Master's Thesis

Adaptive PolyDice Loss with Uncertainty-Based Difficulty Estimation for Medical Image Segmentation

M243422 Tomoya Hiroike

Supervisor Assoc. Prof. Akira Furui

February 3, 2026

Informatics and Data Science Program
Graduate School of Advanced Science and Engineering
Hiroshima University

Copyright © 2026, Tomoya Hiroike.

Contents

1 Introduction

Medical image segmentation is an indispensable technology in diagnostic support and treatment planning, requiring the extraction of regions corresponding to normal or abnormal tissues. Clinical applications are advancing rapidly, particularly in the detection of colorectal polyps [1] and organs at risk in head and neck cancer radiotherapy [2].

However, medical image segmentation presents inherent challenges. A particularly significant issue is class imbalance. In medical images, background regions typically occupy the majority of the image, while target lesions often occupy only a relatively small area. Under these conditions, the Cross-Entropy Loss [3], which is widely used in conventional classification tasks, becomes biased toward learning the background regions, making accurate segmentation of clinically significant small lesions and ambiguous boundaries difficult.

To address this challenge, the Dice Loss [4], which is robust to class imbalance, and its many extensions have been proposed, with high performance reported in CT images [5, 6] and MRI images [7]. However, these loss functions have the constraint of possessing a fixed shape across all images. Medical images exhibit significant diversity due to imaging conditions and individual differences, as well as substantial variations in lesion size and shape, differences in tissue contrast, and boundary ambiguity; consequently, the difficulty of segmentation varies greatly from image to image. Using a fixed loss function applies the same training signal to both easy and difficult images, which may result in insufficient learning for difficult cases. Therefore, an approach that adaptively adjusts the loss function according to the difficulty of each image is promising. Realizing such adaptive learning requires two elements: first, a method to flexibly control the shape of the loss function to strengthen the training signal for high-difficulty images; and second, a difficulty metric to evaluate how challenging each image is for the model during training.

In this study, we propose an adaptive learning framework that integrates these two elements. For controlling the shape of the loss function, we employ PolyDice Loss [8], which is obtained by the polynomial expansion of Dice Loss. PolyDice Loss is suitable for adaptive learning because it allows for continuous control of the loss function's shape using image-specific parameters. To quantify difficulty, we use uncertainty estimation via Monte Carlo （MC） Dropout [9]. MC Dropout enables the efficient estimation of the model's epistemic uncertainty by enabling dropout during inference. This uncertainty reflects the degree of the model's lack of confidence in segmenting the image and can be utilized as an indicator of segmentation difficulty. The proposed method dynamically controls the shape parameters of the PolyDice Loss based on estimated uncertainty metrics. It aims to achieve efficient and robust learning by assigning steeper gradients to images for which the model currently lacks confidence, while providing gentler gradients to those where the model can already produce stable predictions.

The main contributions of this study are as follows:

1. **Introduction of a dynamic quantification method for image difficulty based on uncertainty:** We propose a method to estimate epistemic uncertainty during inference using MC Dropout and quantify it as image-level "learning difficulty." This enables objective evaluation of diverse segmentation difficulties in medical images based on the model's confidence.
2. **Construction of an adaptive control framework for the loss function based on difficulty:** We constructed a learning framework that adaptively controls the shape parameter of the PolyDice Loss based on the quan-

tified difficulty metric. The proposed method automatically adjusts the gradients of the loss function according to the difficulty that changes with the progress of learning, thereby simultaneously achieving a focus on learning for difficult cases and the suppression of gradient dominance by easy cases.

3. **Demonstration of effectiveness and versatility using multiple datasets:** We verified the effectiveness of the proposed method through comparative experiments using medical image datasets. The experimental results demonstrated that the proposed method improves segmentation accuracy compared to conventional loss functions with fixed shapes.

2 Preliminaries

2.1 PolyDice Loss

Dice Loss, which is widely used in medical image segmentation, is robust against class imbalance but is constrained by having a fixed shape for all images. In this study, we adopt PolyDice Loss [8], which extends Dice Loss via polynomial expansion, specifically its practical form, PolyDice-1 Loss. PolyDice-1 Loss allows for controlling the shape of the loss function with a single parameter ϵ, enabling adjustment of the gradient steepness according to the difficulty of the image.

2.1.1 Definition of Dice Loss

Let the image size be $H \times W$ and the pixel position be denoted by (i, j) $(i \in \{1, ..., H\}, j \in \{1, ..., W\})$. In the segmentation task, let the model's predicted probability map be $\hat{\mathbf{Y}} = \{\hat{y}_{i,j}\}_{i,j} \in \mathbb{R}^{H \times W}$ and the ground truth mask for that image be $\mathbf{Y} = \{y_{i,j}\}_{i,j} \in \mathbb{R}^{H \times W}$. The Dice Loss is defined by the following equation:

$$\mathcal{L}_{\text{Dice}}(\hat{\mathbf{Y}}, \mathbf{Y}) = 1 - \frac{2 \sum_{j=1}^{W} \sum_{i=1}^{H} \hat{y}_{i,j} y_{i,j}}{\sum_{j=1}^{W} \sum_{i=1}^{H} (\hat{y}_{i,j}^2 + y_{i,j}^2)}. \tag{1}$$

2.1.2 Geometric Interpretation and Polynomial Expansion

By flattening the predicted probability map $\hat{\mathbf{Y}}$ and the ground truth mask $\mathbf{Y}$ into vectors $\hat{\mathbf{y}}$ and $\mathbf{y}$ of length HW, respectively, Dice Loss can be decomposed as follows:

$$\mathcal{L}_{\text{Dice}} = 1 - s \cos \theta, \tag{2}$$

where $s = \frac{2\langle \hat{\mathbf{y}}, \mathbf{y} \rangle}{\|\hat{\mathbf{y}}\|^2 + \|\mathbf{y}\|^2}$ represents the scale component, and $\theta = \arccos \frac{\langle \hat{\mathbf{y}}, \mathbf{y} \rangle}{\|\hat{\mathbf{y}}\| \|\mathbf{y}\|}$ represents the angle between the two vectors. Through this decomposition, Dice Loss can be understood as the product of the scale component s and $\cos \theta$.

We derive the polynomial representation of PolyDice Loss by applying Taylor expansion to the direction component $\cos \theta$. As training progresses, the prediction

$\hat{\mathbf{y}}$ approaches the ground truth $\mathbf{y}$, so the angle θ between the two vectors approaches 0. Utilizing this property, $\cos\theta$ can be approximated by Taylor expansion around $\theta = 0$ as follows:

$$\cos\theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \cdots. \tag{3}$$

Substituting this into the Dice Loss and simplifying yields the general form of PolyDice Loss:

$$\begin{aligned}\mathcal{L}_{\text{PolyDice}} &= 1 - s\left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \cdots\right) \\ &= (1-s) + s\sum_{k=1}^{\infty} \alpha_k \theta^{2k},\end{aligned} \tag{4}$$

where $\alpha_k = \frac{(-1)^{k-1}}{(2k)!}$ is the sign coefficient for each Taylor term.

2.1.3 PolyDice-1 Loss

PolyLoss [10] achieved practical performance improvements in classification tasks by expanding Cross-Entropy Loss into a polynomial and making only the first term adjustable. PolyDice Loss [8] applies this approach to Dice Loss, and in this study, we adopt PolyDice-1 Loss, which adjusts only the first term.

$$\mathcal{L}_{\text{PolyDice-1}} = (1-s) + s\left(\frac{1}{2} + \epsilon\right)\theta^2, \tag{5}$$

where $\epsilon \in \mathbb{R}$ is a hyperparameter that controls the shape of the loss function. Fig. 1 shows the shape change of PolyDice-1 Loss according to ϵ. When $\epsilon > 0$, the penalty for prediction errors is strengthened, and when $\epsilon < 0$, it is relaxed. This characteristic of flexible shape control plays an important role in the adaptive learning framework of this study. In the proposed method described later, dynamically adjusting this ϵ enables gradient control according to the difficulty of individual images, making it possible to optimize the learning strategy on a per-sample basis.

