both 0.787 at 10%). Therefore, for the rest of our experiments, we selected query-based supervision.

Table 3 reports the results of the fixed 3,000 iterations with constant learning rate experiment (Sec. 2.4.1). All three samplers converged to similar performance (0.778 vs. 0.773 vs. 0.773 mean Dice for episodic, random, and weighted, respectively). Episodic sampling achieved the highest Dice on five of nine classes (ESM, IMAT, PSM, RAM, SM), while weighted led on PEM and QLM, and random on SAT and VAT. For HD95, weighted achieved the best average (15.09 mm), followed by random (15.55 mm) and episodic (15.70 mm), with no single sampler consistently dominating across classes. Fig. 5a shows representative segmentations for each sampling strategy under low-data fixed-iteration training.

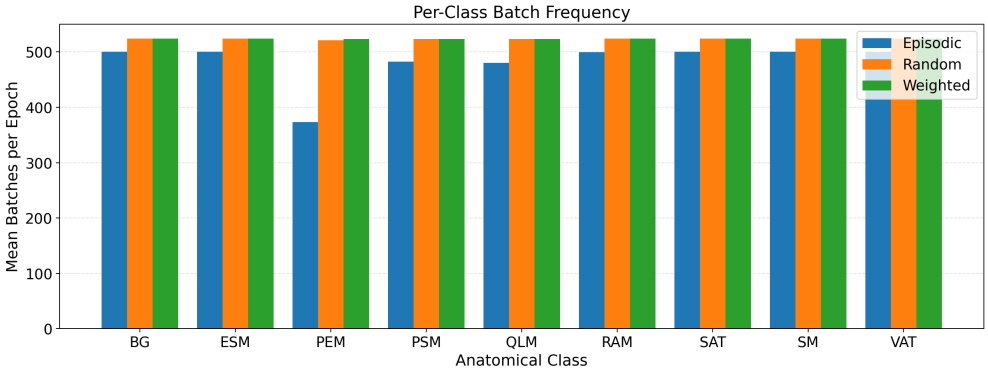
Table 4 reports the results of the iteration-calibrated schedule experiment (Sec. 2.4.2). Iteration calibration substantially closed the gap between samplers. Random sampling improved from 0.758 mean Dice (Table 2) to 0.777, and weighted

Table 2: Performance of episodic, random, and weighted sampling on the held-out test set under full-data (100%) and low-data (10%) training. Best scores are highlighted in bold.

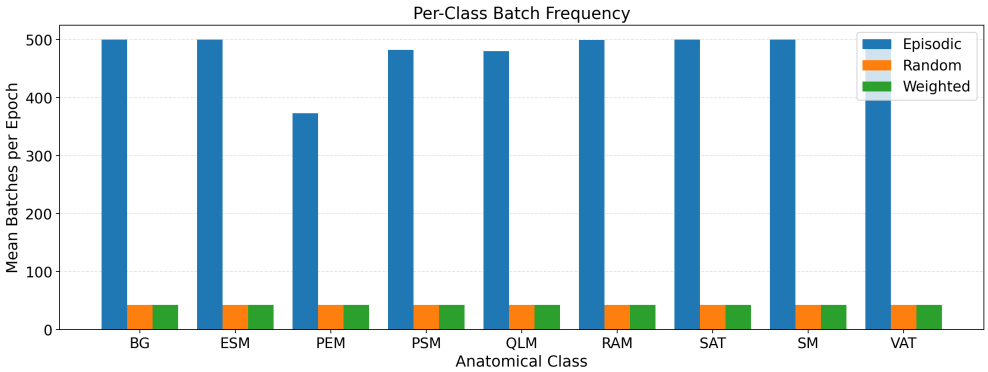
	Dice Scores ↑			HD95 ↓		
Class	Episodic	Random	Weighted	Episodic	Random	Weighted
Full-Data (100%) Training						
AVERAGE	**0.882** ± 0.13	0.878 ± 0.14	0.878 ± 0.14	**6.77** ± 14.07	7.98 ± 22.75	7.80 ± 18.77
ESM	**0.967** ± 0.02	**0.967** ± 0.02	**0.967** ± 0.02	**0.49** ± 0.90	0.52 ± 0.93	0.53 ± 0.92
IMAT	**0.783** ± 0.06	0.782 ± 0.06	0.781 ± 0.06	**8.74** ± 5.58	8.81 ± 5.60	8.79 ± 5.47
PEM	**0.646** ± 0.23	0.621 ± 0.25	0.620 ± 0.25	**32.56** ± 30.90	43.87 ± 57.56	41.16 ± 43.69
PSM	**0.903** ± 0.05	0.901 ± 0.05	**0.903** ± 0.05	**4.67** ± 4.44	5.08 ± 4.78	5.18 ± 4.90
QLM	**0.860** ± 0.06	0.855 ± 0.06	0.853 ± 0.06	6.90 ± 5.80	**6.86** ± 6.27	7.31 ± 7.09
RAM	**0.904** ± 0.03	0.902 ± 0.04	0.901 ± 0.04	4.87 ± 3.79	**4.49** ± 3.28	5.01 ± 3.47
SAT	**0.983** ± 0.01	**0.983** ± 0.01	**0.983** ± 0.01	0.22 ± 0.91	**0.20** ± 0.91	**0.20** ± 0.92
SM	**0.936** ± 0.02	**0.936** ± 0.02	0.935 ± 0.02	**2.85** ± 3.84	2.98 ± 3.73	2.89 ± 3.62
VAT	0.948 ± 0.02	**0.950** ± 0.03	0.949 ± 0.03	**1.34** ± 1.58	1.37 ± 1.69	**1.34** ± 1.74
Low-Data (10%) Training						
AVERAGE	**0.787** ± 0.17	0.758 ± 0.18	0.762 ± 0.18	16.05 ± 27.56	**15.89** ± 20.52	17.02 ± 26.35
ESM	**0.936** ± 0.03	0.923 ± 0.03	0.923 ± 0.03	3.08 ± 4.91	**2.91** ± 2.57	3.04 ± 2.49
IMAT	**0.667** ± 0.11	0.618 ± 0.11	0.617 ± 0.10	**12.99** ± 9.47	14.55 ± 11.27	14.24 ± 10.43
PEM	**0.519** ± 0.22	0.483 ± 0.20	0.495 ± 0.21	62.10 ± 58.44	**50.88** ± 34.00	62.43 ± 52.99
PSM	**0.805** ± 0.10	0.756 ± 0.12	0.773 ± 0.11	**11.93** ± 8.95	14.07 ± 9.96	14.05 ± 9.15
QLM	**0.670** ± 0.13	0.632 ± 0.14	0.646 ± 0.13	**18.75** ± 13.57	18.96 ± 11.48	20.09 ± 11.97
RAM	**0.772** ± 0.07	0.756 ± 0.07	0.751 ± 0.07	**14.14** ± 7.30	17.48 ± 7.25	15.72 ± 6.86
SAT	**0.969** ± 0.02	0.956 ± 0.04	0.955 ± 0.04	**0.82** ± 1.47	0.84 ± 1.47	0.94 ± 1.78
SM	**0.849** ± 0.07	0.833 ± 0.08	0.834 ± 0.08	**11.97** ± 18.39	13.42 ± 18.74	13.23 ± 18.53
VAT	**0.890** ± 0.07	0.860 ± 0.08	0.855 ± 0.09	**10.17** ± 14.06	11.08 ± 16.23	10.94 ± 15.85

