RECI-Q: A CLASS-AWARE PATCH SAMPLING CURRICULUM FOR IMBALANCED 3D MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Class imbalance is a fundamental challenge in medical image segmentation, where sampling strategies have been widely explored to improve the representation of underrepresented classes during training. However, sampling strategies developed for long-tailed classification cannot be directly transferred to segmentation, as dense prediction additionally requires spatial localization to determine where training crops are extracted. Existing sampling strategies for segmentation typically address this issue at a coarse foreground–background level, overlooking the substantially different prevalence of individual foreground classes. More recent adaptive and curriculum-based approaches dynamically adjust sampling during training, but often rely on auxiliary models, online difficulty signals, or additional mechanisms that increase training complexity. To address these limitations, we propose Reci-Q, a dynamic class-aware sampling strategy that explicitly decouples class selection from spatial localization. The sampling distribution is controlled by a single parameter and follows a reciprocal curriculum, transitioning from class-balanced sampling to the natural prevalence distribution and then back to class-balanced sampling. This design provides class-specific rebalancing while fully exploiting the information from prevalent samples, without any auxiliary models or explicit loss redesign. Experiments across multiple 3D medical image segmentation tasks demonstrate that Reci-Q consistently improves segmentation performance, with particularly notable gains for tail classes.

1 INTRODUCTION

Class imbalance is a fundamental challenge in medical image segmentation, where clinically important structures such as lesions and small anatomical regions may occupy only a small fraction of a 3D volume (Ronneberger et al., 2015; Milletari et al., 2016). This imbalance is particularly problematic in patch-based training (Kamnitsas et al., 2017; Isensee et al., 2021), where low-prevalence classes appear less often, so more prevalent classes or background voxels dominate their contribution to the training objective (Sudre et al., 2017; Azad et al., 2023). A straightforward solution is to modify the training distribution to emphasize underrepresented classes. However, in dense 3D segmentation, class imbalance is inherently coupled with spatial localization. Unlike image classification, where each sample has a single class label, a training volume can contain multiple classes simultaneously. Consequently, effective class rebalancing is not only determined by *which class* to sample, but also by *where* to extract the corresponding training patch from the massive 3D imaging data volume.

Existing sampling strategies only partially address this problem. Foreground oversampling, as widely adopted in frameworks like nnU-Net (Isensee et al., 2021) and MONAI (Cardoso et al., 2022), increases the exposure of foreground regions to emphasize the information of interest, but it treats all foreground classes as a single sampling target. Therefore, a rare lesion and a highly prevalent anatomical structure may receive the same sampling priority despite having drastically different voxel prevalences. Strategies developed for long-tailed classification (Cui et al., 2019; Kang et al., 2019) provide more explicit control over class sampling, but their instance-level formulation are not directly applicable to segmentation, where class selection must be converted into a spatial crop.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Alternatively, loss-level methods such as focal loss (Lin et al., 2017) address imbalance through loss reweighting, while adaptive sampling methods modify sample or patch selection using predefined schedules or model-dependent signals such as loss or uncertainty (Jesson et al., 2017; Schmidt-Mengin et al., 2022). These approaches demonstrate the potential of altering the effective training distribution, but they either rely on online sample-difficulty estimators, increasing training complexity or introducing training instability.

In this work, we propose a *class-aware curriculum sampling strategy* for imbalanced 3D medical image segmentation. Given a training volume, we first select an anchor class according to its voxel prevalence and then sample an anchor voxel from the selected class to determine the patch location. Specifically, in our method (see Eq. 4), a single parameter q controls the sampling distribution: $q = 0$ produces class-balanced sampling, while $q = 1$ recovers the natural prevalence distribution, *i.e.*, proportional to classes' voxel counts. Instead of fixing q throughout training, we introduce a reciprocal curriculum, namely **Reci-Q**, that transitions from class-balanced sampling to natural sampling and then back to class-balanced sampling:

$$q\colon 0 \longrightarrow 1 \longrightarrow 0.$$

This design provides stronger exposure to low-prevalence classes at both the beginning and end of training, allowing the model to receive more informative training signals from these minority classes. By restoring the natural distribution during the intermediate stage, it also reduces the dominance of numerous small background voxel errors that individually provide limited learning signals but collectively steer the model toward further fitting the background. More fundamentally, Reci-Q implicitly reweights the per-voxel loss by controlling the frequency with which voxels from different classes are encountered during training, without requiring manually specified class/loss weights, auxiliary models, or online difficulty estimation.

