Rheinische Friedrich-Wilhelms-Universität Bonn

Master Thesis

Abstention for Noise-Robust Learning in Medical Image Segmentation

by
Wesam Moustafa
Mat. Nr. 3410585

First Examiner:
Prof. Dr. Rafet Sifa

Second Examiner:
Prof. Dr. Christian Bauckhage

Supervisors:
Prof. Dr. Rafet Sifa & Dr. Helen Schneider

June 30, 2025

Abstract

The pervasive presence of label noise constitutes a significant impediment to the accuracy and reliability of deep learning models. Although the broader field of machine learning has witnessed extensive research dedicated to developing methods for mitigating the impact of noisy labels in classification tasks, this critical area remains notably under-investigated within the specialized domain of image segmentation. This thesis endeavours to bridge this crucial research gap by exploring the abstention mechanism, a strategy that has demonstrably proven its effectiveness against label noise in classification, and subsequently adapting it for robust application in image segmentation.

Our core contribution rigorously demonstrates the profound versatility of the abstention mechanism. While abstention was originally conceived as an extension specifically for the traditional Cross Entropy loss, we significantly improve and generalize its core definition by incorporating an informed regularization term, and employing a flexible, power-law-based auto-tuning algorithm for the abstention penalty. This refined mechanism is then systematically integrated with other distinct loss functions: Generalized Cross Entropy, Symmetric Cross Entropy, and Dice Loss. This process reveals an emergent pattern: the capability to fundamentally enhance existing loss functions with abstention, creating novel noise-robust variants tailored for segmentation.

Empirical evaluations on CaDIS and DSAD medical image datasets under varying noise levels unequivocally validate the efficacy of our proposed methods. Across diverse network architectures, our abstaining loss functions consistently demonstrated superior performance compared to their respective non-abstaining baselines. This consistent leap in performance, coupled with architectural robustness, underscores the fundamental advantage of enabling models to intelligently disengage from potentially corrupted training samples, leading to enhanced generalization.

This work establishes abstention as a powerful, generalizable extension capable of fundamentally enhancing the noise resistance of a broad spectrum of deep learning models. The demonstrated modularity simplifies the development of robust AI systems, paving the way for more reliable and trustworthy diagnostic decision support in medical imaging.

Acknowledgments

This thesis is the culmination of not only my own work but also the generous support, guidance, and encouragement of many people and organizations, to whom I am deeply grateful.

I would first and foremost like to express my deepest and most sincere gratitude to my supervisors, Prof. Dr. Rafet Sifa and Dr. Helen Schneider. Your invaluable guidance, continuous support, and insightful feedback have been instrumental throughout every stage of this research. Your mentorship has not only been pivotal in shaping this thesis but has also inspired and challenged me, for which I am truly grateful.

I would also like to extend my sincere thanks to Prof. Dr. Christian Bauckhage for graciously agreeing to serve as the second examiner for this thesis. I deeply appreciate the time and expertise you have dedicated to evaluating my work.

This project would not have been possible without the foundational support of the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS). I extend my sincere gratitude for the opportunity to work on such a timely and challenging research topic. The generous provision of high-performance computing resources was essential for the empirical validation of my work, and I am particularly grateful for the expert supervision provided by the institute, which was invaluable.

Lastly, I am deeply thankful for the love and support of my family and friends. Your belief in me has been a constant source of strength. Thank you for everything.

Contents

Chapter 1

Introduction

The remarkable advancements in deep learning have revolutionized numerous fields, from natural language processing to computer vision, largely propelled by the availability of vast, meticulously labelled datasets [18, 50, 8]. However, the pervasive presence of label noise—defined as errors or inaccuracies in the assigned ground truth annotations—constitutes a significant and often unavoidable impediment to the generalizability and predictive accuracy of Deep Neural Networks (DNNs) in real-world applications [11]. Such imperfections can arise from various sources, including human annotation errors, subjective interpretations, sensor malfunctions, or the inherent biases of automated labelling processes [18, 50]. DNNs, with their immense capacity and expressive power, are particularly vulnerable to label noise; they possess a strong propensity to memorize incorrect labels, even those that are contradictory or random, which ultimately harms their ability to generalize to unseen, clean data [22]. This phenomenon can lead to models that perform well on the noisy training set but exhibit substantial degradation in performance on validation and test sets. Label noise can manifest in various forms, from arbitrary (randomly flipped labels) to systematic or structured noise (biased errors correlated with underlying data features), with the latter typically being more challenging to detect and correct.

In the critical domain of medical image segmentation, the impact of label noise is particularly pronounced and carries significant clinical consequences [18, 50]. Medical image segmentation, which involves partitioning an image into distinct anatomical or pathological structures, is indispensable for precise disease diagnosis, effective treatment planning (e.g., tumour delineation for radiotherapy), and facilitating advanced medical research [46]. The impressive success of deep learning models in this field is intrinsically contingent upon the availability of extensive, meticulously annotated pixel-level training datasets [14, 18, 31]. However, obtaining high-quality pixel-level annotations in medical imaging is an inherently challenging and labour-intensive task, constrained by the scarcity of experienced medical annotators, intrinsic visual ambiguities at object boundaries, and practical budgetary limitations [14, 18]. The scarcity and high cost of these annotators limit the availability of meticulously labelled data, often leading to the use of less rigorous methods like crowdsourcing or automated labelling with limited manual refinement, which inherently introduce label noise [18].

The nearly unavoidable presence of label noise in training datasets has a profound and detrimental impact on the performance and generalizability of deep learning models. Train-

ing DNNs with noisy labels impairs their performance because loss function calculations receive 'partially incorrect' gradients, misleading the optimization process and causing the network to learn incorrect patterns or memorize erroneous semantic correlations [28]. Empirical studies have quantitatively demonstrated this, showing that DNN performance consistently degrades as the scale of contamination increases, with degradation being rapid for structured noise like erosion and dilation that directly alter object boundaries [28]. Critically, neural networks not only perform worse but actively learn and internalize annotator biases present in the noisy data, leading to predictable and potentially dangerous failure modes in critical applications if not addressed [28, 50].

To counteract the detrimental impact of label noise, extensive research has been dedicated to developing robust learning methodologies, primarily within the domain of classification tasks [11, 18]. These proposed methods broadly fall into several categories: noise filtering techniques aim to identify and correct mislabelled instances during training [11, 18]; noise-tolerant algorithms modify their loss functions or optimization procedures to be less sensitive to incorrect labels, with notable examples including robust loss functions like Mean Absolute Error (MAE) or modified versions of Cross Entropy [18, 22]; and label correction approaches predict pseudo-labels for instances suspected of being mislabelled [22]. Other sophisticated strategies involve loss reweighting, which dynamically assigns importance weights to training instances to reduce the influence of noisy ones [18], or curriculum learning paradigms that prevent early memorization of noise by progressively shifting supervision [18, 22]. While these techniques have shown considerable promise in classification, many introduce additional computational complexity, require assumptions about noise characteristics (e.g., noise transition matrices), or risk inadvertently removing genuinely clean samples [11].

Among the various explored directions and paradigms for mitigating label noise, one particular approach stands out for several reasons. **Noise-robust loss functions** offer a fundamental and highly accessible approach to mitigating label noise in deep learning, primarily due to their inherent simplicity, model-agnostic nature, and minimal computational overhead, often serving as 'plug-and-play' replacements for standard losses [41]. They directly counteract the problematic tendency of DNNs to memorize erroneous labels by modifying the optimization objective, leveraging properties like boundedness [51] and symmetry [45] to limit the influence of noisy examples and prevent overfitting [43]. This directness and efficiency distinguish them from other paradigms, such as data cleaning/sample selection, loss correction, robust network architectures, and ensemble methods, which typically involve significantly higher implementation complexity, computational demands (e.g., requiring multiple networks or intricate pipelines), or reliance on difficult-to-estimate noise characteristics [7]. While noise-robust loss functions can be susceptible to underfitting due to vanishing gradients, particularly in complex multi-class scenarios, their strong theoretical foundations and straightforward application make them a distinct and often preferred initial strategy for building resilient deep learning models [43, 5].

Despite the critical importance of robust learning, there remains a relative shortage of dedicated research investigating the specific utility and effectiveness of these noise mitigation methods within the domain of image segmentation, or developing new methods tailored for its unique challenges [14, 18]. Image segmentation, particularly in medical contexts, demands not only correct class labels but also precise spatial delineation at a

pixel-wise level, making it exceptionally susceptible to annotation noise. Existing learning-from-noise approaches frequently struggle to adequately address the spatial inaccuracies inherent in segmentation-specific noise, highlighting a significant research gap [14]. This research void underscores an urgent need for robust medical image segmentation techniques capable of effectively handling noisily labelled data.

Within the landscape of noise-robust learning paradigms, the mechanism of **abstention** distinguishes itself as a strategy of significant merit, primarily due to its demonstrated efficacy in mitigating the effects of label noise in classification contexts [18, 42, 36]. Unlike traditional supervised learning paradigms that compel a model to make a definitive prediction for every input, the abstention mechanism empowers a DNN to 'abstain' from classifying confusing or unreliable samples [42, 36]. This is achieved by extending the network's output to include an explicit abstention option, allowing the model to mitigate misclassification loss by incurring a predefined abstention penalty. This approach fundamentally differs from post-inference rejection systems, as abstention is integrated directly into the training process, enabling the model to learn to identify and manage uncertainty as an intrinsic component of its optimization objective [42]. Thulasidasan et al. (2019) exemplifies this, demonstrating that models can continue to learn the true class even while abstaining, progressively reducing their abstention rate as confidence grows [42]. The Deep Abstaining Classifier (DAC) offers advantages in simplicity and does not require assumptions about noise characteristics or the availability of clean data [42]. More recently, Schneider et al. (2024) introduced the Informed Deep Abstaining Classifier (IDAC), which extends the DAC by incorporating noise level estimations directly into the training process, providing a more sophisticated and context-aware mechanism for handling label imperfections [36].

Building upon the demonstrated efficacy of DAC and IDAC in classification, this thesis makes several significant contributions to bridge the aforementioned research gaps and advance the field of noise-robust medical image segmentation:

- Adaptation of Abstention to Segmentation: We initially investigate the applicability of the abstention mechanism to image segmentation by adapting the DAC and IDAC loss functions for this domain. This foundational step demonstrates that the benefits of abstention in mitigating label noise indeed extend beyond classification to pixel-level prediction tasks.

- Enhanced and Generalized Abstention Definition: While abstention was originally conceived as an extension specifically for Cross Entropy loss, our core contribution significantly improves and generalizes its definition. This enhanced framework incorporates an informed regularization term, guided by estimated noise rates $\tilde{\eta}$, and employs a flexible power-law-based α auto-tuning algorithm, offering superior control over abstention behaviour compared to prior approaches.

