

Size Penalty Loss

"What Dice Misses": Size-Stratified Volume Regularization for Ischemic Stroke Lesion Prognostication

Mohit Piyushkumar Joshi[1,2]
IAR/13136

Supervisor:
Dr Kshitij Jadhav[2] (MD, PhD)
Assistant Professor, Koita Centre for Digital Health,
IIT Bombay

Additional Supervisor:
Dr Sarfaraz Alam[1] (PhD)
Assistant Professor, SBB, IAR

[1]School of Biotechnology and Bioengineering, Institute of Advanced Research
[2]Koita Centre for Digital Health, Indian Institute of Technology Bombay

May, 2026

This dissertation is submitted for the degree of Bachelor of Technology

Work done during exchange semester & internship at IIT Bombay

Abstract

"What Dice Misses": Size-Stratified Volume Regularization for Ischemic Stroke Lesion Prognostication

Mohit Piyushkumar Joshi

Small ischemic stroke lesions are difficult targets for deep segmentation models because the positive class can occupy a tiny fraction of the image volume, and because voxel-overlap objectives can remain numerically acceptable while clinically important small lesions are missed. This thesis introduces **Size Penalty Loss**, a continuous size-stratified auxiliary loss for medical image segmentation. The proposed term penalizes relative error between the predicted soft lesion volume and the ground-truth lesion volume, with an exponential weight that gives stronger optimization pressure to smaller lesions. Unlike distribution-based losses such as binary cross entropy and focal loss, region-overlap losses such as Dice and Tversky, boundary losses, and instance-aware wrappers such as blob loss, the proposed loss operates as a sample-level volume regularizer conditioned directly on lesion size.

The loss was evaluated with a 3D Attention U-Net using ADC, DWI, and an ADC-DWI mismatch channel on two ischemic stroke MRI datasets: a proprietary de-identified cohort from Bharati Vidyapeeth Medical College and a public ISLES 2022 working subset. The proprietary evaluation used 5-fold cross-validation over fourteen loss configurations built from Dice, Tversky, Focal, and Size Penalty terms. ISLES 2022 was evaluated on a held-out split with the principal triple-loss configurations. Because small disconnected lesions can be missed while aggregate Dice remains competitive, the main evidence is interpreted through volume error and Panoptica-style instance metrics that decompose lesion recognition from matched-mask quality. The results show that **Size Penalty Loss** is not a standalone segmentation objective; by itself it has no localization signal and fails. When used as an auxiliary term with strong spatial losses, however, it improves the failure modes it was designed to target. On the proprietary cohort, **DTS** improves missed-lesion count, detection, RVE, RQ, PQ, and ASSD compared with DTF, with the clearest gain in sub-milliliter lesions. On ISLES 2022, **FSD** is the strongest size-penalty configuration and improves Dice, volume error, PQ, SQ, RQ, and boundary distance relative to DTF.

The central conclusion is that **Size Penalty Loss** defines a new class of size-stratified volume regularization. Its value is not that it replaces Dice, Tversky, or Focal loss, but that it adds a clinically interpretable pressure toward volume fidelity in the small-lesion regime where conventional voxel and overlap losses can under-represent failure.

Although the experiments are performed on ischemic stroke lesion segmentation, the method is not limited to stroke. The same objective is intended for minute-target segmentation tasks where target size is central, including multiple sclerosis lesions, glomeruli, microscopy structures, material micro-defects, remote-sensing targets, and autonomous-system small-object perception.

Code availability. The project repository and source code are available at https://github.com/mhtjsh/Size-Penalty-Loss-Function/.

Project Page. The project page and interactive visualizations are available at https://mhtjsh.github.io/Size-Penalty-Loss-Function/.

Acknowledgements

This work was supported by the **Indian Institute of Technology Bombay** prestigious Research Internship Awards 2025-26.

I would like to express my sincere gratitude to my supervisor, **Dr Kshitij Jadhav**, Assistant Professor at the Koita Centre for Digital Health, IIT Bombay, for his invaluable guidance, continuous support, and mentorship throughout this research. His expertise in medical imaging and deep learning has been instrumental in shaping this work.

I am also grateful to lab members at AIDE Lab, **Viraj Singh Gaur, Bhavik Kanekar (penultimate year PhD student) and all the members at the office** for supporting me throughout my research work.

I am equally grateful to my additional supervisor, **Dr Sarfaraz Alam**, Assistant Professor at the School of Biotechnology and Bioengineering, Institute of Advanced Research, for his encouragement and academic support.

I am equally grateful to, **Dr Dhara Patel (HOD, SBB, IAR)**, Associate Professor at the School of Biotechnology and Bioengineering, Institute of Advanced Research, for her encouragement and academic support which helped me pursue my time at IIT Bombay.

I extend my heartfelt thanks to the **Koita Centre for Digital Health, IIT Bombay** for providing the computational resources and research environment that made this work possible.

Finally, Thank you to my dad for being my role model, my mom for being my rock and all of my conciousness and experiences that shaped me into who I am. I also want to acknowledge all my ride or dies, each of whom has brought something unique and joyful to my academic and personal journey.

–MoJo

Contents

List of Figures

List of Tables

Chapter 1

Small Lesions, Large Optimization Cost

1.1 Why Small Stroke Lesions Break Overlap Objectives

Ischemic stroke lesion segmentation is a small-target 3D medical imaging problem. In diffusion-weighted MRI, clinically relevant infarcts can range from large territorial lesions to punctate embolic lesions that occupy only a few voxels. This creates two linked optimization problems. First, the foreground class is sparse relative to brain background, so voxel-wise losses can be dominated by easy non-lesion voxels. Second, average overlap metrics can hide lesion-wise failures: one well-segmented large lesion can compensate numerically for several missed small lesions.

Deep segmentation systems usually address this imbalance with distribution-based, region-based, boundary-based, or compound losses [1–3]. These families are useful, but they do not explicitly ask whether the predicted lesion volume is correct as a function of lesion size. **Dice** loss optimizes overlap [4]. Tversky loss changes the false-positive and false-negative tradeoff [5]. Focal loss emphasizes hard examples by down-weighting easy predictions [6]. Boundary losses focus on contours and surfaces [7, 8]. Instance-aware losses such as blob loss address instance imbalance by averaging a base loss over connected components [9]. None of these mechanisms directly implements the prior that a fixed relative volume error should matter more for a small lesion than for a large lesion.

1.2 A Size-Conditioned Volume Signal

This thesis proposes **Size Penalty Loss**, a continuous size-stratified volume regularizer for segmentation. For each lesion-positive sample, the loss computes the relative mismatch between predicted soft foreground volume and true foreground volume,

then weights this error by an exponential function of ground-truth lesion size. The key term is

$$\exp(-V_g/\tau_s)\frac{|V_p - V_g|}{V_g + \varepsilon},$$

where V_g is the ground-truth lesion volume in voxels, V_p is the predicted soft lesion volume, τ_s is a dataset-calibrated size scale, and ε is a numerical stabilizer. The exponential factor is high for small lesions and decays for larger lesions. The loss therefore gives a smooth size curriculum without a hard threshold.

The contributions are:

1. A mathematical definition of a continuous size-stratified volume penalty for segmentation.
2. A taxonomy argument that places the proposed loss outside standard distribution, region, boundary, and instance-wrapper families.
3. A full ischemic stroke segmentation evaluation on a proprietary 5-fold cohort and a public ISLES 2022 split.
4. A size-stratified analysis showing that the benefit is strongest in small-lesion and volume-sensitive regimes.
5. A discussion of when the loss should be used as an auxiliary term and why future gated variants should reduce its weight for larger lesions.

1.3 Evidence at a Glance

The strongest proprietary configuration is Dice+Tversky+Size Penalty, abbreviated **DTS**. Compared with the Dice+Tversky+Focal baseline **DTF**, **DTS** has slightly lower mean **Dice** but better missed-lesion count, detection, **RVE**, **RQ**, **PQ**, and **ASSD**. The most direct validation appears in sub-milliliter proprietary lesions, where **DTS** improves **PQ**, **SQ**, **RQ**, **Dice**, and volume error against **DTF**.