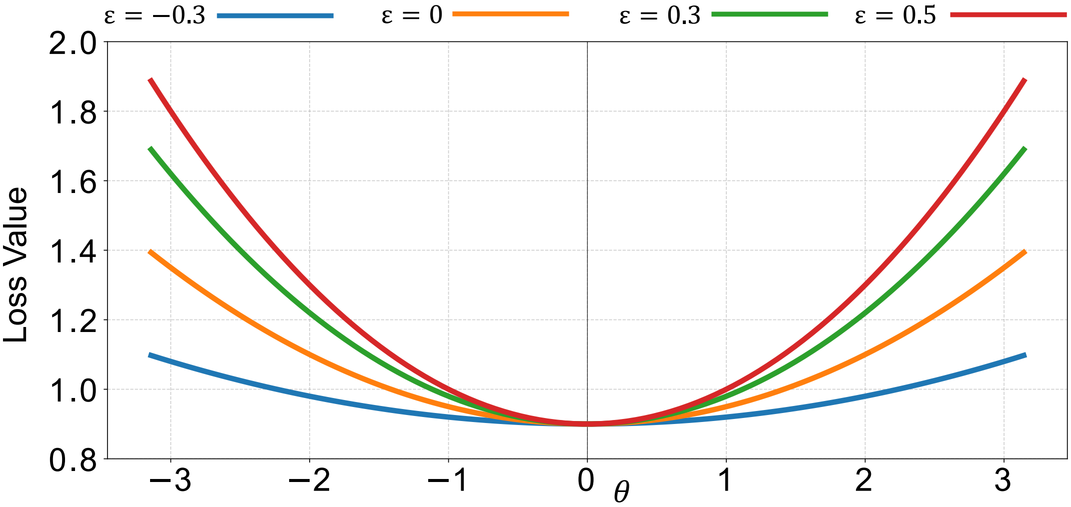

Fig. 1. Effect of the shape parameter ϵ on PolyDice-1 Loss. The loss function is plotted as a function of θ (the angle between prediction and ground truth vectors) with $s = 0.1$. Larger ϵ values result in steeper gradients near $\theta = 0$, amplifying the penalty for prediction errors.

2.2 MC Dropout

2.2.1 Epistemic and Aleatoric Uncertainty

Uncertainty associated with deep learning model predictions is broadly classified into aleatoric uncertainty and epistemic uncertainty based on its source [11].

Aleatoric uncertainty stems from information deficiency intrinsic to the data itself, such as noise caused by imaging devices, low resolution, or physical ambiguity of tissue boundaries. Since this uncertainty is an intrinsic statistical property of the data, it has the characteristic of not being resolved even by adding training data from the same domain.

Epistemic uncertainty stems from unlearned patterns by the model or a lack of knowledge due to insufficient training data. This can be reduced by adding appropriate training data, allowing the model to learn the target distribution in more detail. An image with high epistemic uncertainty indicates that the model has not acquired stable feature representations and the prediction is in an unstable state. In this study, we utilize this epistemic uncertainty as an indicator reflecting the segmentation difficulty of an image.

2.2.2 Principle and Application of MC Dropout

Dropout was proposed as a regularization technique to prevent overfitting in neural networks [12]. It improves the model's generalization performance by randomly inactivating neurons in each layer with probability p during training. Typically, Dropout is disabled during inference, and deterministic prediction is performed with all neurons activated.

MC Dropout [9] is a method that estimates the model's epistemic uncertainty by enabling Dropout not only during training but also during inference. When inference is performed with Dropout enabled, different neurons are inactivated in each inference pass, effectively yielding predictions from different sub-networks. By executing this stochastic inference multiple times for the same input, a distribution of predictions can be obtained.

Gal and Ghahramani [9] showed that training a neural network with Dropout applied is mathematically equivalent to an approximation in Bayesian inference. Through this theoretical framework, the predictive distribution obtained by MC Dropout can be interpreted as an approximation of predictive uncertainty based on the posterior distribution of model parameters, i.e., epistemic uncertainty.

The proposed method utilizes the epistemic uncertainty estimated by MC Dropout as an indicator reflecting the segmentation difficulty of images and employs it for the adaptive control of the loss function.

3 Proposed Method

3.1 Overview

Let $\mathcal{D} = \{(\mathbf{X}_n, \mathbf{Y}_n)\}_{n=1}^{N}$ denote the training dataset, where N is the total number of training images, $\mathbf{X}_n \in \mathbb{R}^{H \times W \times C}$ is the n-th input image (C is the number of channels), and $\mathbf{Y}_n = \{y_{n,i,j}\}_{i,j} \in \mathbb{R}^{H \times W}$ is the corresponding ground truth mask. In uncertainty estimation using MC Dropout, T stochastic inferences are performed for each image. We denote the predicted probability map in the t-th inference ($t \in \{1, \ldots, T\}$) as $\hat{\mathbf{Y}}_n^{(t)} = \{\hat{y}_{n,i,j}^{(t)}\}_{i,j}$.

Fig. 2 illustrates the overview of the proposed method. The design of this method is primarily based on the following two perspectives. First, the difficulty assessment for training samples is dynamically updated during the learning process. Since image difficulty is not absolute but relative, changing with the model's training progress, sequentially re-evaluating difficulty based on the current state of the model allows learning to focus on images where the model currently lacks confidence. Second, epistemic uncertainty is suitable as a quantitative metric for difficulty. Since epistemic uncertainty stems from the model's lack of knowledge, it directly reflects unlearned patterns or regions where the model is hesitant.

Therefore, it allows for the appropriate quantification of how unconfident the model is, which can be improved through learning, without being affected by noise.

In the proposed method, MC Dropout is used every τ epochs during training to perform multiple inferences for each image, and uncertainty is quantified from the variance of the predictions. This uncertainty information reflects the degree of the model's lack of confidence in the segmentation of that image. Subsequently, this uncertainty information is aggregated on an image-wise basis to dynamically control the shape of the PolyDice Loss. This mechanism assigns steeper gradients to images where the model currently lacks confidence, while providing gentler gradients to those for which the model can already produce stable predictions. By applying the updated ϵ to the training in the next τ epochs, we realize adaptive learning that dynamically changes the optimization weighting according to the progress of learning.

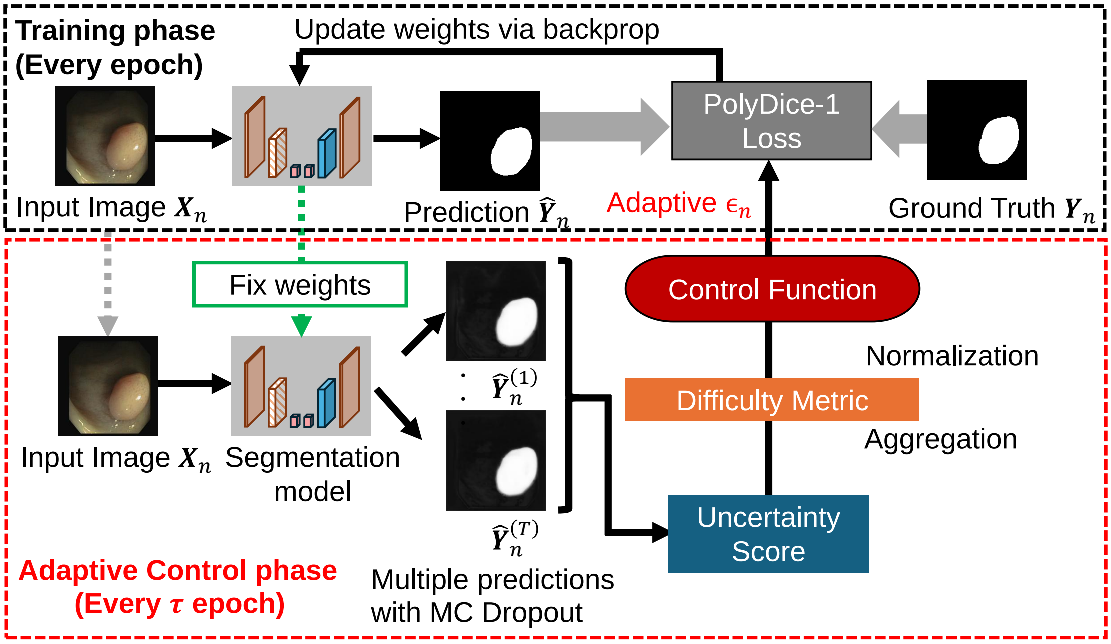

Fig. 2. Overview of the proposed adaptive learning framework. The framework alternates between (1) the adaptive control phase, executed every τ epochs, which estimates image difficulty via MC Dropout and updates the loss shape parameter ϵ for each sample; and (2) the training phase, which optimizes the model using sample-specific ϵ values. This design enables dynamic gradient adjustment based on current model uncertainty.