- Loss-Agnostic Integration and Novel Loss Functions: We rigorously demonstrate the profound versatility of this enhanced abstention mechanism by systematically integrating it with other distinct and widely utilized loss functions, each possessing unique characteristics and inherent noise-robust properties. This includes Generalized Cross Entropy (GCE) [51], Symmetric Cross Entropy (SCE) [45], and Dice Loss [29]. This systematic extension culminates in the introduction of three novel noise-robust loss functions: the Generalized Abstaining Classifier (GAC), the Symmetric

Abstaining Classifier (SAC), and the Abstaining Dice Segmenter (ADS). Notably, ADS introduces specific architectural adaptations for class-wise abstention and class-specific noise rates $\tilde{\eta}_c$, showcasing the mechanism's adaptability to the unique characteristics of Dice Loss for segmentation tasks.

- Empirical Validation of Robustness and Versatility: Through extensive empirical evaluations on challenging medical image datasets (CaDIS [13] and DSAD [3]) under varying noise levels and across distinct network architectures (U-Net [33] and DeepLabV3+ [4]), we provide compelling evidence for the consistent superiority of our proposed abstaining loss functions over their non-abstaining baselines. These results unequivocally establish abstention as a powerful and flexible mechanism to combat label noise, significantly enhancing the robustness and generalization capabilities of deep learning models in medical image segmentation.

This thesis thus contributes to the development of more reliable and trustworthy AI systems for diagnostic decision support, paving the way for robust medical image analysis in real-world scenarios characterized by imperfect data.

To systematically address the research questions posed, this thesis is structured to guide the reader from foundational concepts to novel contributions and empirical validation. The subsequent chapter first lays the theoretical groundwork, elucidating the principles of image segmentation and the baseline loss functions that underpin this research. To contextualize our contributions, Chapter 3 provides a comprehensive review of the existing literature on noise-robust learning, highlighting the current state-of-the-art and identifying the research gaps this work aims to address. Building upon this foundation, Chapter 4 presents the core methodological contribution of this thesis; it delineates the development of our generalized abstention framework and details its novel integration with diverse loss functions, culminating in the formulation of the GAC, SAC, and ADS methods. The theoretical efficacy of these novel methods is then subjected to rigorous empirical scrutiny in Chapter 5, which outlines the design of our experiments, including the datasets, noise simulation protocols, and network architectures employed. Chapter 6 is dedicated to the presentation and in-depth analysis of the experimental outcomes, substantiating the performance of our proposed loss functions through both quantitative metrics and qualitative visual comparisons. Finally, the thesis culminates in Chapter 7, which synthesizes the key findings, discusses the broader implications of this work, and proposes promising directions for future research in robust medical image analysis.

Chapter 2

Theoretical Fundamentals

This chapter establishes the essential theoretical framework required to understand the research presented in this thesis. The discussion begins by introducing the core principles of the supervised learning paradigm, including the concept of Empirical Risk Minimization, which governs how machine learning models are trained. Following this, we detail the fundamentals of Deep Neural Networks, the class of models employed in this work, explaining their basic architecture and learning mechanisms.

With this general theoretical groundwork laid, the chapter then narrows its focus to the specific application domain of this thesis: medical image analysis. We will delineate the fundamental differences between the computer vision tasks of image classification, object detection, and semantic segmentation, with a particular emphasis on the distinct challenges posed by the latter. Subsequently, the chapter elucidates the critical role of loss functions within the training process. Finally, we will review several baseline loss functions prevalent in the field, providing the necessary context for the novel contributions introduced in Chapter 4.

2.1 Supervised Learning and Risk Minimization

Supervised learning represents a fundamental paradigm in machine learning where an algorithm is tasked with learning a mapping from input data to desired output labels [1, 12]. This learning process relies on datasets comprising example input-output pairs, with the ultimate objective of approximating an unknown underlying function. The aim is for the trained algorithm to accurately predict outputs for novel, unseen data [12]. Models are thus trained on datasets containing both input features and their corresponding 'ground truth' or 'target' labels [1]. For instance, an input could be a set of patient characteristics, and the output could be a diagnosis. The model's primary challenge lies in discerning the intricate relationship between these inputs and outputs, thereby achieving robust generalization—the capacity to yield accurate predictions on data not encountered during its training [15]. While the availability of labelled datasets is ideal, their creation often presents practical difficulties, as high-quality, detailed annotations can be time-consuming and demand specialized expertise. This inherent challenge underscores a recurring theme in the application of supervised learning across various domains.

The training of a machine learning model is fundamentally an optimization problem,

meticulously guided by the principle of Empirical Risk Minimization (ERM) [30, 15]. ERM endeavours to identify a model that demonstrates optimal performance on the observed data [30]. Given that the true error a model might incur across the entire, unobserved data distribution is inherently unknown, the concept of "empirical risk" is employed. This serves as an estimate of the true error, derived from the average error the model exhibits on its finite, known training dataset [30]. The principle of ERM dictates that the learning process should select the model that minimizes this empirical risk [30].

However, the exclusive minimization of errors on the training data does not invariably guarantee robust performance on new, unseen data, particularly if the model's complexity is excessive. As illustrated in Fig. 2.1, such a scenario can lead to **overfitting** , a condition where a model learns the training data, including any idiosyncratic patterns specific to that set, with undue fidelity [15]. This results in superior performance on the training data but diminished accuracy on novel data, thus compromising generalization. Conversely, **underfitting** describes a model that is insufficiently complex to capture the underlying patterns within the training data [15], leading to suboptimal performance on both training and unseen datasets. Effective model training therefore necessitates a judicious balance: the model must possess sufficient complexity to discern genuine patterns without merely memorizing specific data points [15]. Regularization techniques are frequently employed to mitigate overfitting by imposing penalties on overly intricate models [12].

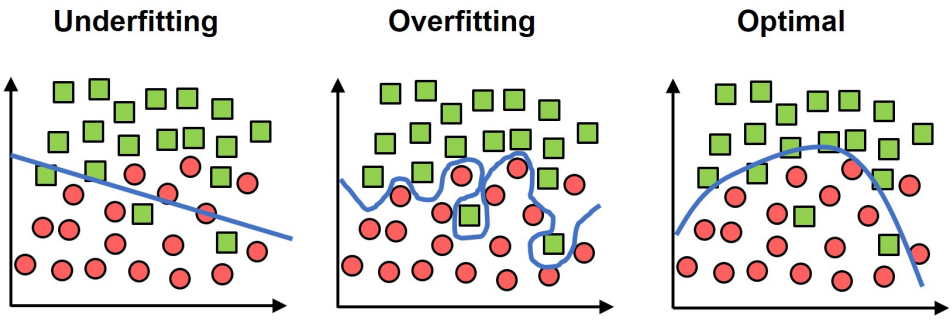


Figure 2.1: A conceptual illustration of the trade-off between model complexity and generalization. The underfitting model (left) is too simple to capture the underlying data trend. The overfitting model (center) is overly complex, memorizing the training data including its noise, which will lead to poor performance on unseen data. The optimal model (right) captures the general pattern of the data and is expected to generalize well.

Training machine learning models entails the continuous refinement of its internal parameters to diminish the divergence between its predictions and the actual ground truth values [1, 12]. This process typically employs optimization methods that leverage the concept of gradients to identify the direction of steepest error reduction [1, 2, 12]. The objective is to systematically adjust the model's parameters, thereby navigating towards a state of minimal error. The mathematical function employed to quantify this discrepancy, known as the loss function, is pivotal, as it dictates how errors are measured and, consequently, how the model's parameters are updated [1, 2, 12]. Distinct loss functions possess unique properties, influencing the training trajectory by penalizing different types of errors with varying magnitudes [12].

2.2 Fundamentals of Neural Networks

Deep Neural Networks (DNNs), commonly referred to as neural networks, represent a powerful computational methodology within artificial intelligence, drawing inspiration from the structural and functional organization of the human brain [12, 35]. These networks fall under the umbrella of deep learning and are characterized by an interconnected arrangement of computational units, termed 'neurons', organized into distinct layers [12]. Detailed in Fig. 2.2, a typical neural network comprises three principal types of layers: an input layer, one or more hidden layers, and an output layer. The **input layer** receives the raw data, which is subsequently processed and transmitted to successive layers. **Hidden layers** perform intermediate computations, relaying their transformed results onward. Deep neural networks are distinguished by the presence of multiple hidden layers, occasionally encompassing millions of these artificial neurons [12]. The **output layer** yields the network's final result, with the number of nodes in this layer being contingent upon the specific task, such as classifying data into various categories [12]. Each individual neuron receives inputs, performs a calculation involving a weighted sum followed by an activation function, and subsequently generates an output [12].

Neural networks are engineered to enable computational systems to learn from data, facilitate intelligent decision-making, and continuously refine their performance [12]. They exhibit remarkable proficiency in learning and representing complex, non-linear relationships inherent within data [12]. Their capability to process vast quantities of data and discern intricate patterns has catalysed significant advancements in diverse fields, including image recognition, natural language processing, and various forms of data analysis [20]. The capacity of neural networks to model non-linear and complex relationships directly stems from the incorporation of specialized 'activation functions' that introduce non-linearity into the network's computations [12, 9]. Without these non-linear transformations, a multi-layered network would effectively operate as a single-layer linear model, thereby lacking the requisite capacity to solve intricate real-world problems [9]. Consequently, the integration of non-linearity at each processing step is indispensable for neural networks to effectively comprehend complex data patterns.

Central to a neural network's learning efficacy are its **weights** and **biases**. These constitute the primary adjustable parameters within the network that are refined during the learning process [12]. Their continuous tuning during training is essential for minimizing prediction errors and enabling the network to accurately represent underlying data patterns [12]. Weights are numerical values assigned to the connections between neurons in distinct layers. They modulate the 'strength' or 'importance' of an input signal as it propagates from one neuron to the next [12]. Each connection possesses an associated weight that scales the signal traversing it; a higher weight implies a greater influence on the receiving neuron's output [12]. During the prediction (forward) process, inputs are multiplied by their respective weights, and these weighted inputs are then aggregated before transmission to the subsequent layer [12]. Biases are constant values that are added to the weighted sum of inputs prior to the application of an activation function [12]. In contrast to weights, which scale inputs, biases provide an offset, enabling a neuron to adjust its output independently of its input features. Biases are critical as they allow a neuron to produce a non-zero output even if all its inputs are zero, thereby preventing the network from becoming inoperative [12]. Typically, each neuron is associated with its own bias, providing flexible adjustments for different features [12].