Table 3: Performance of episodic, random, and weighted sampling on the held-out test set under low-data (10%) fixed-iteration training (Sec. 2.4.1). Best scores are highlighted in bold.

	Dice Scores ↑			Hausdorff Scores (95th) ↓		
Class	Episodic	Random	Weighted	Episodic	Random	Weighted
Low-Data (10%) Training						
AVERAGE	**0.778** ± 0.17	0.773 ± 0.17	0.773 ± 0.17	15.70 ± 26.50	15.55 ± 20.80	**15.09** ± 19.12
ESM	**0.932** ± 0.03	0.926 ± 0.03	0.927 ± 0.03	3.12 ± 5.05	3.66 ± 4.60	**3.04** ± 2.76
IMAT	**0.657** ± 0.10	0.650 ± 0.10	0.637 ± 0.11	**13.03** ± 9.20	13.29 ± 8.70	13.28 ± 9.02
PEM	0.514 ± 0.22	0.517 ± 0.22	**0.523** ± 0.21	56.21 ± 57.91	52.40 ± 35.53	**48.95** ± 30.63
PSM	**0.783** ± 0.11	0.773 ± 0.11	0.782 ± 0.10	**12.80** ± 9.43	13.13 ± 8.34	13.35 ± 8.99
QLM	0.648 ± 0.14	0.650 ± 0.13	**0.657** ± 0.13	20.35 ± 14.28	20.10 ± 11.66	**19.29** ± 11.27
RAM	**0.777** ± 0.07	0.754 ± 0.07	0.757 ± 0.07	**14.20** ± 7.05	15.45 ± 7.90	15.85 ± 7.36
SAT	0.960 ± 0.03	**0.963** ± 0.03	0.960 ± 0.03	0.85 ± 1.54	**0.73** ± 1.34	0.84 ± 1.56
SM	**0.845** ± 0.07	0.838 ± 0.07	0.833 ± 0.08	12.91 ± 18.49	**12.71** ± 18.45	13.08 ± 18.55
VAT	0.875 ± 0.07	**0.878** ± 0.07	0.871 ± 0.08	**9.15** ± 14.30	9.70 ± 14.34	10.34 ± 14.92