3.2 Quantification of Image Difficulty Based on Uncertainty

3.2.1 MC Dropout Inference During Training

In the proposed method, the learning process is divided into two phases: an initial training phase and an adaptive training phase. The period from epoch 1 to $E_0 - 1$ is defined as the initial training phase, where training is performed with the loss shape parameter ϵ fixed at 0. This period is established because the model's feature representation is immature in the early stages of learning, and uncertainty at this stage depends more on the model's initialization than on the intrinsic difficulty of the image. E_0 is set as the number of epochs sufficient for the model to acquire basic segmentation capabilities.

In the adaptive training phase ($e \geq E_0$), the re-evaluation of uncertainty and the update of ϵ are performed every period τ. That is, the update is executed before the start of learning for epochs satisfying $e \in \{E_0, E_0 + \tau, E_0 + 2\tau, \ldots\}$. During the update, the model parameters $\mathbf{W}$ at that point are fixed, and T stochastic inferences are performed for each image $\mathbf{X}_n$ in the training data with a dropout rate $p \in (0, 1)$. Let the obtained set of predictions be $\{\hat{\mathbf{Y}}_n^{(t)}\}_{t=1}^{T}$:

$$\hat{\mathbf{Y}}_n^{(t)} = f_{\mathbf{W}}(\mathbf{X}_n; \mathbf{z}^{(t)}), \quad \mathbf{z}^{(t)} \sim \text{Bernoulli}(1 - p), \tag{6}$$

where $\mathbf{z}^{(t)}$ is the Dropout mask for the t-th inference, and $\hat{\mathbf{Y}}_n^{(t)}$ is the resulting predicted probability map. The model's epistemic uncertainty is quantified from the variance of predictions obtained through this stochastic inference.

3.2.2 Calculation of Pixel-wise Uncertainty Metrics

For the T predicted images obtained by MC Dropout, we calculate the Mutual Information $I_{n,i,j}$, which can directly capture epistemic uncertainty, as a pixel-wise uncertainty metric.

$$I_{n,i,j} = \underbrace{H\left(\frac{1}{T}\sum_{t=1}^{T}\hat{y}_{n,i,j}^{(t)}\right)}_{\text{Entropy of Mean}} - \underbrace{\frac{1}{T}\sum_{t=1}^{T}H\left(\hat{y}_{n,i,j}^{(t)}\right)}_{\text{Mean of Entropy}}, \tag{7}$$

where $H(p)$ is the binary entropy function for binary classification, defined as follows:

$$H(p) = -p\log p - (1-p)\log(1-p). \tag{8}$$

Mutual Information is widely used as a metric to evaluate the uncertainty accompanying model predictions and to isolate its underlying factors. It quantifies epistemic uncertainty by subtracting the expected entropy (the mean of the entropy of individual inferences), which indicates aleatoric uncertainty derived from data-inherent noise and ambiguity, from the predictive entropy, which is the uncertainty of the predictive distribution after averaging T inference results (indicating uncertainty from both data and model). High values in a region suggest that the model has not sufficiently learned that area.

3.2.3 Aggregation to Image Level

The mean value of the pixel-wise mutual information, after outlier removal, is quantified as the difficulty metric for the entire image. In medical image segmentation, there is an extreme class imbalance where background regions occupy the majority of the image, while the lesion areas of interest are extremely small. Background regions are generally easy to infer, and their uncertainty tends to take extremely low values. Therefore, if the average uncertainty is calculated over the entire image, the low values from the massive number of background pixels may dominate the overall difficulty metric, failing to properly quantify the local difficulty of the lesion that should be captured.

Therefore, to sensitively reflect the difficulty of lesion detection, the proposed method calculates the average mutual information restricted to the lesion region in the ground truth mask. Let Ω_n be the entire image domain, and $\mathcal{P}_n = \{(i,j) \in \Omega \mid y_{i,j} = 1\}$ be the set of pixels in the positive region of the ground truth mask. Here, the calculated mutual information may contain sporadic noise or extreme outliers, which can destabilize the quantification of the difficulty metric. Therefore, prior to calculating the score, statistical outlier removal is performed. Specifically, let $\mu_{\mathcal{P}_n}$ be the mean and $\sigma_{\mathcal{P}_n}$ be the standard deviation of the mutual information within the region $\mathcal{P}_n$. The valid pixel set $\mathcal{P}'_n$ is defined as follows:

$$\mathcal{P}'_n = \{(i,j) \in \mathcal{P}_n \mid \mu_{\mathcal{P}_n} - 2\sigma_{\mathcal{P}_n} \leq I_{n,i,j} \leq \mu_{\mathcal{P}_n} + 2\sigma_{\mathcal{P}_n}\}\,. \tag{9}$$

The rationale for adopting $2\sigma_{\mathcal{P}_n}$ as the threshold is based on the concept of statistical confidence intervals. Assuming the distribution of mutual information approximates a normal distribution, approximately 95% of all data falls within the range of $\pm 2\sigma$ centered on the mean. Therefore, by rejecting data outside this range, statistically singular extreme values (outliers) are effectively removed, enabling robust difficulty estimation that reflects the main features of the lesion. Using this valid set $\mathcal{P}'_n$, the difficulty score D_n for the entire image is calculated as:

$$D_n = \frac{1}{|\mathcal{P}'_n|} \sum_{(i,j) \in \mathcal{P}'_n} I_{n,i,j}, \tag{10}$$

where $|\mathcal{P}'_n|$ represents the number of pixels in the positive region after outlier removal. Note that for images with no positive region, $D_n = 0$.

Next, to determine the relative difficulty of each sample, normalization is performed based on the difficulty score distribution of the entire dataset. The purpose here is to absorb numerical scale differences between images and evaluate how relatively difficult each image is within the overall distribution. Given the set of scores $\{D_n\}_{n=1}^{N}$ for the entire dataset, let D_q be the q-th percentile value and σ_D be the standard deviation. The normalized score is calculated as follows:

$$D_n^{\text{norm}} = \frac{D_n - D_q}{\sigma_D + \delta}, \tag{11}$$

where $\delta > 0$ is a small constant for numerical stability. Subtraction by D_q serves to center the input for the control function described later. By using the q-th percentile instead of the mean of the distribution, the baseline for difficulty can be flexibly set without being affected by outliers, even in distributions dominated by easy samples. Division by σ_D unifies the scale, serving to adjust the sensitivity of the control function so that it does not depend on the scale of uncertainty specific to each dataset.

3.3 Adaptive Loss Shape Control

3.3.1 Design of the Control Function

Based on the obtained difficulty metric D_n^{norm}, the shape parameter ϵ of the PolyDice-1 Loss is dynamically updated. We use the following sigmoid-based control function for the update equation.

$$\epsilon = \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min})\sigma(k \cdot D_n^{\text{norm}}) \tag{12}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ is the standard sigmoid function, $k > 0$ is a parameter, and $\epsilon_{\min}, \epsilon_{\max}$ represent the variation range of ϵ.

The reason for adopting the sigmoid function in this method lies in its boundedness and smoothness. Abrupt switching, such as that with a step function, may compromise training stability, while a linear function may cause the parameter ϵ to deviate from the appropriate range $[\epsilon_{\min}, \epsilon_{\max}]$. By using the sigmoid function, it is possible to smoothly transition from low to high difficulty regions while strictly constraining the output value within a predetermined range. The parameter k controls the response sensitivity of the function; as shown in Fig. 3, a larger value results in a steeper boundary for difficulty judgment, while a smaller value results in a smoother transition.

Through this control, a large ϵ is assigned to difficult images (large D_n^{norm}), making the gradient of the loss function steeper. This implies giving a larger loss value and gradient for the same prediction error, resulting in the relative strengthening of the learning signal from difficult images. On the other hand, a small ϵ is assigned to easy images that have already been sufficiently learned, preventing overfitting while concentrating learning resources on difficult images. The updated ϵ is applied to training, enabling the model to focus on learning difficult images.