On ISLES 2022, the strongest size-penalty configuration is Focal+Size Penalty+Dice, abbreviated **FSD**. It improves **Dice**, **RVE**, **PQ**, **SQ**, **RQ**, and **ASSD** compared with **DTF**. The ISLES result also shows an important limitation: **DTF** retains a better missed-lesion count on that single split. The correct claim is therefore not universal dominance. The claim is that **Size Penalty Loss** is a useful auxiliary regularizer for size-sensitive segmentation, especially when the evaluation includes volume and lesion-wise metrics rather than **Dice** alone.

Chapter 2

From Overlap Losses to Size-Aware Regularization

2.1 Where Current Segmentation Losses Help and Stop

Recent surveys organize segmentation losses into distribution-based, region-based, boundary-based, and compound families [1, 2]. Distribution-based losses, including binary cross entropy and focal loss, optimize voxel-wise classification. Region-based losses, including Dice and Tversky, optimize overlap or overlap-derived tradeoffs. Boundary-based losses use surface, contour, or distance information. Compound losses combine losses to capture multiple error modes. The large-scale analysis by Ma et al. [1] is important for this thesis because it shows that no single loss family is consistently best across medical segmentation tasks, while compound objectives are often robust.

This taxonomy leaves a gap for losses whose defining variable is the physical or voxel size of the target object. **Size Penalty Loss** is motivated by this gap. It is not a distribution loss because it does not score each voxel as an independent label. It is not an overlap loss because it does not use an intersection term. It is not a boundary loss because it does not use contour or distance information. It is also not an instance-wrapper loss because it does not decompose the mask into connected components. It is a size-conditioned sample-level volume regularizer.

2.2 Voxel Difficulty and Region Overlap

Binary cross entropy treats segmentation as voxel-wise Bernoulli classification. It is simple and stable, but class imbalance can make the background term dominate. Focal

loss was introduced to address dense detection imbalance by down-weighting easy examples and focusing training on hard examples [6]. In segmentation, focal-style losses can help when lesion voxels are rare, but the modulation is based on prediction difficulty rather than lesion size.

Dice loss directly optimizes overlap and is widely used for imbalanced medical segmentation [4]. Generalized Dice variants further address class imbalance by weighting classes [10]. Tversky loss generalizes Dice by assigning different weights to false positives and false negatives [5]. This is relevant for stroke segmentation because false negatives are often more costly than false positives. Focal Tversky further applies focal modulation to the Tversky index and has been used for small lesion segmentation [11]. These methods remain overlap-centered. They can change the precision-recall balance, but they do not explicitly apply a continuous size prior to the volume error.

2.3 Contours, Components, and Their Limits

Boundary losses were proposed for highly imbalanced segmentation because regional integrals can be unstable when foreground and background sizes differ by orders of magnitude [7]. Boundary Difference over Union similarly focuses on boundary-region mismatch and adapts attention using target geometry [8]. These losses address contour quality and boundary alignment. For minute lesions, however, a boundary signal can become unstable because a one-voxel erosion or dilation can strongly change the contour while the lesion mass remains clinically important. Boundary losses do not directly penalize absolute or relative volume mismatch.

Blob loss targets a different limitation of Dice: instance imbalance within the foreground class [9]. It applies a base loss to each connected component and averages over instances so that large foreground objects do not dominate smaller ones. This is related to the clinical goal of detecting many small lesions, but the mechanism is different. Blob loss equalizes instances after connected-component decomposition. **Size Penalty Loss** directly reweights relative volume error as a continuous function of lesion size. It does not require instance labels or connected-component decomposition during training.

2.4 From Size Bounds to Volume Fidelity

The most closely related size-based work is the constrained-CNN loss for weakly supervised segmentation [12]. That method imposes lower and upper bounds on the predicted foreground size and adds a penalty only when the prediction violates the

feasible interval. This is an interval-satisfaction constraint. It is designed for weak supervision with partial labels and external size bounds.

Size Penalty Loss is different in three ways. First, it uses the exact ground-truth lesion volume during supervised training and minimizes deviation from that value rather than from a broad interval. Second, the penalty is a weighted relative volume regression term rather than a quadratic feasibility constraint. Third, the weighting is explicitly size-stratified through $\exp(-V_g/\tau_s)$, which gives stronger emphasis to smaller lesions. This makes the proposed term a volume-fidelity regularizer rather than a weak-supervision constraint.

2.5 Metrics That Expose Small-Lesion Failure

2.5.1 Why Dice Alone Is Insufficient

Dice remains necessary because it is the standard global overlap check for lesion masks. It is insufficient as the only primary lens because it aggregates all foreground voxels into one score. In fragmented stroke, a well-segmented large lesion can keep aggregate Dice competitive while separate small lesions are missed, false-positive lesion islands are added, or the predicted lesion volume is biased. These are different failure modes: missed lesions and false positives are recognition errors, while poor overlap among already matched lesions is a mask-quality error.

This thesis therefore reports Dice alongside RVE, missed lesions, false positives, lesion F1, volume-weighted Dice, concordance correlation coefficient, ASSD, and Panoptica-style instance metrics. This follows the broader validation principle that biomedical segmentation metrics should be chosen to match the domain question rather than selected as a single default score [13, 14].

2.5.2 Panoptic Decomposition

Panoptic quality was introduced to combine semantic segmentation quality with instance recognition in one interpretable score [15]. Panoptica adapts instance-wise evaluation to 2D and 3D biomedical segmentation maps and can complement the original overlap-based panoptic formulation with alternative matched-instance similarities [16]. For stroke lesion evaluation, let R denote a reference lesion instance, P denote a predicted lesion instance, and $f(R, P)$ denote the similarity score assigned to a matched

reference-prediction pair. Then

$$\mathrm{PQ} = \frac{\sum_{(R,P)\in TP} f(R,P)}{|TP| + 0.5|FP| + 0.5|FN|} = \underbrace{\frac{\sum_{(R,P)\in TP} f(R,P)}{|TP|}}_{\mathrm{SQ}} \cdot \underbrace{\frac{|TP|}{|TP| + 0.5|FP| + 0.5|FN|}}_{\mathrm{RQ}}.$$

SQ measures the quality of masks that were already matched; **RQ** measures whether lesion instances were recognized at all; **PQ** combines the two. This decomposition is the methodological reason for treating **Dice** as necessary but not sufficient in the small, disconnected-lesion setting.

Chapter 3

Size Penalty Loss: A Size-Aware Volume Regularizer

3.1 Volumes, Masks, and Soft Predictions

Let $z_j^{(i)}$ denote the model logit at voxel j for sample i, and let

$$p_j^{(i)} = \sigma(z_j^{(i)})$$

be the predicted soft foreground probability after the sigmoid. Let $y_j^{(i)} \in \{0, 1\}$ be the binary ground-truth lesion mask. The ground-truth and predicted soft lesion volumes are

$$V_g^{(i)} = \sum_{j=1}^{N} y_j^{(i)}, \qquad V_p^{(i)} = \sum_{j=1}^{N} p_j^{(i)}.$$

Both are expressed in voxels during training. Physical volume in milliliters is obtained by multiplying voxel count by the dataset-specific voxel volume.

3.2 The Size-Weighted Relative Volume Error

For lesion-positive samples, the proposed **Size Penalty Loss** is

$$\ell_{\text{size}}^{(i)} = \mathbf{1}\left[V_g^{(i)} > 0\right] \min\left(c, \exp\left(-\frac{V_g^{(i)}}{\tau_s}\right) \frac{\left|V_p^{(i)} - V_g^{(i)}\right|}{V_g^{(i)} + \varepsilon}\right),$$

where $\tau_s = 600$ voxels in the experiments, $\varepsilon = 10^{-5}$, and $c = 10$ caps extreme penalty values. For a mini-batch $\mathcal{B}$, only lesion-positive samples contribute:

$$\mathcal{L}_{\text{size}} = \frac{1}{|\mathcal{B}_+|} \sum_{i \in \mathcal{B}_+} \ell_{\text{size}}^{(i)}, \qquad \mathcal{B}_+ = \{i \in \mathcal{B} : V_g^{(i)} > 0\}.$$

The empty-mask exclusion is intentional. If $V_g = 0$, a relative volume error term would be ill-conditioned and could create unstable gradients. Empty samples remain supervised by the spatial losses.