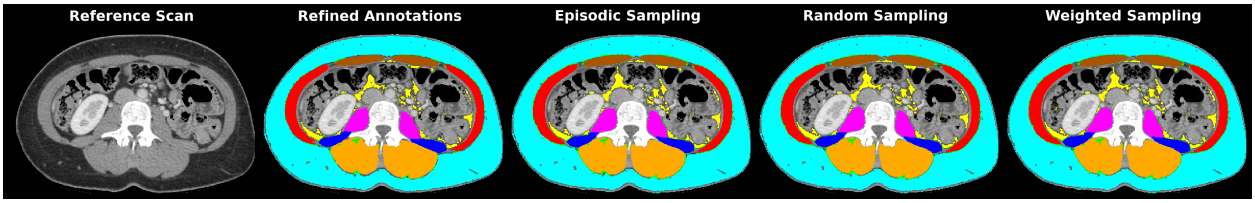


(a) Full-Data (100%) Training

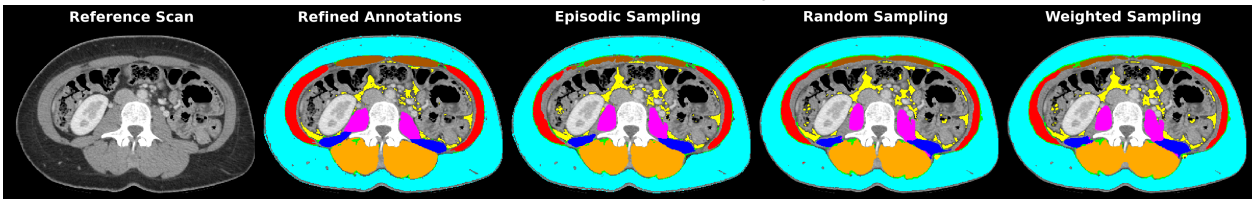


(b) Low-Data (10%) Training

Figure 3: Per-class batch frequency per epoch for episodic (blue), random (orange), and weighted (green) sampling for a representative fold (fold 0) under full-data (a) and low-data (b). Random and weighted sampling scale down proportionally with reduced dataset size. Episodic sampling maintains $\sim$500 batches per class per epoch in both regimes.



(a) Full-Data (100%) Training



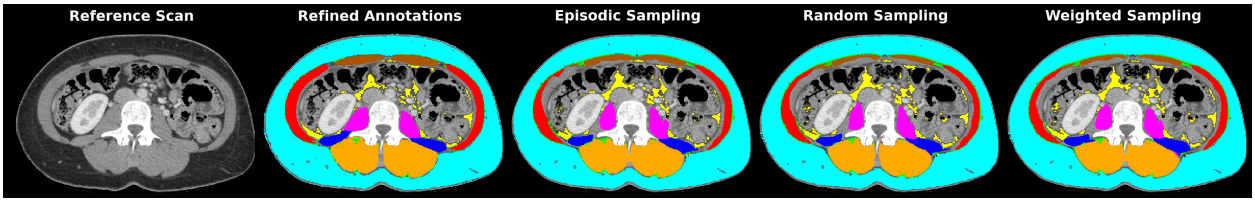
(b) Low-Data (10%) Training

Figure 4: Representative segmentations at the L3 vertebral level for a test case under standard epoch-based full-data (a) and low-data (b) training. Each column (left to right) shows the reference CT scan, refined annotations, and predictions from episodic, random, and weighted sampling, respectively.

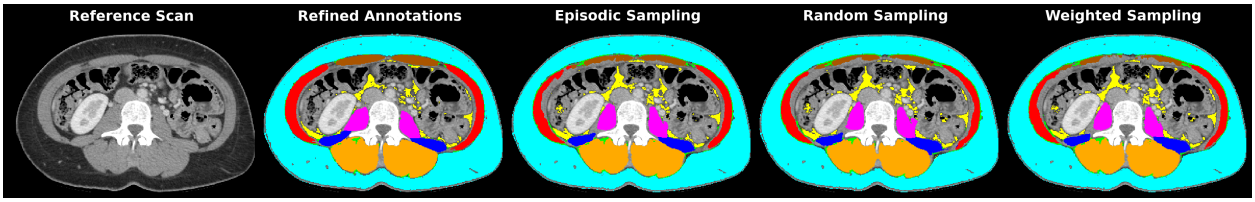
from 0.762 to 0.778. The largest margins between episodic and random samplings appeared on RAM (+3.0 pp), QLM (+2.0 pp), and PSM (+1.3 pp). The iteration-calibrated schedule also narrowed the HD95 gap, with episodic achieving the lowest HD95 on seven of nine classes, yet random still led on average (15.95 mm vs. 16.05 mm for episodic). This was attributed largely to PEM performance (62.10 mm for episodic vs. 55.77 mm for random). Fig. 5b shows representative segmentations for each sampling strategy under low-data iteration-calibrated training.

Table 4: Performance of episodic, random, and weighted sampling on the held-out test set under low-data (10%) iteration-calibrated training (Sec. 2.4.2). Best scores are highlighted in bold.

	Dice Scores ↑			Hausdorff Scores (95th) ↓		
Class	Episodic	Random	Weighted	Episodic	Random	Weighted
Low-Data (10%) Training						
AVERAGE	**0.787** ± 0.17	0.777 ± 0.17	0.778 ± 0.17	16.05 ± 27.56	**15.95** ± 24.38	16.23 ± 23.64
ESM	**0.936** ± 0.03	0.929 ± 0.03	0.930 ± 0.03	**3.08** ± 4.91	3.48 ± 4.16	4.37 ± 5.42
IMAT	**0.667** ± 0.11	0.659 ± 0.10	0.658 ± 0.11	**12.99** ± 9.47	13.18 ± 8.59	13.28 ± 9.19
PEM	0.519 ± 0.22	**0.521** ± 0.20	0.516 ± 0.21	62.10 ± 58.44	**55.77** ± 49.25	57.23 ± 44.97
PSM	**0.805** ± 0.10	0.792 ± 0.10	0.795 ± 0.10	**11.93** ± 8.95	12.24 ± 8.78	12.06 ± 7.93
QLM	**0.670** ± 0.13	0.650 ± 0.14	0.661 ± 0.13	**18.75** ± 13.57	19.84 ± 12.95	19.39 ± 12.58
RAM	**0.772** ± 0.07	0.742 ± 0.08	0.743 ± 0.08	**14.14** ± 7.30	16.38 ± 8.66	16.19 ± 8.00
SAT	**0.969** ± 0.02	0.968 ± 0.03	0.968 ± 0.03	0.82 ± 1.47	**0.81** ± 1.50	0.86 ± 1.59
SM	**0.849** ± 0.07	0.840 ± 0.08	0.839 ± 0.08	**11.97** ± 18.39	12.65 ± 18.52	12.70 ± 18.42
VAT	**0.890** ± 0.07	0.882 ± 0.08	0.882 ± 0.08	**10.17** ± 14.06	10.53 ± 15.24	11.34 ± 15.50



(a) Low-Data (10%) Fixed-Iteration Training



(b) Low-Data (10%) Iteration-Calibrated Training

Figure 5: Representative segmentations at the L3 vertebral level for a test case under low-data fixed-iteration (a) and iteration-calibrated (b) training. Each column (left to right) shows the reference CT scan, refined annotations, and predictions from episodic, random, and weighted sampling, respectively.

