A Unified Loss for Handling Inter-Class and Intra-Class Imbalance in Medical Image Segmentation

Feilong Xu[1,2], Feiyang Yang[1,2], Xiongfei Li[1,2], Xiaoli Zhang[1,2]*

[1]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
[2]College of Computer Science and Technology, Jilin University, China
{xufl23, yangfy21}@mails.jlu.edu.cn, {lxf, zhangxiaoli}@jlu.edu.cn

Abstract

In utilizing deep learning techniques for medical image segmentation, two types of imbalance issues are observed: inter-class imbalance between majority and minority classes and intra-class imbalance between easy and hard samples. However, existing loss functions typically confuse these issues, leading to enhancements that cater to only one aspect. Moreover, loss functions optimized for specific tasks often exhibit limited generalizability. To address these issues, we propose Inter-class and Intra-class Balance loss, as well as a unified loss termed Balance loss. The Inter-class Balance loss controls the extent of hard sample mining for majority class samples by considering the frequency of minority classes present in each input image. This approach requires no manual adjustment weights and adapts automatically to different datasets. The Intra-class Balance loss enhances the network's ability to learn from hard samples by performing mining on hard samples within each class. We evaluate our loss functions on five segmentation tasks with varying degrees of class imbalance. The experimental results show that our proposed Balance loss enhances segmentation performance compared with the current loss functions and exhibits superior robustness.

Introduction

Most machine learning algorithms assume that training samples are approximately uniformly distributed across classes (Haixiang et al. 2017; He and Garcia 2009). However, this assumption is often invalid in practical applications, particularly medical image segmentation. This discrepancy is primarily due to the prevalent issue of class imbalance in medical image datasets. Firstly, the regions of interest (ROIs) are relatively small in clinical data, with normal tissue occupying most of the image. For instance, in a widely used pancreas segmentation dataset (Roth et al. 2015), the ROIs are below 1%. Moreover, ROIs typically exhibit complex features such as irregular shapes, blurred boundaries, and heterogeneity, which make the accurate classification of edges more challenging than that of the interior (Li et al. 2023). Therefore, given the distinctive attributes of medical images, deep learning techniques for segmentation must address two types of imbalance issues: **i) Inter-class imbalance:** Inter-class imbalance refers to the difference in the number of samples between different classes. Because of the numerical superiority of the majority class samples, the model tends to learn and predict the majority class better during training, while frequently neglecting or misclassifying the minority class samples. **ii) Intra-class imbalance:** Intra-class imbalance refers to the difference in the characteristics of different regions or structures within the same class. This is reflected in the disparity between the number of easy and hard samples during training. Because the number of simple samples is larger, the model is often dominated by these during training. In contrast, hard samples may contain more noise, blur, or other complex features, which pose challenges for accurate classification.

Typical approaches to address class imbalance in medical image segmentation include re-sampling (including under-sampling and over-sampling) and re-weighting. Based on these approaches, various training strategies (Yan, Yang, and Cheng 2019), data augmentation techniques (Dai et al. 2022), and methods for designing loss functions (Abraham and Khan 2019) have been proposed. A prevalent under-sampling technique involves maintaining a manageable balance between foreground and background by randomly removing some background samples. Alternatively, a two-stage training strategy is employed (Yu et al. 2018; Zhang et al. 2018). In contrast, over-sampling aims to achieve inter-class balance by synthesizing or duplicating minority class samples (Hamghalam and Simpson 2024). In addition to specifically designed deep network architectures or training strategies, the loss function to be minimized during training plays a crucial role (Li et al. 2020). Weighted Cross Entropy (WCE) loss (Ronneberger, Fischer, and Brox 2015) and Tversky loss (Salehi, Erdogmus, and Gholipour 2017) address the inter-class imbalance by assigning higher weights to the minority class. For intra-class imbalance, the issue is typically addressed through hard example mining or re-weighting. For instance, Focal loss (Lin et al. 2017) and TopK loss (Wu, Shen, and Hengel 2016) force the network to focus on hard samples during training.

An optimal segmentation method should address inter-class and intra-class imbalances while maintaining robust generalization across diverse datasets and scenarios. This entails demonstrating high performance across region seg-

*Corresponding author

mentation tasks of various scales. While under-sampling can mitigate inter-class imbalance, the random deletion of majority class samples and using localization-based constraints can lead to the loss of valuable information. On the other hand, over-sampling, which involves duplicating minority class samples, may result in overfitting due to the model memorizing these duplicated samples. Re-weighted loss functions show effectiveness in addressing imbalance issues; however, their performance is influenced by hyper-parameter settings, which typically cannot be shared across various segmentation tasks. Furthermore, the lack of agreement on suitable loss functions is mainly attributed to varying degrees of class imbalance. Loss functions tailored for highly imbalanced tasks typically perform poorly on slightly imbalanced tasks. As our experiments will demonstrate, on the CVC-ClinicDB dataset (Bernal et al. 2015), the performance of most loss functions is even inferior to the classic cross entropy (CE) loss.