3.3 Implementation as a Differentiable Auxiliary Term

Callout: PyTorch implementation of the Size Penalty term

```
class ContSizePenaltyLoss(nn.Module):
    """
    Continuous size-penalty loss (v5) - no hard thresholds.

    L_size = exp(-V_g / tau_s) * |V_p - V_g| / (V_g + eps)

      V_g  = GT lesion volume (voxel count)
      V_p  = predicted volume (soft sum of sigmoid outputs)
      tau  = SIZE_TAU (600) - characteristic decay scale

    The loss is fully differentiable before the clamp and only fires
    for samples that have a lesion.
    """
    def __init__(self, tau: float = 600.0, smooth: float = 1e-5):
        super().__init__()
        self.tau = tau
        self.smooth = smooth

    def forward(self, pred_logits: torch.Tensor,
                binary_mask: torch.Tensor) -> torch.Tensor:
        p = torch.sigmoid(pred_logits)
        eps = self.smooth

        V_g = binary_mask.sum(dim=(1, 2, 3, 4))   # (B,)
        V_p = p.sum(dim=(1, 2, 3, 4))             # (B,)

        # Only penalize samples that have a lesion.
        # Empty masks remain handled by the spatial loss terms.
        has_lesion = (V_g > 0).float()            # (B,)
```

```
        # Continuous exponential weighting: smooth decay, no threshold.
        w = torch.exp(-V_g / self.tau)

        # Relative volume error.
        rel_error = torch.abs(V_p - V_g) / (V_g + eps)

        # Hard cap for extreme errors; saturated samples have zero
        # gradient through this auxiliary term.
        per_sample = torch.clamp(w * rel_error, max=10.0) * has_lesion

        n_with_lesion = has_lesion.sum().clamp(min=1.0)
        return per_sample.sum() / n_with_lesion
```

3.4 Exponential Weighting as a Small-Lesion Curriculum

The defining function is

$$w(V_g) = \exp(-V_g/\tau_s).$$

It is monotone decreasing in lesion size. For two samples with the same relative volume error, the smaller lesion receives the larger penalty. At $V_g = \tau_s$, the weight is $e^{-1} \approx 0.368$. With $\tau_s = 600$ voxels, the proprietary voxel volume of 0.0048 mL per voxel gives $\tau_s \approx 2.9060$ mL. Under an approximate 2.0000 mm $\times$ 2.0000 mm $\times$ 2.0000 mm ISLES voxel assumption, $\tau_s \approx 4.8000$ mL.

3.5 Why Relative Error, Clamping, and Gradients Matter

The relative error term

$$E(V_p, V_g) = \frac{|V_p - V_g|}{V_g + \varepsilon}$$

is minimized only when the predicted soft lesion volume matches the ground-truth volume. Expressed as a ratio $r = V_p/V_g$, it is approximately $|r - 1|$, which explains the V-shaped behavior in Fig. 3.1. The size adaptivity does not come from this relative-error term by itself; it comes from multiplying the relative error by $w(V_g)$.

The clamp in Section 3.2 is a hard cap on the penalty:

$$\ell_{\text{size}} = \min(c, wE).$$

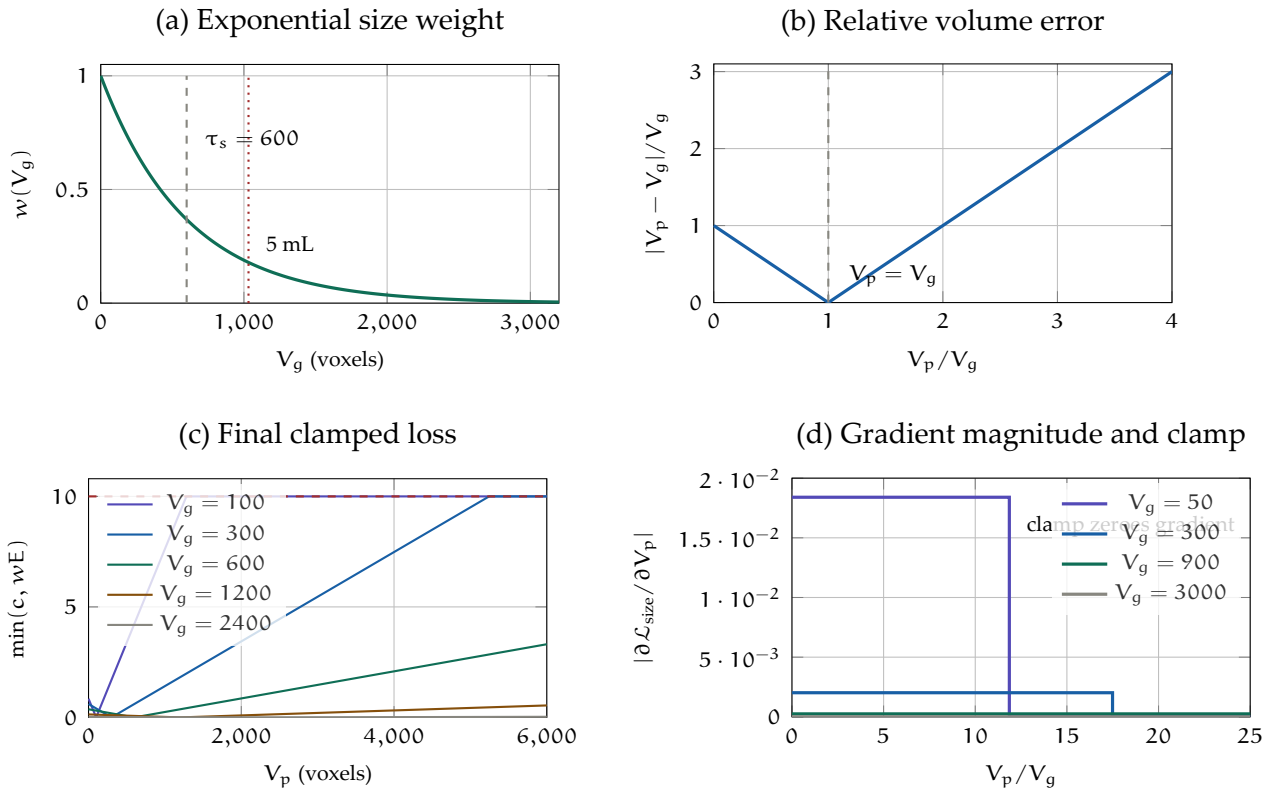


Figure 3.1: Component behavior of **Size Penalty Loss**. Panel (a) shows the exponential size weight. Panel (b) shows the relative volume error term, which is minimized when the predicted and ground-truth volumes match. Panel (c) shows the final clamped loss for different ground-truth lesion sizes. Panel (d) shows the pre-clamp gradient magnitude with respect to predicted volume and the point where the hard cap makes the gradient zero for extreme over-prediction.

Before the cap is reached and away from the exact equality point $V_p = V_g$, the derivative with respect to predicted volume is

$$\frac{\partial \ell_{\text{size}}}{\partial V_p} = \exp(-V_g/\tau_s)\frac{\text{sign}(V_p - V_g)}{V_g + \varepsilon}.$$

Thus smaller lesions receive larger volume-correction gradients. Once $wE \geqslant c$, the clamp saturates and the derivative through the clamp becomes zero. This prevents extreme over-segmentation cases from dominating the optimization. At exact volume equality, the usual subgradient of the absolute-error term is zero.