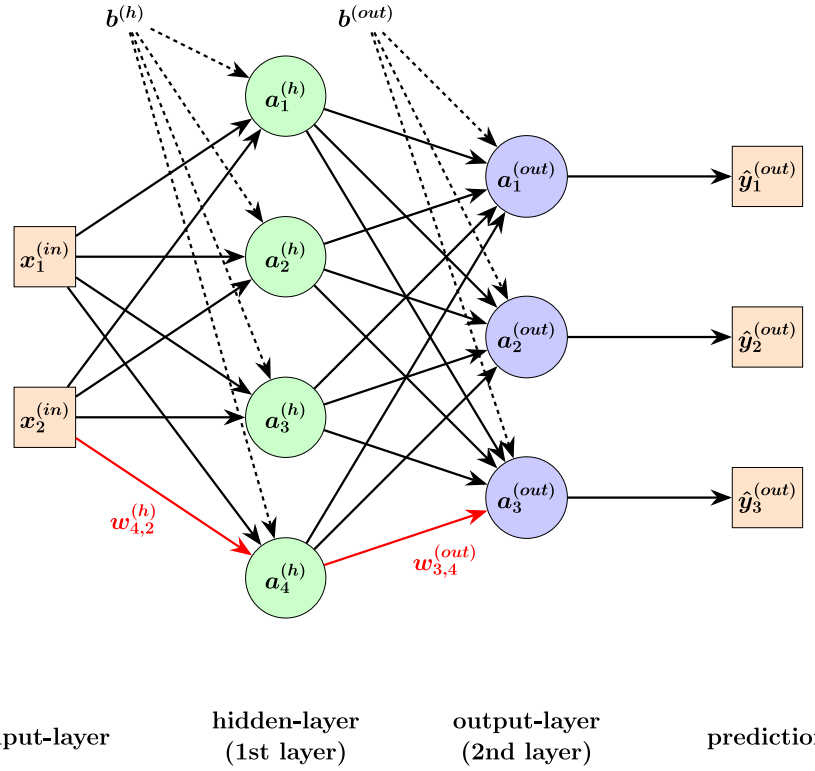
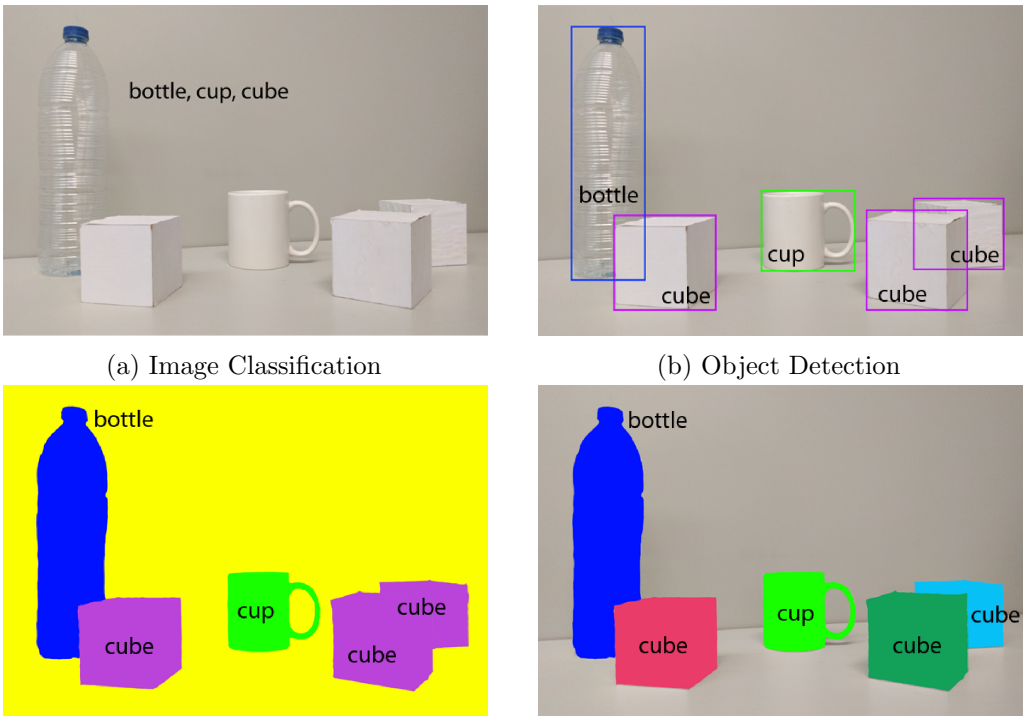


Figure 2.2: A conceptual diagram of a simple feedforward neural network, illustrating the flow of information and the role of its fundamental components. The network consists of three main parts: an **input layer** $x^{(\text{in})}$ which receives the raw data, a single **hidden layer** of neurons $a^{(h)}$ which performs intermediate, non-linear computations, and an **output layer** $a^{(\text{out})}$ which produces the final model predictions $\hat{y}^{(\text{out})}$. Each circular node $a_j^{(l)}$ represents the activation, or output value, of neuron j in a given layer l. The connections between neurons are modulated by **weights**, which are the primary learnable parameters of the model. The notation $w_{j,k}$ represents the weight of the connection from neuron k of the preceding layer to neuron j of the current layer. The figure highlights two such connections: $w_{4,2}^{(h)}$ is the weight connecting the second input unit $x_2^{(\text{in})}$ to the fourth hidden neuron $a_4^{(h)}$, while $w_{3,4}^{(\text{out})}$ connects the fourth hidden neuron to the third output neuron $a_3^{(\text{out})}$. In addition to weights, each neuron in the hidden and output layers has an associated **bias** (represented conceptually by $b_j^{(h)}$ and $b_j^{(\text{out})}$). During the training process, the network systematically adjusts all of these weights via backpropagation to minimize a loss function, thereby learning the optimal mapping from inputs to outputs.

The principal objective of training is to fine-tune these weights and biases to reduce the quantified error (loss) to its minimum possible value. This adjustment is predominantly accomplished through a process known as **backpropagation**, which efficiently computes the contribution of each weight and bias to the overall error [12, 34]. The calculated contribution, or gradient, precisely indicates the magnitude and direction of change required for each weight and bias to reduce the error. Optimization algorithms then utilize this information to update the parameters in the direction that decreases the error [12]. This iterative cycle—encompassing prediction, error quantification, and subsequent parameter adjustment—empowers the network to learn from its discrepancies and enhance its accuracy [12]. This continuous feedback loop, where the loss function signals error, backpropagation translates it into necessary adjustments, and optimizers apply those changes, allows the network to minimize its errors and acquire the desired mapping [12]. This dynamic process of adjusting internal representations based on error signals shows how changes to loss functions influence learning to achieve goals such as robustness to noise.

2.3 Image Analysis: from Classification to Segmentation



(a) Image Classification

(b) Object Detection

(c) Semantic Segmentation

(d) Instance Segmentation

Figure 2.3: A comparison of computer vision tasks, from coarse to fine-grained inference. (a) Image classification identifies classes present. (b) Object localization provides approximate locations with bounding boxes. (c) Semantic segmentation provides a precise, pixel-level mask for each class. (d) Instance segmentation distinguishes between individual objects of the same class. [8]

Within the field of computer vision, a variety of tasks exist, distinguished by the type and detail of their output. The most fundamental of these is **image classification**, where the goal is to assign a single, descriptive label to an entire image. In a medical setting, a classification model might analyse an X-ray of a bone and determine if it is 'Fractured' or 'Not Fractured'. The model processes the entire image and produces a single conclusion about its primary content. The modern era of deep learning-based classification was arguably launched by the performance of AlexNet, which demonstrated a significant leap in accuracy on general-purpose image recognition tasks and established the deep convolutional neural network (CNN) as the state-of-the-art approach [19].

The progression from high-level understanding to more granular, spatially-aware tasks is well illustrated in Fig. 2.3. At the highest level of abstraction is image classification. As shown in Fig. 2.3a, this task identifies the classes present in the scene—'bottle', 'cup', and 'cube'—without providing any information about their specific location, number, or boundaries. A step beyond this global analysis is **object detection**, which introduces a crucial layer of spatial awareness. As depicted in Fig. 2.3b, object detection not only identifies the classes but also provides their coarse spatial localization, typically in the form of bounding boxes. This provides an answer not just to *what* is in the image, but also *where* it is located.

In contrast, this thesis is concerned with **semantic segmentation**, a task of significantly greater detail and precision. Instead of the approximate localization offered by a bounding box, the objective of semantic segmentation is to classify every single pixel in the image, effectively partitioning it into semantically meaningful regions (Section 2.5.2). **Instance segmentation** takes this a step further by not only classifying each pixel but also differentiating between individual objects of the same class, as illustrated by the uniquely coloured cubes in Fig. 2.3d. In a clinical setting, semantic segmentation would be the equivalent of producing a precise map outlining all tumorous tissue within a brain scan, separating its pixels from those of healthy tissue. The introduction of Fully Convolutional Networks (FCNs) was a landmark achievement for this task, providing an elegant end-to-end framework for producing these dense prediction maps [26]. The move from the approximate bounding boxes of object detection to the dense pixel map of segmentation introduces significant challenges that make the latter a more demanding task. A primary challenge is the need to preserve and utilize fine-grained spatial information to generate exact outlines. This requires specialized network architectures capable of processing both the high-level context required for recognition and the low-level detail essential for precise boundary delineation [8].

Perhaps the most significant challenge, and a central theme in medical image analysis, is the immense burden of data annotation. The process of generating segmentation masks is not only extraordinarily time-consuming and expensive but is also subject to significant inter-observer and intra-observer variability, where different experts (or the same expert at different times) may produce inconsistent boundaries for the same anatomical structure [17]. This difficulty in acquiring large-scale, high-quality labelled datasets is frequently cited as a major bottleneck for the application of deep learning in medicine [23]. This reality, where ground truth labels can be scarce and inherently noisy, motivates a strong research focus on developing models that are robust to such imperfections, which is the central aim of our work.

2.4 Loss Functions in Machine Learning

The process of training a machine learning model is fundamentally an optimization problem. At its core is the loss function, also known as a cost or objective function, which serves as the primary guide for the learning process. A loss function is a mathematical formulation that quantifies the discrepancy, or 'loss', between a model's prediction and the corresponding ground truth label. This scalar value represents a measure of the model's error; the higher the loss, the worse the prediction. The ultimate goal of training is to iteratively adjust the model's internal parameters (weights and biases) to find a configuration that minimizes the value of this loss function [12, 6].

This minimization is typically achieved via a gradient-based optimization algorithm, such as stochastic gradient descent (SGD) or its variants. The loss function defines a high-dimensional 'loss landscape', and the optimizer's task is to navigate this landscape to find its lowest point. In each training iteration, the gradient of the loss with respect to the model's parameters is computed. This gradient indicates the direction of steepest ascent, so the parameters are updated in the opposite direction, taking a small step 'downhill'. The choice of loss function is therefore critical, as it directly shapes this landscape and dictates what the model considers an 'error', thereby influencing all subsequent parameter updates and having a significant impact on the model's final accuracy [12].