Our contributions are threefold:

- We propose a new **class-aware patch sampling strategy** that unifies class-balanced and natural prevalence-based sampling through a single parameter q. The strategy implicitly reweights class influence through sampling frequency, without modifying the loss or requiring model-dependent sampling estimation.
- We introduce Reci-Q, a **reciprocate sampling curriculum** that dynamically transitions between class-balanced and natural sampling, providing class-specific exposure at the beginning and end of training while reducing persistent distributional bias towards the background class in the intermediate stage.
- We conduct experiments on five severely imbalanced 3D medical image segmentation tasks, demonstrating that the proposed method more effectively improves segmentation performance compared with the widely used default nnU-Net sampling strategy and focal loss (Lin et al., 2017).

2 Related Work

2.1 Sampling Strategies for Imbalanced Classification

Sampling is a direct way to modify the effective training distribution for imbalanced learning (Chawla et al., 2002; He & Garcia, 2009; Buda et al., 2018; Wang et al., 2019; Ren et al., 2020). In long-tailed classification, Kang et al. (2019); Ju et al. (2022) studied several sampling schemes, including instance-balanced, class-balanced, square-root, and progressively-balanced sampling. These strategies highlight that the sampling distribution can substantially affect representation learning and classifier adaptation (Cui et al., 2019; Galdran et al., 2021), motivating methods that progressively alter the sampling distribution during optimization rather than fixing it throughout training (Cao et al., 2019; Zhou et al., 2020).

However, these formulations operate at the *instance level*: each training instance has a single class label, so class selection directly determines which instances are sampled. This formulation does not transfer directly to dense volumetric segmentation, where a single 3D volume contains multiple semantic classes simultaneously. The fundamental challenge in segmentation is therefore not only *which class to sample*, but also *where to crop* within the continuous volume coordinates.

2.2 Sampling Strategies for Imbalanced Segmentation

Class imbalance is particularly severe in 3D medical image segmentation (Havaei et al., 2017; Kamnitsas et al., 2017; Guerrero et al., 2018; Tappeiner et al., 2022), where small anatomical structures and rare lesions may occupy only a tiny fraction of a volume. A widely used practical solution is the foreground oversampling strategy in nnU-Net (Isensee et al., 2021) and MONAI (Cardoso et al., 2022), which forces a fixed fraction of training patches to contain foreground regions while sampling the remaining patches randomly. Although effective for background–foreground imbalance, this approach treats all foreground classes as a single uniform target and therefore does not explicitly account for their drastically different voxel-level prevalences.

Recent work has further explored adaptive patch sampling and curriculum strategies (Hoyer et al., 2022; Katharopoulos & Fleuret, 2018). Some methods progressively alter the sampling schedule or patch dimensions over the course of training (Bengio et al., 2009; Jesson et al., 2017; Fischer et al., 2025), while others leverage model-derived loss, uncertainty, or prediction error to dynamically sample informative patches (Shrivastava et al., 2016; Schmidt-Mengin et al., 2022; Park et al., 2026). These approaches demonstrate the benefit of adapting the training distribution, but typically operate at the patch level and rely on either predefined spatial schedules or computationally heavy, model-dependent sample-difficulty signals.

Our approach instead formulates imbalance at the *class level within each volume*. Class prevalence directly determines the probability of selecting an anchor class, and the selected class is then converted into a spatial sampling decision through an anchor voxel. This explicitly connects class rebalancing with spatial localization, without requiring an auxiliary model or online sample-difficulty estimator. Finally, our training distribution is not fixed: it follows a two-phase curriculum from class-balanced to natural sampling and then back to class-balanced sampling. This differs from existing curricula that primarily define progression through sample difficulty or patch-level model feedback, providing a clear theoretical and practical motivation for the reciprocal schedule studied in our method.

3 Method

3.1 Preliminaries

Let $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N}$ denote a training dataset with N samples, where $\mathbf{X}_i$ denotes a 3D medical image volume and $\mathbf{Y}_i$ the corresponding voxel-wise segmentation map. Here, we assume the dataset contains $\mathcal{C}$ segmentation classes, including the background class. Dropping the sample index i for clarity, we calculate the number of voxels belonging to class $c \in \mathcal{C}$ for a single training sample $\mathbf{X}$ as

$$n_c = \sum_{v \in \mathbf{X}} \mathbb{1}[\mathbf{Y}_v = c], \tag{1}$$

where v denotes a voxel in the image volume and $\mathbf{Y}_v$ denotes v's corresponding segmentation label, $\mathbb{1}[.]$ denotes the indicator function. We further define the set of classes present in $\mathbf{X}$ as

$$\mathcal{C}_{\mathbf{X}} = \text{set}(\mathbf{Y}), \tag{2}$$

where the $\text{set}(\cdot)$ function computes the unique label values.