3.6 How the Size Term Partners With Spatial Losses

The size term has no localization signal. Any permutation of predicted probabilities that preserves the total predicted volume gives the same $\mathcal{L}_{\text{size}}$. This is not a weakness when the term is used correctly; it defines its role. **Size Penalty Loss** should be an auxiliary term paired with spatial losses:

$$\mathcal{L}_{\text{total}} = \lambda_D \mathcal{L}_{\text{Dice}} + \lambda_T \mathcal{L}_{\text{Tversky}} + \lambda_F \mathcal{L}_{\text{Focal}} + \lambda_S \mathcal{L}_{\text{size}},$$

with inactive terms omitted from each experiment. The experiments used $\lambda_D = \lambda_T = \lambda_F = 1.0$ and $\lambda_S = 0.3$ when combined with other losses. When S was trained alone, $\lambda_S = 1.0$.

3.7 Positioning Against Existing Loss Families

Table 3.1: Conceptual distinction between **Size Penalty Loss** and major loss families.

Loss family	Optimized quantity	Difference from **Size Penalty Loss**
Distribution-based	Voxel-wise label probability, as in BCE or focal loss	Reweights voxels by class probability or difficulty, not by lesion size.
Region-based	Foreground overlap, as in Dice or Tversky	Uses intersection and overlap counts, while Size Penalty uses only volume totals.
Boundary-based	Surface, contour, or distance-to-boundary agreement	Optimizes boundary alignment rather than direct lesion-volume fidelity.
Instance-aware	Per-connected-component base losses	Requires instance decomposition; Size Penalty does not.
Size-constrained weak supervision	Predicted size inside an allowed interval	Penalizes only outside a feasible interval; Size Penalty regresses toward exact ground-truth volume.
Proposed term	Size-weighted relative volume mismatch	Directly prioritizes smaller lesions through a continuous exponential weight.

Chapter 4

Experimental Pipeline for Stroke Lesion Segmentation

4.1 Stroke MRI Cohorts and Lesion Scale

Two ischemic stroke MRI datasets were used. The public dataset was a working subset of ISLES 2022, a multicenter, multi-vendor acute-to-subacute stroke lesion segmentation dataset [17]. The proprietary dataset was a de-identified MRI cohort from Bharati Vidyapeeth Medical College, Pune, developed over eight years and annotated by a board-certified consulting neurologist.

4.2 ADC-DWI Normalization and Mismatch Encoding

ADC and DWI volumes were normalized per volume by percentile clipping between the first and ninety-ninth percentiles, followed by min-max scaling to $[0, 1]$. A third mismatch channel was computed as

$$\text{Mismatch}(x) = \text{DWI}(x)\,[1 - \text{clip}(\text{ADC}(x), 0, 1)]\,.$$

This channel encodes the acute ischemia pattern of high DWI signal with low ADC signal. It also suppresses patterns such as T2 shine-through, where both DWI and ADC are high.

4.3 3D Attention U-Net Backbone

All experiments used a 3D Attention U-Net style model with attention gates, ASPP bottleneck, and dropout. U-Net provides the encoder-decoder structure and skip connections for biomedical segmentation [18]; 3D U-Net extends this idea to volumetric

Table 4.1: Dataset and input summary used in the experiments.

Property	ISLES 2022 working subset	Proprietary cohort
Cases	185 usable MRI volumes from the public training data	210 de-identified MRI volumes
Split	80/10/10 train/validation/test	80/10/10 within cross-validation folds
Input size	$112 \times 112 \times 72$	$256 \times 256 \times 25$, zero-padded to 32 slices
Slice thickness	2.0000 mm	6.0000 mm
Voxel volume used for volume conversion	Approximately 0.0080 mL under a 2.0000 mm $\times$ 2.0000 mm $\times$ 2.0000 mm assumption	0.0048 mL
Modalities	DWI, ADC, FLAIR available; DWI/ADC used in this model pipeline	ADC and DWI used in this model pipeline
Lesion distribution	Heterogeneous acute/subacute infarcts with multiple disconnected lesions and strong imbalance	Mean scan infarct volume 32.7990 mL, range 0.1310 mL to 370.4920 mL; mean disconnected ischemias 5.900 per scan

data [19]; Attention U-Net adds attention gates that suppress irrelevant regions and improve sensitivity to target structures [20]. The implemented network used three input channels, one output channel, base filters of 32, dropout of 0.20, attention gates enabled, ASPP enabled, and deep supervision disabled. The model contained approximately 14.54 million trainable parameters.

4.4 Loss Ablations and Optimization Protocol

The proprietary ablation used 5-fold evaluation summaries for fourteen loss combinations built from **Dice** (D), Tversky (T), Focal (F), and Size Penalty (S). The training description used a 10-fold splitting rule during training, with sequential CSV order preserved and identical folds across loss combinations. The principal reported proprietary analysis aggregates five evaluated test folds. ISLES 2022 was evaluated on one held-out split and therefore supports external replication, not fold-level consistency claims.

Training patches used large patches of $32 \times 128 \times 128$ and small patches of $32 \times 64 \times 64$, with multi-resolution small-patch sampling probability 0.2. Positive sampling was used to address sparse lesions: the large-patch positive ratio was 0.7 and the small-

patch positive ratio was 1.0. Each training and validation volume contributed ten patches. Augmentations included spatial flips, depth flips, elastic deformation, background cutout, intensity scale/shift, noise, and gamma augmentation. The mismatch channel was recomputed from augmented ADC and DWI rather than augmented directly.

Optimization used AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, gradient clipping at 1.0, mixed precision, batch size 18, and a maximum of 400 epochs. AdamW was selected because decoupled weight decay is better matched to adaptive optimizers than classical L2 regularization [21]. OneCycleLR was used with `pct_start=0.2`, `div_factor=25`, `final_div_factor=1000`, and cosine annealing inside the cycle. One-cycle scheduling was chosen because sparse lesion segmentation benefits from an early exploratory phase followed by a low-learning-rate refinement phase [22]; cosine annealing and warm restart schedules provide a useful baseline for learning-rate decay [23], but monotonic early decay can become conservative before lesion-specific filters are fully shaped.

4.5 Sliding-Window Inference and Validation-Selected Thresholds

Inference used 3D sliding windows with Gaussian blending. ISLES used patch size $32 \times 112 \times 112$, while the proprietary cohort used $32 \times 128 \times 128$. Standard sweep models used 50% overlap, while continuous size-penalty models used 75% overlap. Eight-flip test-time augmentation averaged predictions across flips over depth, height, and width axes.

Probability thresholding used the candidate set $\{0.3, 0.4, 0.5, 0.6, 0.7\}$. For each fold or dataset split, the operating threshold was selected on the validation split by maximizing aggregate validation **Dice**, then that validation-selected threshold was applied once to the independent test split. Test-set metrics were not used to choose or tune the threshold. Connected components smaller than 5 voxels were removed using 3D 18-connectivity.

4.6 Overlap, Volume, Boundary, and Instance Metrics

The main reported metrics were **Dice**, mean false-positive lesion count, mean missed-lesion count, detection rate, **RVE**, **PQ**, **SQ**, **RQ**, and **ASSD**. Additional metrics included volume-weighted **Dice**, lesion F1, concordance correlation coefficient, and absolute volume error in milliliters.

For instance-level evaluation, each thresholded semantic lesion mask was converted into lesion instances by 3D connected-component labeling after the same small-component filtering described in Section 4.5. Let $\mathcal{R}$ be the set of reference lesion instances and $\mathcal{P}$ be the set of predicted lesion instances. Panoptica-style matching assigns predicted instances to reference instances, producing matched pairs TP, unmatched predicted instances FP, and unmatched reference instances FN. The false-positive and missed-lesion counts reported in the tables are derived from these unmatched instance sets.