Training dynamics across all protocols were consistent with these results (Appendix A). In the full-data regime (Fig. 6), all three samplers converged comparably, consistent with their near-matched iteration budgets (Sec. 2.4). In the uncalibrated low-data regime (Fig. 7), random and weighted sampling terminated at ∼2,500 iterations via early stopping while episodic ran to ∼30,000. Under the fixed 3,000-iteration protocol (Fig. 8), all three samplers reached similar Dice but episodic's rare-class curves climbed later, indicating incomplete class coverage at this budget. Under the iteration-calibrated schedule (Fig. 9), random and weighted peaked at 5,100 and 7,800 iterations and overfit thereafter, while episodic peaked at ∼15,000 with validation loss remaining below the other two throughout the budget.

4 Discussion

In this work, we evaluated episodic sampling as a model-agnostic, plug-and-play batch construction strategy for class-imbalanced muscle and adipose tissue segmentation in CT. We decoupled episodic batch construction from metric-based

learning [Snell et al., 2017, Ouyang et al., 2020] and compared it against random and weighted sampling, under both full- and low-data training regimes (Sec. 2.4). Our work establishes episodic sampling as a feasible supervised-training mechanism, while exposing the intrinsic relationships between number of epochs, training iterations, learning rate milestones, and early stopping patience as an under-recognized confound.

We showed that sampling strategies are heavily influenced by this confound in epoch-based training. When iteration counts per epoch differ, epoch-tied scheduling decisions translate into substantially different effective training budgets. Once the confound is controlled, the apparent benefit of class-aware sampling diminishes and performance gains are mostly attributable to increased iteration budget. Under a matched fixed-iteration budget, the three samplers performed comparably (Table 3), and even under the iteration-calibrated schedule, episodic sampling's advantage narrowed from 2.9 to 1.0 percentage points (Table 4). The residual advantage coincided with delayed overfitting: random and weighted sampling reached their best checkpoints earlier and showed rising validation loss thereafter, while episodic sampling continued improving for approximately three times as many iterations. This pattern is consistent with an implicit regularization effect of class-balanced batch construction, though the mechanism remains unclear without direct measurements of gradient variance or feature-space geometry.

Notably, weighted sampling performed comparably to random sampling across all conditions, with neither consistently outperforming the other. Both strategies adjust slice-level sampling probabilities but leave the pixel-level class composition within each selected slice unchanged. This suggests that slice-level frequency rebalancing alone is insufficient when the dominant source of imbalance is within image rather than across-image.

Yet these observations raise concerns about epoch-based training and comparisons, which usually attribute gains to the sampling mechanism or the network architecture [Arazo et al., 2021, Shwartz-Ziv et al., 2023]. The state-of-the-art nnU-Net framework circumvents this confound by fixing the training iterations to a high number ($250{,}000$) regardless of dataset size [Isensee et al., 2020]. Here, we identify that the confound is not specific to the sampling mechanism, but applies to any epoch-based scheduler. Therefore, our work recommends scaling the training schedule to the dataset size by adjusting the iterations of the schedule. In practice, this means disentangling the sampling mechanism from the training budget, a step that can also enhance reproducibility [Li et al., 2020].

Episodic sampling has direct implications for body composition analysis, where accurate segmentation of small muscle subgroups, like psoas and quadratus lumborum, and intermuscular adipose tissue is highly challenging. Body composition pipelines typically operate on a single 2D slice or collapse the problem to coarse tissue labels [Blankemeier et al., 2023, Haubold et al., 2024], leaving rare classes unaddressed. Thus, episodic sampling offers a low-cost, model-agnostic improvement directly at the input level, without modifying the loss or requiring additional annotations.

Finally, three methodological choices are worth noting regarding their impact on our results. Reference annotations were refined using automated body segmentation models [Haubold et al., 2024], which may introduce label noise warranting multi-seed cross-validation [Bouthillier et al., 2021]. The calibrated schedule was anchored to the 500 iterations per epoch of episodic sampling, without exploring alternative reference budgets. Additionally, our evaluation was limited to a single task, baseline model, and loss configuration, though since the budget decomposition is mechanism-driven rather than task-specific, the findings are likely to transfer to other class-imbalanced settings. Future work could assess why episodic sampling remains advantageous at matched iteration budgets, through gradient-variance analysis [Zhao and Zhang, 2014, Katharopoulos and Fleuret, 2018], feature-space geometry [You et al., 2023], or calibration metrics. Furthermore, combining episodic sampling with random or weighted sampling and loss-based rebalancing strategies could yield complementary gains.