Curriculum learning augments the training process by dynamically changing the training sampling prior distribution (*i.e.*, image crops in our case) rather than optimizing over a fixed one. Let $\mathcal{P}_t$ denote the effective training distribution at training step t, a curriculum learning algorithm progressively changes the distribution over the course of training so that

$$\mathcal{P}_t = \{p_c(t) | c \in \mathcal{C}_{\mathbf{X}}\}, \tag{3}$$

where $p_c(t)$ is a function that controls the sampling probability at step t for a class c.

In the conventional curriculum-learning formulation (Bengio et al., 2009; Soviany et al., 2022), the changing distribution is often associated with a progression in sample difficulty, usually from easy to hard to gradually warm up the training process.

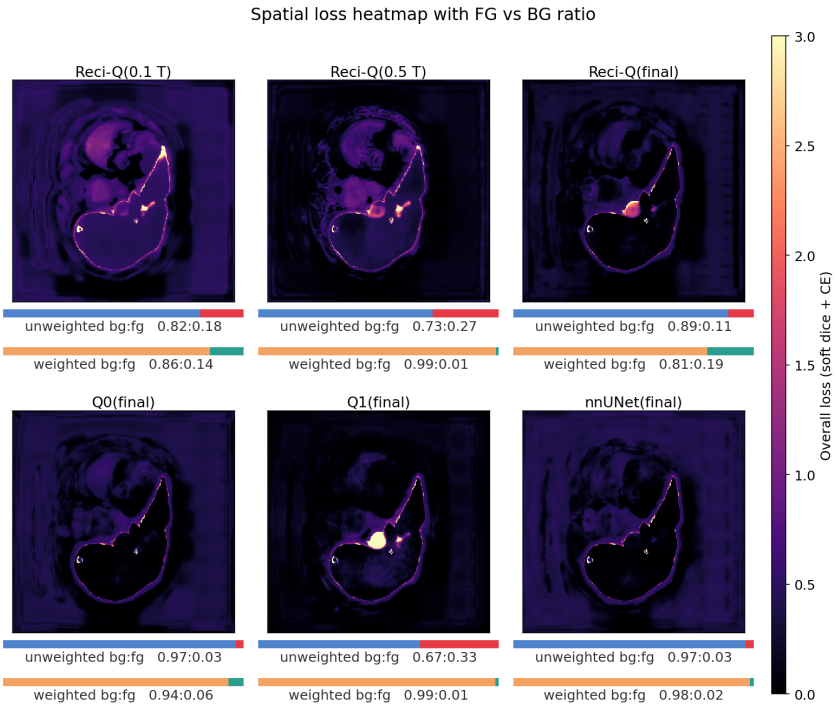


Figure 1: An example slice-level training sample loss heatmap from the MSD Task3 Liver segmentation task. **Top row:** the proposed Reci-Q at $0.1T$, $0.5T$, and the final checkpoints. **Bottom row:** final-checkpoint comparisons among $q = 0$, $q = 1$, and the nnU-Net default sampling strategy ($q \approx 0.2$). Brighter regions indicate larger errors relative to the ground-truth labels, whereas darker regions indicate correct and highly confident predictions. The two colored bars beneath each panel show the unweighted and sampling-weighted background (bg) to foreground (fg) loss ratios, with the latter accounting for the sampling probability of each class.

3.2 Class-aware Sampling Strategy

Our goal is to construct a dynamic sampling strategy that is customized for the highly imbalanced spatial distribution of classes in 3D medical image segmentation. Our sampling strategy consists of two stages: *sampling probability calculation* and *patch sampling and cropping*.

Sampling probability calculation. Given a training sample $\mathbf{X}$ and a class $c \in \mathcal{C}_{\mathbf{X}}$, we determine the probability of selecting c for training as:

$$p_c(t) = \frac{n_c^{q^{(t)}}}{\sum_{c' \in \mathcal{C}_{\mathbf{X}}} n_{c'}^{q^{(t)}}}, \qquad q \in [0, 1]. \tag{4}$$

The exponent $q^{(t)}$ controls the extent to which the sampling probability reflects the observed voxel prevalence of each class at step t:

- When $q^{(t)} = 0$, all classes present in the sample are assigned equal sampling probabilities, yielding a class-balanced sampling distribution.
- When $q^{(t)} = 1$, the sampling probability is proportional to the number of voxels in $\mathbf{X}$ belonging to each class, recovering the natural class prevalence distribution within the sample.
- Intermediate values of $q^{(t)}$ provide a continuous interpolation between these two sampling extremes.