PQ, **SQ**, and **RQ** were computed using Panoptica-style instance matching and evaluation [16], following the panoptic quality decomposition introduced by Kirillov et al. [15]. With R denoting a reference lesion instance, P denoting a predicted lesion instance, and $f(R, P)$ denoting the matched-instance similarity score,

$$\mathrm{PQ} = \frac{\sum_{(R,P)\in \mathrm{TP}} f(R,P)}{|\mathrm{TP}| + 0.5|\mathrm{FP}| + 0.5|\mathrm{FN}|} = \underbrace{\frac{\sum_{(R,P)\in \mathrm{TP}} f(R,P)}{|\mathrm{TP}|}}_{\mathrm{SQ}} \cdot \underbrace{\frac{|\mathrm{TP}|}{|\mathrm{TP}| + 0.5|\mathrm{FP}| + 0.5|\mathrm{FN}|}}_{\mathrm{RQ}}.$$

Thus **SQ** summarizes mask quality among matched lesion instances, while **RQ** summarizes lesion-instance recognition after penalizing false positives and missed lesions.

4.7 Reproducibility Companion

The code repository for **Size Penalty Loss** is available at https://github.com/mhtjsh/Size-Penalty-Loss-Function [24]. The repository should be treated as the implementation companion to the mathematical definition and the PyTorch class shown in Chapter 3. The project page for graphs and behaviour of the loss function is available at https://mhtjsh.github.io/Size-Penalty-Loss-Function/.

Chapter 5

Evidence for Size Penalty Loss as Volume Regularization

5.1 Proprietary Cohort: Recognition and Volume Trade-offs

Table 5.1: Proprietary 5-fold mean performance across all loss combinations. Lower is better for FP, missed lesions, RVE, and ASSD. Higher is better for Dice, detection, PQ, SQ, and RQ.

Loss	Dice	FP	Missed	Det.	RVE	PQ	SQ	RQ	ASSD
D	0.7277	1.59	6.28	0.5135	0.2726	0.2972	0.6589	0.4001	4.0142
D+T	0.7155	1.96	6.32	0.5375	0.2936	0.2921	0.6352	0.3966	4.2491
DTF	0.7315	1.93	6.15	0.5293	0.2777	0.2903	0.6552	0.3873	3.9814
DTS	0.7268	2.07	6.03	0.5368	0.2745	0.2940	0.6424	0.4022	3.5665
F	0.7144	2.41	6.38	0.5227	0.2778	0.2590	0.6110	0.3429	4.7535
F+D	0.7229	1.62	6.16	0.5309	0.3573	0.2883	0.6584	0.3813	3.6349
F+S	0.6290	1.89	7.59	0.4389	0.3642	0.1607	0.4838	0.2330	3.3632
FSD	0.7204	1.79	6.24	0.5359	0.3095	0.2763	0.6407	0.3730	3.9299
S	0.0067	112.97	3.20	0.7823	3441.0637	0.0000	0.0055	0.0000	63.3154
S+D	0.7310	1.75	6.26	0.5234	0.3014	0.2958	0.6583	0.3916	3.7925
S+T	0.7334	1.99	5.97	0.5488	0.2954	0.2873	0.6501	0.3923	3.3583
T	0.7142	2.16	5.85	0.5576	0.3677	0.2754	0.6398	0.3755	3.8228
T+F	0.7351	1.60	5.94	0.5544	0.2973	0.2985	0.6578	0.4017	3.3165
TFS	0.7212	2.21	5.78	0.5463	0.3105	0.2740	0.6391	0.3705	4.2837

Raw Dice ranks T+F first, S+T second, and DTF third. This ranking is useful but incomplete. The clinically relevant comparison for the proposed loss is **DTS** against DTF, since DTF is the strongest non-size triple baseline. **DTS** gives a small Dice de-

crease, from 0.7315 to 0.7268, but improves missed lesions, detection, RVE, PQ, RQ, and ASSD. The panoptic improvement is mainly recognition-side: RQ increases from 0.3873 to 0.4022, while SQ decreases from 0.6552 to 0.6424. The claim is therefore better lesion recognition and volume-sensitive behavior, not universally better matched-mask quality. The main tradeoff is a higher false-positive count and lower SQ.

Table 5.2: Core proprietary triple-combo comparison. **DTS** improves recognition-side and volume-sensitive metrics but not raw Dice or false positives.

Metric	DTF	DTS	Direction for DTS
Dice	0.7315	0.7268	−0.0047
Mean FP	1.93	2.07	+0.14
Mean missed	6.15	6.03	−0.12
Detection	0.5293	0.5368	+0.0075
RVE	0.2777	0.2745	−0.0032
PQ	0.2903	0.2940	+0.0037
SQ	0.6552	0.6424	−0.0128
RQ	0.3873	0.4022	+0.0149
ASSD	3.9814	3.5665	−0.4149

5.2 ISLES 2022: Public-Dataset Generalization

Table 5.3: ISLES 2022 held-out split results. **FSD** is the best size-penalty configuration on this public dataset.

Loss	Dice	FP	Missed	Det.	RVE	PQ	SQ	RQ	ASSD	Epochs
DTF	0.7503	0.84	6.16	0.6839	0.2618	0.3486	0.6569	0.4712	1.6417	205
DTS	0.7514	0.42	7.42	0.6091	0.2757	0.3389	0.6596	0.4440	1.5659	271
FSD	0.7774	0.68	7.21	0.6475	0.2478	0.3882	0.6801	0.5074	1.4994	336
TFS	0.7472	0.79	7.21	0.6543	0.2562	0.3600	0.6791	0.4999	1.6486	216

On ISLES, **FSD** improves over DTF in Dice by 0.0271, RVE by 0.0140, PQ by 0.0396, SQ by 0.0232, RQ by 0.0362, and ASSD by 0.1423. All size-penalty combinations reduce false positives relative to DTF, with **DTS** reducing FP from 0.84 to 0.42. However, DTF has fewer missed lesions than the size-penalty combinations on this single split. The public-dataset conclusion is therefore that **FSD** improves segmentation, volume, panoptic, and false-positive behavior, while missed-lesion behavior remains a limitation.

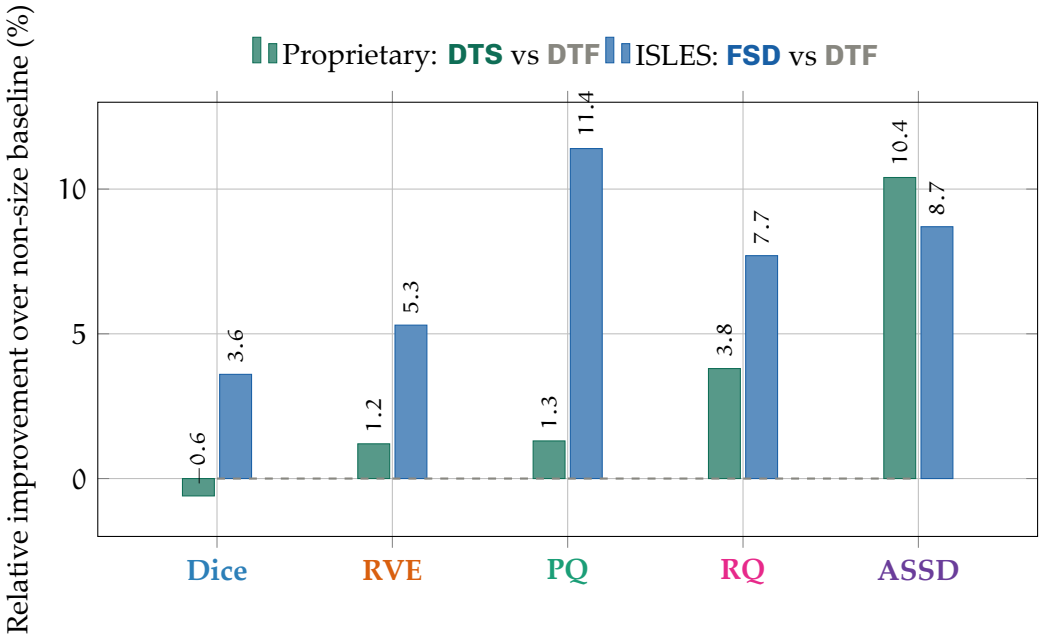


Figure 5.1: Compact relative-improvement view of the main result. Positive values mean better performance. For RVE and ASSD, improvement is plotted as relative reduction. The plot shows why raw Dice alone is insufficient: on the proprietary cohort, **DTS** slightly reduces Dice but improves volume-sensitive and recognition-sensitive metrics; on ISLES, **FSD** improves the full plotted metric set against DTF.