In this study, we propose the Inter-class Balance loss (Inter-CBL) and Intra-class Balance loss (Intra-CBL) based on hard sample mining technique. Additionally, we propose a unified loss function called the Balance loss. These methods mitigate both inter-class and intra-class imbalances while also demonstrating robust generalization capabilities, rendering them suitable for segmentation tasks with varying degrees of class imbalance. Specifically, first, we provide a detailed analysis of how inter-class imbalance affects segmentation performance. We identify that the hidden false positives (FP) and false negatives (FN) calculation bias between testing and training is the main reason for model performance degradation. Based on this conclusion, we propose Inter-CBL. It controls the extent of hard sample mining for majority class samples by considering the frequency of minority classes present in each input image. This approach requires no manual adjustment weights and adapts automatically to different datasets, ensuring consistently high performance in segmentation tasks with varying degrees of class imbalance. The Intra-CBL, on the other hand, is an improvement over TopK loss, enhancing the network's ability to learn from hard samples by performing mining on hard samples within each class. We conduct experimental evaluations on five segmentation tasks with different levels of class imbalance and compare our loss functions with currently popular ones. The results demonstrate that Balance loss achieves significant performance improvements across tasks with various degrees of imbalance, proving its effectiveness and practicality in the field of medical image segmentation. Additionally, Balance loss offers a versatile solution for region segmentation tasks of different scales due to its simple design and robust flexibility.

To summarize, our contributions are shown as follows:

- 1) We provide new insights into how class imbalance harms model performance, specifically the bias in FP and FN calculations between the training and testing scenarios.
- 2) Based on 1, we explain why the weights of many weighted loss functions are difficult to generalize.
- 3) We propose Inter-class and Intra-class Balance loss. We use a two-stage strategy and a linear weighting method to integrate Inter-CBL and Intra-CBL into the Balance loss. This unified loss explicitly controls both inter-class and intra-class balance.
- 4) We perform experiments on segmentation tasks with varying degrees of class imbalance, validating the effectiveness of our method and its robustness to class imbalance.

Related Work

Re-sampling Method

Re-sampling method primarily includes over-sampling and under-sampling (Garcea et al. 2023; Bali and Mahara 2023). **Over-sampling** aims to enhance the proportion of minority class samples by synthesizing or duplicating them (Elyan, Moreno-Garcia, and Jayne 2021). For instance, Soft-CP (Dai et al. 2022) and TumorCP (Yang et al. 2021) utilized copy-paste operations to duplicate ROIs, thereby creating new samples. While this technique addresses data scarcity, it does not introduce any new information to the dataset and may lead to overfitting. **Under-sampling** achieves data balance by reducing the number of majority class samples (Xie et al. 2021; Prusa et al. 2015). Two-stage segmentation approach known as coarse-to-fine segmentation is utilized as an under-sampling technique. For instance, Zhou et al. (2017) proposed a fixed-point pancreatic segmentation model, and Yu et al. (2018) proposed a recurrent saliency transformation network. Another two-stage segmentation method employs detection-assisted segmentation. For instance, Man et al. (2019) utilized deep Q-learning to dynamically locate lesion areas, generating localization bounding boxes, and subsequently segmenting the identified regions. However, the accuracy of the fine segmentation stage depends on the precision of the first stage's segmentation or localization, and missed detections in the first stage pose challenges for recovery in the subsequent stage. Consequently, the fine segmentation sometimes yields lower accuracy than the coarse stage.

Re-weighting Method

Re-weighting method mitigates the class imbalance issue by balancing the weights of different class samples in the loss function. It is typically divided into **class re-weighting** and **hard example mining**. WCE (Ronneberger, Fischer, and Brox 2015) assigned higher weights to minority class samples in the loss function to achieve inter-class balance. Tversky loss function (Salehi, Erdogmus, and Gholipour 2017) aimed to strike a balance between FP and FN by introducing a weight parameter. Focal loss (Lin et al. 2017) reduced the weight of easy examples using a weighting factor, allowing the model to concentrate on difficult or minority examples. TopK loss (Wu, Shen, and Hengel 2016) discarded easy samples by setting a threshold or percentage, forcing the network to focus on hard samples. However, these methods often blur the two types of imbalance issues. The most advanced solutions typically combine different losses. Combo loss (Taghanaki et al. 2019) was a weighted sum of Dice loss

and WCE. The Unified Focal loss (Yeung et al. 2022) integrated multiple loss functions within a unified framework. However, its authors indicated that due to its complexity, how to optimize hyperparameters was unclear.

Method

Balance loss is designed to address both inter-class and intra-class imbalance issues. We introduce our loss starting from WCE loss and TopK loss.

Weighted Cross Entropy Loss

To address the issue of inter-class imbalance, a common approach is introducing different weights for different classes. WCE is an extension of CE, defined as follows:

$$\mathrm{WCE} = -\frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N} w_c y_i^c \log(p_i^c), \tag{1}$$

where y_i^c represents the ground truth label for pixel i belonging to class c, and p_i^c is the corresponding network prediction. The indices c and i iterate over all classes and pixels and $w_c \in [0,1]$ is the weighting factor. Typically, w_c is inversely proportional to the class frequency to increase the weight of minority classes. It can also be manually adjusted as a hyperparameter.

TopK Loss

TopK loss forces the network to focus on hard samples during training, primarily addressing the issue of intra-class imbalance. In practice, there are two implementation methods. One method is to retain only the pixels with probability values below a given threshold, defined as follows:

$$\mathrm{TopK_{thr}} = -\frac{\sum_{c=1}^{C}\sum_{i=1}^{N}\mathbb{I}\{y_i = c \ \ and \ \ p_i^c < t\}\log(p_i^c)}{\sum_{c=1}^{C}\sum_{i=1}^{N}\mathbb{I}\{y_i = c \ \ and \ \ p_i^c < t\}}, \tag{2}$$

where, $t \in (0,1]$ is the threshold, and $\mathbb{I}$ is the binary indicator function. Put pixels with a probability greater than t are discarded because the model easily classifies them. Another implementation method is to retain the loss of the worst $k\%$ of pixels, defined as follows:

$$\mathrm{TopK} = \frac{-1}{N}\sum_{c=1}^{C}\sum_{i\in K} y_i^c \log(p_i^c), \tag{3}$$

where K is the set of the worst $k\%$ of pixels.