5 Conclusion

We evaluated episodic sampling as a plug-and-play strategy for class imbalance in medical image segmentation, comparing it against standard random and weighted sampling. We showed that the apparent advantage of episodic sampling stemmed primarily from epoch-based scheduling, rather than the sampling mechanism itself. Even at matched training budgets, episodic sampling retained a small advantage, with random and weighted sampling overfitting earlier under extended training. Our findings exposed a confound in sampling-strategy comparisons, motivating iteration-aware protocols on small datasets to distinguish true algorithmic improvements from incidental compute differences. Future work could explore curriculum frameworks that combine episodic sampling with random or weighted sampling, or with loss-based rebalancing strategies, and develop systematic heuristics for dataset-adaptive iteration budgets.

Acknowledgments

This work is supported by the Artillery Consortium and European Union's Horizon Europe under Grant No. 101080983.

References

Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *CoRR*, abs/1707.03237, 2017. URL http://arxiv.org/abs/1707.03237.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018. URL https://arxiv.org/abs/1708.02002.

Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L. Martel. Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71:102035, 2021. ISSN 1361-8415. doi: 10.1016/j.media.2021.102035. URL http://dx.doi.org/10.1016/j.media.2021.102035.

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, December 2020. ISSN 1548-7105. doi: 10.1038/s41592-020-01008-z. URL http://dx.doi.org/10.1038/s41592-020-01008-z.

Haibo He and E.A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. ISSN 1041-4347. doi: 10.1109/tkde.2008.239. URL http://dx.doi.org/10.1109/TKDE.2008.239.

Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. *CoRR*, abs/1908.03195, 2019. URL http://arxiv.org/abs/1908.03195.

Burhaneddin Yaman, Tanvir Mahmud, and Chun-Hao Liu. Instance-aware repeat factor sampling for long-tailed object detection, 2023. URL https://arxiv.org/abs/2305.08069.

Konstantinos Kamnitsas, Christian Ledig, Virginia F.J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, February 2017. ISSN 1361-8415. doi: 10.1016/j.media.2016.10.004. URL http://dx.doi.org/10.1016/j.media.2016.10.004.

SeyedEhsan Roshan, Jafar Tanha, Mahdi Zarrin, Alireza Fakhim Babaei, Haniyeh Nikkhah, and Zahra Jafari. A deep ensemble medical image segmentation with novel sampling method and loss function. *Computers in Biology and Medicine*, 172:108305, April 2024a. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2024.108305. URL http://dx.doi.org/10.1016/j.compbiomed.2024.108305.

Peilin Zhao and Tong Zhang. Accelerating minibatch stochastic gradient descent using stratified sampling. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 829–837, 2014. URL https://arxiv.org/abs/1405.3080.

Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. *CoRR*, abs/1803.00942, 2018. URL http://arxiv.org/abs/1803.00942.

Chenyu You, Weicheng Dai, Yifei Min, Fenglin Liu, David A. Clifton, S Kevin Zhou, Lawrence Hamilton Staib, and James S Duncan. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective, 2023. URL https://arxiv.org/abs/2302.01735.

Jonathon Byrd and Zachary C. Lipton. What is the effect of importance weighting in deep learning?, 2019. URL https://arxiv.org/abs/1812.03372.

Ravid Shwartz-Ziv, Micah Goldblum, Yucen Lily Li, C. Bayan Bruss, and Andrew Gordon Wilson. Simplifying neural network training under class imbalance, 2023. URL https://arxiv.org/abs/2312.02517.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1910.09217.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9719–9728, 2020. URL https://arxiv.org/abs/1912.02413.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning, 2017. URL https://arxiv.org/abs/1703.05175.

Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. *Self-supervision with Superpixels: Training Few-Shot Medical Image Segmentation Without Annotation*, page 762–780. Springer International Publishing, 2020. ISBN 9783030585266. doi: 10.1007/978-3-030-58526-6_45. URL http://dx.doi.org/10.1007/978-3-030-58526-6_45.

Erjian Guo, Zicheng Wang, Zhen Zhao, and Luping Zhou. Imbalanced Medical Image Segmentation With Pixel-Dependent Noisy Labels. *IEEE Transactions on Medical Imaging*, 44(5):2016–2027, May 2025. ISSN 1558-254X. doi: 10.1109/TMI.2024.3524253. URL https://ieeexplore.ieee.org/document/10818702/.

SeyedEhsan Roshan, Jafar Tanha, Mahdi Zarrin, Alireza Fakhim Babaei, Haniyeh Nikkhah, and Zahra Jafari. A deep ensemble medical image segmentation with novel sampling method and loss function. *Computers in Biology and Medicine*, 172:108305, April 2024b. ISSN 00104825. doi: 10.1016/j.compbiomed.2024.108305. URL https://linkinghub.elsevier.com/retrieve/pii/S0010482524003895.

Chunna Tian, Zhenxi Zhang, Xinbo Gao, Heng Zhou, Ran Ran, and Zhicheng Jiao. An Implicit-Explicit Prototypical Alignment Framework for Semi-Supervised Medical Image Segmentation. *IEEE Journal of Biomedical and Health Informatics*, 28(2):929–940, February 2024. ISSN 2168-2208. doi: 10.1109/JBHI.2023.3330667. URL https://ieeexplore.ieee.org/document/10310088/.

Mengtian Li, Ersin Yumer, and Deva Ramanan. Budgeted training: Rethinking deep neural network training under resource constraints, 2020. URL https://arxiv.org/abs/1905.04753.

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. How important is importance sampling for deep budgeted training?, 2021. URL https://arxiv.org/abs/2110.14283.

Louis Blankemeier, Arjun Desai, Juan Manuel Zambrano Chaves, Andrew Wentland, Sally Yao, Eduardo Reis, et al. Comp2Comp: open-source body composition assessment on computed tomography. arXiv:2302.06568, 2023.