Patch sampling and cropping. After selecting a class, we determine the corresponding location for image patch cropping. For a foreground class, an anchor voxel is sampled uniformly from all voxels belonging to that class:

$$w \sim \begin{cases} \text{Uniform}\left(\{v \mid v \in \mathbf{X},\ \mathbf{Y}_v = c\}\right), & \text{if } c = \text{foreground}, \\ \text{Uniform}(\mathbf{X}), & \text{otherwise.} \end{cases} \tag{5}$$

For a foreground class, this procedure explicitly selects an anchor voxel from the corresponding class, thereby increasing its spatial exposure during training. For the background class, however, explicitly identifying background voxels is unnecessary and computationally inefficient, as background typically occupies more than 95% of the volume in 3D medical image segmentation tasks. Thus, we sample w uniformly from the entire image as an efficient approximation.

Finally, an image patch $\mathbf{x}$ is extracted around the selected anchor voxel using the default patch size determined by nnU-Net preprocessing. The anchor voxel determines the initial crop center location, while the crop boundaries are adjusted when necessary to ensure that the entire patch remains within the valid image region. Specifically, when the anchor voxel is close to an image boundary, the crop is shifted along the corresponding dimension to maintain the prescribed patch size, following the default cropping strategy of nnU-Net (Isensee et al., 2021). The class label c of the selected anchor voxel is then assigned to the resulting patch.

3.3 Q-BASED RECIPROCAL (RECI-Q) TRAINING SCHEDULER

In general, sampling difficulty can be defined directly by voxel prevalence: classes occupying fewer voxels are considered harder to sample, whereas more prevalent classes are considered easier. Rather than fixing a single sampling distribution, we allow the class sampling distribution to traverse the full spectrum of p_c induced by $q \in [0, 1]$, forming a reciprocal sampling schedule:

$$q^{(t)} = \begin{cases} q_{\text{start}} + (q_{\text{mid}} - q_{\text{start}})\,\dfrac{t}{\lambda T}, & 0 \le t \le \lambda T, \\ q_{\text{mid}} + (q_{\text{end}} - q_{\text{mid}})\,\dfrac{t - \lambda T}{(1-\lambda)T}, & \lambda T < t \le T, \end{cases} \tag{6}$$

where T denotes the total number of training steps and $\lambda \in [0, 1]$ defines the timing where q_{mid} is reached. We set $q_{\text{start}} = 0$, $q_{\text{mid}} = 1$, $q_{\text{end}} = 0$, and $\lambda = \frac{1}{2}$ so that:

$$q : 0 \xrightarrow{\frac{1}{2}T} 1 \xrightarrow{\frac{1}{2}T} 0. \tag{7}$$

As a result, the two $q = 0$ endpoints increase the exposure of low-prevalence classes, whereas the intermediate $q = 1$ phase restores sampling according to the natural class prevalence. This reciprocal schedule provides stronger class-specific learning signals at the beginning and end of training, while temporarily returning to the natural data distribution to mitigate the artificially set sampling bias and potential overfitting of positive foreground classes associated with persistent class rebalancing (*i.e.*, $q_{\text{start/mid/end}} = 0$).

3.4 THE IMPLICIT EFFECT OF SAMPLING FREQUENCY ON LOSS WEIGHTING

Within each training patch $\mathbf{x}$ centered at w, the loss is computed as the unweighted average of the voxel-wise loss terms $\ell(v)$ over all voxels $v \in \mathbf{x}$ (*i.e.*, soft Dice and cross-entropy losses in our case):

$$\ell(\mathbf{x}) = \frac{1}{|\mathbf{x}|} \sum_{v \in \mathbf{x}} \ell(v). \tag{8}$$

We note that, for each training sample $\mathbf{X}_i$, a single patch $\mathbf{x}_i$ is sampled according to $q^{(t)}$, and t is updated once at the end of each epoch.

Assuming M patches are drawn in one epoch, the accumulated loss of the epoch can be written as:

$$L_t = \frac{1}{M} \sum_{i=1}^{M} l(\mathbf{x}_i) = \sum_{c \in \mathcal{C}} \frac{\sum_{i=1}^{M} \mathbb{1}[\mathbf{Y}_{w_i} = c]}{M} \ell(\mathbf{x}_i), \tag{9}$$

where $\frac{1}{M}\sum_{i=1}^{M}\mathbb{1}[\mathbf{Y}_{w_i}=c] \approx p_c(t)$ and therefore varies with t. From this perspective, the sampling scheduler can be interpreted as implicitly reweighting the loss contributions of the foreground and background classes over the course of training without explicitly assigning the class weights.