Inter-class Balance Loss

Although the WCE method can achieve inter-class balance, the adjustment of its weights lacks theoretical basis and standardization, relying entirely on experience and practice. Moreover, the weighting factors in the WCE method are difficult to share between tasks with different degrees of inter-class imbalance. Therefore, when addressing the inter-class imbalance problem, a natural question is how to determine a general method to identify appropriate weights and ensure that these weighting factors can be shared or transferred across different tasks.

To address the issues above, we first discuss how inter-class imbalance affects network training.

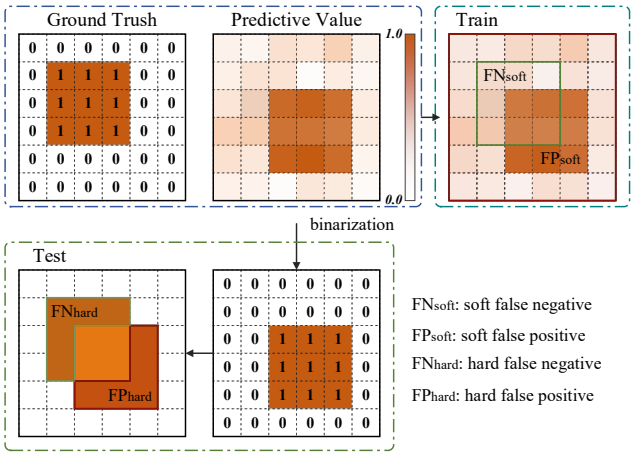

Figure 1: Illustration of FP and FN calculation bias during training and testing.

FP and FN Calculation Bias. As shown in Fig. 1, we identify that inter-class imbalance leads to FP and FN calculation bias between training and testing scenarios. This bias results in a sub-optimal optimization objective during model training, ultimately degrading model performance in tasks with inter-class imbalance. Additionally, we find that the magnitude of the bias is positively correlated with the difference between the majority and minority classes, rather than their ratio. This explains why WCE has the aforementioned issues. For convenience, the following discussion is conducted in the context of binary medical image 2D segmentation.

During testing, researchers typically binarize the predicted probability values into 0 and 1 based on a threshold T (usually $T = 0.5$) to classify pixels, and then calculate FP and FN. We refer to this method as hard calculation, and the FP and FN are called hard FP and hard FN. The formulas are as follows:

$$\mathrm{FP_{hard}} = \sum_{i\in bg}\mathbb{I}\{p_i > T\}, \tag{4}$$

$$\mathrm{FN_{hard}} = \sum_{i\in fg}\mathbb{I}\{(1-p_i) \geq T\}, \tag{5}$$

where $i \in bg$ indicates that pixel i belongs to the background class, and $i \in fg$ indicates that pixel i belongs to the foreground class.

However, the network calculates FP and FN during training based on the probability values. We refer to this calculation method as soft calculation. The FP and FN are called soft FP and soft FN. The formulas are: $\mathrm{FP_{soft}} = \sum_{i\in bg} p_i$, $\mathrm{FN_{soft}} = \sum_{i\in fg}(1-p_i)$. Specifically, $\mathrm{FP_{soft}}$ and $\mathrm{FN_{soft}}$ can be decomposed into two terms based on the threshold T:

$$\mathrm{FP_{soft}} = \sum_{i\in bg} p_i\mathbb{I}\{p_i > T\} + \sum_{i\in bg} p_i\mathbb{I}\{p_i \leq T\}, \tag{6}$$

$$\begin{aligned}\mathrm{FN_{soft}} = &\sum_{i\in fg}(1-p_i)\mathbb{I}\{(1-p_i) \geq T\} \\ &+ \sum_{i\in fg}(1-p_i)\mathbb{I}\{(1-p_i) < T\}.\end{aligned} \tag{7}$$

The soft computation method may yield unequal results when the hard computation method results in equal FP and FN.

Assume the number of foreground pixels is f, and the number of background pixels is b, to simplify the discussion, we assume that the probability value of the predicted foreground region is p, and the probability value of the background region is $1-p$.

According to Eq. (4), Eq. (5), Eq. (6) and Eq. (7), we have:

$$\mathrm{FP}_{\mathrm{soft}} = p\mathrm{FP}_{\mathrm{hard}} + (1-p)(b - \mathrm{FP}_{\mathrm{hard}}), \tag{8}$$

$$\mathrm{FN}_{\mathrm{soft}} = p\mathrm{FN}_{\mathrm{hard}} + (1-p)(f - \mathrm{FN}_{\mathrm{hard}}). \tag{9}$$

Obviously, when $\mathrm{FP}_{\mathrm{hard}}$ and $\mathrm{FN}_{\mathrm{hard}}$ are equal, $\mathrm{FP}_{\mathrm{soft}}$ and $\mathrm{FN}_{\mathrm{soft}}$ are not equal, and:

$$\mathrm{FP}_{\mathrm{soft}} - \mathrm{FN}_{\mathrm{soft}} = (1-p)(b-f). \tag{10}$$

Our discussion indicates that inter-class imbalance does not directly cause prediction bias in the network. Instead, it indirectly affects network performance by causing FP and FN calculation bias during training. This calculation bias implicitly introduces optimization bias through the loss function, making the network more inclined toward the majority class. Simply put, during training on tasks with inter-class imbalance, FP naturally carries higher weight. A more detailed discussion is in the appendix.