Table 5.4: Proprietary size-gated results for selected configurations. The < 5 mL bin is the target regime for **Size Penalty Loss**.

Loss	Bin	n	Dice	RVE	FP	Missed	PQ	SQ	RQ
DTF	< 5 mL	37	0.6316	0.3626	0.59	1.46	0.3857	0.5342	0.5207
DTS	< 5 mL	37	0.6301	0.3253	0.65	1.51	0.3807	0.5535	0.5329
S+T	< 5 mL	37	0.6476	0.3594	0.59	1.35	0.3917	0.5616	0.5491
T+F	< 5 mL	37	0.6455	0.3581	0.41	1.43	0.3943	0.5585	0.5422
DTF	$\geqslant 5$ mL	64	0.7890	0.2299	2.69	8.91	0.2348	0.7253	0.3099
DTS	$\geqslant 5$ mL	64	0.7828	0.2450	2.89	8.69	0.2431	0.6935	0.3259

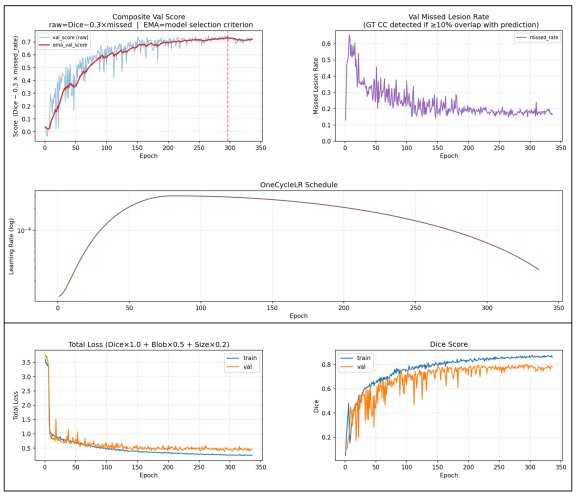
5.3 Where Size Penalty Loss Helps Most

For proprietary lesions below 5 mL, **DTS** reduces RVE from 0.3626 to 0.3253 relative to **DTF** and improves SQ from 0.5342 to 0.5535 and RQ from 0.5207 to 0.5329. Dice is essentially tied. The same table also shows why a future gate is justified: above 5 mL, **DTF** has better Dice and RVE, while **DTS** retains recognition-side advantages. This is consistent with the exponential size weight, whose influence is already reduced near and above 5 mL.

Table 5.5: Sub-milliliter proprietary lesion results for triple-loss combinations. **DTS** is the strongest result in the smallest lesion bin.

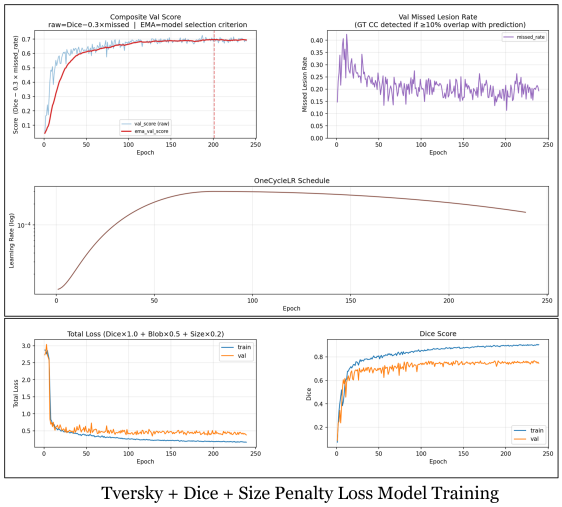
Loss	n	PQ	SQ	RQ	Dice
DTF	17	0.2885	0.4140	0.4033	0.4948
DTS	17	0.2961	0.4355	0.4286	0.5119
TFS	17	0.2501	0.3669	0.3549	0.4919
FSD	17	0.2255	0.3474	0.3034	0.4428

5.4 Training Dynamics and Case-Level Behavior



Focal + Dice + Size Penalty Loss Model Training

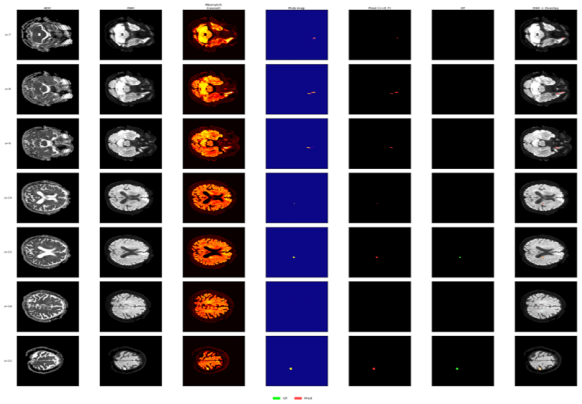
(a) ISLES **FSD** training summary.



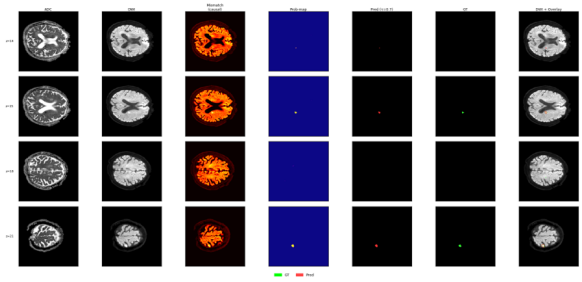
Tversky + Dice + Size Penalty Loss Model Training

(b) Proprietary **DTS** training summary.

Figure 5.2: Training curves for the best size-penalty configurations on the two datasets.

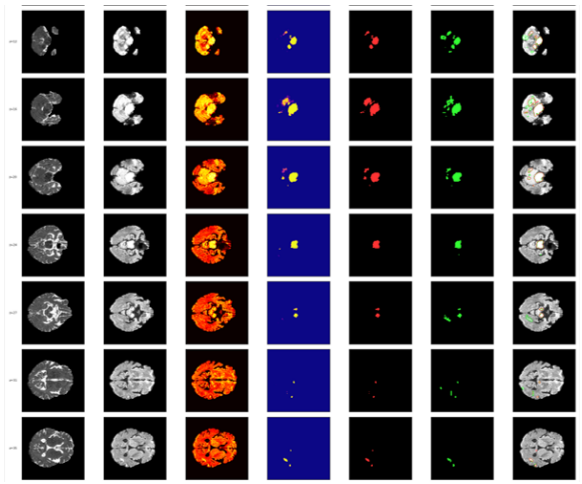


(a) **DTF**: GT 0.63 mL, predicted 1.49 mL, **Dice** 0.5434.

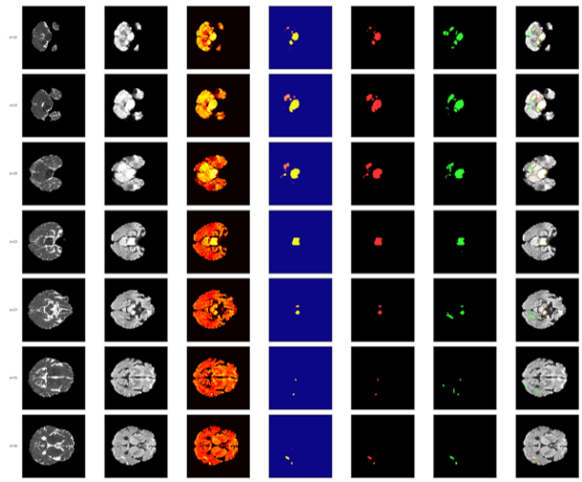


(b) **DTS**: GT 0.63 mL, predicted 0.73 mL, **Dice** 0.7330.

Figure 5.3: Proprietary case comparison on the same sub-milliliter lesion. Size Penalty improves volume fidelity and overlap for this case.



