This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Rebalancing Lesion Exposure: Instance-Aware Sampling for msTBI Lesion Segmentation

Seungha Noh[1][0009−0006−7800−6666], Cheolhee Yu[1][0009−0003−6891−9048], Hyun Yang[2][0009−0006−3395−3918], and Jin Tae Kwak[1,2][0000−0003−0287−4097]

[1] Korea University, Seoul 02841, Republic of Korea
{shnoh164, songdam99, jkwak}@korea.ac.kr
[2] DeepClue Inc., Seoul 04790, Republic of Korea
hyang05@deepclue.io

Abstract. Automated segmentation of moderate-to-severe traumatic brain injury (msTBI) lesions on T1-weighted MRI remains challenging because multiple lesions with substantial inter-instance size imbalance may coexist within a scan. Conventional nnU-Net foreground oversampling draws anchors from pooled foreground voxels, making lesion sampling probability approximately proportional to lesion volume and consequently underexposing small lesions. We adapt instance-aware foreground sampling to nnU-Net v2 by decomposing the ground-truth mask into connected lesion instances, uniformly selecting an instance, and then sampling an anchor voxel from that instance. This rebalances lesion exposure during training without modifying the network architecture. We evaluated the approach on the AIMS-TBI 2026 dataset using a 3D full-resolution Residual-Encoder U-Net. On the official AIMS-TBI validation set, averaged across five fold-trained models, instance-aware sampling improved mean Dice from 0.592 to 0.601, while reducing HD95 from 22.864 to 21.731 and ASSD from 7.356 to 5.888 compared with conventional foreground sampling. Lesion-wise analysis further showed improved performance for small lesions, with recall increasing from 0.490 to 0.517, F1 from 0.530 to 0.555, and Matched Dice from 0.654 to 0.667 for lesions with volumes ≤ 1 mL. These results indicate that balancing foreground sampling at the lesion-instance level can improve msTBI lesion segmentation, particularly for small lesions that are underrepresented by voxel-based sampling.

Keywords: Traumatic brain injury · Lesion segmentation · MRI · Class imbalance · Foreground sampling.

1 Introduction

Moderate-to-severe traumatic brain injury (msTBI) causes primary injuries (e.g., contusions, hemorrhages, and haematomas) and initiates secondary pathological changes (e.g., gliosis and encephalomalacia), whose patient-specific accumulation produces highly heterogeneous lesion profiles and structural brain deformation [1]. Unlike lesions typically encountered in stroke, multiple sclerosis, or

brain tumours, msTBI lesions may be focal or diffuse and vary widely in size, number, shape, and laterality, and extend across multiple tissue types. Accurate automated lesion segmentation is therefore essential for reliable and reproducible downstream neuroimaging analysis.

Modern 3D segmentation frameworks such as nnU-Net [2] rely on patch-based training and employ foreground oversampling to mitigate foreground–background imbalance. In msTBI, however, a single scan often contains multiple lesions with substantial variation in size. Under conventional foreground oversampling, all lesion voxels are pooled into a single foreground set without distinguishing individual lesion instances. Consequently, the probability that a particular lesion provides the anchor for a foreground patch is approximately proportional to its volume. Large lesions therefore dominate foreground patch selection, whereas small lesions may receive substantially fewer training exposures.

Prior work has recognized that conventional foreground–background balancing does not explicitly account for lesion–level imbalance. From a data-sampling perspective, [3] introduced instance-balanced patch sampling for whole-body lesion segmentation, in which lesion–centered positive patches are generated offline and each lesion instance is presented once per epoch. Their results demonstrate the benefit of balancing lesion instances in large-field-of-view images containing numerous lesions of highly variable size. From a loss-design perspective, [4] proposed Blob Loss, which decomposes the foreground into individual connected components and balances their contributions to the training objective, thereby reducing domination by large lesion instances. These approaches address instance imbalance at different stages of the training pipeline. Loss-level methods rebalance the contributions of lesion instances after they are included in a sampled patch, whereas sampling-level methods control which lesion instances are exposed during training. In contrast to the offline lesion-centered patch construction of [3], we investigate instance balancing within the native on-the-fly foreground-oversampling pathway of nnU-Net.