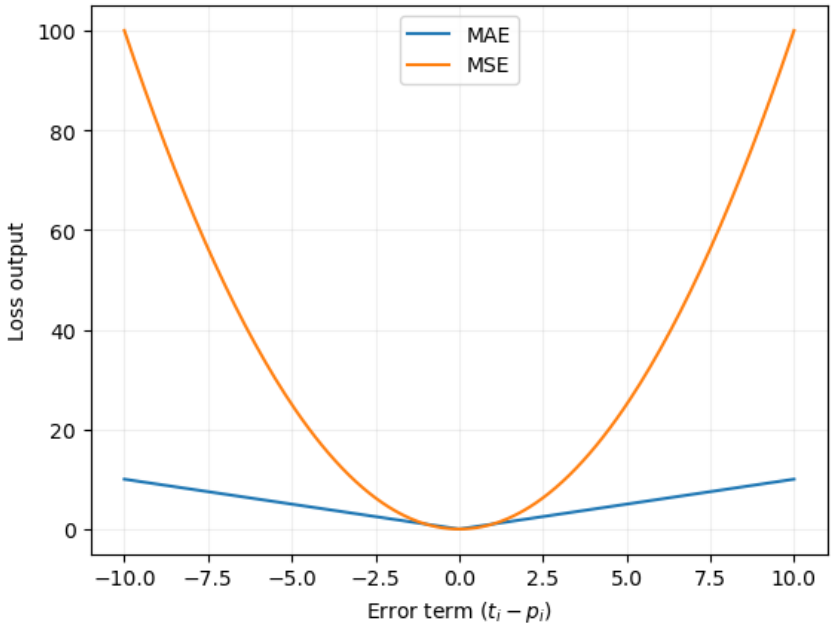


Figure 2.4: A visual comparison of the Mean Absolute Error (MAE) and Mean Squared Error (MSE) loss functions. The plot illustrates the differing penalties applied based on the magnitude of the prediction error. MSE's quadratic curve imposes a disproportionately large penalty on significant errors (outliers), while MAE's penalty increases linearly.

Different loss functions embody different mathematical properties and, consequently, prioritize penalizing different types of errors. This concept is best illustrated with two standard and relatively simple loss functions: Mean Absolute Error (MAE) and Mean Squared Error (MSE) [12, 6].

- Mean Absolute Error (MAE), or L1 Loss, is the average of the absolute differences between the predicted values p and the true values t:

$$\mathcal{L}_{MAE} = \frac{1}{n}\sum_{i=1}^{n}|t_i - p_i| \tag{2.1}$$

- Mean Squared Error (MSE), or L2 Loss, is the average of the squared differences between the predicted and true values:

$$\mathcal{L}_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(t_i - p_i)^2 \tag{2.2}$$

The mathematical difference between these two functions leads to distinct model behaviours, as depicted in Fig. 2.4. MSE, by squaring the error term, penalizes large errors far more severely than small ones. For example, an error of 10 contributes 100 times more to the total loss than an error of 1. This makes models trained with MSE highly sensitive to outliers, as the optimizer will be strongly incentivized to adjust the model to reduce these few, large errors. Conversely, MAE's penalty scales linearly with the error's magnitude, making it inherently more robust to outliers as anomalous data points do not disproportionately dominate the loss signal [6]. This trade-off between the stability of MAE and the strong penalization of MSE is a foundational concept in selecting

a loss function [12]. In the context of this thesis, where the data is presumed to contain noise, understanding these priorities is paramount, as the choice of loss function directly determines how the model will react to incorrect or ambiguous labels during training.

2.5 Baseline Loss Functions

To properly contextualize the contributions of this work, this section reviews the four foundational loss functions upon which our methods are built. Our analysis begins with Categorical Cross Entropy, the de facto standard for classification, whose properties and inherent vulnerabilities motivate much of the subsequent research. We then explore two direct responses to Cross Entropy's fallibility in the presence of label noise: Generalized Cross Entropy and Symmetric Cross Entropy. The discussion concludes with a paradigm shift, moving from distribution-based losses to the region-based Dice loss, whose unique formulation for measuring spatial overlap is particularly suited to segmentation tasks.

2.5.1 Cross Entropy Loss

The concept of Cross Entropy (CE) itself originates from the field of information theory, a discipline pioneered by Claude Shannon in his groundbreaking work 'A Mathematical Theory of Communication' [37]. Within this theory, 'entropy' measures the degree of randomness or uncertainty inherent in a system or an event. For instance, if an outcome is highly predictable, its entropy is low; if it's highly uncertain, entropy is high. Building upon this, CE extends the idea to measure the dissimilarity between two different probability distributions. In the context of machine learning, it quantifies how inefficiently one probability distribution (the model's predicted probabilities) encodes another (the true probability distribution of the labels). A lower cross entropy value indicates that the model's predicted probabilities are a closer and more accurate representation of the true underlying distribution.

As a loss function in machine learning, CE is predominantly utilized for classification tasks. It measures the performance of a classification model by evaluating how well its predicted probabilities align with the actual class labels. The objective during model training is to minimize this CE value, which directly translates to reducing the error between the model's predictions and the true outcomes. This function is particularly effective because it penalizes confident but incorrect predictions more severely than less confident ones, providing a strong signal for the model to learn from its mistakes.

For classification problems involving only two possible outcomes (e.g., 'yes' or 'no', 'spam' or 'not spam'), Cross Entropy is specifically referred to as Binary Cross Entropy (BCE) loss, also known as log loss. For a single data point x, its mathematical formulation is:

$$\mathcal{L}_{BCE}(x) = -\Big[t\log(p) + (1-t)\log(1-p)\Big] \tag{2.3}$$

In this context, t denotes the true binary label, taking a value of either 0 or 1. Conversely, p signifies the model's predicted probability that the sample belongs to the positive class (class 1), with $p \in [0, 1]$. When the true label t is 1, the term $(1-t)\log(1-p)$ evaluates to zero, simplifying the loss expression to $-\log(p)$. Under this condition, a high predicted probability $p \approx 1$ (approaching 1) for the true class results in a minimal loss. Conversely, a low predicted probability $p \approx 0$ for the true class yields a substantial loss, indicative of a

significant prediction error. Conversely, if the true label t is 0, the term $t\log(p)$ becomes zero, and the loss simplifies to $-\log(1-p)$. In this scenario, a high predicted probability $p \approx 1$ for the positive class leads to a large loss. This effectively penalizes the model for confidently predicting the incorrect class.

When a classification problem involves more than two possible outcomes (e.g., classifying images as 'cat', 'dog', or 'bird'), the concept extends to Categorical Cross Entropy loss (CCE). For a single data point, the formula is:

$$\mathcal{L}_{CCE}(x_j) = \sum_{i=1}^{k} t_i \log p_i \tag{2.4}$$

In this formula, k is the total number of classes. t_i is a binary indicator, which is 1 if the true class of the sample is i, and 0 otherwise. This is typically represented using a technique called 'one-hot encoding', where the true class is marked with a 1 and all other classes with a 0. p_i is the predicted probability that the sample belongs to class i. These probabilities are usually generated by a *softmax* activation function in the model's final layer, ensuring that all predicted probabilities sum up to 1. The objective remains the same: to minimize this loss, thereby making the model's predicted probability distribution as close as possible to the true distribution of the labels. In the scope of this work we are only interested in Categorical Cross Entropy and will henceforth refer to it as Cross Entropy loss (CE).

The dominating popularity of CE as the loss function of choice for classification tasks stems from several key strengths. Firstly, its direct connection to information theory provides a robust theoretical foundation for measuring the dissimilarity between predicted and true probability distributions. Secondly, CE effectively penalizes models that make confident but incorrect predictions more severely than those that are merely uncertain, providing a strong and clear signal for optimization. This characteristic is particularly beneficial for training Deep Neural Networks (DNNs), as it yields effective learning gradients that facilitate efficient weight adjustments during back-propagation, especially when paired with activation functions like *sigmoid* (for binary classification) or *softmax* (for multi-class classification). The probabilistic nature of its output aligns seamlessly with the goal of classification models to output probabilities for each class, making it an intuitive and powerful metric for evaluating how well a model's predictions align with the actual labels. Consequently, its ability to guide model improvement by iteratively adjusting parameters to reduce loss has cemented its role as a fundamental component in the training of robust classification models across various machine learning applications.

While CE is a cornerstone of modern classification tasks due to its probabilistic interpretation and strong performance with clean data, it exhibits a critical vulnerability in the presence of label noise. The CE loss operates by maximizing the log-likelihood of the ground truth class, which effectively encourages the model to produce high-confidence predictions (i.e., probabilities approaching 1.0) for the provided target labels [37]. When a label is incorrect due to annotation error, this mechanism becomes detrimental. The model is presented with a corrupted supervision signal and is thus penalized for predicting the true, underlying class. Due to the unbounded nature of the logarithm function, as the model becomes more confident in the correct class and assigns a correspondingly low probability to the incorrect label, the CE loss for that sample approaches infinity.

This generates large, disruptive gradients that force the model to memorize the erroneous annotations, leading to severe overfitting on the noisy samples [51, 45, 42]. Ultimately, this compromises the model's ability to learn robust, generalizable features, resulting in significantly degraded performance on clean, unseen data. This inherent vulnerability of CE loss to label noise has motivated the development of alternative loss functions designed to enhance robustness during DNN training.

2.5.2 Generalized Cross Entropy

To address the aforementioned challenges, Zhang and Sabuncu (2018) proposed the Generalized Cross Entropy (GCE) loss function as a theoretically grounded approach to improve the robustness of DNNs against noisy labels [51]. GCE can be conceptualized as a generalization that encompasses both the Cross Entropy loss and the Mean Absolute Error (MAE) loss. The motivation behind this generalization stems from the observation that while MAE has been theoretically shown to be robust to label noise under certain assumptions, its practical application with DNNs on complex datasets can lead to suboptimal performance and slower convergence. GCE aims to combine the noise robustness properties of MAE with the favourable training characteristics of CE, thereby offering a more versatile and effective solution for learning in the presence of label noise [51].

The mathematical formulation of the Generalized Cross Entropy loss for a single sample is given by:

$$\mathcal{L}_{GCE}(x_j) = \frac{1 - (p_{true})^q}{q} \tag{2.5}$$

where p_{true} represents the predicted probability of the true class. The crucial role of the parameter $q \in (0, 1]$ lies in its ability to control the behaviour of the loss function, effectively interpolating between CE and MAE. As q approaches 0, the GCE loss converges to the standard CE loss. Conversely, when q is set to 1, the GCE loss reduces to the MAE loss. This parameter thus provides a continuous spectrum of loss functions, allowing for fine-grained control over the penalty incurred for misclassifications. By adjusting q, the model can be configured to exhibit varying degrees of sensitivity to the predicted probabilities and, consequently, to the presence of label noise.

The noise-robustness of GCE is primarily attributed to the Box-Cox transformation implicitly applied to the predicted probabilities through the parameter q [51]. Unlike the logarithmic function in standard CE, which assigns an unbounded penalty to incorrect predictions, the power function $(p_{true})^q$ in GCE bounds the loss. This bounding mechanism is critical for robustness against label noise. When a label is incorrect, standard CE would impose a very large gradient, strongly pulling the model towards fitting that erroneous label [51]. In contrast, GCE, particularly with smaller values of q, mitigates the impact of these large gradients from mislabeled samples. This property prevents the model from overfitting severely to noisy labels, as the penalty for highly confident incorrect predictions is attenuated. By controlling the influence of individual misclassified samples, GCE enables DNNs to learn more effectively from corrupted datasets, leading to improved generalization performance in noisy environments [51].

2.5.3 Symmetric Cross Entropy

In an effort to address the same challenges as Zhang and Sabuncu (2018) and mitigate the adverse effects of label noise, Wang et al. (2019) proposed the Symmetric Cross Entropy (SCE) loss function as a robust alternative [45]. SCE was introduced to address the observed limitations of standard CE loss when training DNNs with noisy labels. Specifically, training with CE often leads to models that overfit to erroneous labels in 'easy' classes while simultaneously under-learning 'hard' classes [45]. This class-biased learning behaviour compromises the generalization performance of the model. SCE addresses these challenges by introducing a symmetrical formulation designed to simultaneously tackle both the overfitting to noisy labels and the under-learning of difficult classes [45].