Felix O. Hofmann, Christian Heiliger, Tengis Tschaidse, Stefanie Jarmusch, Liv A. Auhage, Ughur Aghamaliyev, et al. Validation of body composition parameters extracted via deep learning-based segmentation from routine computed tomographies. *Sci. Rep.*, 15(1):11909, April 2025. doi: 10.1038/s41598-025-96238-6.

Sven Koitka, Giulia Baldini, Lennard Kroll, Natalie Van Landeghem, Olivia B. Pollok, Johannes Haubold, Obioma Pelka, Moon Kim, Jens Kleesiek, Felix Nensa, and René Hosch. SAROS: A dataset for whole-body region and organ segmentation in CT imaging. *Scientific Data*, 11(1):483, May 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03337-6. URL https://www.nature.com/articles/s41597-024-03337-6.

Johannes Haubold, Giulia Baldini, Vicky Parmar, Benedikt Michael Schaarschmidt, Sven Koitka, Lennard Kroll, Natalie van Landeghem, Lale Umutlu, Michael Forsting, Felix Nensa, et al. Boa: a ct-based body and organ analysis for radiologists at the point of care. *Investigative Radiology*, 59(6):433–441, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1), August 2015. ISSN 1471-2342. doi: 10.1186/s12880-015-0068-x. URL http://dx.doi.org/10.1186/s12880-015-0068-x.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks, 2021. URL https://arxiv.org/abs/2103.03098.

A Appendix

Table 5: Performance of episodic query-based and support-based training on the held-out test set under full-data (100%) and low-data (10%). Best scores are highlighted in bold.

	Dice Scores ↑		Hausdorff Scores (95th) ↓	
Class	Queries	Supports	Queries	Supports
	Full-Data (100%) Training			
AVERAGE	**0.882** ± 0.13	0.881 ± 0.13	**6.77** ± 14.07	8.28 ± 20.79
ESM	**0.967** ± 0.02	**0.967** ± 0.02	**0.49** ± 0.90	**0.49** ± 0.91
IMAT	**0.783** ± 0.06	0.782 ± 0.06	**8.74** ± 5.58	8.98 ± 5.87
PEM	**0.646** ± 0.23	0.638 ± 0.24	**32.56** ± 30.90	44.76 ± 48.02
PSM	**0.903** ± 0.05	**0.903** ± 0.05	**4.67** ± 4.44	4.73 ± 4.57
QLM	**0.860** ± 0.06	0.859 ± 0.06	**6.90** ± 5.80	7.28 ± 6.51
RAM	**0.904** ± 0.03	**0.904** ± 0.04	**4.87** ± 3.79	4.94 ± 3.84
SAT	**0.983** ± 0.01	**0.983** ± 0.01	**0.22** ± 0.91	0.23 ± 0.91
SM	**0.936** ± 0.02	**0.936** ± 0.02	**2.85** ± 3.84	2.93 ± 4.00
VAT	0.948 ± 0.02	**0.949** ± 0.02	**1.34** ± 1.58	1.39 ± 1.59
	Low-Data (10%) Training			
AVERAGE	**0.787** ± 0.17	**0.787** ± 0.17	16.05 ± 27.56	**15.08** ± 23.92
ESM	**0.936** ± 0.03	0.933 ± 0.03	**3.08** ± 4.91	3.16 ± 3.68
IMAT	**0.667** ± 0.11	0.664 ± 0.11	12.99 ± 9.47	**12.87** ± 8.91
PEM	0.519 ± 0.22	**0.522** ± 0.21	62.10 ± 58.44	**54.74** ± 49.65
PSM	0.805 ± 0.10	**0.811** ± 0.09	11.93 ± 8.95	**11.44** ± 8.26
QLM	0.670 ± 0.13	**0.675** ± 0.13	**18.75** ± 13.57	18.93 ± 13.37
RAM	**0.772** ± 0.07	0.767 ± 0.07	**14.14** ± 7.30	14.76 ± 7.54
SAT	**0.969** ± 0.02	0.968 ± 0.03	**0.82** ± 1.47	0.84 ± 1.57
SM	0.849 ± 0.07	**0.850** ± 0.08	**11.97** ± 18.39	12.30 ± 18.58
VAT	**0.890** ± 0.07	0.888 ± 0.07	10.17 ± 14.06	**9.32** ± 13.40