In this work, we modify the foreground anchor-selection step of nnU-Net v2 to perform instance-level balancing for msTBI lesion segmentation. Connected-component analysis decomposes the binary ground-truth mask into individual lesion instances, from which a lesion is selected uniformly whenever a foreground-anchored patch is requested, independently of its volume. An anchor voxel is then sampled from the selected instance. This decouples lesion size from foreground-anchor selection probability while preserving the remainder of the nnU-Net training pipeline. The strategy requires no architectural modification and is implemented within an nnU-Net v2 3D full-resolution Residual-Encoder U-Net [5]. We evaluate the proposed strategy on the AIMS-TBI 2026 lesion segmentation task using models independently trained on all five folds, together with lesion-wise analyses stratified by lesion volume, and compare it directly against conventional nnU-Net foreground sampling.

2 Method

2.1 Dataset

We evaluated our method on the AIMS-TBI 2026 challenge dataset, curated by the ENIGMA consortium msTBI working groups (Pediatric and Adult) [1, 6]. The segmentation target is a binary mask of msTBI lesions on T1-weighted MRI. The released training set comprised 552 cases, including 322 lesion-containing and 230 lesion-free cases. Additional modalities, including diffusion MRI, T2-weighted MRI, FLAIR, and SWI, were available during training, but we used only T1-weighted MRI as input for all experiments. The training set was partitioned into five folds at the patient level, stratified by case label (lesion-free vs. lesion-containing) and lesion size among lesion-containing cases.

2.2 Preprocessing

Before applying the nnU-Net preprocessing pipeline, each T1w volume and its corresponding ground-truth mask were reoriented to the RAS coordinate system. The reoriented image-mask pairs were then cropped to the non-zero image region and resampled to 1-mm isotropic voxel spacing. T1w intensities were normalized independently for each case using z-score normalization. The mean μ and standard deviation σ were computed over the non-zero image region, and each voxel within this region was normalized as $x' = (x - \mu)/\sigma$. Voxels outside the non-zero region were set to zero. This case-wise normalization reduces differences in absolute intensity scales across scanners and acquisition protocols in the multi-site AIMS-TBI dataset.

2.3 Baseline Architecture

We adopt the nnU-Net v2 3D full-resolution Residual-Encoder U-Net (the ResEnc-M) [5] as the baseline architecture. As illustrated in Fig. 1, the network follows an encoder-decoder structure with skip connections between corresponding resolution levels. The encoder comprises six stages in which the standard convolutional blocks are replaced by residual blocks, while the decoder progressively restores spatial resolution through five decoding stages. The network receives a single-channel T1-weighted (T1w) volume as input. The segmentation head produces two-channel logits corresponding to background and lesion. During inference, voxel-wise softmax probabilities are computed, and the class with the highest probability at each voxel is selected to obtain the final binary segmentation mask.

2.4 Instance-aware Sampling

We modify only the foreground anchor-selection step, using lesion instances rather than individual foreground voxels as the sampling unit. After preprocessing and resampling, connected-component analysis with 26-connectivity in 3D is applied to the binary ground-truth mask to decompose the foreground

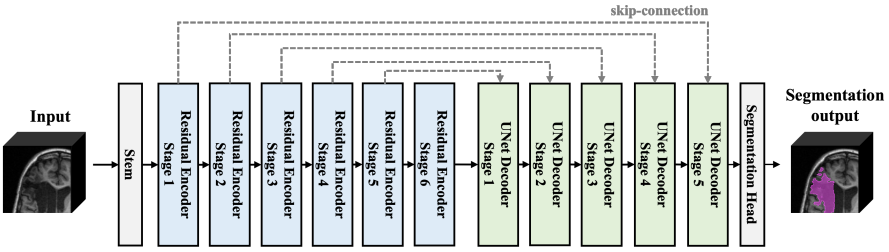

Fig. 1. Overview of the nnU-Net v2 3D full-resolution Residual-Encoder U-Net used as the baseline segmentation network.

into lesion instances, and the candidate coordinates belonging to each instance are stored in a separate coordinate pool. When the coordinate set is subsampled for memory efficiency, at least one candidate coordinate is retained for every non-empty instance. Across the five folds, each training set contained 441–442 cases, with 1,623–1,741 lesion instances identified using 26-connectivity. Lesion-size distributions were similar across folds, with approximately 86% of instances smaller than 1 mL.