The implementation of SCE loss for a single sample is defined as the sum of the standard Cross Entropy loss and a novel noise-tolerant Reverse Cross Entropy (RCE) loss. The RCE loss is defined as:

$$\mathcal{L}_{RCE}(x_j) = -\sum_{i=1}^{k} p_i \log(t_i) \tag{2.6}$$

While the standard CE term penalizes the model when its predicted probability for the true class is low, the RCE term operates distinctly. It essentially quantifies a form of inverse dissimilarity, guiding the model away from confident incorrect predictions [45]. Specifically, the RCE term imposes a penalty when the model assigns a high probability p_i to a class i for which t_i is 0 (i.e., it is not the true class). This is handled in practical implementations by considering only the true class's probability, or by employing a formulation that effectively penalizes confident incorrect classifications. The primary role of RCE is to provide a corrective signal towards the true label, even when the CE term might be misled by label corruption [45].

The authors introduce two weighting parameters, α and β, to control the contribution of the CE and RCE terms, respectively. The full formulation of the SCE loss is given by:

$$\mathcal{L}_{SCE}(x_j) = \alpha\mathcal{L}_{CE}(x_j) + \beta\mathcal{L}_{RCE}(x_j) \tag{2.7}$$

These hyperparameters allow for fine-tuning the balance between the two components of the loss function. By adjusting α and β, researchers and practitioners can control the relative importance of both the CE and the RCE penalties during training. This flexibility is crucial for optimizing performance across different datasets and varying levels or types of label noise. For instance, a higher β value would place more emphasis on the ability of RCE to prevent confident incorrect predictions, potentially beneficial in scenarios with high levels of symmetric label noise. Conversely, adjusting α allows for control over the CE's contribution [45]. The optimal values for α and β are typically determined through empirical experimentation and validation on a given dataset, as they are dataset-dependent and can significantly impact the model's ability to learn robustly from noisy labels.

The inherent noise-robustness of SCE is derived from this dual perspective and symmetrical penalty mechanism. The standard CE term, by design, strongly encourages the model to assign a high probability to the given label, rendering it susceptible to label noise. In contrast, the RCE term functions as a crucial counter-balance by penalizing highly confident erroneous predictions. In instances of label corruption, while the CE term might induce overfitting to the noisy label, the RCE term, actively mitigates the model's

propensity to become overly confident in misclassifications, by evaluating the predicted probabilities against the true (albeit noisy) labels [45]. This balanced, symmetrical combination ensures that the model is neither excessively penalized for minor deviations from the true label nor unduly influenced by the presence of noise. Consequently, this approach fosters a more stable and accurate learning trajectory, enabling DNNs to achieve superior generalization performance even when trained on datasets containing a substantial proportion of corrupted labels. In experiments across MNIST, CIFAR-10, and CIFAR-100, with various levels of label noise, SCE consistently and significantly outperformed CE and GCE. This enhanced performance is reflected in higher test accuracies and more stable learning processes, especially at high noise levels [45].

2.5.4 Dice Loss

The Dice Loss function, derived from the Dice Similarity Coefficient (DSC), constitutes a pivotal metric and optimization objective specifically built for segmentation tasks within the specialized field of image analysis, notably in medical image segmentation [29]. The DSC is a statistical measure employed to quantify the spatial overlap and similarity between two sets, frequently utilized to compare a predicted segmentation mask against its corresponding ground truth mask at a pixel-wise resolution. Diverging from conventional classification loss functions that evaluate individual pixel predictions (e.g., CE), The Dice loss prioritizes the volumetric or areal overlap between the predicted and true regions, rendering it exceptionally pertinent for tasks where the precise delineation of anatomical structures or pathological entities is paramount [29]. Its application is particularly pervasive in medical imaging, where accurate segmentation is indispensable for diagnostic precision, therapeutic planning, and scientific inquiry.

In the context of image segmentation, the mathematical formulation of the class-wise Dice Coefficient, can be expressed as:

$$\mathcal{DSC}(x_j) = \frac{2\sum_{i=1}^{N} p_i t_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} t_i} \tag{2.8}$$

Here, N represents the total number of pixels in the image. For each pixel i, $t_i \in \{0, 1\}$ denotes the ground truth label, and $p_i \in [0, 1]$ signifies the model's predicted probability (or binary prediction) for that pixel belonging to the foreground class. The term $\sum_{i=1}^{N} p_i t_i$ effectively calculates the sum of true positives (intersection) by element-wise multiplication, while $\sum_{i=1}^{N} t_i$ and $\sum_{i=1}^{N} p_i$ represent the sum of all ground truth positive pixels and predicted positive pixels, respectively. A small smoothing constant, ϵ, is typically added to the denominator to prevent division by zero, particularly in cases where both the predicted and ground truth masks are entirely empty. In the context of an optimization objective, the Dice loss is conventionally defined as $\mathcal{L}_{Dice}(x_j) = 1 - \mathcal{DSC}(x_j)$ with the training objective being to minimize this value, thereby maximizing the spatial overlap.

A pervasive challenge in numerous real-world classification and segmentation tasks, particularly pronounced in medical imaging, is class imbalance. Standard loss functions such as CE frequently exhibit suboptimal performance in such scenarios. CE loss, by virtue of its emphasis on overall pixel-wise accuracy, tends to disproportionately allocate

learning resources to the numerically dominant classes (e.g., background pixels in an image) while concurrently under-representing or neglecting the less frequent, yet often clinically critical, minority classes (e.g., a small tumor or lesion) [29]. This propensity can lead to models that, despite achieving high overall accuracy, demonstrate poor performance on the minority classes, effectively overfitting to the abundant background pixels and failing to accurately delineate the target object. For instance, if a model produces an imprecise segmentation of a small object within a large image, the standard CE loss might remain low due to the correct classification of the vast majority of background pixels, thereby providing a misleading indication of satisfactory model performance [29].

The class-wise nature of the Dice loss allows for a a significantly higher degree of robustness against class imbalance when compared to CE [29]. In contrast to CE, which assigns uniform importance to all pixels, the Dice loss intrinsically prioritizes the overlap of the foreground (or target) class. This characteristic imparts greater significance to the smaller, foreground class during the training process, effectively mitigating the overwhelming influence of background pixels. By directly optimizing for the similarity between the predicted and true positive regions, the Dice loss ensures that the model incurs a substantial penalty for failing to detect or for inaccurately segmenting the minority class, even if the overall pixel accuracy remains high due to accurate background predictions. This attribute renders the Dice loss exceptionally well-suited for medical image segmentation, where the structures of clinical interest (e.g., organs, tumors) often constitute a minute fraction of the total image area, yet their precise delineation is of paramount clinical importance [29].

Chapter 3

Related Work

This chapter delves into the advanced methodologies developed to enhance the resiliency of Deep Neural Networks (DNNs) against label noise in image segmentation. Moving beyond rudimentary adaptations of classification strategies, these sophisticated approaches encompass a spectrum of innovations. We explore techniques ranging from granular pixel-wise and image-level noise handling, including graph-based corrections and uncertainty estimation, to spatially-aware noise modelling that explicitly accounts for boundary distortions. Furthermore, the chapter examines methods that exploit the intrinsic learning dynamics of deep networks for adaptive label correction, alongside multi-network and co-training paradigms designed to enhance robustness through collective intelligence. The discussion extends to the development of inherently robust loss functions tailored for segmentation, and finally, to emerging paradigms of pretraining and bootstrapping that leverage large-scale noisy labels as a valuable resource for feature learning.

3.1 Pixel-wise and Image-level Noise Handling

Early advancements in mitigating label noise for image segmentation involved strategies that operate at different granularities of annotation. Some methods focused on addressing noise at the individual pixel level. For instance, approaches have been proposed to learn spatially adaptive weight maps, which dynamically adjust the contribution of each pixel to the loss based on a meta-reweighting framework [38]. Other pixel-wise strategies involve training multiple networks simultaneously, such as a tri-network system, where pairs of networks collaboratively select reliable pixels to guide the learning of a third network, extending the co-teaching method [38]. Furthermore, disagreement strategies have been employed to develop label-noise-robust methods, where models are updated only on pixel-wise predictions where two models exhibit divergence [38].

Complementing pixel-level approaches, a second category of methods concentrated on image-level noise estimation and learning. These techniques introduce label quality evaluation strategies to assess the overall quality of image-level annotations. Based on this assessment, the loss function is re-weighted to fine-tune the network [38, 54]. For example, Zhu et al. (2019) proposed an automatic quality evaluation module coupled with an overfitting control module. This system assesses the relative quality of labels within the training set and leverages these insights to re-weight the loss, thereby prioritizing training on more reliable annotations [54].

A more comprehensive understanding of segmentation noise necessitates considering both pixel-level and image-level information, as they capture distinct yet complementary aspects of annotation quality and noise distribution. Many existing methods have historically focused predominantly on either pixel-wise noise estimation or image-level quality evaluation [38]. However, a truly robust solution for label noise in image segmentation benefits from integrating both perspectives. This is because a complete assessment of label noise degree in a segmentation task involves not only judging whether image-level labels are noisy but also identifying which specific pixels within an image possess noisy labels [38].

Shi et al. (2021) introduced PINT (Pixel-wise and Image-level Noise Tolerant learning), a novel two-phase framework for medical image segmentation with noisy labels, designed to distil effective supervision from both pixel and image levels [38]. In its first phase, PINT employs a pixel-wise noise estimation method that explicitly quantifies the uncertainty of every pixel, for example, by using the entropy of predictions under various perturbations. This phase then guides robust learning by combining original noisy labels with generated pseudo labels. The second phase extends this concept to image-level robust learning, where image-level uncertainty is derived from the aggregated pixel-wise uncertainties. This image-level approach is crucial as it accommodates additional information and serves as a complement to pixel-level learning, particularly in scenarios where clean pixels, such as those lying in boundary regions, might exhibit high uncertainty [38]. The integration of these multi-granular perspectives leads to a more holistic and effective mitigation strategy.

Adding to the pixel-level noise handling, Yi et al. (2021) propose a novel perspective for semi-supervised semantic segmentation by formulating it as a problem of learning with pixel-level label noise [49]. They observe that pixel-level labels generated from weak supervisions, such as Class Activation Maps (CAM) [52], inevitably contain noise [49]. To address this, they introduce a graph-based label noise detection and correction framework [49]. This framework first trains a clean segmentation model using a small set of strong pixel-level annotations to detect reliable labels from the noisy CAM-generated labels based on cross entropy loss [49]. Subsequently, a superpixel-based graph is constructed to represent the spatial adjacency and semantic similarity between pixels within an image [49]. Finally, a Graph ATtention network (GAT) [44] is employed to correct the noisy labels, supervised by the detected clean labels [49]. This approach is notable for being one of the first to explicitly tackle pixel-level label noise in semi-supervised semantic segmentation by modelling the spatial relationships between pixel labels [49].