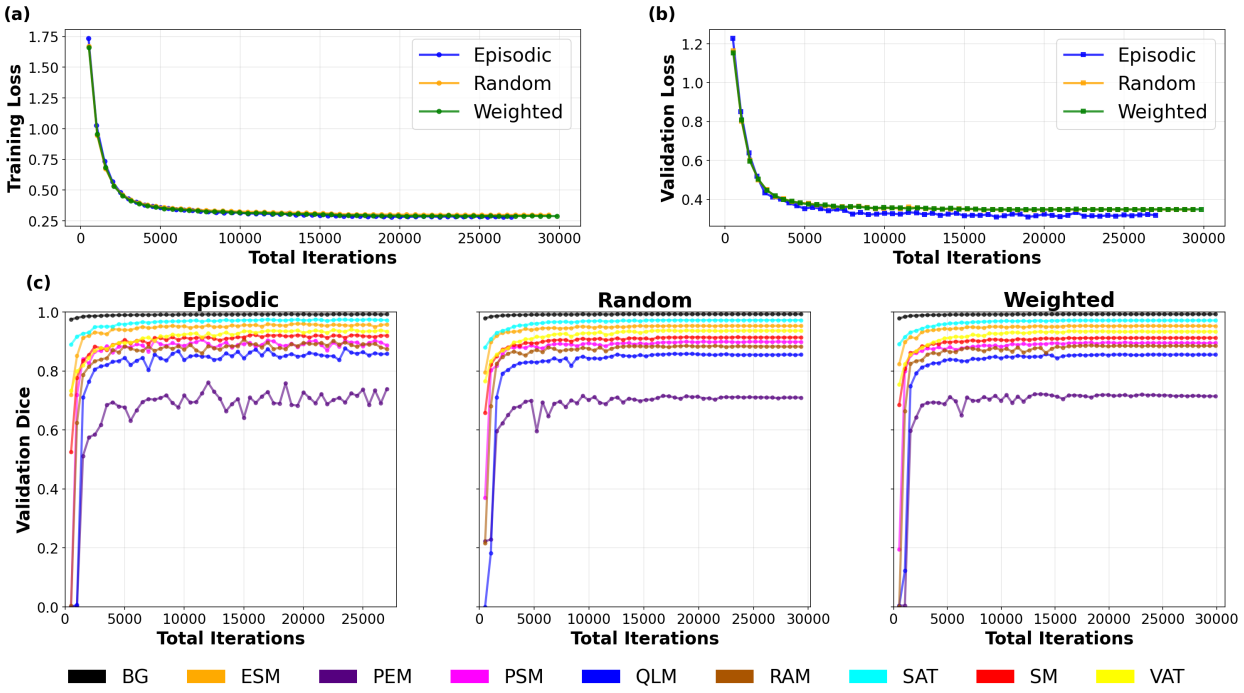


Figure 6: Full-data 100% regime. (a; top–left) training Loss, (b; top–right) validation Loss, (c; bottom) per-class validation Dice performance for episodic (left), random (middle), and weighted (right).

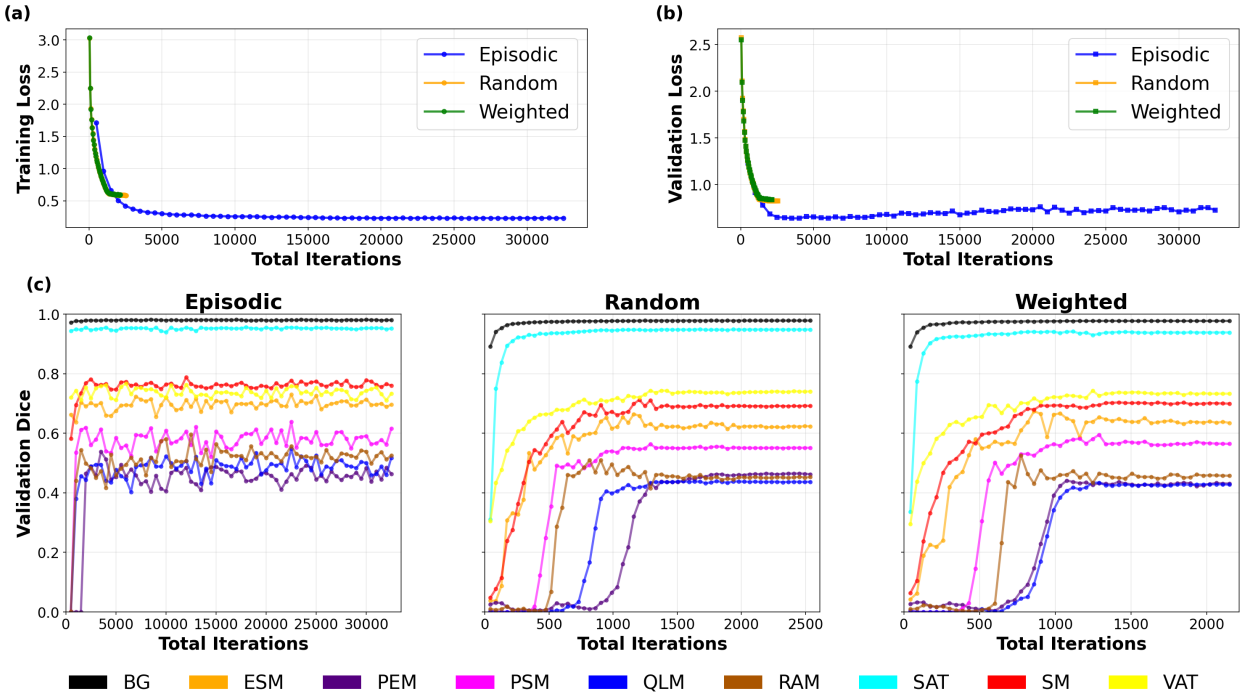


Figure 7: Low-data 10% regime. (a; top–left) training Loss, (b; top–right) validation Loss, (c; bottom) per-class validation Dice performance for episodic (left), random (middle), and weighted (right).