When a foreground patch is requested, we first select a lesion instance uniformly at random, independently of its size, and then draw an anchor coordinate uniformly from within that instance. For lesion-free cases, no valid foreground instance is available; therefore, the foreground-anchored pathway falls back to the standard random patch-sampling procedure. The final crop location is determined using the original nnU-Net bounding-box routine, subject to the constraint that the selected anchor lies within the extracted patch.

The left panel of Fig. 2 contrasts conventional voxel-wise foreground sampling with the proposed instance-level sampling. Under conventional sampling, the probability that a given lesion anchors a foreground patch is proportional to its voxel count; under the proposed strategy, every lesion instance within the selected case is equally likely to be chosen, regardless of size. The right panel shows the resulting sampled lesion-size distributions, demonstrating that smaller lesion instances are selected more frequently as foreground anchors. By intervening only at the data-exposure stage, the proposed strategy rebalances lesion exposure without introducing additional network modules or modifying the loss function.

2.5 Training Strategy

Rather than processing entire volumes, we train the network using $128 \times 128 \times 128$ voxel patches. At the target 1-mm isotropic spacing, each patch corresponds to a physical field of view of $128 \times 128 \times 128$ mm^3. The same patch dimensions are used for sliding-window inference. Training patches are generated using a mixture of random sampling and foreground-anchored sampling. Following the nnU-Net default, the foreground oversampling ratio is set to 0.33. The proposed

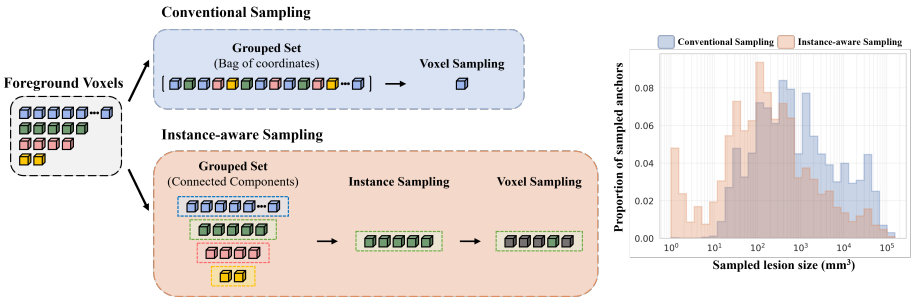

Fig. 2. Overview of the proposed instance-aware sampling strategy. Left: conventional voxel-wise foreground sampling and the proposed instance-level sampling. Right: lesion-size distributions of foreground anchors obtained with each strategy. Lesion size is displayed on a logarithmic scale, with histogram bins equally spaced in log space.

instance-aware strategy is applied only to foreground-anchored patches, while the random-sampling pathway remains identical under conventional and instance-aware sampling.

We optimize a compound objective combining cross-entropy, Dice, and Focal losses:

$$\mathcal{L} = 0.3\,\mathcal{L}_{\text{CE}} + 0.4\,\mathcal{L}_{\text{Dice}} + 0.3\,\mathcal{L}_{\text{Focal}}, \tag{1}$$

The loss weights are selected based on the fold-0 ablation study reported in Table 1. The objective is applied with deep supervision, where the loss is evaluated at multiple decoder resolutions. The unnormalized weights assigned to each progressively coarser output are halved relative to the preceding output $(1, \frac{1}{2}, \frac{1}{4}, \dots)$ and the weights are then normalized to sum to one.

Models used for the final sampling comparison were trained from scratch for 1,000 epochs with 250 iterations per epoch and a batch size of 8. Earlier architecture and patch-size ablations were conducted with a batch size of 2, as indicated in Table 1. Optimization was performed using SGD with Nesterov momentum of 0.99, weight decay of 3×10^{-5}, and an initial learning rate of 1×10^{-2}. The learning rate was reduced during training using the polynomial learning-rate decay schedule.