(a) **DTF**: GT 24.33 mL, predicted 22.98 mL, **Dice** 0.8287.



(b) **FSD**: GT 24.33 mL, predicted 23.84 mL, **Dice** 0.8412.

Figure 5.4: ISLES case comparison. **FSD** improves **Dice** on this case despite a larger absolute volume difference than **DTF**, showing why case-level visualization should accompany aggregate metrics.

5.5 Claims Supported by the Evidence

The evidence supports five main results. First, **Size Penalty Loss** fails as a standalone loss because it has no localization signal. Second, it works as an auxiliary regularizer when paired with spatial losses. Third, **DTS** is the most defensible proprietary size-penalty configuration because it improves missed lesions, detection, **RVE**, **PQ**, **RQ**, and **ASSD** relative to **DTF**, with the strongest direct evidence in sub-milliliter lesions. Fourth, **FSD** is the best ISLES size-penalty configuration. Fifth, a gated or reduced-weight version is justified for larger lesions, but it should be presented as future work unless directly implemented and evaluated.

Chapter 6

Interpreting the Size Penalty Loss Signal

6.1 What the Volume Signal Adds

Size Penalty Loss adds a scalar volume-fidelity signal that conventional segmentation losses do not provide directly. Dice and Tversky care about overlap. Focal loss cares about hard examples. Boundary losses care about contour alignment. Blob loss cares about instance imbalance. The proposed loss cares about whether the total predicted lesion mass matches the true lesion mass, and it makes that question more important when the lesion is small.

This distinction is the main novelty. The loss is not proposed as a better Dice loss. It is orthogonal to Dice because it is permutation-invariant over voxels. Two predictions with identical soft volume receive the same size loss even if one is spatially correct and the other is not. For that reason, **Size Penalty Loss** must be paired with spatial losses. The experiments confirm this: S alone collapses, while **DTS** and **FSD** are competitive or improved in the settings where size and instance metrics matter.

6.2 Why the Best Loss Partner Depends on the Dataset

The best size-penalty partner is dataset-dependent. In the proprietary cohort, **DTS** is the most defensible thesis configuration. It is not the raw-Dice winner, but it improves recognition-side and volume-sensitive metrics over DTF. This matters because the proprietary cohort contains a wide lesion-size range, sub-milliliter lesions, and multiple disconnected ischemias per scan. The size term appears to help most when the model already has strong spatial overlap and false-negative pressure from Dice and Tversky.

In ISLES 2022, **FSD** is the cleanest result. ISLES is multicenter and scanner-heterogeneous [17]. Focal loss can help with hard examples in this setting, while Dice

provides overlap gradients and **Size Penalty Loss** adds volume control. The result is a model that improves Dice, RVE, PQ, SQ, RQ, and ASSD over DTF. The limitation is that DTF still has fewer missed lesions on the single ISLES split, so the public-dataset claim must remain specific to overall segmentation and panoptic quality rather than universal lesion detection.

6.3 Reading PQ Through Recognition and Segmentation Quality

The reviewer-facing metric argument is not that Dice is wrong. Dice is necessary for confirming global voxel overlap, but it is not sufficient for the clinical question in small, fragmented stroke lesions. That question has two parts: whether separate lesion instances are recognized, and how good the masks are once they have been matched. PQ is useful because it keeps those parts connected while RQ and SQ expose them separately.

The proprietary gain is mostly recognition-side. **DTS** improves RQ more clearly than SQ, and the small PQ gain is driven by better lesion instance recognition rather than better masks among already matched lesions. This is clinically meaningful because missing separate lesions changes interpretation even if aggregate Dice remains acceptable.

The SQ decrease is also important. It shows that the proposed loss can encourage the model to recognize or preserve more lesion instances without always improving matched-mask quality. This argues for reporting PQ, SQ, and RQ separately. Reporting only PQ would hide the mechanism. Reporting only Dice would miss the main benefit.

6.4 The 5 mL Gate as a Future Control Mechanism

The experimental results support a gating hypothesis. A 5 mL lesion corresponds to approximately 1032 proprietary voxels and approximately 625 ISLES voxels under the ISLES spacing approximation. With $\tau_s = 600$, the size weight at 5 mL is about 0.179 for the proprietary cohort and about 0.353 for ISLES. The loss is not zero above 5 mL, but its influence is reduced.

The proprietary results show why this matters. In the < 5 mL bin, **DTS** improves RVE and recognition-side panoptic metrics relative to DTF. In the $\geqslant 5$ mL bin, DTF is safer for Dice and RVE, while **DTS** retains some recognition advantages. A future gated version could estimate lesion scale from a coarse prediction or soft foreground

mass, keep **Size Penalty Loss** active below 5 mL, and reduce λ_S above 5 mL. This thesis should describe that as a future mechanism, not as a completed trained gate.

6.5 Beyond Stroke: Minute-Target Segmentation

The experiments in this thesis validate **Size Penalty Loss** on ischemic stroke MRI, but the loss is not intrinsically tied to stroke. Its defining assumption is broader: when the target object is minute, sparse, and clinically or operationally important, global overlap metrics can understate the cost of missing or under-sizing that object. Any segmentation task with that structure is a candidate for size-stratified volume regularization.

In medical imaging, the most direct future applications are lesion and microstructure tasks where small regions carry high diagnostic value. Examples include multiple sclerosis plaque segmentation, glomeruli segmentation in renal pathology, cell and organelle segmentation in microscopy, small vessel or retinal lesion segmentation, micro-metastasis segmentation, and other pathology tasks where small target completeness matters more than the average foreground overlap. In these settings, the value of **Size Penalty Loss** would be tested by whether it improves **RVE**, instance recognition, and small-object recall without introducing unacceptable false positives.

The same principle can extend outside clinical imaging. Candidate non-medical domains include satellite arterial or fine-road-network imaging, material micro-defect segmentation, semiconductor inspection, remote-sensing small-object segmentation, and autonomous-system perception tasks such as small obstacle or small-object avoidance. These domains differ in scale and cost function, but they share the same failure mode: a small object can be practically important while contributing little to a global overlap objective.

The main research step is therefore not simply to reuse $\tau_s = 600$. Future work should estimate τ_s from each dataset's object-size distribution, report size-stratified metrics, and compare fixed, learned, and gated size-penalty schedules. A mature version of the method should also test whether the size term should be switched on only below a task-specific volume threshold, smoothly annealed with predicted object size, or combined with instance-aware penalties when object count is also clinically important.

6.6 Implementation Availability

The implementation repository for this project is available at https://github.com/mhtjsh/Size-Penalty-Loss-Function. The repository link is included to make the loss definition, training scripts, and experiment material easier to audit and extend. The

project page for graphs and behaviour of the **Size Penalty Loss** is available at https://mhtjsh.github.io/Size-Penalty-Loss-Function/.

6.7 Boundaries of the Current Evidence

The first limitation is statistical power. The proprietary evidence is based on five evaluated folds, while ISLES uses one held-out split. Directional consistency can be described, but strong statistical significance claims should not be made without additional testing.

The second limitation is false-positive behavior. **Size Penalty Loss** does not directly penalize disconnected false-positive islands. On ISLES, all size-penalty combinations reduce false positives compared with **DTF**. On the proprietary cohort, **DTS** increases false positives compared with **DTF**. This is not a contradiction; it shows that volume regularization, thresholding, and lesion morphology interact.

The third limitation is that $\tau_s = 600$ was not ablated. This value is defensible for the tested datasets and maps to clinically meaningful small-lesion scales, but it is not proven globally optimal.

The fourth limitation is calibration and threshold dependence. The reported metrics depend on the threshold sweep. Because size regularization changes predicted soft volume, it can shift the operating point. Future work should test calibration-aware thresholding and fixed-threshold deployment.

The fifth limitation is localization. **Size Penalty Loss** is intentionally global and volume-based. It should not be used alone, and it should not be described as boundary-aware or instance-aware. Its strength comes from complementing spatial losses, not replacing them.