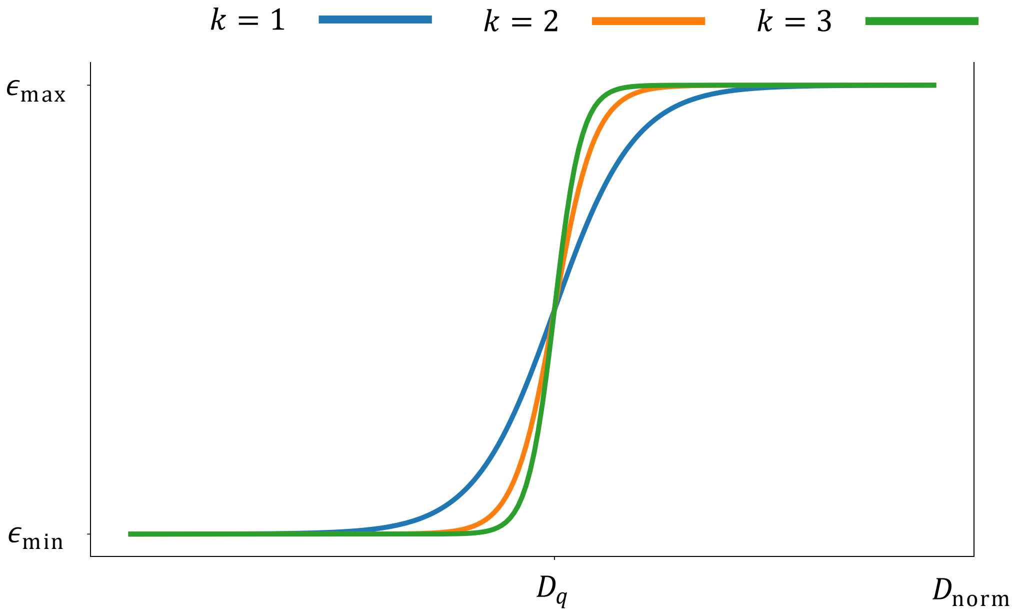

Fig. 3. Sigmoid-based control function mapping normalized difficulty scores to the loss shape parameter ϵ. The slope parameter k controls the sharpness of the transition: larger k yields a more binary assignment, while smaller k produces a gradual mapping. The function is bounded within $[\epsilon_{\min}, \epsilon_{\max}]$ to ensure stable training.

3.3.2 Learning Algorithm

Algorithm 1 presents the detailed algorithm. The learning process consists of two components: adaptive control, which performs difficulty evaluation and loss parameter updates, and the training phase, which actually updates the model parameters.

At the start of training, the model parameters $\mathbf{W}$ are initialized, and the loss shape parameter ϵ for all images is initialized to 0. In the case where $E_0 > 0$, the period from epoch 1 to $E_0 - 1$ is the initial training phase, and training is performed with the standard PolyDice-1 Loss fixing $\epsilon = 0$. Through this period, the model acquires the fundamental feature representations necessary for uncertainty estimation. If $E_0 = 0$, adaptive control begins from the first epoch.

In the adaptive training phase ($e \geq E_0$), adaptive control is executed every τ epochs. Here, following the procedure described in Section 3.2, ϵ for each image is updated through uncertainty metric estimation via MC Dropout inference, aggregation to image units, and normalization. Note that during the adaptive control stage, the model parameters $\mathbf{W}$ are fixed, and only the update of the loss function shape parameter ϵ is performed.

In the subsequent training phase (Epoch e), mini-batch learning is performed using the updated ϵ. Specifically, let $\mathcal{B} \subset \{1, \dots, N\}$ be the set of indices of images constituting a randomly sampled mini-batch. The objective function $\mathcal{L}$ to be optimized is defined as the average of the PolyDice-1 Loss using the shape parameter ϵ_n individually assigned to each image $n \in \mathcal{B}$ in the batch, as follows:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} \mathcal{L}_{\text{PolyDice-1}}(\hat{\mathbf{Y}}_n, \mathbf{Y}_n; \epsilon_n) \tag{13}$$

where $\mathcal{L}_{\text{PolyDice-1}}(\cdot; \epsilon_n)$ denotes the loss function for a single image to which the parameter ϵ_n is applied. By assigning a distinct ϵ_n to each image within a mini-batch, this enables simultaneously application of stronger learning signals to uncertain samples and weaker signals to well-learned samples. Through the iteration of this cycle, the training process can progressively prioritize challenging images in accordance with the model's learning progress.

Algorithm 1 Training procedure for the proposed uncertainty-based adaptive PolyDice-1 Loss framework

```
Require: Training dataset $\mathcal{D} = \{(\mathbf{X}_n, \mathbf{Y}_n)\}_{n=1}^{N}$, Model $f_{\mathbf{W}}$, Max epochs $E$
Require: Hyperparameters: Dropout probability $p$, Start epoch $E_0$, Interval $\tau$, MC
    iterations $T$, Normalization percentile $q$, Slope $k$, Range $[\epsilon_{\min}, \epsilon_{\max}]$
Initialize model parameters $\mathbf{W}$; set $\epsilon_n \leftarrow 0$ for all $n \in \{1, \dots, N\}$
for $e = 1$ to $E$ do
    if $e \geq E_0$ and $(e - E_0) \pmod{\tau} = 0$ then
        Set model to evaluation mode (enable Dropout)
        for $n = 1$ to $N$ do
            for $t = 1$ to $T$ do
                $\hat{\mathbf{Y}}^{(t)} = f_{\mathbf{W}}(\mathbf{X}_n; \mathbf{z}^{(t)}), \quad \mathbf{z}^{(t)} \sim \text{Bernoulli}(1 - p)$
            end for
            Calculate pixel-wise Mutual Information (Eq. 7):
            $I_{n,i,j} = H\left(\frac{1}{T}\sum_{t=1}^{T} \hat{y}_{n,i,j}^{(t)}\right) - \frac{1}{T}\sum_{t=1}^{T} H\left(\hat{y}_{n,i,j}^{(t)}\right)$
            if positive region $\mathcal{P}_n \neq \emptyset$ then
                Compute $\mu_{\mathcal{P}_n}, \sigma_{\mathcal{P}_n}$ from $\{I_{n,i,j} \mid (i,j) \in \mathcal{P}_n\}$
                Identify valid pixels: $\mathcal{P}'_n = \{(i,j) \in \mathcal{P}_n \mid |I_{n,i,j} - \mu_{\mathcal{P}_n}| \leq 2\sigma_{\mathcal{P}_n}\}$
                $D_n = \frac{1}{|\mathcal{P}'_n|} \sum_{(i,j) \in \mathcal{P}'_n} I_{n,i,j}$
            else
                $D_n \leftarrow 0$                                  ▷ Handle negative samples
            end if
        end for
        Step 2: Normalization & $\epsilon$ Update
        Compute $q$-percentile $D_q$ and std $\sigma_D$ from $\{D_n\}_{n=1}^{N}$
        for $n = 1$ to $N$ do
            Normalize score (Eq. 11): $D_{n,\text{norm}} = \frac{D_n - D_q}{\sigma_D + \delta}$
            Update $\epsilon_n$ (Eq. 12): $\epsilon_n \leftarrow \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min})\sigma(k \cdot D_{n,\text{norm}})$
        end for
    else
                              ▷ Keep current $\epsilon_n$ (Note: $\epsilon_n = 0$ if $e < E_0$)
    end if

    Set model to training mode (disable MC Dropout)
    for each minibatch $\mathcal{B} \subset \{1, \dots, N\}$ do
        Compute batch loss with sample-specific $\epsilon_n$ (Eq. 13):
            $\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} \mathcal{L}_{\text{PolyDice-1}}(\hat{\mathbf{Y}}_n, \mathbf{Y}_n; \epsilon_n)$
        Update parameters: $\mathbf{W} \leftarrow \mathbf{W} - \eta \nabla_{\mathbf{W}} \mathcal{L}$
    end for
end for
return Trained parameters $\mathbf{W}$
```

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets

We conducted experiments using the CVC-ClinicDB dataset [13] and the Kvasir-SEG dataset [14]. The CVC-ClinicDB dataset consists of 612 colonoscopy images (384×288 pixels), and the Kvasir-SEG dataset consists of 1000 colonoscopy images. Both datasets are comprised of colonoscopy images and their corresponding ground truth polyp masks. The data was split using 5-fold cross-validation. For the CVC-ClinicDB dataset, we used GroupKFold splitting to ensure that frames from the same video sequence did not span across different folds.

4.1.2 Implementation Details

We adopted U-Net [15] with the architecture shown in Table 1 as the segmentation model. We used the Adam optimizer [16] for training, with the batch size set to 32 and the learning rate set to 10^{-3}. As preprocessing, all images were resized to $W = 224$ pixels and $H = 224$ pixels. During training, we applied horizontal/vertical flips and brightness/contrast adjustments with a probability of 50%. The maximum number of epochs E was set to 200.

Uncertainty estimation via MC Dropout was performed every $\tau = 10$ epochs. Dropout layers were placed in the final block of the encoder and the final block of the decoder. During each evaluation, based on previous work [9], we performed $T = 10$ stochastic inferences with a dropout rate of $p = 0.5$. Additionally, the percentile for normalizing the difficulty metric within the dataset was set to $q = 25$ considering the skewness of the difficulty distribution. Based on the results of preliminary experiments, the start epoch for adaptive learning was set to $E_0 = 10$, with $\epsilon_{\min} = 0$, $\epsilon_{\max} = 0.5$, and $k = 2$.