Inter-class Balance Loss. Based on the above discussion and Eq. (10), we find that this bias is positively correlated with the difference between the majority and minority classes, rather than the ratio. Traditional WCE addresses inter-class imbalance by adjusting the ratio between foreground and background using weighting factors. However, as shown in Eq. (8) and Eq. (9), it is challenging to avoid this bias solely by adjusting the ratio. Therefore, the determination of its weights relies more on experience and practice rather than concrete theoretical guidance.

To solve the problems above, we propose an improved loss function, Inter-CBL, which is an improvement over the WCE method, addressing the lack of theoretical basis and standardization in weight adjustment. Inter-CBL aims to offer a more generalizable and transferable mechanism for weight determination. It is defined as follows:

$$\begin{aligned}\text{Inter-CBL} = \frac{-1}{f}\sum_{i\in fg}\log(p_i) + \frac{-1}{f}\sum_{j\in bg_{wf}}\log(1-p_j) \\ + \frac{-1}{N}\sum_{k\in bg\setminus bg_{wf}}\log(1-p_k),\end{aligned} \tag{11}$$

where, bg_{wf} represents the set of the f worst pixels in the background region, and $bg\backslash bg_{wf}$ represents the set of background pixels excluding bg_{wf}. Inter-CBL achieves inter-class balance by reducing the loss weight of $b-f$ simple pixels within the majority class. Specifically, we identify the top f pixels with the highest losses in the majority class and assign them weights equal to those of the minority class pixels. In contrast, other pixels are given smaller weights. This not only effectively alleviates the FN and FP calculation bias but also improves intra-class imbalance within the majority class. Formally, Inter-CBL is a variant of WCE, and when the dataset is balanced, i.e., the minority and majority classes are equal, this loss is equivalent to twice the CE.

Task	Dataset	Tr.	Val.	Test	%FG
Pancreas	NIH	60	6	16	0.4
Kidney	KiTS19	152	16	42	3.1
Kidney Tumor	KiTS19	152	16	42	1.8
Polyp	CVC-ClinicDB	392	98	122	9.3
Breast Tumor	BUSI	415	104	128	9.4

Table 1: Overview of tasks and datasets. %FG represents the proportion of foreground (FG) pixels.

Loss	CVC-ClinicDB		BUSI	
	HD↓	DSC↑	HD↓	DSC↑
CE	11.7±1.1	87.8±0.4	31.3±5.4	73.5±0.9
Focal	13.4±1.8	86.6±1.3	29.2±2.4	74.7±0.7
Dice	16.1±3.5	87.6±0.9	29.4±3.4	71.1±0.6
Tversky	15.1±0.4	87.2±0.7	34.1±4.1	71.9±3.3
Combo	12.6±1.1	87.9±0.8	27.2±1.4	73.7±1.5
Unified Focal	13.5±0.9	88.7±0.8	28.5±1.3	74.2±0.6
RCE	11.4±1.7	88.0±0.4	28.9±1.2	73.1±0.6
Balance	**8.8±0.2**	**90.7±0.3**	**26.8±0.9**	**76.9±0.2**

Table 2: Results on the CVC-ClinicDB and BUSI datasets (with standard deviations over three independent runs), with the best results highlighted in bold.

Intra-class Balance Loss

As training progresses, a significant portion of the loss is attributed to easily distinguishable pixels. However, continual learning from these simple pixels results in minimal enhancement in segmentation performance. Therefore, focusing the training on these challenging samples can enhance the network's performance. While TopK loss focuses on learning from difficult samples and can mitigate intra-class imbalance, it applies the same division of difficult and easy samples across all classes. In scenarios with highly imbalanced segmentation tasks, this strategy may discard a significant amount of loss from rare class, thereby causing unstable training. As our experiments will show, TopK loss requires adjustment of the threshold or percentage based on the degree of class imbalance; otherwise, it can cause the model to collapse.

To address this issue, we propose Intra-CBL, an improvement over TopK loss. The specific definition is as follows:

$$\begin{aligned}\text{Intra-CBL} = -\sum_{c=1}^{C}\Bigg(\frac{\sum_{i=1}^{N}\mathbb{I}\{y_i=c \ and \ p_i^c \le t\}\log(p_i^c)}{\sum_{i=1}^{N}\mathbb{I}\{y_i=c \ and \ p_i^c \le t\}} \\ + \frac{\sum_{i=1}^{N}\mathbb{I}\{y_i=c \ and \ p_i^c > t\}\log(p_i^c)}{\sum_{i=1}^{N}\mathbb{I}\{y_i=c \ and \ p_i^c > t\}}\Bigg).\end{aligned} \tag{12}$$

Specifically, we do not discard easy pixels. As training progresses, the number of easy pixels increases, causing the weight of easy pixels, $\sum_{i=1}^{N}\mathbb{I}\{y_i = c \ and \ p_i^c > t\}$, to decrease, while the weight of difficult pixels increases correspondingly. When $t=0$, this loss is equivalent to WCE, where the weight factor is inversely proportional to the class frequency.

Loss	Kidney		Tumor	
	HD↓	DSC↑	HD↓	DSC↑
CE	6.7±0.5	91.3±0.2	26.1±10.6	59.8±1.8
Focal	6.8±0.7	91.0±0.2	29.8±0.9	56.8±0.5
Dice	6.7±1.3	91.3±0.7	31.2±5.1	62.7±1.0
Tversky	7.8±0.8	91.6±0.3	31.2±5.1	62.7±1.0
Combo	6.9±0.5	91.5±0.1	23.0±1.0	61.4±0.5
Unified Focal	7.5±0.5	91.5±0.2	23.7±1.9	61.4±2.5
RCE	6.5±0.6	91.5±0.5	**19.3±3.0**	63.9±0.7
Balance	**6.4±1.2**	**92.4±0.3**	21.2±3.1	**66.8±2.0**

Table 3: Results on the KITS19 dataset.