How does the reciprocal scheduling affect the training? The nature of highly class-imbalanced 3D medical segmentation tasks, such as small lesions against large backgrounds, raises concerns that training may be dominated by background information. A strict class-balancing strategy emphasizes foreground information but may cause overfitting, as training is repeatedly focused on the limited foreground regions. In Fig. 1, we show an example loss map of a CT slice from MSD Task3 (Antonelli et al., 2022), where the task is to segment the liver and liver tumor. The losses are computed from the final training checkpoints for $q = 0$, $q = 1$, nnU-Net (effective $q \approx 0.2$), and our $q = 0 \to 1 \to 0$ schedule. We illustrate two additional quantities for each setting:

- **Unweighted loss composition bar (background: blue; foreground: red).** This is computed as $\sum_{v,\mathbf{Y}_v=c}\ell(v)$, illustrating the relative contributions of foreground and background voxel losses to the training loss.

- **Weighted loss composition bar (background: yellow; foreground: green).** This is computed as $\sum_{v,\mathbf{Y}_v=c}p_c(t)\ell(v)$, illustrating how the reciprocal scheduler dynamically adjusts the effect of the losses on the model training.

For $q = 0$, the loss is concentrated along the thin liver–background and liver–tumor boundaries, showing that class balancing provides strong supervision for correctly distinguishing foreground voxels. Notably, the proportion of foreground loss remains small in both the unweighted and weighted loss composition bars, as the foreground loss has already been substantially reduced.

For $q = 1$, the loss map has a much smaller overall footprint in the background, suggesting that majority-driven sampling produces highly confident and correct background predictions. Yet, the bright loss regions remain around the liver and liver tumor (near the center of the image), suggesting that training is less effective at correctly identifying these foreground tissues. Consequently, the unweighted loss composition bar shows a large foreground loss component. However, weighting by the natural class prevalence substantially reduces the contribution of the foreground loss, to the extent that the prominent incorrect predictions near the image center remain unresolved at the end of training.

Finally, our proposed $q = 0 \to 1 \to 0$ reciprocal sampling schedule provides a balance between the $q = 0$ and $q = 1$ extremes. In the second row of Fig. 1, we show the loss distributions at three checkpoints, $t = \{0.1T, 0.5T, T\}$. Comparing the final-epoch result with those of $q = 0$ and $q = 1$, the weighted loss composition is substantially more balanced than under either constant setting. This can be attributed to the use of $q_{\text{mid}} = 1$, which allows the majority of the background loss to be resolved first, followed by a gradual transition back to $q_{\text{end}} = 0$, which increases the relative emphasis on the remaining unresolved foreground loss and gives the network a greater opportunity to correct these errors.

As a comparison, the nnU-Net example, with an average effective $q \approx 0.2$, produces a loss pattern close to the final result when setting $q = 0$. Intuitively, the fixed q value might overlook the individual prevalence, thus failing to generalize. Thus, nnU-Net would still yield relatively higher loss on the background areas compared to Reci-Q. Note that the approximated q value is obtained by averaging the 13 $q_c^{\text{nnU-Net}}$ values in Table 1.

Illustration. To illustrate the relationship between q and the resulting sampling probability p, we visualize different scheduling settings on the same MSD liver segmentation task in Fig. 2. Although q is linearly interpolated within each phase, the induced sampling probability p is generally changing nonlinearly. Finally, from left to right in Fig. 2, we progressively delay the training point at which $q = 1$ is reached, from q_{mid} toward q_{end}, illustrating different variants of the reciprocal sampling schedule evaluated in the ablation study in Table 3.

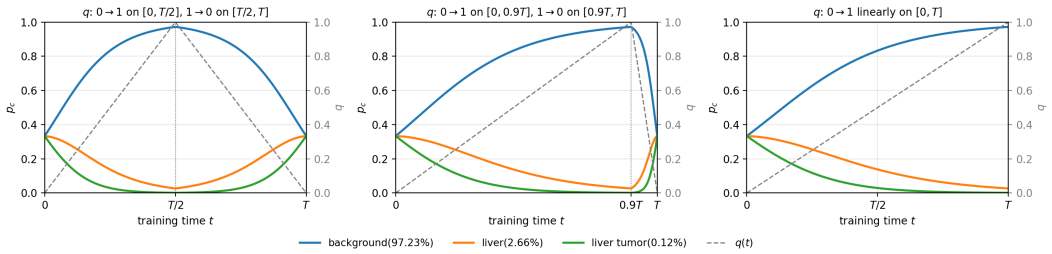


Figure 2: Relationship between the interpolation coefficient q and the induced sampling probability p for MSD Task3 Liver under different timings at which q reaches 1. **Left:** $q = 0 \to 1 \to 0$ [5:5]; **middle:** $q = 0 \to 1 \to 0$ [9:1]; and **right:** $q = 0 \to 1$.

Table 1: MSD Dataset statistics (Antonelli et al., 2022) and the per-class q_c value that reproduces nnU-Net's default foreground-oversampling behavior under our q-parameterized sampling formula. Because nnU-Net's (Isensee et al., 2021) default draws foreground classes *uniformly* rather than proportionally to voxel count, q_c generally differs across classes within the same dataset; values are reported for reference only.