4.1.3 Comparison Conditions

To evaluate the effectiveness of adaptive learning, we compare the proposed method with the following methods:

- Dice Loss [4]: A standard loss function in medical image segmentation, adopted as the baseline.
- Focal Loss [17]: A method that addresses class imbalance by down-weighting the loss of easy samples. It shares a common motivation with the proposed method in terms of "weighting according to sample difficulty". However, the difference lies in that Focal Loss uses static weighting based on prediction confidence, whereas the proposed method uses dynamic weighting based on model uncertainty. The rate for down-weighting easy samples was set to $\gamma = 2$.
- PolyDice-1 Loss ($\epsilon = 0$): The standard form of PolyDice-1 Loss, which theoretically approximates the standard Dice Loss. It was adopted as a reference to verify the pure effect of manipulating the parameter ϵ in comparison with the proposed method and the Optimal setting described below.
- PolyDice-1 Loss (optimal): To evaluate the theoretical upper bound of performance with a fixed ϵ, we retrospectively searched for the ϵ value that maximizes the Dice coefficient on the test data and included this as an ideal setting for comparison. Specifically, we exhaustively evaluated ϵ in the range of $\{-0.3, -0.2, \ldots, 0.5\}$ and determined the value ϵ that achieved the highest accuracy for each dataset. Although this setting is impossible to realize in practical operation, it represents the performance under the ideal condition where the "optimal fixed value is known beforehand".

Through these comparisons, we verify: (1) whether the proposed method is superior to standard loss functions, (2) whether adaptive ϵ control is more effective than fixed $\epsilon = 0$, and (3) whether the proposed method can achieve performance comparable to or exceeding the optimal setting.

4.1.4 Evaluation Metrics

To evaluate segmentation performance, we used the Dice coefficient and IoU (Intersection over Union) to measure region overlap, and precision and recall to measure detection accuracy. Let $\tilde{Y}_n = \{\tilde{y}_{n,i,j}\}$ be the predicted mask obtained by binarizing the model's prediction map $\hat{\mathbf{Y}}_n$ for image n with a threshold $\theta_{\text{th}} = 0.5$. The numbers of true positive (TP), false positive (FP), and false negative (FN) pixels for image n are defined as follows:

$$TP_n = \sum_{j=1}^{W} \sum_{i=1}^{H} \tilde{y}_{i,j} y_{i,j}, \tag{14}$$

$$FP_n = \sum_{j=1}^{W} \sum_{i=1}^{H} \tilde{y}_{i,j} (1 - y_{i,j}), \tag{15}$$

$$FN_n = \sum_{j=1}^{W} \sum_{i=1}^{H} (1 - \tilde{y}_{i,j}) y_{i,j}, \tag{16}$$

- Dice coefficient: Evaluates the overlap between the ground truth and predicted regions directly. It was adopted as the main metric because it can appropriately reflect the extraction accuracy of minute objects even in unbalanced images like medical images.

$$\text{Dice}_n = \frac{2TP_n}{2TP_n + FP_n + FN_n}. \tag{17}$$

- IoU: Evaluates the intersection over union of the predicted and ground truth regions. It is widely used as a general evaluation metric in segmentation tasks.

$$\text{IoU}_n = \frac{TP_n}{TP_n + FP_n + FN_n}. \tag{18}$$

- Precision: Evaluates the accuracy of the region extracted by the model. It was adopted to quantify the performance in suppressing excessive detection.

$$\text{Precision}_n = \frac{TP_n}{TP_n + FP_n}. \tag{19}$$

- Recall: Evaluates the extent to which the ground truth region is detected. It was adopted specifically to verify the performance in preventing missed lesions.

$$\text{Recall}_n = \frac{TP_n}{TP_n + FN_n}. \tag{20}$$

For each metric, we report the average value over the entire test dataset.

Table 1. Architecture of the U-Net model used in experiments. Each layer shows the output spatial resolution and number of channels.

Layer	**Output Size**
— *Encoder* —	
Input	$224 \times 224 \times 3$
inc (DoubleConv)	$224 \times 224 \times 64$
down1 (MaxPool + DoubleConv)	$112 \times 112 \times 128$
down2 (MaxPool + DoubleConv)	$56 \times 56 \times 256$
down3 (MaxPool + DoubleConv)	$28 \times 28 \times 512$
down4 (MaxPool + DoubleConv)	$14 \times 14 \times 512$
— *Decoder* —	
up1 (Upsample + DoubleConv)	$28 \times 28 \times 256$
up2 (Upsample + DoubleConv)	$56 \times 56 \times 128$
up3 (Upsample + DoubleConv)	$112 \times 112 \times 64$
up4 (Upsample + DoubleConv)	$224 \times 224 \times 64$
outc (Conv2d)	$224 \times 224 \times 2$

4.2 Results and Discussion

4.2.1 Performance Comparison with Existing Methods

Table 2 and Table 3 show the performance comparison with existing methods on each dataset. On both datasets, the proposed method achieved performance surpassing all comparative methods. Particularly noteworthy is that the proposed method outperformed PolyDice-1 Loss (optimal), which retrospectively searched for the optimal fixed ϵ for the test data. Fig. 4 illustrates the performance comparison between the proposed adaptive method and PolyDice-1 Loss with various fixed ϵ values. On both datasets, the proposed method consistently outperforms all fixed ϵ settings, including the retrospectively optimized values. With a fixed ϵ, the same gradient characteristics are applied to all images, which can simultaneously lead to overfitting on easy images and underfitting on difficult images. Furthermore, although the relative difficulty of each image changes as learning progresses, a fixed ϵ cannot follow this change. The proposed method avoids these problems by re-evaluating the difficulty every τ epochs and dynamically updating ϵ.

Focal Loss resulted in lower performance than Dice Loss on both datasets. Note that while Focal Loss addresses class imbalance by reducing the loss of easy samples, this imbalance occurs on a pixel-wise basis rather than an image-wise basis in segmentation tasks. Therefore, it is considered that weighting based solely on pixel prediction uncertainty could not appropriately reflect the segmentation difficulty of the image as a whole.

4.2.2 Performance Analysis by Difficulty Level

To verify the effectiveness of the proposed adaptive learning strategy, which aims to assign steeper gradients to challenging cases and gentler gradients to easier ones, we conducted a stratified analysis based on case difficulty. As an independent criterion for classification, we utilized the Dice coefficients of the baseline Dice

Loss model on the test data. Specifically, cases were categorized as “Hard” if their Dice coefficients were in the bottom 33rd percentile, “Medium” for the 33rd to 66th percentile, and “Easy” for the top 33rd percentile.

Fig. 5 illustrates the performance comparison across these difficulty levels for both datasets. In the Hard category, the proposed method achieved a Dice coefficient improvement of 0.32 for CVC-ClinicDB and 0.13 for Kvasir-SEG compared to the baseline Dice Loss. This improvement suggests that by assigning a larger ϵ to cases identified as difficult due to high uncertainty, the resulting steeper gradients enabled the model to prioritize learning the features of these challenging images. In contrast, the Easy and Medium categories showed minimal changes in accuracy, with slight decreases in some instances. These results reflect the design intent of relatively suppressing gradients from samples that have been sufficiently learned, which can be interpreted as a mechanism to prevent overfitting to simpler patterns. Overall, these findings confirm that the improvement in total performance is primarily driven by the substantial gains in the Hard category.

4.2.3 Analysis of the Adaptive Control Mechanism

To understand the underlying mechanism of the proposed method, we analyzed the transition of the parameter ϵ for the training data across different difficulty levels. Fig. 6 illustrates the evolution of ϵ for each difficulty category on the CVC-ClinicDB dataset. The results indicate that there was no distinct separation between the distributions of ϵ for each group. Notably, a characteristic behavior was observed during the early stages of training (around epochs 10–40), where the ϵ values for images eventually categorized as “Easy” tended to be higher than those in the “Hard” group.

This seemingly contradictory result can be interpreted as a consequence of the difference between the criteria for difficulty classification and the calculation of

ϵ. Difficulty classification in this study is a post-hoc metric based on the final performance of the baseline model using Dice Loss. In contrast, ϵ is dynamically calculated from the model's uncertainty at each point in time. This discrepancy in definition likely caused the different behaviors of ϵ for each group depending on the training stage.