3.2 Spatially-Aware Noise Modelling and Correction

Recognizing that segmentation label noise is inherently spatially correlated and often biased, recent research has focused on developing techniques that explicitly model these characteristics. A significant contribution in this area is the introduction of novel noise models tailored to the unique properties of segmentation annotations. Yao et al. (2023) proposed a Markov model for segmentation noisy annotations, which explicitly encodes both spatial correlation and inherent bias [47]. This model is designed to simulate realistic annotation scenarios where human annotators delineate object boundaries, conceptualizing the noisy boundary as a random yet continuous distortion of the true boundary. The model employs Bernoulli variables at each step to control decisions of expansion or shrinkage and their spatially-dependent strength along the boundary. Furthermore, it incorporates a

random flipping noise to account for sparse mislabels that might occur in regions distant from the boundary [47]. The emphasis on simulating real annotation processes and continuous boundary distortions underscores the understanding that accurate noise models are crucial for effective mitigation. If the underlying model of noise is inaccurate, any subsequent correction mechanism built upon it may be misdirected or suboptimal. This highlights the importance of developing and utilizing noise models that precisely reflect the complex, spatially correlated, and biased nature of real-world annotation processes in segmentation, as such realistic models can lead to more fundamentally sound and precisely targeted noise mitigation strategies.

Building upon these spatially-aware noise models, methods for iterative label correction and bias removal have been developed, often leveraging small clean validation sets. Yao et al. (2023) introduced Spatial Correction (SC), an algorithm that progressively recovers true labels by removing the bias inherent in noisy annotations [47]. This approach acknowledges that correcting model bias without any ground truth reference is challenging, thus necessitating a small, clean validation set of well-curated annotations to estimate and rectify the noise-induced bias. Theoretical guarantees support the method's correctness, demonstrating that even a minimal amount of validation data, such as a single image annotation, can be sufficient for bias correction in practice [47]. SC operates as an iterative process, repeatedly training a segmentation model and correcting labels until convergence. A key advantage of SC is its independence from the specific DNN training process, allowing it to be agnostic to the backbone DNN architecture and compatible with various segmentation models [47]. The iterative nature of SC implies a self-refinement process, where the model's improving predictions are continuously used to enhance label quality, leading to a more robust learning outcome.

Further contributing to spatially-aware noise mitigation, Shu et al. (2019) introduced LVC-Net for medical image segmentation, specifically addressing the susceptibility of CNNs to annotation noise due to a lack of semantic guidance [39]. Their approach focuses on automatic label error correction by leveraging intrinsic visual information. LVC-Net captures Local Visual Cues (LVCs), which are local visual saliency regions, from low-level feature channels, recognizing that the front-end of a network is less affected by supervised signals [39]. A deformable spatial transformation module is integrated into the network to establish visual connections between the model's predictions and these LVCs, providing extra freedom for local receptive fields to navigate classification boundaries away from incorrect labels [39]. This is complemented by a novel loss function that combines noisy labels with image LVCs, exploiting their intrinsic spatial relationship to effectively suppress the influence of label noise through potential visual guidance during the learning process [39].

3.3 Exploiting Learning Dynamics and Adaptive Correction

A significant area of advancement in combating label noise involves capitalizing on the intrinsic learning dynamics of DNNs. A key observation, initially noted in classification, is the "early-learning" phenomenon: deep segmentation networks tend to first accurately fit clean pixel-level labels before eventually memorizing and overfitting to false annotations [25, 48]. This phenomenon is particularly nuanced in semantic segmentation, as memoriza-

tion does not occur simultaneously for all semantic categories [25]. This understanding transforms a potential vulnerability of DNNs (memorization) into a valuable diagnostic signal. The ability to detect the transition point from early learning to memorization offers a precise moment for intervention, allowing for adaptive strategies that exploit the "clean" phase of learning. This implies that monitoring the internal learning dynamics provides an intrinsic and powerful signal for identifying and managing label noise, enabling the design of adaptive intervention strategies.

Based on these learning dynamics, adaptive correction mechanisms have been developed that adjust interventions based on category-specific memorization and performance. Liu et al. (2022) proposed ADELE (ADaptive Early-Learning correction), a method specifically designed for segmentation with noisy annotations [25]. ADELE detects the onset of the memorization phase independently for each semantic category during training, which allows it to adaptively correct noisy annotations to leverage the early-learning phase for individual classes. This detection is achieved by monitoring the deceleration of the Intersection over Union (IoU) curve (between model output and noisy annotations) through the fitting of an exponential parametric model [25].

Similarly, Rong et al. (2023) advocate for a shift in focus for Weakly Supervised Semantic Segmentation (WSSS) from merely generating pseudo-labels to emphasizing robust learning with noisy labels [32]. Their Boundary-enhanced Co-training (BECO) method incorporates a co-training paradigm to improve the learning of uncertain pixels and a boundary-enhanced strategy to boost predictions in challenging boundary areas. This approach is motivated by the observation that mislabeled pixels are predominantly concentrated on boundaries, making their accurate prediction critical for overall performance [32].

Furthermore, multi-scale consistency regularization plays a vital role in enhancing the robustness of these adaptive correction strategies. Liu et al. (2022) integrated a regularization term into ADELE that enforces consistency across multiple scales, thereby bolstering robustness against annotation noise [25]. This multi-scale consistency regularization contributes to more accurate corrected annotations by providing an additional supervision signal. This mechanism is crucial as it helps prevent the network from exclusively training on and overfitting to the noisy segmentation annotations, stabilizing the model's internal representations and making them more robust before label correction is applied [25].

3.4 Multi-Network and Co-Training Paradigms

A prominent and effective trend in mitigating label noise involves the deployment of multi-network architectures and co-training paradigms. The underlying principle is to leverage the collective intelligence or disagreement among multiple models to filter out erroneous signals and enhance learning on uncertain data. If individual models are prone to overfitting to noise, their consensus or divergence can serve as a more reliable indicator for identifying and correcting noisy labels, thereby shifting robustness from a single model's internal mechanism to a systemic property [21, 32].

Several approaches exemplify this paradigm. Shi et al. (2021) describe prior works that train three networks concurrently, where each pair of networks collaboratively identifies and selects reliable pixels to guide the learning of the third network, extending the concept of

co-teaching [38]. This cooperative learning environment helps to cross-validate predictions and reduce the impact of individual model errors.

Rong et al. (2023) proposed a co-training framework within their BECO method, consisting of two parallel deep networks [32]. These networks are designed to interactively teach each other about potentially noisy pixels. By imposing consistency constraints on their predictions, the framework aims to rectify the semantic information for uncertain pixels, which are often characterized by low confidence in their initial pseudo-labels. In this setup, high-confidence pixels continue to be supervised by their original pseudo-labels, while the online predictions from a peer network guide the learning for low-confidence areas [32].

To further enhance learning and reduce confirmation bias, Li et al. (2023) introduced an approach for semi-supervised semantic segmentation that employs two diverse learning groups, each equipped with different network architectures [21]. This diversity is crucial as it encourages complementary learning paths and reduces the likelihood of both groups making the same errors [21]. Each learning group comprises a teacher network, a student network, and a novel filter module [21]. The filter module of one learning group utilizes pixel-level features from its teacher network to detect incorrectly labelled pixels [21]. To explicitly mitigate confirmation bias—where models might reinforce their own prediction errors—the labels cleaned by the filter module from one learning group are used to train the other learning group [21]. This sophisticated mutual knowledge distillation mechanism fosters a more robust learning environment [21].

A specialized strategy within this multi-network context, particularly relevant in weakly-supervised settings, is boundary-enhanced learning. Rong et al. (2023) proposed a boundary-enhanced co-training (BECO) framework to improve prediction accuracy in challenging boundary regions [32]. This is achieved by assigning a higher weight to these areas in the loss function. A key innovation is the construction of artificial boundaries with accurate labels. This process involves copying and pasting high-confidence areas from one image to another, leveraging the observation that high-confidence pixels typically reside within objects and are often correctly predicted, even if incomplete. These artificially constructed boundaries provide reliable ground truth signals for the model to learn precise object contours, which are notoriously difficult to segment accurately under noisy supervision [32].

3.5 Robust Loss Functions Tailored for Segmentation

An alternative, and often more fundamental, approach to mitigating label noise involves designing loss functions that are inherently robust to outliers and noise in segmentation masks. This strategy contrasts with methods that rely on architectural modifications or complex training procedures, offering a potentially simpler and more elegant solution. The core idea is that if the loss function itself possesses inherent robustness, it can naturally down-weight the influence of noisy samples during optimization without requiring explicit noise detection or correction modules, thereby streamlining the overall learning framework [10, 48].

Gonzalez-Jimenez et al. (2023) introduced the T-Loss, a novel robust loss function specifically designed for medical image segmentation [10]. The T-Loss is based on the

negative log-likelihood of the Student-t distribution, which is known for its heavier tails compared to the Gaussian distribution, making it effective in handling outliers in data [10]. A key advantage of the T-Loss is its ability to control sensitivity to noise with a single parameter that is adaptively updated during the backpropagation process [10]. This eliminates the need for additional computation or prior information about the level and spread of noisy labels [10]. The authors demonstrated that while many traditional robust loss functions are vulnerable to memorizing noisy labels, the T-Loss maintains performance even under high noise contamination [10]. The dynamic adjustment of this parameter during the early stages of training, irrespective of its initial value, automates a critical aspect of noise handling and reduces reliance on manual hyperparameter tuning or external noise estimation [10].

Similarly, Ye et al. (2024) proposed the Active Negative Loss (ANL) framework, which incorporates Normalized Negative Loss Functions (NNLFs) as a new class of robust passive loss functions [48]. They argue that conventional passive loss functions, such as Mean Absolute Error (MAE), treat clean and noisy samples equally, potentially hindering convergence [48]. NNLFs are designed to address this by focusing more on "memorized clean samples," implicitly leveraging the early learning phase through the loss function's design [48]. By replacing MAE with NNLFs within the ANL framework, they demonstrated improved robustness across various types of label noise, extending its application to image segmentation tasks [48]. Ye et al. (2024) also address label imbalance in non-symmetric noise scenarios by proposing an entropy-based regularization technique. This technique aims to encourage more balanced model output marginal probabilities, highlighting the necessity for specialized handling of different, complex noise types [48]. The successful extension of ANL to segmentation tasks further demonstrates its versatility and potential impact.

3.6 Pretraining and Bootstrapping with Noisy Labels

An emerging and significant paradigm in mitigating label noise involves leveraging large-scale, readily available noisy labels for pretraining segmentation models. This approach represents a crucial shift in perspective, transforming label noise from a mere problem into a valuable resource. Given the high cost of acquiring perfectly clean annotations and the abundance of noisy data, this strategy offers a scalable solution for enhancing feature learning, particularly in data-scarce domains like remote sensing [24]. If models can learn robust, generalizable features from noisy pretraining, it significantly reduces the annotation burden for downstream tasks.