Loss	HD↓	DSC↑
CE	6.2±0.4	74.7±1.2
Focal	8.6±0.5	66.4±2.1
Dice	5.8±0.3	76.8±0.6
Tversky	6.7±0.4	76.3±1.2
Combo	5.6±0.4	76.5±0.2
Unified Focal	6.6±1.6	76.3±1.1
RCE	6.0±0.7	74.5±0.7
Balance	**5.3±0.5**	**77.6±0.3**

Table 4: Results on the NIH dataset.

Balance Loss

Inter-CBL and Intra-CBL address inter-class and intra-class imbalance issues, respectively. Our Balance loss is defined as follows:

$$\mathrm{BL} = \alpha \mathrm{Intra\text{-}CBL} + (1-\alpha)\mathrm{Inter\text{-}CBL}. \tag{13}$$

However, the Inter-CBL is not suitable for training models from scratch. This is because, during the early stages of training, the predicted probabilities for all pixels are typically close to 0.5. Consequently, a very small number of difficult pixels with large losses dominate the gradient calculation, while the majority of pixels contribute minimally, leading to an unstable training process. As training progresses and gradually converges, the use of Inter-CBL becomes more stable. Therefore, we employ a two-stage training strategy. In the initial stage, only the Intra-CBL is used. Once the network converges, the Balance loss is applied to continue guiding the network training. Specifically, we consider the network to have converged to an appropriate level when more than half of the images within the same batch have their f-th best probability value in the foreground and background greater than t. At this point, we switch to using the Balance loss for further training. The more detailed related explanations are in the appendix.

Experiments

Tasks and Datasets

We evaluate our loss function on five segmentation tasks with varying degrees of class imbalance, using four public datasets: the CVC-ClinicDB (Bernal et al. 2015), the Breast Ultrasound Images (BUSI) (Al-Dhabyani et al. 2020), the 2019 Kidney Tumor Segmentation (KiTS19) challenge (kidney and tumor) (Heller et al. 2019), and the NIH pancreas segmentation dataset (Roth et al. 2015). The CVC-ClinicDB dataset comprises 612 static images extracted from colonoscopy videos for polyp segmentation. We shuffle the data randomly and resize the images to [224, 224]. The BUSI dataset consists of 780 images, categorized into three types: normal, benign, and malignant. We use 487 benign and 210 malignant images, following preprocessing steps consistent with those for the CVC-ClinicDB dataset. The KiTS19 dataset contains 300 arterial-phase abdominal CT scans, with 210 cases having publicly available labels. We truncate the original image intensities to [-79, 304], rescale them to [0, 255], resize the images to [224, 224], and repeat the process three times to form three channels. The NIH dataset includes 82 contrast-enhanced abdominal CT scan volumes. We truncate the original image intensities to [-100, 240], and the remaining steps are consistent with the KiTS19 dataset. Table 1 provides an overview of the tasks and datasets.

Implementation Details

All experiments use a U-Net architecture with batch normalization and dropout for the 2D automatic segmentation of medical images. This framework is implemented in PyTorch, and computations are performed on an NVIDIA GeForce RTX 4090 GPU. The network weights are optimized using the Stochastic Gradient Descent (SGD) algorithm, where the initial learning rate is 0.1 and the batch size is 24. For the convergence criteria, the learning rate is halved if the validation performance does not improve within 10 epochs, and training is terminated if there is no improvement within 20 epochs. We perform the data augmentation to avoid overfitting, including random rotations, flips, and elastic deformations.

We evaluate the following loss functions: cross entropy-based losses including CE loss and Focal loss; dice-based losses including Dice loss and Tversky loss; and compound losses including Combo loss, asymmetric variants of the Unified Focal loss and region-size regularizers cross entropy (RCE) loss (Liu et al. 2024). We use the optimal parameters reported in the original studies for losses with hyperparameters for each loss function. Specifically, we set the Focal loss with $\alpha = 0.25$, $\gamma = 2$ (Lin et al. 2017), Tversky loss with $\alpha = 0.3$, $\beta = 0.7$ (Salehi, Erdogmus, and Gholipour 2017), Combo loss with $\alpha = 0.5$, $\beta = 0.5$ (Taghanaki et al. 2019), Unified Focal loss with $\lambda = 0.5, \sigma = 0.6$, and $\gamma = 0.5$ (Yeung et al. 2022), and RCE loss with $\lambda = 1$ (Liu et al. 2024). For the Balance loss, we set $t = 0.9, \alpha = 0.5$.

Evaluation Metrics. We use the Dice Similarity Coefficient (DSC) as the primary performance metric and additionally calculate the Hausdorff Distance (HD) for a more comprehensive evaluation (Wang, Wang, and Zhu 2020). All ablation studies (in Sec.) are conducted on CVC-ClinicDB, KiTS19 (tumor), and NIH datasets.