Specifically, in the early stages of training, images in the "Easy" category are in an unlearned state for the model, leading to high predictive variance and temporarily elevated uncertainty (and thus ϵ). Conversely, for images in the "Hard" group, the model has yet to acquire the feature extraction capabilities necessary to recognize their complexity, leading to "overconfidence" —where incorrect regions are predicted with high certainty. Consequently, the uncertainty for the "Hard" group is underestimated, resulting in relatively suppressed ϵ values. As training progresses, the "Easy" group quickly acquires stable feature representations, causing predictive variance to converge and ϵ to decrease rapidly. In contrast, for the "Hard" group, the model gradually begins to recognize the complexity of the cases, and the uncertainty—and subsequently the value of ϵ—begins to rise appropriately.

These analyses suggest that the proposed method functions not as a static weighting based on difficulty labels, but rather as a dynamic gradient adjustment corresponding to the model's current learning state. Rather than simply "assigning large gradients to difficult images," the mechanism can be interpreted as "suppressing gradients for images where learning is already established and promoting learning for unlearned images." This dynamic adaptive control is considered to contribute effectively to the enhanced learning of challenging cases.

Subsequently, to verify the impact of the dynamic control of ϵ on the training process, we analyzed the transition of Dice coefficients on the training data for each difficulty group. The results are shown in Fig. 7. For the Easy and Medium groups, no significant difference was observed between the proposed method and the baseline, with both methods converging rapidly to high accuracy from the early stages of training. In contrast, for the Hard group, the proposed method consistently outperformed the baseline, with the performance gap tending to widen further in the later stages of training.

This result is consistent with the transition of ϵ shown in Fig. 6. In the early stages of training, the temporarily elevated ϵ for the Easy group facilitates learning, after which the decrease in ϵ suppresses overfitting. Meanwhile, for the Hard group, ϵ increases as training progresses, providing continuous reinforcement of the learning signal. This dynamic gradient adjustment is considered to have effectively promoted the learning of difficult cases.

Table 2. Performance comparison with existing loss functions on CVC-ClinicDB dataset

Method	Dice	IoU	Precision	Recall
Dice Loss	0.5408	0.4347	0.6359	0.5819
Focal Loss ($\gamma = 2$)	0.5007	0.4133	0.7078	0.4707
PolyDice-1 ($\epsilon = 0$)	0.5825	0.4808	0.6817	0.6167
PolyDice-1 (Opt. ϵ)	0.6145	0.5145	0.7090	0.6372
Adaptive PolyDice-1	**0.6924**	**0.5262**	**0.7827**	**0.7113**

Table 3. Performance comparison with existing loss functions on Kvasir-SEG dataset

Method	Dice	IoU	Precision	Recall
Dice Loss	0.7895	0.7021	0.8281	0.8154
Focal Loss ($\gamma = 2$)	0.7192	0.6082	0.8634	0.6769
PolyDice-1 ($\epsilon = 0$)	0.8095	0.7198	0.8461	0.8268
PolyDice-1 (Opt. ϵ)	0.8095	0.7198	0.8461	0.8268
Adaptive PolyDice-1	**0.8272**	**0.7440**	**0.8707**	**0.8397**

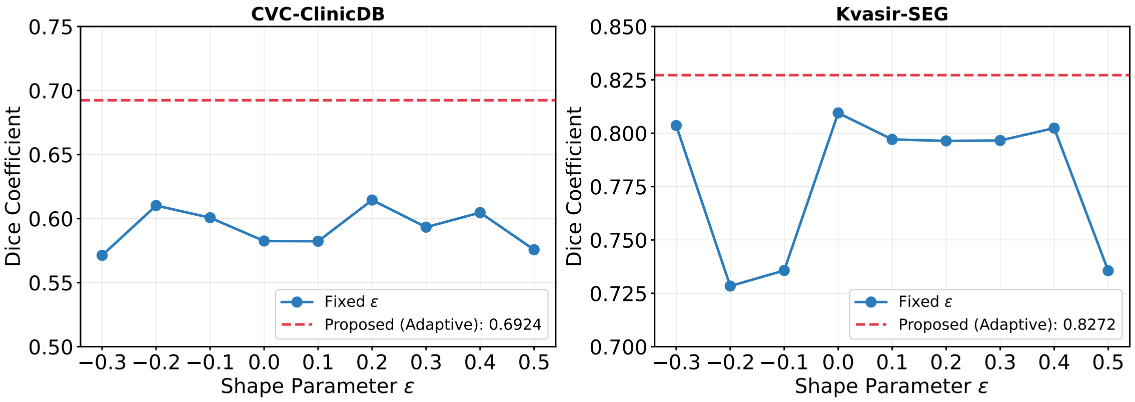

Fig. 4. Comparison of segmentation performance (Dice coefficient) between the proposed adaptive method and PolyDice-1 Loss with various fixed ϵ values. The left graph shows the results for the CVC-ClinicDB dataset, and the right graph shows the results for the Kvasir-SEG dataset. The red dashed line represents the performance of the proposed method.

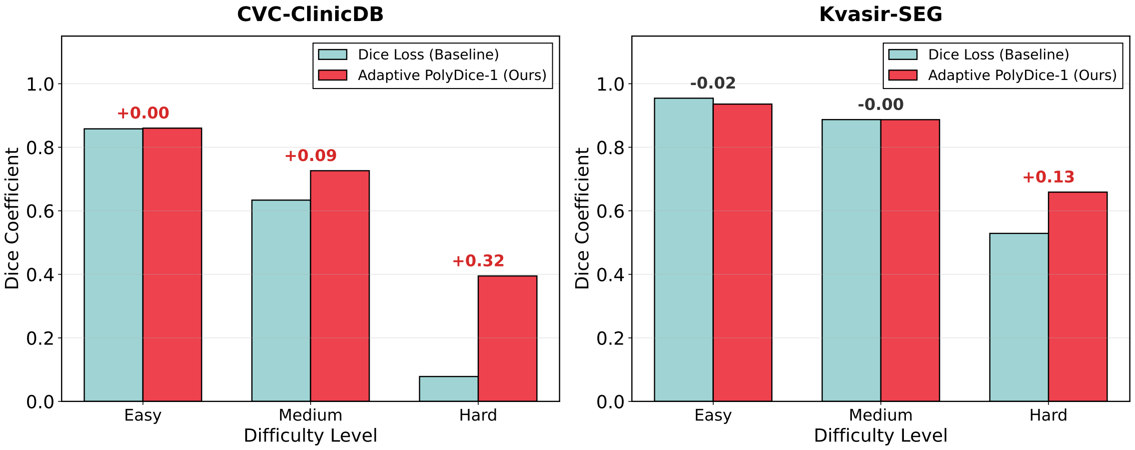

Fig. 5. Stratified performance comparison of the Dice coefficient across three difficulty levels (Easy, Medium, Hard) on the CVC-ClinicDB (left) and Kvasir-SEG (right) datasets. The proposed Adaptive PolyDice-1 Loss (red bars) demonstrates consistent improvements over the baseline Dice Loss (blue bars), particularly in the “Hard” category, where significant performance gains (+0.32 and +0.13) are observed.