Dataset	**Class**	**#Samples**	n_c	$p_c^{\text{nnU-Net}}$	$q_c^{\text{nnU-Net}}$
Task3 Liver	background	105	97.23%	0.670	0.291
	liver		2.66%	0.165	0.436
	liver tumor		0.12%	0.165	0.179
Task6 Lung	background	50	99.96%	0.670	0.091
	lung tumor		0.04%	0.330	0.091
Task7 Pancreas	background	225	99.78%	0.670	0.198
	pancreas		0.20%	0.165	0.239
	pancreas tumor		0.03%	0.165	0.161
Task8 Hepatic	background	242	99.54%	0.670	0.226
	vessel		0.11%	0.165	0.196
	liver tumor		0.35%	0.165	0.259
Task10 Colon	background	101	99.94%	0.670	0.096
	colon tumor		0.06%	0.330	0.096

4 Experiments

4.1 Datasets and Implementation

Datasets. We evaluate on five tumor-related datasets from the Medical Segmentation Decathlon (MSD) benchmark (Antonelli et al., 2022): Liver, Lung, Pancreas, Hepatic Vessel, and Colon, covering both binary and multi-class segmentation. Each dataset is randomly split into training and test sets, with all training samples used for training and no separate validation set. We report the average per-class voxel ratios across training volumes in Table 1.

Implementation. We use nnU-Net (Isensee et al., 2021) as the segmentation baseline and retain its default training configuration, modifying only the sampling strategy. Models are trained for 1000 epochs using stochastic gradient decent (SGD) with momentum 0.99, weight decay 3×10^{-5}, an initial learning rate of 10^{-2}, and polynomial decay to approximately 2×10^{-5}. We use the standard equally weighted soft Dice and cross-entropy loss.

The default nnU-Net sampling strategy oversamples the foreground with a probability of 0.33, selecting uniformly among the foreground classes present in the sampled case. For reference, we map this behavior to per-class q_c values by matching the empirical sampling probability in Eq. 4; the resulting q_c values are reported in Table 1. Since the two-stage foreground/background sampling strategy of nnU-Net cannot, in general, be represented by a single global q, these values are pro-

Table 2: Dice score (%) for each class under different sampling strategies. The tail class of each task is marked by the **bold** font in the header. Best performance is in **bold** font for each column. For nnUNet + PGPS (Fischer et al., 2025), training failed to converge on Task 6, Task 7 and Task 8, yielding near-zero Dice scores; these entries are marked with a dash and excluded from the Average and Tail Class AVG for this row.

	Task3 Liver		Task6 Lung	Task7 Pancreas		Task8 Hepatic		Task10 Colon	Average	
Sampling strategy	liver	**liver tumor**	**lung tumor**	pancreas	**pancreas tumor**	**vessel**	liver tumor	**colon tumor**	**All classes**	**Tail Classes**
nnUNet (Isensee et al., 2021)	95.66	71.21	69.99	82.93	52.48	64.45	**73.24**	46.77	69.59	60.98
nnUNet + Focal loss (Lin et al., 2017)	95.65	70.96	63.56	82.89	52.09	64.60	73.09	47.56	68.80	59.75
nnUNet + PGPS (Fischer et al., 2025)	93.14	2.29	–	–	–	–	–	**51.23**	–	–
SMOTE($q = 0$) (Chawla et al., 2002)	95.61	71.93	60.36	83.16	53.01	64.75	72.71	44.11	68.20	58.83
$q = 1$	95.66	71.99	66.15	82.92	**54.23**	64.90	73.03	49.70	69.82	61.39
$q = 0 \rightarrow 1$	**95.87**	**73.81**	70.91	83.14	53.23	64.69	72.87	47.29	70.23	61.99
$q = 1 \rightarrow 0$	95.67	72.29	69.05	**83.35**	53.73	**65.43**	72.88	49.12	70.19	61.92
$q = 0 \rightarrow 1 \rightarrow 0$ [5:5]	95.56	73.03	**72.37**	83.06	52.76	65.05	72.98	49.85	**70.58**	**62.61**

Table 3: Ablation on the mixing schedule of the two-phase $q = 0 \rightarrow 1 \rightarrow 0$ sampling strategy (Dice score %). The ratio in square brackets denotes the ratio of the "$q : 0 \rightarrow 1$" phase versus the "$q : 1 \rightarrow 0$" phase. Best performance in **bold** for each column.