Chapter 7

Conclusion: A New Axis for Loss Design

This thesis introduced **Size Penalty Loss**, a continuous size-stratified volume regularizer for medical image segmentation. The proposed loss penalizes relative mismatch between predicted and true lesion volume, weighted by an exponential function of lesion size. This creates a smooth optimization bias toward small lesions without introducing a hard threshold.

The experiments show that the loss defines a useful auxiliary objective, not a standalone segmentation loss. On the proprietary ischemic stroke cohort, **DTS** improves missed lesions, detection, RVE, RQ, PQ, and ASSD compared with **DTF**, with the clearest evidence in sub-milliliter lesions. On ISLES 2022, **FSD** is the strongest size-penalty configuration and improves Dice, volume error, panoptic metrics, and boundary distance relative to **DTF**. The results also establish important boundaries: **Size Penalty Loss** does not always reduce false positives, does not always improve SQ, and should not be claimed as a universal replacement for established compound losses.

The main contribution is therefore a new loss-function category: size-stratified volume regularization. It addresses a gap in the current taxonomy of segmentation losses by making lesion size an explicit conditioning variable in the objective. For stroke segmentation, this gives the model a direct pressure to preserve small lesion volume, which Dice-centered evaluation can understate. Future work should test the same principle beyond stroke in other minute-target segmentation problems, including multiple sclerosis lesions, glomeruli, microscopy structures, material microdefects, remote-sensing targets, and autonomous-system small-object perception. The strongest methodological next step is a gated version that keeps the size penalty active below a task-specific volume threshold and reduces its influence for larger targets.

Appendix A

Appendix: Abbreviations and Metric Formulae

A.1 Loss Combination Key

Table A.1: Loss abbreviations used throughout the thesis.

Symbol	Loss function
D	Dice loss
T	Tversky loss, $\alpha = 0.3$, $\beta = 0.7$
F	Focal loss, $\gamma = 2.0$, $\alpha = 0.25$
S	Continuous **Size Penalty Loss**, $\tau_s = 600$ voxels
DTF	Dice + Tversky + Focal
DTS	Dice + Tversky + Size Penalty
FSD	Focal + Size Penalty + Dice
TFS	Tversky + Focal + Size Penalty

A.2 Metric Formulae Used in Evaluation

Dice is reported as the standard soft or binary overlap coefficient after thresholding for inference:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|}.$$

RVE is

$$\text{RVE} = \frac{|V_P - V_G|}{V_G + \varepsilon}.$$

PQ follows

$$\mathrm{PQ} = \frac{\sum_{(R,P)\in TP} f(R,P)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} = \underbrace{\frac{\sum_{(R,P)\in TP} f(R,P)}{|TP|}}_{\mathrm{SQ}} \cdot \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\mathrm{RQ}},$$

where R is a reference lesion instance, P is a predicted lesion instance, $f(R,P)$ is the matched-instance similarity score, SQ measures the average quality of matched instances, and RQ measures instance recognition quality.

Bibliography

[1] 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. doi:10.1016/j.media.2021.102035. URL https://doi.org/10.1016/j.media.2021.102035.

[2] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7, 2020. doi:10.1109/CIBCB48159.2020.9277638. URL https://doi.org/10.1109/CIBCB48159.2020.9277638.

[3] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95:102026, 2022. doi:10.1016/j.compmedimag.2021.102026. URL https://doi.org/10.1016/j.compmedimag.2021.102026.

[4] 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)*, pages 565–571, 2016. doi:10.1109/3DV.2016.79. URL https://doi.org/10.1109/3DV.2016.79.

[5] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *Machine Learning in Medical Imaging*, pages 379–387. Springer, 2017. doi:10.1007/978-3-319-67389-9_44. URL https://doi.org/10.1007/978-3-319-67389-9_44.

[6] 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*, pages 2999–3007, 2017. doi:10.1109/ICCV.2017.324. URL https://doi.org/10.1109/ICCV.2017.324.

[7] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Éric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. *Medical*

Image Analysis, 67:101851, 2021. doi:10.1016/j.media.2020.101851. URL https://doi.org/10.1016/j.media.2020.101851.

[8] Fan Sun, Zhiming Luo, and Shaozi Li. Boundary difference over union loss for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 292–301. Springer, 2023. doi:10.1007/978-3-031-43901-8_28. URL https://doi.org/10.1007/978-3-031-43901-8_28.

[9] Florian Kofler, Suprosanna Shit, Ivan Ezhov, Lucas Fidon, Izabela Horvath, Rami Al-Maskari, Hongwei Bran Li, Harsharan Bhatia, Timo Loehr, Marie Piraud, Ali Ertürk, Jan Kirschke, Jan C. Peeken, Tom Vercauteren, Claus Zimmer, Benedikt Wiestler, and Bjoern Menze. blob loss: Instance imbalance aware loss functions for semantic segmentation. In *Information Processing in Medical Imaging*, pages 755–767. Springer, 2023. doi:10.1007/978-3-031-34048-2_58. URL https://doi.org/10.1007/978-3-031-34048-2_58.

[10] 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 *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248. Springer, 2017. doi:10.1007/978-3-319-67558-9_28. URL https://doi.org/10.1007/978-3-319-67558-9_28.

[11] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 683–687, 2019. doi:10.1109/ISBI.2019.8759329. URL https://doi.org/10.1109/ISBI.2019.8759329.

[12] Hoel Kervadec, Jose Dolz, Meng Tang, Éric Granger, Yuri Boykov, and Ismail Ben Ayed. Constrained-cnn losses for weakly supervised segmentation. *Medical Image Analysis*, 54:88–99, 2019. doi:10.1016/j.media.2019.02.009. URL https://doi.org/10.1016/j.media.2019.02.009.

[13] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, et al. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 21:195–212, 2024. doi:10.1038/s41592-023-02151-z. URL https://doi.org/10.1038/s41592-023-02151-z.

[14] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15:29, 2015. doi:10.1186/s12880-015-0068-x. URL https://doi.org/10.1186/s12880-015-0068-x.

[15] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. doi:10.1109/CVPR.2019.00963. URL https://doi.org/10.1109/CVPR.2019.00963.

[16] Florian Kofler, Hendrik Möller, Josef A. Buchner, Ezequiel de la Rosa, Ivan Ezhov, Marcel Rosier, Isra Mekki, Suprosanna Shit, Moritz Negwer, Rami Al-Maskari, Ali Ertürk, Shankeeth Vinayahalingam, Fabian Isensee, Sarthak Pati, Daniel Rueckert, Jan S. Kirschke, Stefan K. Ehrlich, Annika Reinke, Bjoern Menze, Benedikt Wiestler, and Marie Piraud. Panoptica – instance-wise evaluation of 3d semantic and instance segmentation maps, 2023. URL https://arxiv.org/abs/2312.02608.

[17] Moritz R. Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, David Robben, Alexander Hutton, Tassilo Friedrich, Teresa Zarth, Johannes Bürkle, The Anh Baran, Bjoern Menze, Gabriel Broocks, Lukas Meyer, Claus Zimmer, Tobias Boeckh-Behrens, Maria Berndt, Benno Ikenberg, Benedikt Wiestler, and Jan S. Kirschke. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific Data*, 9(1):762, 2022. doi:10.1038/s41597-022-01875-5. URL https://doi.org/10.1038/s41597-022-01875-5.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer, 2015. doi:10.1007/978-3-319-24574-4_28. URL https://doi.org/10.1007/978-3-319-24574-4_28.

[19] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432. Springer, 2016. doi:10.1007/978-3-319-46723-8_49. URL https://doi.org/10.1007/978-3-319-46723-8_49.

[20] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018. URL https://arxiv.org/abs/1804.03999.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[22] Leslie N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay, 2018. URL https://arxiv.org/abs/1803.09820.

[23] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

[24] Mohit Joshi. Size penalty loss function code repository. https://github.com/mhtjsh/Size-Penalty-Loss-Function, 2026. URL https://github.com/mhtjsh/Size-Penalty-Loss-Function.