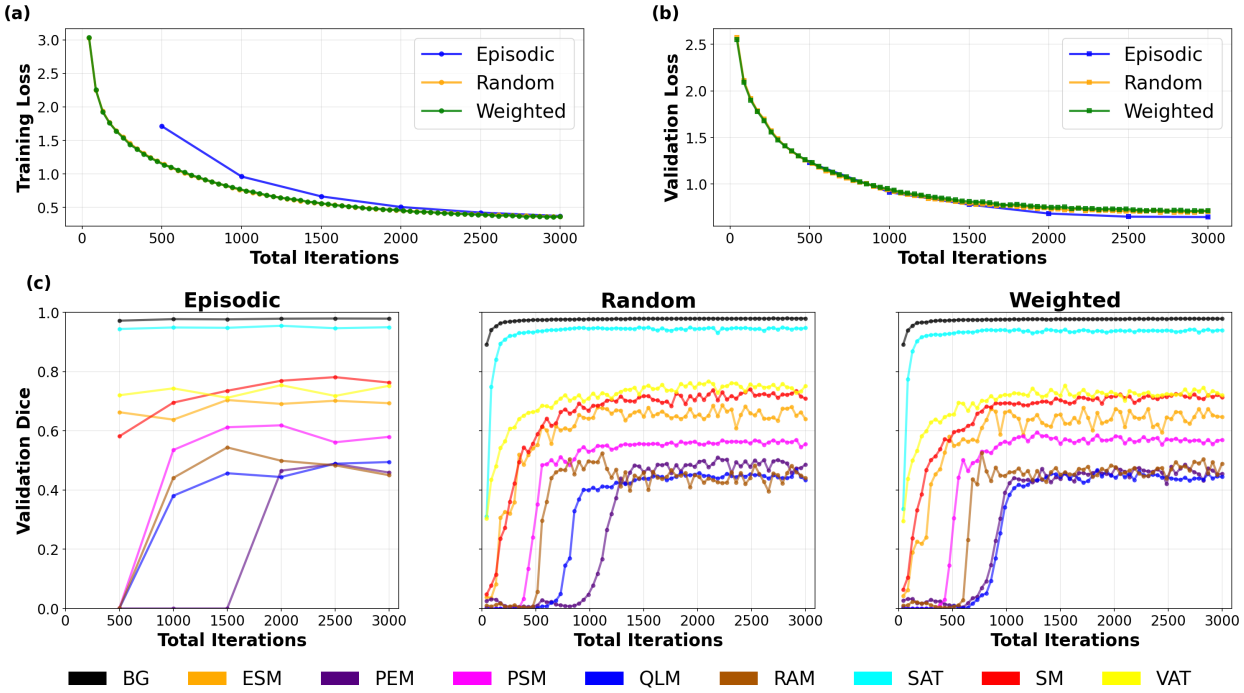


Figure 8: Low-data 10% fixed 3,000 iterations with constant learning rate regime. (a; top–left) training Loss, (b; top–right) validation Loss, (c; bottom) per-class validation Dice performance for episodic (left), random (middle), and weighted (right).

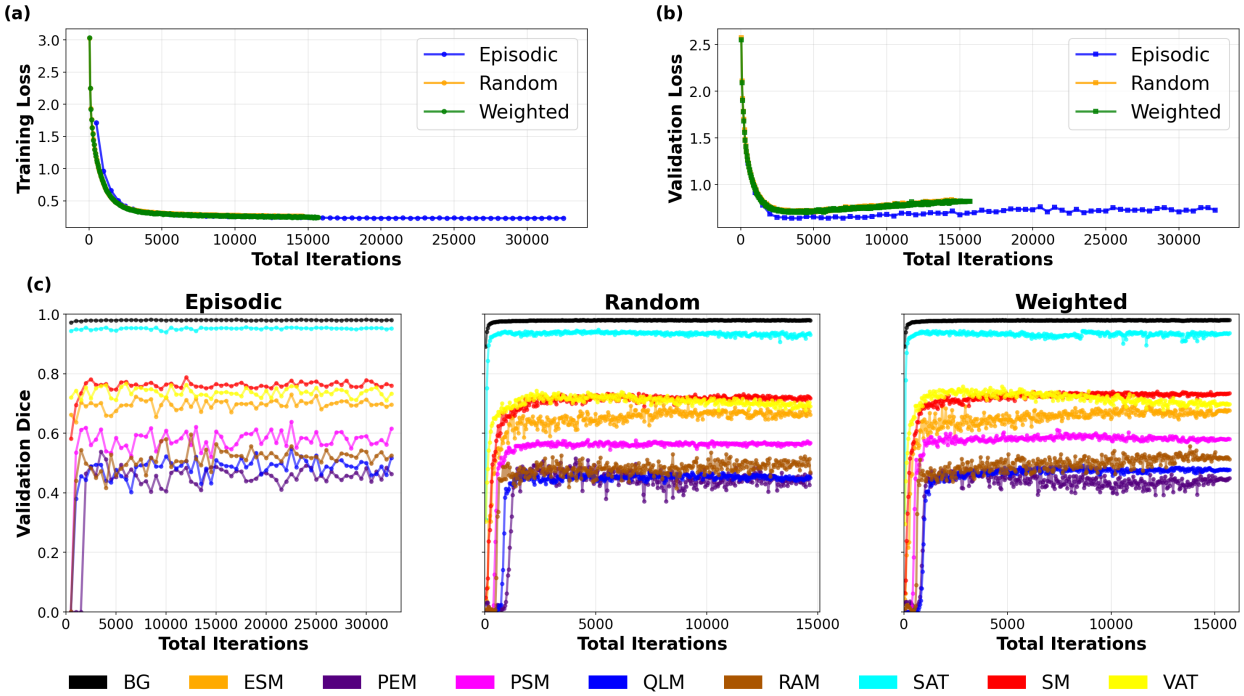


Figure 9: Low-data 10% iteration-calibrated regime. (a; top–left) training Loss, (b; top–right) validation Loss, (c; bottom) per-class validation Dice performance for episodic (left), random (middle), and weighted (right).