	Task3 Liver		Task6 Lung	Task7 Pancreas		Task8 Hepatic		Task10 Colon		
Sampling strategy	liver	**liver tumor**	**lung tumor**	pancreas	**pancreas tumor**	**vessel**	liver tumor	**colon tumor**	**Average**	**Tail Class AVG**
nnUNet (Isensee et al., 2021)	95.66	71.21	69.99	82.93	52.48	64.45	**73.24**	46.77	69.59	60.98
$q = 0 \rightarrow 1 \rightarrow 0$ [1:9]	**95.67**	72.68	69.30	**83.18**	**53.28**	**65.49**	72.40	45.74	69.72	61.30
$q = 0 \rightarrow 1 \rightarrow 0$ [9:1]	95.58	72.34	70.39	83.16	52.09	64.68	72.94	46.90	69.76	61.28
$q = 0 \rightarrow 1 \rightarrow 0$ [5:5]	95.56	**73.03**	**72.37**	83.06	52.76	65.05	72.98	**49.85**	**70.58**	**62.61**

vided only to conceptually align the default nnU-Net strategy with our q-parameterized sampling probability framework.

4.2 Performance and Ablation

Dynamic scheduling outperforms static sampling. Table 2 compares our reciprocal sampling scheduler with static choices of q and the nnU-Net baseline. Static class-balanced sampling ($q = 0$) underperforms the nnU-Net baseline in both overall Dice (68.20 vs. 69.59) and tail-class Dice (58.83 vs. 60.98), indicating that persistent rebalancing alone is not sufficient. Here, the tail class of each task is defined as the least represented foreground class, which are the liver tumor (Task 3), lung tumor (Task 6), pancreas tumor (Task 7), vessel (Task 8), and colon tumor (Task 10) classes. In contrast, natural sampling ($q = 1$) achieves 69.82 overall and 61.39 tail-class Dice. The two monotonic schedules, $q : 0 \rightarrow 1$ and $q : 1 \rightarrow 0$, further improve the overall Dice to 70.23 and 70.19, respectively. These results suggest that changing the sampling distribution during training is more effective than fixing it at either extremes.

The complete $q : 0 \rightarrow 1 \rightarrow 0$ schedule with a symmetric $[5{:}5]$ phase allocation achieves the best performance among the completed runs, reaching 70.58 overall and 62.61 tail-class Dice. Compared with nnU-Net, this corresponds to gains of +0.99 and +1.63 points, respectively. At the individual-class level, the largest improvements occur for lung tumor (+2.38) and colon tumor (+3.08).

Comparison with focal loss and PGPS. Replacing cross-entropy with focal loss does (Lin et al., 2017) not improve nnU-Net performance, with the average Dice decreasing from 69.59% to 68.80% and the tail-class average from 60.98% to 59.75%. We believe this is due to the Dice loss already emphasizing target-region overlap, leaving limited additional benefit from focal reweighting. PGPS (Fischer et al., 2025) progressively increases the patch size from small to large during training and shows stable performance on the binary tasks (Task 6 and Task 10), but is highly unstable on multi-class tasks. In particular, the training collapses on Task 7 and Task 8, with the foreground classes achieving near-zero Dice. We believe the reason could be that PGPS follows the original nnU-Net sampling strategy, which treats all foreground classes equally despite potentially large differences in their voxel prevalence. When multiple foreground classes have highly unequal prevalence, the small patches used in the early training stage may provide insufficient and inconsistent exposure to rare classes, making the training process unstable.

Phase shift. We show ablation of the training duration for each of the two phases in the $q : 0 \rightarrow 1 \rightarrow 0$ schedule in Table 3. Allocating most of the training time to either the first transition $[9{:}1]$

or the second transition $[1{:}9]$ yields similar intermediate performance, with overall Dice of 69.76 and 69.72, respectively. The symmetric $[5{:}5]$ schedule achieves 70.58 overall and 62.61 tail-class Dice, outperforming both skewed schedules. Within the evaluated configurations, this indicates that maintaining a balanced training duration for the two phases is most beneficial.

Discussion. Together, these results support the use of a dynamic, two-phase sampling schedule rather than a fixed sampling distribution. Neither sampling according to the natural class prevalence nor applying a static class-balancing strategy alone necessarily yields the best performance. The 5:5 $q : 0 \to 1 \to 0$ schedule achieves the strongest aggregate performance across the benchmarked 3D medical image segmentation tasks, suggesting that early class balancing, an intermediate natural-distribution phase, and a final return to class balancing provide complementary training signals.

5 Conclusion and Future Work

In this paper, we presented **Reci-Q**, a class-aware curriculum sampling strategy for imbalanced 3D medical image segmentation. By leveraging Eq. 4, Reci-Q can be readily adapted to tasks with different numbers and compositions of foreground classes.