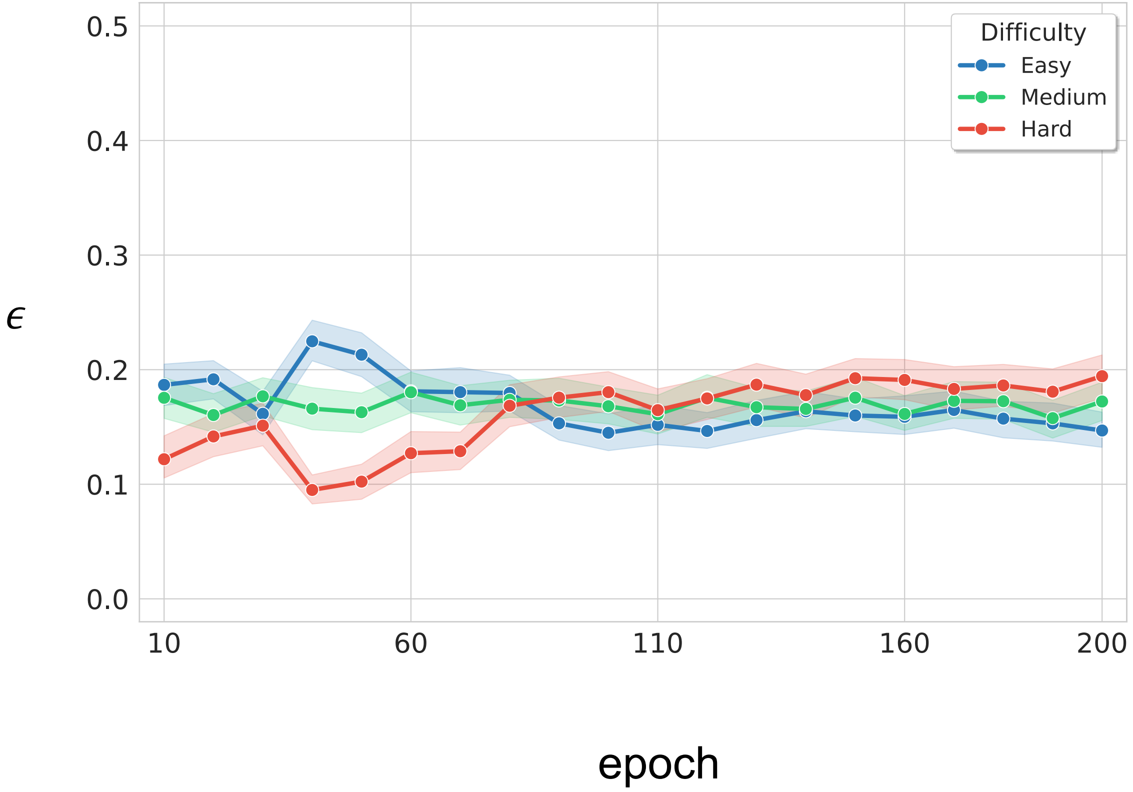

Fig. 6. Evolution of the adaptive parameter ϵ throughout the training process on the CVC-ClinicDB dataset. Cases are categorized into "Easy," "Medium," and "Hard" based on their baseline segmentation performance. Solid lines represent the mean values, and shaded regions denote the 99% confidence intervals.

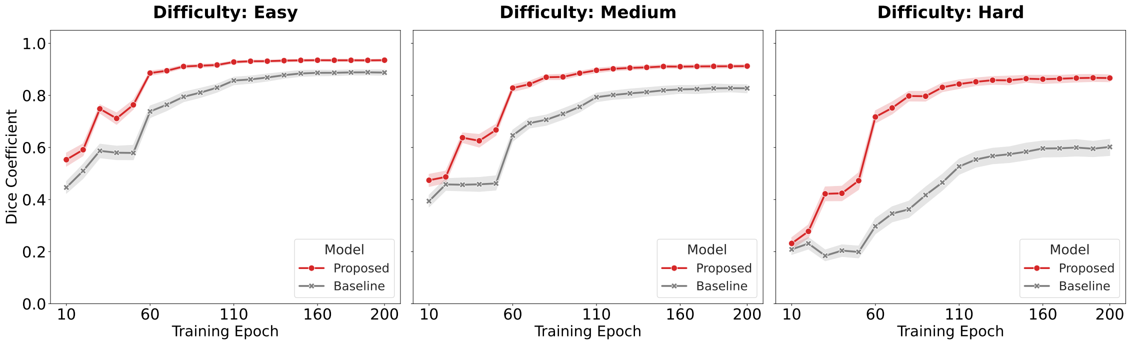

Fig. 7. Dice coefficient transitions throughout training for different difficulty levels on CVC-ClinicDB. Curves show the mean values of the proposed (red) and baseline (gray) methods, with shaded areas indicating 99% confidence intervals.

4.2.4 Impact of Hyperparameters

In this section, we conduct a sensitivity analysis of the main hyperparameters in the proposed method to verify the validity of the design choices.

Impact of ϵ Range: As shown in Table 4 and Table 5, the best segmentation performance was obtained on both datasets when the variation range of ϵ was set to $[0, 0.5]$. Expanding the upper limit $\epsilon_{\max}$ from 0.3 to 0.5 allowed for assigning a larger ϵ to difficult cases, which likely appropriately emphasized the loss gradient and promoted the learning of lesion features. On the other hand, the setting where the lower limit $\epsilon_{\min}$ was extended to -0.3 ($\epsilon_{\min} = -0.3, \epsilon_{\max} = 0.5$) did not improve performance compared to the case of $\epsilon_{\min} = 0$. This is likely because the excessive expansion of the range increased the sensitivity to minute noise contained in the uncertainty estimation more than necessary, thereby impairing training stability.

Impact of Sensitivity Parameter k: Table 6 and Table 7 show the results of varying the slope k of the sigmoid function with respect to the difficulty score. On the Kvasir-SEG dataset, the accuracy was highest when $k = 2$. On the CVC-ClinicDB dataset, although $k = 3$ showed slightly higher values, $k = 2$ achieved comparable performance. When the slope is too gentle, such as $k = 1$, the difference in difficulty is not sufficiently reflected in the loss shape. Conversely, when it is too steep, such as $k = 3$, the loss shape switches binarily, which may cause unstable learning. Based on the above, it is considered that $k = 2$ is an appropriate setting for balancing smooth adaptation according to difficulty and stable learning.

Impact of the Adaptation Start Epoch E_0: As illustrated in Fig. 8, an earlier timing for the commencement of adaptive learning yielded superior results, with the highest performance recorded at the earliest setting, $E_0 = 10$. As demonstrated in Section 4.2.3, distinct differences in the learning progress among cases are already evident during the early stages of training (see Fig. 7). It is considered that initiating adaptive control from this stage maximized the effect of promoting learning for unlearned cases.

Table 4. Impact of the adaptive ϵ range $[\epsilon_{\text{min}}, \epsilon_{\text{max}}]$ on segmentation performance (CVC-ClinicDB dataset).

ϵ_{min}	ϵ_{max}	Dice	IoU	Precision	Recall
0	0.3	**0.7017**	**0.6050**	**0.7875**	**0.7234**
0	0.5	0.6924	0.5913	0.7827	0.7113
-0.3	0.5	0.7000	0.5993	0.7818	0.7220

Table 5. Impact of the adaptive ϵ range $[\epsilon_{\text{min}}, \epsilon_{\text{max}}]$ on segmentation performance (Kvasir-SEG dataset).

ϵ_{min}	ϵ_{max}	Dice	IoU	Precision	Recall
0	0.3	0.8239	0.7391	0.8608	0.8410
0	0.5	**0.8272**	**0.7440**	**0.8707**	0.8397
-0.3	0.5	0.8189	0.7313	0.8537	**0.8454**

Table 6. Sensitivity analysis of the sigmoid slope parameter k in the control function (CVC-ClinicDB dataset).

k	Dice	IoU	Precision	Recall
1	0.6935	0.5962	0.7853	0.7175
2	0.6924	0.5913	0.7827	0.7113
3	**0.7063**	**0.6084**	**0.8010**	**0.7178**

Table 7. Sensitivity analysis of the sigmoid slope parameter k in the control function (Kvasir-SEG dataset).

k	Dice	IoU	Precision	Recall
1	0.8235	0.7404	0.8589	**0.8440**
2	**0.8272**	**0.7440**	**0.8707**	0.8397
3	0.8229	0.7375	0.8663	0.8370

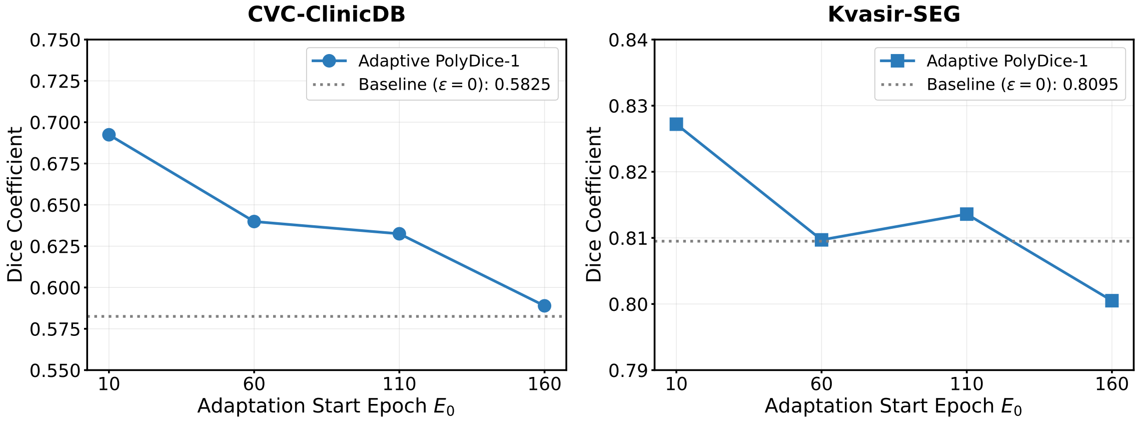

Fig. 8. Effect of adaptation start epoch E_0 on segmentation performance (Dice coefficient) for the CVC-ClinicDB (left) and Kvasir-SEG (right) datasets.