Liu et al. (2024) introduced CromSS (Cross-modal Sample Selection), a weakly supervised pretraining strategy designed to exploit massive amounts of noisy and easily obtainable labels for improving feature learning in remote sensing image segmentation [24]. This approach is motivated by the observation that encoders, even when pretrained with noisy labels, can still acquire robust features [24]. CromSS investigates the optimization of multi-modal pretraining architectures, including both middle and late fusion strategies, by utilizing complementary modalities such as Sentinel-1 Synthetic Aperture Radar (SAR) and Sentinel-2 optical data [24]. A core component of CromSS is its cross-modal sample selection module, which employs a cross-modal entangling strategy to refine estimated confidence masks within each modality, thereby guiding the sampling process [24]. Additionally, a spatial-temporal label smoothing technique is incorporated to counteract overconfidence

and enhance robustness during pretraining [24].

Complementing pretraining, bootstrapping methods dynamically adjust the influence of observed and pseudo-labels, enabling implicit relabelling during training. Zhou et al. (2024) introduced L2B (Learning to Bootstrap), a method that empowers models to self-bootstrap using their own predictions without being negatively impacted by erroneous pseudo-labels [53]. L2B achieves this by dynamically adjusting the importance weight between real observed labels and generated pseudo-labels, as well as between different samples, through a meta-learning framework [53]. This approach offers a new, versatile objective that facilitates implicit relabelling concurrently, without the need to explicitly generate new training targets [53]. This subtle yet crucial distinction from prior bootstrapping methods, which often used fixed weighted combinations, allows for more flexible and powerful adaptive mechanisms for noise mitigation [53]. L2B's dynamic, meta-learned weighting, which allows weights to not sum to one, contributes to its ability to implicitly relabel, thereby refining the training process more effectively.

Chapter 4

Abstaining Loss Functions for Robust Learning

This chapter builds upon the concept of abstention, a noise-robust mechanism for deep learning introduced by the Deep Abstaining Classifier (DAC) and later extended by the Informed Deep Abstaining Classifier (IDAC). These foundational methods established a paradigm where a model can refrain from making predictions on unreliable data, thereby mitigating the negative impact of label noise during training. The primary contribution of this thesis is to significantly expand upon this concept, demonstrating that abstention is a highly versatile and modular mechanism. We present a generalized abstention framework and integrate it with three prominent and diverse loss functions: Generalized Cross Entropy (GCE), Symmetric Cross Entropy (SCE), and Dice Loss. This results in three novel loss functions: the Generalized Abstaining Classifier (GAC), the Symmetric Abstaining Classifier (SAC), and the Abstaining Dice Segmenter (ADS). Our approach introduces key improvements to the original definition, making the abstention mechanism more versatile and powerful, including specific architectural adaptations to ensure its compatibility with class-wise loss functions like Dice Loss. This work provides a robust and adaptable tool set for developing more reliable models, particularly for challenging medical image segmentation tasks.

4.1 Deep Abstaining Classifier

Deep Neural Networks (DNNs) trained in supervised settings are highly susceptible to label noise, a common impediment in large-scale datasets. Conventional classification models are compelled to assign a definitive class prediction to every input, even when the provided label is erroneous. This forced commitment to potentially corrupted labels can lead to overfitting of the noise, thereby degrading the model's generalization performance on clean data. To counteract this inherent vulnerability, Thulasidasan et al. (2019) proposed the Deep Abstaining Classifier (DAC) loss function and thereby introduced a paradigm shift [42]. The DAC loss function enables DNNs to abstain from making a prediction on samples deemed ambiguous or unreliable, rather than enforcing a potentially incorrect classification. This approach is motivated by the principle that a model's uncertainty or the inherent unreliability of a label should not necessarily result in a high misclassification penalty, thus providing a mechanism to circumvent learning from potentially erroneous

information [42].

The abstention mechanism is a foundational element of the DAC. It is implemented by extending the network's output layer from k conventional class units to $k+1$ units, where the $(k+1)$-th unit explicitly represents the abstention option. During inference, the network's output, typically obtained via a *softmax* activation, yields a probability distribution over these $k+1$ options: $\boldsymbol{p} = \{p_0, p_1, \ldots, p_k, p_{k+1}\}$, where $p_i : i \in \{1, \ldots, k\}$ denotes the probability of assigning the sample to class i and p_{k+1} denotes the probability of abstaining. The core of the DAC's training objective is encapsulated in its unique definition:

$$\mathcal{L}_{DAC}(x_j) = (1 - p_{k+1}) \left(- \sum_{i=1}^{k} t_i \log \frac{p_i}{1 - p_{k+1}} \right) + \alpha \log \frac{1}{1 - p_{k+1}} \tag{4.1}$$

Here, t_i is a binary indicator for the true class i, and α represents the abstention penalty. The first term, $(1 - p_{k+1}) \left(- \sum_{i=1}^{k} t_i \log \frac{p_i}{1-p_{k+1}} \right)$, is a modified CE loss applied over the k non-abstaining classes. The crucial factor $\frac{p_i}{1-p_{k+1}}$ re-normalizes the class probabilities by effectively distributing the remaining probability mass $1 - p_{k+1}$ among them. This ensures that the model's discriminative learning capacity is focused on the subset of data it deems classifiable, compelling it to make a definitive 'hard' decision only if it has chosen not to abstain. The second term $\alpha \log \frac{1}{1-p_{k+1}}$, called the 'regularization' term, functions as a direct penalty for abstention, with $\alpha \geq 0$ controlling its severity. A higher α discourages abstention, pushing p_{k+1} towards zero and recovering the standard Cross Entropy loss (CE), while a lower α encourages abstention.

The interplay between these two terms is central to DAC's functionality. The model's optimization process continuously seeks to minimize both the misclassification error on samples it chooses to classify and the penalty associated with abstention. This implies that for any given sample, the optimal value for p_{k+1} is achieved when the marginal benefit of reducing classification error (by increasing p_{k+1} is precisely balanced by the marginal cost of the abstention penalty [42]. This sophisticated balancing act allows the DAC to adaptively manage its confidence and uncertainty. Empirical observations indicate that the DAC progressively reduces its abstention rate as its classification performance on true classes improves, ultimately learning to abstain only on the most genuinely confusing samples. Furthermore, a pivotal analytical result guarantees that the learning process for the true classes persists during gradient descent, even when the model is actively abstaining, ensuring that the underlying objective of correctly classifying the true class is not undermined [42].

To circumvent the challenge of manually tuning α and to guide he learning process, the DAC framework incorporates an adaptive α auto-tuning algorithm, detailed in Algorithm 1. This algorithm dynamically adjusts the value of α during training. The process involves three phases: an initial warm-up period of L epochs with abstention-free training, during which a smoothed moving average $\tilde{\beta}$ of an implicit α threshold is maintained. Then, at the start of epoch $L+1$, α is initialized to a value significantly smaller than $\tilde{\beta}$ (specifically, $\alpha := \tilde{\beta}/\rho$, where ρ is an initialization factor), encouraging initial broad abstention. Subsequently, α is linearly increased over the remaining epochs until it reaches a final value α_{final}. This gradual increase in α progressively raises the penalty associated with abstention, compelling the model to incrementally reduce its abstention rate as it gains confidence and refines its classification abilities on the true classes.

Algorithm 1 α auto-tuning

Input: total iter. (T), current iter. (t), total epochs (E), abstention-free epochs (L), current epoch (e), α init factor (ρ), final α (α_{final}), mini-batch cross entropy over true classes ($H_c(P^M_{1...K})$)

α_{set} = False
for $t := 0$ to T **do**
  if $e < L$ **then**
    $\beta = (1 - P^M_{k+1})H_c(P^M_{1...K})$
    if $t = 0$ **then**
      $\tilde{\beta} = \beta$ {// initialize moving average}
    end if
    $\tilde{\beta} \leftarrow (1 - \mu)\tilde{\beta} + \mu\beta$
  end if
  if $e = L$ and not α_{set} **then**
    $\alpha := \tilde{\beta}/\rho$ {// initialize α at start of epoch L}
    $\delta_\alpha := \frac{\alpha_{final} - \alpha}{E - L}$
    $update_{epoch} = L$
    α_{set} = True
  end if
  if $e > update_{epoch}$ **then**
    $\alpha \leftarrow \alpha + \delta_\alpha$ {// then update α once every epoch}
    $update_{epoch} = e$
  end if
end for

The inherent noise-robustness of the DAC is directly attributable to this abstention capability. By empowering the model to abstain on ambiguous or noisy samples, the DAC effectively mitigates the detrimental impact of label noise during the training process [42]. When confronted with corrupted labels, the model is not compelled to internalize these errors; instead, it can learn to identify and effectively 'filter' such instances by opting for abstention. For structured noise, the DAC learns features associated with unreliable labels, while for unstructured noise, it functions as an implicit data cleaning mechanism by identifying and effectively disregarding noisy training data [42]. This selective learning, guided by the abstention penalty and its adaptive tuning, allows the DNN to focus its learning capacity on reliable data, leading to improved generalization performance. A significant advantage of the DAC approach is its remarkable architectural agnosticism; it can be seamlessly integrated with virtually any existing DNN architecture by simply adding a single $(k + 1)$-th output neuron—or adding an output channel in the case of segmentation—to the network's final layer [42].

The DAC distinguishes itself from traditional rejection classification systems, which typically operate in a post-processing setting, by integrating abstention directly into the training process [42]. This allows the model to learn to identify and manage uncertainty as an intrinsic component of its optimization objective. Furthermore, unlike some other label noise handling methods, DAC does not require detailed modelling of label flipping probabilities, assume a specific amount of noise, or necessitate the existence of a trusted, clean dataset. It is also notably simpler than approaches based on mentor-student networks or complex graphical models, offering a practical and effective solution for robust learning

in noisy environments [42].

Benchmark results on FashionMNIST, CIFAR-10, and CIFAR-100 have demonstrated the efficacy of the DAC loss function in image classification. By allowing the model to abstain from making predictions on confusing samples, DAC significantly improves classification performance compared to CE and GCE. This makes DAC a valuable tool for training DNNs in noisy environments, ensuring more accurate and reliable model predictions [42].

4.2 Informed Deep Abstaining Classifier

Building upon the foundational principles of the DAC loss, the Informed Deep Abstaining Classifier (IDAC) loss, proposed by Schneider et al. (2024), emerges as a recent extension specifically designed to further enhance noise robustness in diagnostic decision support systems, particularly within the challenging domain of medical image analysis [36]. While DAC introduced the crucial ability for a DNN to abstain from classifying ambiguous or noisy samples, IDAC addresses the need for a more sophisticated and context-aware mechanism to handle label imperfections. The primary motivation behind IDAC is to provide a more robust training framework for models operating in environments where label noise is prevalent and can significantly impact diagnostic accuracy and reliability [36].

IDAC extends the DAC loss function by fundamentally altering how the model perceives and reacts to noisy data during training. The key innovation lies in its incorporation of noise level estimations directly into the learning process. Unlike DAC, which primarily relies on the model's internal uncertainty to trigger abstention, IDAC leverages an explicit understanding or estimation of the noise associated with each training sample [36]. This integration of estimated noise levels is precisely what renders IDAC more 'informed'. By providing the loss function with an additional signal about the probable corruption of a label, the model can make more judicious decisions regarding when to classify and when to abstain, moving beyond mere uncertainty to a more targeted response to noise [36].