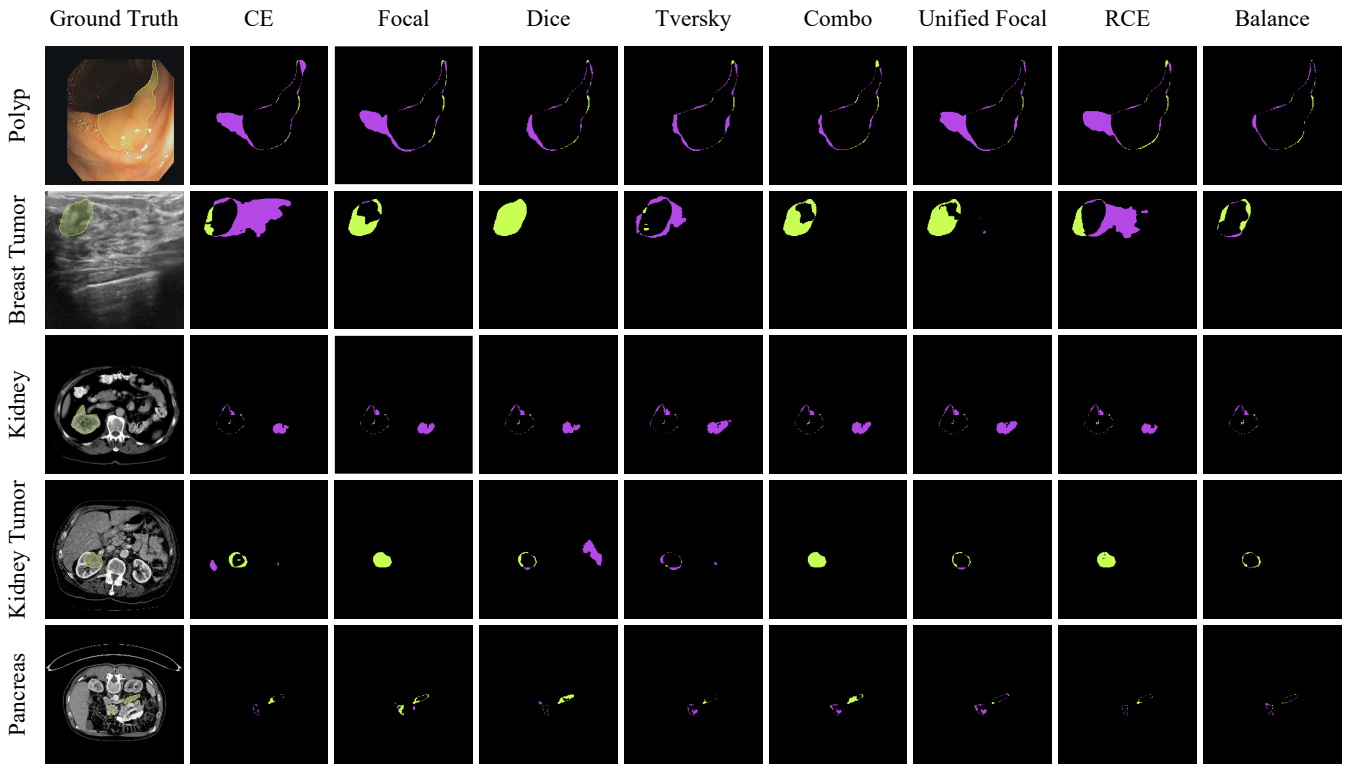

Figure 2: Qualitative comparison of Balance loss with other state-of-the-art methods. In the visualized segmentation results, purple indicates over-segmentation, and green indicates under-segmentation.

Comparison with Other Loss Functions

We conduct two sets of experiments. The first set aims to assess the accuracy and stability of the Balance loss under varying degrees of class imbalance. According to Table 1, we consider polyp and breast tumor segmentation as tasks with slight imbalance, and kidney, kidney tumor, and pancreas segmentation as tasks with high imbalance. Subsequently, we validate the effectiveness of Inter-CBL and Intra-CBL separately and conduct ablation studies on parameters t and α.

Slight Imbalance Segmentation. Table 2 reports the segmentation performance of all loss functions on the CVC-ClinicDB and BUSI datasets. The Balance loss consistently achieves the best results in terms of DSC and HD metrics. Notably, CE loss achieves high DSC performance on both datasets. This observation corroborates our earlier discussion that loss functions designed for highly imbalanced tasks typically perform less effectively on slightly imbalanced tasks.

High Imbalance Segmentation. Table 3 and Table 4 report the segmentation performance of all loss functions on the KiTS19 (kidney and tumor) and the NIH datasets. Our method consistently achieves the best results across all datasets. The improvement in the KiTS19 tumor segmentation is the most significant. In contrast, the improvement in the KiTS19 kidney segmentation is the smallest. This discrepancy likely stems from the varying difficulty of the tasks. The kidney organ has clearer boundaries and less class imbalance. A simpler task means fewer hard samples, thus leading to minimal improvement with the Balance loss, which is based on the hard sample mining method. Conversely, kidney tumors have more ambiguous boundaries and more severe class imbalance issues, resulting in a larger gain from the hard sample mining method.

Dataset	Loss	DSC ↑
CVC-ClinicDB	Unified Focal	88.7±0.8
	Inter-CBL	**89.3±0.7**
NIH	Dice	76.8±0.6
	Inter-CBL	**77.1±0.4**
KiTS19 (tumor)	RCE	63.9±0.7
	Inter-CBL	**65.5±1.0**

Table 5: Comparison of Inter-CBL and sub-optimal methods.

Notably, the sub-optimal approach is different across different segmentation tasks. This finding further supports our theoretical argument that hyperparameters fine-tuned for a particular task often fail to generalize effectively to different segmentation tasks. The Balance loss maintains more stable and higher performance across different datasets, indicating better adaptability to targets of varying sizes.