We further introduced a reciprocal sampling curriculum that transitions between class-balanced and natural prevalence sampling, providing stronger class-specific training signals while eliminating the need to manually specify class-specific weights. Through this dynamic sampling process, the frequency of training patches from different classes is implicitly controlled, effectively reweighting their contributions to training without modifying the per-voxel loss or introducing model-dependent estimates of sample difficulty. Experiments across multiple 3D medical imaging tasks demonstrate consistent improvements in segmentation performance, with particularly notable gains for the least represented classes, which often correspond to the primary targets of clinical interest motivating the imaging study.

Our current framework uses a shared sampling parameter $q^{(t)}$ across all classes, providing a simple and general sampling strategy. However, the optimal sampling trajectory may differ across classes and tasks due to differences in class prevalence and learning characteristics. Future work will therefore explore finer-grained class- and task-specific sampling schedules by monitoring the training dynamics and adaptively adjusting the corresponding $q_c^{(t)}$ trajectories. This may provide further insights into the effects of class balancing in 3D medical image segmentation and potentially lead to further improvements in segmentation performance.

AI Use Statement

In this work, we used generative AI tools to assist with polishing and improving the clarity and readability of the manuscript. We reviewed all AI-assisted text and verified its accuracy and consistency with the research. We take responsibility for the final content of this work, including all text and claims produced with the aid of generative AI.

Reproducibility Statement

Our experiments are conducted within the publicly available nnU-Net framework, providing a standardized training and evaluation pipeline. The main paper describes the proposed class-aware sampling strategy and training schedule in detail. To facilitate future reproduction, we will make the train/test splits used in our experiments and the implementation of the proposed sampling strategy and data loader publicly available upon acceptance.

References

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Reza Azad, Moein Heidary, Kadir Yilmaz, Michael Hüttemann, Sanaz Karimijafarbigloo, Yuli Wu, Anke Schmeink, and Dorit Merhof. Loss functions in the era of semantic segmentation: A survey and outlook. *arXiv preprint arXiv:2312.05391*, 2023.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9260–9269. IEEE, 2019.

Stefan M Fischer, Johannes Kiechle, Laura Daza, Lina Felsner, Richard Osuala, Daniel M Lang, Karim Lekadir, Jan C Peeken, and Julia A Schnabel. Progressive growing of patch size: Curriculum learning for accelerated and improved medical image segmentation. *arXiv preprint arXiv:2510.23241*, 2025.

Adrian Galdran, Gustavo Carneiro, and Miguel A González Ballester. Balanced-mixup for highly imbalanced medical image classification. In *International conference on medical image computing and computer-assisted intervention*, pp. 323–333. Springer, 2021.

R Guerrero, C Qin, O Oktay, C Bowles, L Chen, R Joules, R Wolz, MC Valdés-Hernández, DA Dickie, J Wardlaw, et al. White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. neuro-image: Clin 17: 918–934, 2018.

Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9914–9925. IEEE, 2022.

Fabian Isensee, Paul F Jaeger, Simon AA 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, 2021.

Andrew Jesson, Nicolas Guizard, Sina Hamidi Ghalehjegh, Damien Goblot, Florian Soudan, and Nicolas Chapados. Cased: curriculum adaptive sampling for extreme data imbalance. In *International conference on medical image computing and computer-assisted intervention*, pp. 639–646. Springer, 2017.

Lie Ju, Yicheng Wu, Lin Wang, Zhen Yu, Xin Zhao, Xin Wang, Paul Bonnington, and Zongyuan Ge. Flexible sampling for long-tailed skin lesion classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 462–471. Springer, 2022.

Konstantinos Kamnitsas, Christian Ledig, Virginia FJ 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, 2017.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *International conference on machine learning*, pp. 2525–2534. PMLR, 2018.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

Junha Park, Arthur Cho, and Hae-Jeong Park. Adaptive patch sampling and location-aware reasoning for whole body pet-ct multi-organ segmentation. *Scientific Reports*, 2026.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Marius Schmidt-Mengin, Théodore Soulier, Mariem Hamzaoui, Arya Yazdan-Panah, Benedetta Bodini, Nicholas Ayache, Bruno Stankoff, and Olivier Colliot. Online hard example mining vs. fixed oversampling strategy for segmentation of new multiple sclerosis lesions from longitudinal flair mri. *Frontiers in Neuroscience*, 16:1004050, 2022.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761–769, 2016.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022.

Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *International Workshop on Deep Learning in Medical Image Analysis*, pp. 240–248. Springer, 2017.

Elias Tappeiner, Martin Welk, and Rainer Schubert. Tackling the class imbalance problem of deep learning-based head and neck organ segmentation. *International journal of computer assisted radiology and surgery*, 17(11):2103–2111, 2022.

Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. Dynamic curriculum learning for imbalanced data classification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5016–5025. IEEE, 2019.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9716–9725. IEEE, 2020.