2.6 Evaluation Metrics

We evaluated segmentation performance using the three metrics specified by the AIMS-TBI challenge: the Dice similarity coefficient, the 95th-percentile Hausdorff distance (HD95), and the average symmetric surface distance (ASSD). The Dice score measures volumetric overlap, whereas HD95 and ASSD quantify surface discrepancies between the predicted and reference lesion masks.

To assess lesion-level performance across lesion sizes, the reference and predicted masks were decomposed into individual lesion instances. Ground-truth

and predicted instances were matched when their intersection-over-union (IoU) was at least 0.20. Lesion-wise recall, precision, and F1 were computed from matched and unmatched instances, while Matched Dice was calculated only for matched ground-truth–prediction pairs.

2.7 Implementation Details

All experiments were implemented within the nnU-Net v2 framework and conducted on a single NVIDIA RTX A6000 48GB. Preprocessing, network configuration, and inference follow the framework defaults described above. During inference, we apply overlapping sliding-window prediction using the training patch size with a window overlap of 0.5 along each spatial dimension. The final validation and test submissions were generated by ensembling the five fold-specific models.

3 Results

3.1 Baseline Configuration Study

To efficiently determine the baseline configuration, architecture, patch size, and loss-function experiments were conducted on the held-out fold-0 subset of the released training set. This internal fold-0 validation was used only for configuration selection, while the proposed sampling strategy was subsequently evaluated using five fold-specific models. Table 1 summarizes the fold-0 results.

Among the evaluated architectures, ResEnc-M achieved the highest Dice score of 0.598, outperforming both the standard nnU-Net and ResEnc-L. Although ResEnc-L yielded slightly lower HD95 and ASSD, its Dice score was substantially lower at 0.511. We therefore selected ResEnc-M as the baseline architecture.

Next, we investigated the effect of patch size. The 128×128×128 configuration achieved the best overall performance, with a Dice score of 0.603, an HD95 of 26.915, and an ASSD of 7.495. Reducing the patch size to 64×64×64 may provide insufficient anatomical context, whereas increasing it to $160 \times 192 \times 128$ may dilute the contribution of small lesion regions by including a greater proportion of background. We therefore adopted the 128×128×128 patch size for subsequent experiments.

Finally, we evaluated different training objectives. Adding a weakly weighted Focal term to the CE–Dice objective decreased Dice by 0.039 and increased HD95 by 3.2%, despite reducing ASSD by 13.4%. In contrast, the weighted objective $0.3L_{CE} + 0.4L_{Dice} + 0.3L_{Focal}$ reduced HD95 and ASSD by 11.6% and 41.6%, respectively, while increasing Dice by 0.007. Based on these fold-0 experiments, ResEnc-M with a $128 \times 128 \times 128$ patch size, a batch size of 8, and the weighted CE–Dice–Focal objective was fixed for the subsequent five-fold comparison of the sampling strategies.

Table 1. Internal fold-0 validation results from the training set for baseline configuration selection. Bold indicates the best result within each comparison block.

Configuration	Patch size	Batch	Loss	Dice↑	HD95↓	ASSD↓
nnU-Net	128×160×112	2	$L_{CE} + L_{Dice}$	0.592	30.432	8.114
ResEnc-M	128×160×112	2	$L_{CE} + L_{Dice}$	0.598	30.748	7.969
ResEnc-L	128×160×112	2	$L_{CE} + L_{Dice}$	0.511	29.914	7.607
ResEnc-M	64×64×64	2	$L_{CE} + L_{Dice}$	0.552	57.649	14.013
ResEnc-M	128×128×128	2	$L_{CE} + L_{Dice}$	0.603	26.915	7.495
ResEnc-M	160×192×128	2	$L_{CE} + L_{Dice}$	0.570	29.898	9.551
ResEnc-M	128×128×128	8	$L_{CE} + L_{Dice}$	0.588	28.859	9.489
ResEnc-M	128×128×128	8	$L_{CE} + L_{Dice} + 0.1L_{Focal}$	0.549	29.781	8.221
ResEnc-M	128×128×128	8	$0.3L_{CE} + 0.4L_{Dice} + 0.3L_{Focal}$	**0.595**	**25.501**	**5.540**

Table 2. Performance on the official AIMS-TBI validation and final-phase test sets. Validation results are reported as mean ± standard deviation across the five fold-specific models, while the final-phase result corresponds to the five-model ensemble submitted to the challenge leaderboard.