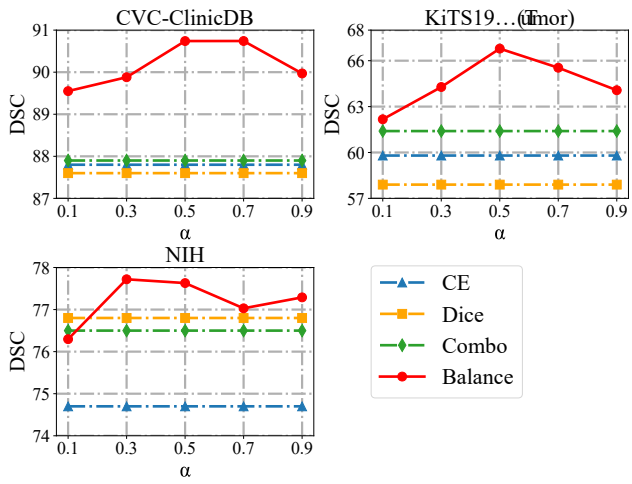

Figure 3: Ablation study on the weight α. The solid lines represent the Balance loss; for reference, the dashed lines represent the DSC performance of the other loss functions.

Qualitative Evaluation. The qualitative results are shown in Fig. 2. Using the Balance loss, the predicted organ masks are improved, closely matching the ground truth labels and significantly outperforming other results. For instance, in kidney and kidney tumor segmentation, as shown in the third and fourth rows of Fig. 2, other methods exhibit large areas of FN or FP, which are corrected by the proposed approach. Visual inspection of the predicted masks indicates that our method has great potential for clinical automatic segmentation.

Ablation Study

Effectiveness of Inter-class Balance Loss. To verify the accuracy and robustness of the Inter-CBL, we set α to 0 in the Balance loss and compare it with the best loss function other than Balance loss across three segmentation tasks. As shown in Table 5, our Inter-CBL maintains the highest performance even without incorporating the Intra-CBL.

Effectiveness of Intra-class Balance Loss. We compare the performance of our Intra-CBL with the two TopK loss variants. As shown in Table 6, the threshold variant may fail during training or produce poor segmentation results, indicating that Intra-CBL is more robust. The TopK (10%) loss performed poorly on the CVC-ClinicDB dataset. A possible reason is that using only 10% of the pixels during training is insufficient because the polyp is very large. This suggests that the percentage variant of TopK loss requires parameter selection based on the degree of class imbalance in the segmentation task, whereas our Intra-CBL is robust to varying levels of class imbalance.

Impact of α. Fig. 3 shows the segmentation performance of Balance loss with different α values across three datasets. The experiments reveal that our loss consistently exhibits strong performance within the range of $\alpha \in [0.1, 0.9]$ on these datasets with varying degrees of class imbalance. This indicates that although our Balance loss benefits from adjusting the balance between the two types of losses, even a sub-optimal α can still yield performance improvements. The performance curves show that the network's performance improves when the Inter-CBL and Intra-CBL are nearly balanced (with α close to 0.5). As previously discussed, addressing only one type of class imbalance typically does not yield the optimal performance boost. Therefore, when applying Balance loss in practice, setting $\alpha = 0.5$ can be used as a default, with slight adjustments within the range of [0.3, 0.7].

Loss	CVC -ClinicDB	KiTS19 (tumor)	NIH
TopK(10%)	34.7±4.0	60.3±1.4	74.4±1.9
TopK_{thr}(0.1)	20.4±3.8	-	-
Intra-CBL(0.1)	60.8±9.7	42.0±3.2	55.6±3.0
TopK(30%)	88.4±0.5	60.4±2.9	74.0±1.6
TopK_{thr}(0.3)	22.1±4.8	-	-
Intra-CBL(0.3)	69.9±2.8	39.4±0.9	39.0±10.5
TopK(50%)	88.7±1.2	59.3±3.9	74.8±7.7
TopK_{thr}(0.5)	73.1±1.1	61.2±1.9	74.4±7.8
Intra-CBL(0.5)	81.9±1.9	62.2±0.9	75.8±2.2
TopK(70%)	88.6±0.8	60.0±1.0	75.0±1.3
TopK_{thr}(0.7)	87.2±0.2	58.7±3.3	73.7±2.5
Intra-CBL(0.7)	**89.3±0.8**	**65.0±2.4**	75.6±2.0
TopK(90%)	88.9±0.6	61.2±2.1	73.5±2.5
TopK_{thr}(0.9)	88.4±2.8	63.7±1.3	76.1±1.1
Intra-CBL(0.9)	89.1±0.5	64.4±3.5	**76.4±0.7**

Table 6: Average DSC of Intra-CBL and different TopK loss variants on three segmentation tasks. The "-" denotes that the results are unavailable because the training process failed with these loss functions.

Conclusion

In this paper, we present a novel Balance loss to mitigate intra-class and inter-class imbalance issues in medical image segmentation. First, we provide a detailed analysis of how inter-class imbalance affects segmentation performance. We identify that the hidden FP and FN calculation bias between testing and training is the main reason for model performance degradation. Based on this conclusion, we propose Inter-CBL. Additionally, we propose the Intra-CBL, which performs hard example mining within each class, enhancing the network's learning ability for hard samples. We then combine these two losses into the Balance loss using a two-stage strategy and a linear weighting method. We evaluate our approach on five segmentation tasks, demonstrating higher accuracy and robustness than other loss functions. For future works, we will explore a single-stage training approach to simplify the training process and attempt to extend it to multi-class segmentation tasks.

Acknowledgments

This work was supported by the Natural Science Foundation of Jilin Province (20220101108JC).

References

Abraham, N.; and Khan, N. M. 2019. 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)*, 683–687.

Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; and Fahmy, A. 2020. Dataset of breast ultrasound images. *Data in Brief*, 28: 104863.

Bali, M.; and Mahara, T. 2023. Comparison of Affine and DCGAN-based Data Augmentation Techniques for Chest X-Ray Classification. *Procedia Computer Science*, 218: 283–290. International Conference on Machine Learning and Data Engineering.

Bernal, J.; Sánchez, F. J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; and Vilariño, F. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43: 99–111.