4.2.5 Qualitative Evaluation

Fig. 9 shows the segmentation results of the comparative methods and the proposed method for representative cases of each difficulty level in the CVC-ClinicDB dataset. For easy and medium cases where the boundaries are clear, both existing methods and the proposed method achieve high-precision segmentation. On the other hand, in difficult cases where the contrast between the polyp and the mucosal wall is low and the boundaries are unclear, the baseline Dice Loss and Focal Loss failed to capture the polyp region, resulting in significant under-detection. Even PolyDice-1 Loss (optimal) shows incomplete detection. In contrast, the proposed method appropriately captures the polyp boundaries and achieves good segmentation results. This result suggests that the proposed method determined the case as a difficult image via MC Dropout and set a large shape parameter ϵ for the loss function, thereby strengthening the gradient during training and improving the feature extraction capability in low-contrast regions.

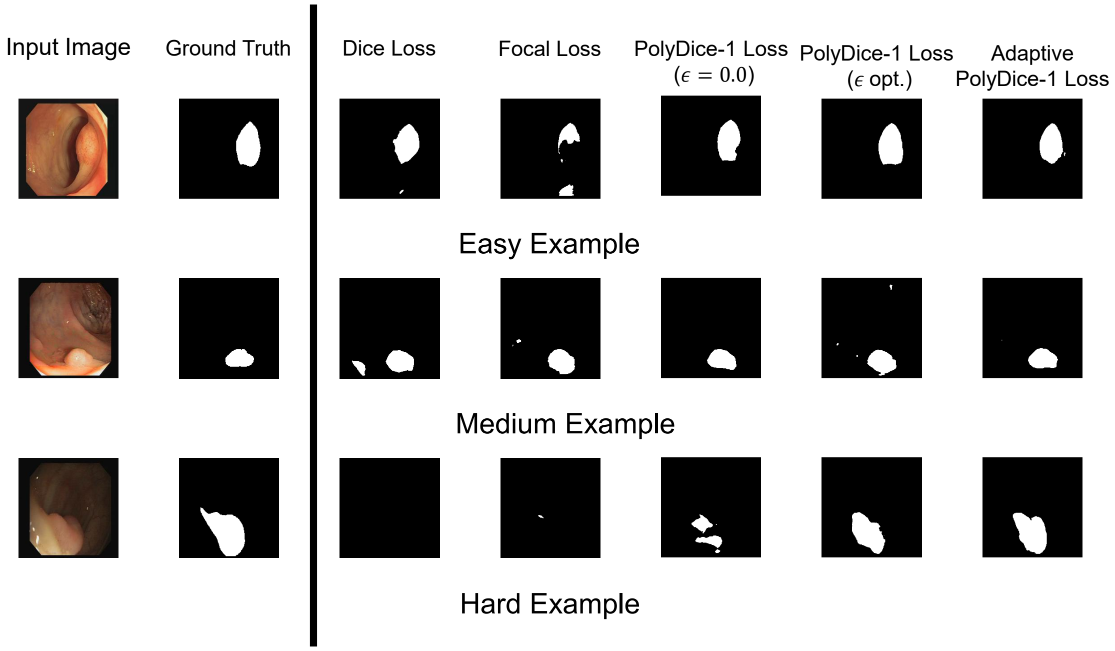

Fig. 9. Qualitative comparison of segmentation results on the CVC-ClinicDB dataset. Each row represents a different difficulty level (easy, medium, hard), categorized by baseline Dice Loss performance.

5 Conclusion

In this paper, we proposed an adaptive learning method for medical image segmentation. The proposed method quantifies the epistemic uncertainty of the model using MC Dropout and utilizes it as an indicator of segmentation difficulty for each image. By dynamically controlling the shape parameters of the PolyDice-1 Loss based on the obtained difficulty metrics, we realized adaptive gradient adjustment tailored to the model's learning state. Our results confirmed that this dynamic control facilitates the training of challenging cases.

In evaluation experiments using the CVC-ClinicDB and Kvasir-SEG datasets, the proposed method achieved performance surpassing not only existing methods such as Dice Loss and Focal Loss but also the fixed parameter setting optimized for the test data. In particular, for cases that were difficult to segment with conventional methods, the Dice coefficient improved by 0.32 on the CVC-ClinicDB dataset, demonstrating the effectiveness of the proposed adaptive learning strategy for detecting challenging lesions.

One limitation of the proposed method is the increased computational cost compared to standard training due to the multiple inferences required for uncertainty estimation. Additionally, the current framework relies on empirically determined hyperparameters, and establishing a mechanism for their automatic optimization remains a challenge. Future work will focus on improving computational efficiency and extending the framework to 3D medical images, such as CT and MRI, to verify its versatility across a broader range of clinical tasks.

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Associate Professor Akira Furui, for his constant and enthusiastic guidance and great encouragement throughout the course of this research. Associate Professor Furui provided me with detailed and careful instruction on every aspect of my work, ranging from discussions on the direction of the research to thesis writing and presentation techniques. Although I had mixed feelings of anticipation and anxiety when I was first assigned to the laboratory, thanks to his advice—at times kind and at times strict—I was able to push forward with my research without hesitation.

I would like to express my deep appreciation to Professor Hiroaki Mukaidani and Associate Professor Zu Soh, who served as my sub-supervisors, for their valuable advice during the interim presentation. The insightful comments and sharp feedback I received from them served as extremely important guidelines for improving the quality of this research and deepening the discussion.

I am also deeply grateful to Assistant Professor Hiroaki Aizawa for providing valuable opinions from multiple perspectives during our collaborative research. He taught me how the fundamental knowledge gained in undergraduate lectures connects to cutting-edge research issues, giving me a rare and invaluable opportunity to reaffirm the depth and fascination of academia.

The presence of the members of the Intelligent Biological Information Systems Laboratory was a great emotional support for me in my daily research life. The time spent not only in active discussions during the general seminars but also sharing meals between research activities and laughing over casual daily conversations are irreplaceable memories for me.

Finally, I would like to dedicate my heartfelt thanks to my parents, who willingly

approved my pursuit of graduate studies and continued to support me both financially and mentally throughout my enrollment. The privileged environment that allowed me to devote myself to research without any inconvenience was realized solely because of the devoted support of my family.

I hereby express my deepest gratitude.

References

[1] G.-P. Ji, G. Xiao, Y.-C. Chou, D.-P. Fan, K. Zhao, G. Chen, and L. Van Gool, "Video polyp segmentation: A deep learning perspective," *Machine Intelligence Research*, vol. 19, no. 6, pp. 531–549, 2022.

[2] F. Maleki, W. T. Le, T. Sananmuang, S. Kadoury, and R. Forghani, "Machine learning applications for head and neck imaging," *Neuroimaging Clinics*, vol. 30, no. 4, pp. 517–529, 2020.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[4] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571. IEEE, 2016.

[5] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, and X. Xie, "AnatomyNet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Medical Physics*, vol. 46, no. 2, pp. 576–589, 2019.

[6] G. Wang, X. Liu, C. Li, Z. Xu, J. Ruan, H. Zhu, T. Meng, K. Li, N. Huang, and S. Zhang, "A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from ct images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2653–2663, 2020.

[7] S. Kato and K. Hotta, "Adaptive t-vMF dice loss: An effective expansion of dice loss for medical image segmentation," *Computers in Biology and Medicine*, vol. 168, pp. 107695, 2024.

[8] H. Aizawa, "Polynomial dice loss for medical image segmentation," (submit-

ted).

[9] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on Machine Learning*, vol. 48, pp. 1050–1059, 20–22 Jun 2016.

[10] Z. Leng, M. Tan, C. Liu, E. D. Cubuk, J. Shi, S. Cheng, and D. Anguelov, "Polyloss: A polynomial expansion perspective of classification loss functions," in *International Conference on Learning Representations*, 2022.

[11] L. Smith and Y. Gal, "Understanding measures of uncertainty for adversarial example detection," *arXiv preprint arXiv:1803.08533*, 2018.

[12] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

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

[14] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. d. Lange, D. Johansen, and H. D. Johansen, "Kvasir-SEG: A segmented polyp dataset," in *Proceedings of the 26th International Conference on Multimedia Modeling (MMM 2020), Part II*, vol. 11962 of *Lecture Notes in Computer Science*, pp. 451–462, Springer. 2020.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

[16] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.