Sampling strategy	Dice↑	HD95↓	ASSD↓
Validation phase			
Conventional sampling	0.592 ± 0.012	22.864 ± 3.713	7.356 ± 1.994
Instance-aware sampling	$\mathbf{0.601 \pm 0.009}$	$\mathbf{21.731 \pm 3.668}$	$\mathbf{5.888 \pm 2.193}$
Final phase			
Instance-aware sampling	0.536	29.948	10.314

3.2 Effect of Instance-aware Sampling

We evaluated the proposed instance-aware sampling against conventional sampling under otherwise identical training conditions. Table 2 summarizes segmentation performance on both the official AIMS-TBI validation set and the final-phase test set. Evaluation was performed on the 55 lesion-containing cases among the 103 released validation cases, with results averaged across five fold-trained models. Instance-aware sampling improved Dice from 0.592 to 0.601, while reducing HD95 from 22.864 to 21.731 and ASSD from 7.356 to 5.888, corresponding to relative reductions of 4.96% and 19.96%, respectively. These results show improvements in both volumetric overlap and surface agreement. For the final challenge submission, the five fold-specific instance-aware models were ensembled, achieving a Dice score of 0.536, an HD95 of 29.948, and an ASSD of 10.314 on the official final-phase test set.

To examine whether increased small-lesion exposure translates into improved lesion-level performance, Table 3 reports results stratified by lesion volume on the same 55 lesion-containing validation cases. The validation set comprised 551 lesion instances, of which 480 were < 1 mL, 51 were 1–10 mL, 10 were 10–20 mL, and 10 were ≥ 20 mL. For lesions with volumes ≤ 1 mL, instance-aware sampling increased recall from 0.490 to 0.517, precision from 0.577 to 0.598, F1 from 0.530 to 0.555, and matched Dice from 0.654 to 0.667. For 1–10 mL lesions, recall, F1, and matched Dice increased slightly, whereas precision decreased from

Table 3. Lesion-wise performance stratified by lesion volume on the official AIMS-TBI validation set, averaged across five fold-trained models. Bold indicates the strictly better result; ties are not highlighted.

Sampling Strategy	Recall↑				Precision↑				F1↑				Matched Dice↑			
	≤ 1 (mL)	1–10 (mL)	10–20 (mL)	20≤ (mL)	≤ 1 (mL)	1–10 (mL)	10–20 (mL)	20≤ (mL)	≤ 1 (mL)	1–10 (mL)	10–20 (mL)	20≤ (mL)	≤ 1 (mL)	1–10 (mL)	10–20 (mL)	20≤ (mL)
Conventional	0.490	0.745	0.900	1.000	0.577	**0.919**	1.000	1.000	0.530	0.823	0.947	1.000	0.654	0.727	0.785	0.778
Instance-aware	**0.517**	**0.765**	0.900	1.000	**0.598**	0.897	1.000	1.000	**0.555**	**0.826**	0.947	1.000	**0.667**	**0.738**	**0.787**	**0.791**