Dai, P.; Dong, L.; Zhang, R.; Zhu, H.; Wu, J.; and Yuan, K. 2022. Soft-cp: A credible and effective data augmentation for semantic segmentation of medical lesions. *arXiv preprint arXiv:2203.10507*.

Elyan, E.; Moreno-Garcia, C. F.; and Jayne, C. 2021. CDSMOTE: class decomposition and synthetic minority class oversampling technique for imbalanced-data classification. *Neural computing and applications*, 33: 2839–2851.

Garcea, F.; Serra, A.; Lamberti, F.; and Morra, L. 2023. Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, 152: 106391.

Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; and Bing, G. 2017. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73: 220–239.

Hamghalam, M.; and Simpson, A. L. 2024. Medical image synthesis via conditional GANs: Application to segmenting brain tumours. *Computers in Biology and Medicine*, 170: 107982.

He, H.; and Garcia, E. A. 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9): 1263–1284.

Heller, N.; Sathianathen, N.; Kalapara, A.; Walczak, E.; Moore, K.; Kaluzniak, H.; Rosenberg, J.; Blake, P.; Rengel, Z.; Oestreich, M.; et al. 2019. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*.

Li, X.; Li, X.; Pan, D.; and Zhu, D. 2020. On the Learning Property of Logistic and Softmax Losses for Deep Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 4739–4746.

Li, Z.; Kamnitsas, K.; Ouyang, C.; Chen, C.; and Glocker, B. 2023. Context Label Learning: Improving Background Class Representations in Semantic Segmentation. *IEEE Transactions on Medical Imaging*, 42(6): 1885–1896.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Liu, B.; Dolz, J.; Galdran, A.; Kobbi, R.; and Ben Ayed, I. 2024. Do we really need dice? The hidden region-size biases of segmentation losses. *Medical Image Analysis*, 91: 103015.

Man, Y.; Huang, Y.; Feng, J.; Li, X.; and Wu, F. 2019. Deep Q Learning Driven CT Pancreas Segmentation With Geometry-Aware U-Net. *IEEE Transactions on Medical Imaging*, 38(8): 1971–1980.

Prusa, J.; Khoshgoftaar, T. M.; Dittman, D. J.; and Napolitano, A. 2015. Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data. In *2015 IEEE International Conference on Information Reuse and Integration*, 197–202.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–241. Cham: Springer International Publishing.

Roth, H. R.; Lu, L.; Farag, A.; Shin, H.-C.; Liu, J.; Turkbey, E. B.; and Summers, R. M. 2015. DeepOrgan: Multi-level Deep Convolutional Networks for Automated Pancreas Segmentation. In Navab, N.; Hornegger, J.; Wells, W. M.; and Frangi, A., eds., *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 556–564. Cham: Springer International Publishing.

Salehi, S. S. M.; Erdogmus, D.; and Gholipour, A. 2017. Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In Wang, Q.; Shi, Y.; Suk, H.-I.; and Suzuki, K., eds., *Machine Learning in Medical Imaging*, 379–387. Cham: Springer International Publishing. ISBN 978-3-319-67389-9.

Taghanaki, S. A.; Zheng, Y.; Kevin Zhou, S.; Georgescu, B.; Sharma, P.; Xu, D.; Comaniciu, D.; and Hamarneh, G. 2019. Combo loss: Handling input and output imbalance in multi-organ segmentation. *Computerized Medical Imaging and Graphics*, 75: 24–33.

Wang, Z.; Wang, E.; and Zhu, Y. 2020. Image segmentation evaluation: a survey of methods. *Artificial Intelligence Review*, 53(8): 5637–5674.

Wu, Z.; Shen, C.; and Hengel, A. v. d. 2016. Bridging category-level and instance-level semantic image segmentation. *arXiv preprint arXiv:1605.06885*.

Xie, X.; Liu, H.; Zeng, S.; Lin, L.; and Li, W. 2021. A novel progressively undersampling method based on the density peaks sequence for imbalanced data. *Knowledge-Based Systems*, 213: 106689.

Yan, Z.; Yang, X.; and Cheng, K.-T. 2019. A Three-Stage Deep Learning Model for Accurate Retinal Vessel Segmentation. *IEEE Journal of Biomedical and Health Informatics*, 23(4): 1427–1436.

Yang, J.; Zhang, Y.; Liang, Y.; Zhang, Y.; He, L.; and He, Z. 2021. TumorCP: A Simple but Effective Object-Level

Data Augmentation for Tumor Segmentation. In de Bruijne, M.; Cattin, P. C.; Cotin, S.; Padoy, N.; Speidel, S.; Zheng, Y.; and Essert, C., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, 579–588. Cham: Springer International Publishing.

Yeung, M.; Sala, E.; Schönlieb, C.-B.; and Rundo, L. 2022. Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95: 102026.

Yu, Q.; Xie, L.; Wang, Y.; Zhou, Y.; Fishman, E. K.; and Yuille, A. L. 2018. Recurrent Saliency Transformation Network: Incorporating Multi-Stage Visual Cues for Small Organ Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Z.; Tang, M.; Cobzas, D.; Zonoobi, D.; Jagersand, M.; and Jaremko, J. L. 2018. End-to-end detection-segmentation network with ROI convolution. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 1509–1512.

Zhou, Y.; Xie, L.; Shen, W.; Wang, Y.; Fishman, E. K.; and Yuille, A. L. 2017. A Fixed-Point Model for Pancreas Segmentation in Abdominal CT Scans. In Descoteaux, M.; Maier-Hein, L.; Franz, A.; Jannin, P.; Collins, D. L.; and Duchesne, S., eds., *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*, 693–701. Cham: Springer International Publishing.