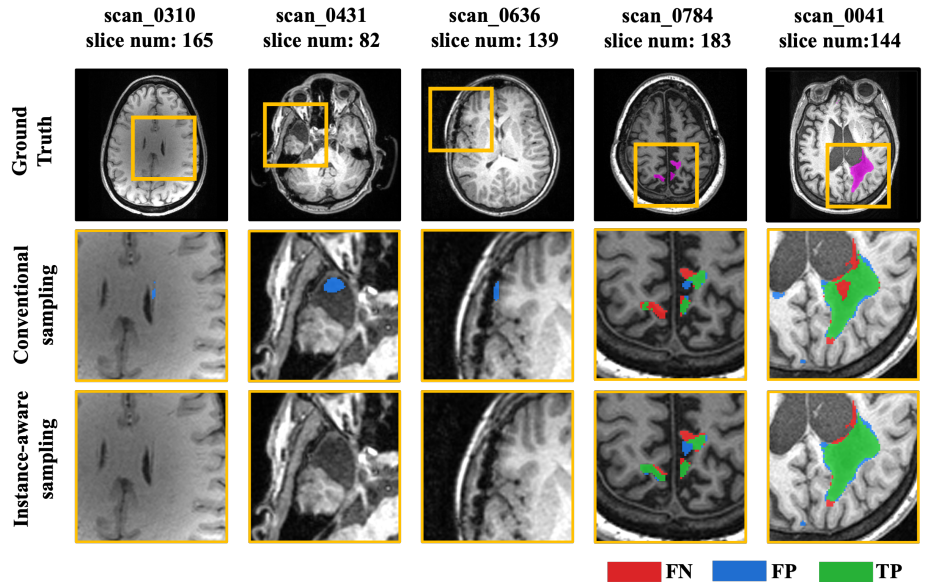

Fig. 3. Visualization of segmentation results between conventional sampling and instance-aware sampling across five different scans. Red, blue, and green indicate false-negative, false-positive, and true-positive regions, respectively.

0.919 to 0.897. For lesion groups of 10 mL or larger, detection metrics were largely unchanged, with only small differences in matched Dice. Overall, the most consistent improvements were observed in the smallest lesion group, where recall, precision, F1, and matched Dice all increased, whereas changes in larger lesion groups were smaller or mixed.

3.3 Qualitative Analysis

Fig. 3 presents qualitative comparisons between conventional foreground sampling and the proposed instance-aware sampling. In the first three cases, conventional sampling produced isolated false-positive predictions in lesion-free regions, whereas the proposed method reduced these spurious responses. In these examples, increased exposure to small lesion instances did not result in visibly

increased false-positive predictions. In the fourth case, both methods identified the general locations of the spatially disconnected lesion components. However, instance-aware sampling captured a larger portion of the reference lesions, increasing the true-positive overlap and reducing false-negative regions. In the fifth case, which contains a relatively large lesion, instance-aware sampling reduced both false-negative regions and peripheral false-positive predictions. Overall, these examples show reduced false-positive and false-negative regions with instance-aware sampling.

4 Conclusion

In this paper, we introduced an instance-aware foreground-sampling strategy within the native nnU-Net sampling pipeline for msTBI lesion segmentation. By anchoring foreground patches on lesion instances selected independently of their size, our approach increases the training exposure of small lesions. Experimental results on the AIMS-TBI 2026 validation set showed improved overall segmentation performance compared with conventional foreground sampling, with the most consistent lesion-wise gains observed for the smallest lesions. These findings highlight the importance of explicitly addressing lesion-level imbalance at the data-exposure stage during patch-based training. Future work will explore extending instance-aware sampling to multi-class lesion subtypes and incorporating auxiliary MRI modalities to further improve small-lesion segmentation.

References

1. Dennis, E.L., Baron, D., Bartnik-Olson, B., et al.: ENIGMA brain injury: Framework, challenges, and opportunities. Human Brain Mapping **43**(1), 149-166 (2022)
2. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203-211 (2021)
3. Wuts, J., Ceranka, J., Vandemeulebroucke, J., Lecouvet, F.: Instance-balanced patch sampling for whole-body lesion segmentation. In: Data Engineering in Medical Imaging (DEMI 2025), LNCS, **16191**, pp. 180–189. Springer, Cham (2026)
4. Kofler, F., Shit, S., Ezhov, I., Fidon, L., Horvath, I., Al-Maskari, R., Li, H., Bhatia, H., Loehr, T., Piraud, M., Erturk, A., Kirschke, J., Peeken, J.C., Vercauteren, T., Zimmer, C., Wiestler, B., Menze, B.: Blob loss: Instance imbalance aware loss functions for semantic segmentation. In: Information Processing in Medical Imaging (IPMI) 2023, LNCS, **13939**, pp. 755–767. Springer, Cham (2023)
5. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K.H., Jaeger, P.F.: nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI) 2024, LNCS, **15009**, pp. 488-498. Springer, Cham (2024)
6. Thompson, P.M., Jahanshad, N., Ching, C.R.K., et al.: ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. Translational Psychiatry **10**(1), 100 (2020)