The core of IDAC's extension lies in its modified abstention regularization term. For a given sample x_j, the IDAC loss function is defined as:

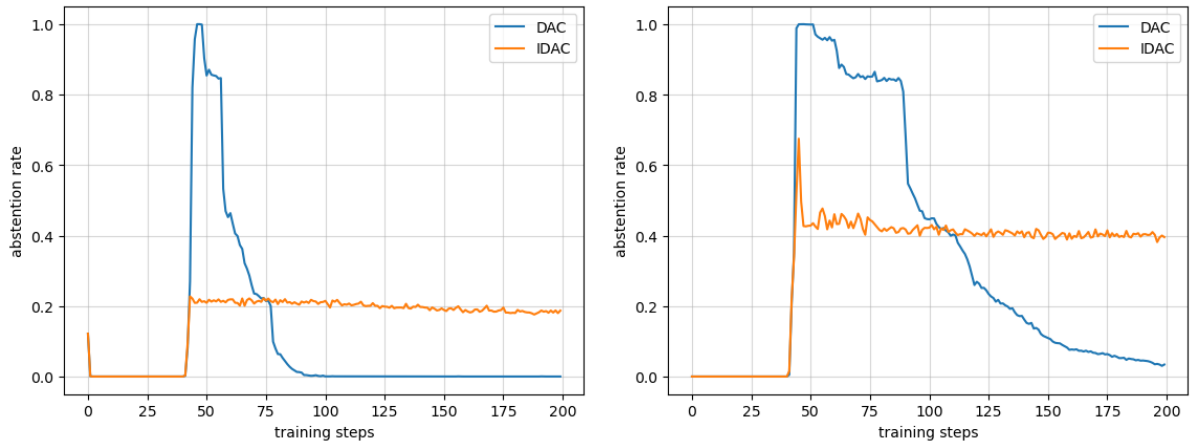
$$\mathcal{L}_{IDAC}(x_j) = (1 - p_{k+1}) \left(- \sum_{i=1}^{k} t_i \log \frac{p_i}{1 - p_{k+1}} \right) + \alpha(\tilde{\eta} - \hat{\eta})^2 \tag{4.2}$$

The first term remains consistent with the modified classification term of DAC, ensuring that the model's discriminative learning focuses on non-abstained samples. The critical difference resides in the regularization term, which now incorporates two key noise-related parameters: $\tilde{\eta}$ and $\hat{\eta}$. The parameter $\tilde{\eta}$ represents a fixed, prior estimation of the dataset's expected label noise, acting as a hyperparameter that is not adapted during training. Conversely, $\hat{\eta}$ denotes the currently applied abstention rate of the classifier per batch, approximated by summing the *softmax* outputs of the abstaining neuron across the batch and dividing by the batch size N (i.e., $\hat{\eta} = \sum_{l=1}^{N} \frac{p_{l,k+1}}{N}$). The hyperparameter α in IDAC, unlike in DAC, remains constant throughout training and is independent of the number of training epochs. This modified regularization term penalizes the squared difference between the expected noise level $\tilde{\eta}$ and the model's current approximated abstention rate

$\hat{\eta}$, thereby encouraging the model's abstention behaviour to align with the prior knowledge of noise in the dataset. If the model abstains on too many samples compared to $\tilde{\eta}$, or too few, this term increases, leading to a stronger penalization that guides the model towards an abstention rate consistent with the estimated noise, ultimately aiming to exclude more noisy samples and minimize overfitting.

This informed approach allows IDAC to achieve superior noise robustness compared to its predecessor, DAC, and other state-of-the-art loss functions [36]. While DAC learns to abstain on samples that are generally confusing or difficult to classify, IDAC's explicit consideration of noise levels enables it to more precisely identify and mitigate the impact of genuinely noisy labels. This means that the model's abstention mechanism is not solely driven by its internal confidence, but also by an external, estimated measure of label quality. Consequently, IDAC can more effectively prevent the network from overfitting to erroneous labels, as it is guided by a more direct assessment of data fidelity [36]. The practical implication of this extension is a more resilient training process for DNNs, particularly critical in medical imaging where accurate segmentation and classification are paramount for clinical outcomes. By actively accounting for estimated noise during training, IDAC facilitates the learning of more robust feature representations and decision boundaries. This leads to improved generalization performance on clean, unseen data, as the model's parameters are less influenced by the misleading signals from corrupted labels [36].

The temporal dynamics of the abstention rates, as illustrated in Fig. 4.1, highlight the distinct strategies employed by DAC and IDAC. The DAC curve follows a path of exploration and subsequent convergence: it first launches to a near-maximal abstention rate, allowing the model to freely disregard uncertain samples, before initiating a steady, monotonic descent towards zero as the regularization term gains influence. In contrast, the IDAC curve represents a trajectory of rapid stabilization. After an initial activation, its abstention rate quickly settles into a stable equilibrium, maintaining a consistent, non-zero level that is directly influenced by the expected noise rate $\tilde{\eta}$. This sustained plateau demonstrates a fundamentally different, steady-state approach to noise management.



(a) 10% noise (b) 20% noise

Figure 4.1: Abstention behaviour in DAC and IDAC at 10% (a) and 20% (b) label noise on the CaDIS Dataset.

4.3 Abstention beyond Cross Entropy

Building upon the demonstrated efficacy of DAC and IDAC in mitigating label noise through abstention, this thesis takes a significant step further to explore the broader applicability and versatility of the abstention mechanism. Our contribution extends this powerful concept beyond its original formulations by integrating abstention capabilities into three other prominent loss functions: GCE, SCE, and Dice, which we presented in Section 2.5. By systematically adapting these diverse loss functions to include an abstention option, we aim to rigorously demonstrate that the abstention mechanism is a highly versatile and modular extension. This comprehensive investigation underscores that abstention can be effectively incorporated into a wide array of loss functions, fundamentally enhancing their resistance to label noise and thereby improving the robustness and reliability of deep learning models across various challenging medical image segmentation tasks. To this end, we introduce three novel loss functions that combine the abstention mechanism with GCE, SCE, and Dice loss, naming these Generalized Abstaining Classifier (GAC), Symmetric Abstaining Classifier (SAC), and Abstaining Dice Segmenter (ADS). [1]

While a straightforward substitution of CE with another loss function could yield an abstaining variant, our approach employs two critical enhancements to the loss definition. These modifications are designed to provide greater flexibility and adaptability to varying noise levels, ultimately leading to improved model performance. We establish an enhanced and universal definition of the abstention mechanism that can be readily adapted to virtually any underlying loss function, formulated as:

$$\mathcal{L}_{abstention}(x_j) = (1 - p_{k+1})\mathcal{L}_X(x_j) + \alpha \left| \log \frac{1-\tilde{\eta}}{1-p_{k+1}} \right| \tag{4.3}$$

Here, $\mathcal{L}_X(x_j)$ represents the underlying loss function, which in our specific contributions refers to GCE, SCE, or Dice Loss. This generalized formulation introduces two key improvements over the DAC loss definition, enhancing its capacity to manage label noise effectively.

The first improvement lies in the **regularization term** $\alpha \left| \log \frac{1-\tilde{\eta}}{1-p_{k+1}} \right|$. This term draws inspiration from the Informed Deep Abstaining Classifier (IDAC) by explicitly incorporating the expected noise rate $\tilde{\eta}$ to guide the abstention behaviour. Unlike DAC's regularization, which primarily encourages the abstention probability p_{k+1} towards zero regardless of the dataset's noise characteristics, this new term incentives the model to maintain p_{k+1} in proximity to the estimated noise rate $\tilde{\eta}$. This implies that the model is encouraged to abstain precisely on samples where it perceives the labels to be noisy, rather than simply abstaining on all uncertain samples irrespective of their true noise status. This informed regularization allows for a more nuanced and targeted response to label corruption, enabling the model to differentiate between inherent ambiguity and genuine label errors. It is crucial to note that the efficacy of this regularization term does not strictly depend on having access to an impeccably accurate estimation of the noise rate $\tilde{\eta}$. In case an accurate $\tilde{\eta}$ is not available or feasible to estimate, setting $\tilde{\eta}$ to 0 effectively reduces this regularization term to its original form in DAC, which has already demonstrated its strength and effectiveness in combating label noise. This inherent flexibility makes

[1] We chose to name ADS this way to convey the fact that similarly to Dice loss, ADS is mainly intended for segmentation tasks, while the other 'Classifier' losses can be used for classification.

the generalized abstention mechanism robust to varying levels of prior knowledge about dataset noise, ensuring its applicability across diverse real-world scenarios.

The second and more significant enhancement concerns the **α auto-tuning algorithm**. The original algorithm proposed by DAC, detailed in Algorithm 1, employed a linear ramp-up strategy for α after a warm-up phase, which, while effective, offered limited flexibility in controlling the learning trajectory. Our refined approach replaces this with a simpler yet more powerful and flexible method. For every epoch after the initial warm-up phase, α is calculated dynamically using the following power-law formulation:

$$\alpha = \alpha_{final} * \left(\frac{e - L}{E - L}\right)^{\gamma} \tag{4.4}$$

In this equation, $\gamma > 0$ serves as a growth factor that precisely controls the rate at which α increases throughout the abstention period. The parameters $e, L,$ and E denote the current epoch, the number of warm-up epochs, and the total number of training epochs, respectively, while α_{final} is the target maximum value for α. The behaviour of α is modulated by γ: if $\gamma > 1$, α exhibits a sublinear growth, increasing slowly at the beginning of the abstention period and accelerating its growth towards the end of training. This behaviour intensifies with larger values of γ. Conversely, if $\gamma < 1$, α experiences superlinear growth early in the abstention phase, with its rate of increase slowing down as training progresses. Setting $\gamma = 1$ yields a linear increment, akin to DAC's approach, with the minor difference of skipping the initialization step of α to a very small value at the beginning of the abstention phase. This sophisticated formulation provides unparalleled flexibility in penalizing and guiding the abstention behaviour, enabling a more optimal balance between the model's learning from clean data and its strategic abstention from noisy or ambiguous samples. This dynamic tuning of α can be conceptualized as a form of curriculum learning, where the model is initially allowed to abstain more freely on challenging samples and is gradually compelled to learn from increasingly difficult examples as its confidence and representational capacity improve. This prevents early overfitting to noise and promotes the development of robust feature representations. Fig. 4.2 portrays how the growth of α can be modulated by changing the value of γ, where each curve represents the change in the value of α over the span of 100 abstention epochs under a specific value of γ, when $\alpha_{final} = 1$. Changing α_{final} affects the amplitude of the curves, but not the indicated growth behaviour.

4.3.1 Adapting Abstention to Loss Functions

The generalized abstention mechanism, with its enhanced regularization and flexible α tuning, is applied to three distinct loss functions, resulting in novel noise-robust variants tailored for medical image segmentation.

Generalized Abstaining Classifier (GAC)

This novel loss function integrates the universal abstention mechanism with the Generalized Cross Entropy (GCE) loss. GCE, as previously discussed in Section 2.5.2, offers inherent robustness to label noise by interpolating between standard CE and MAE through its tunable parameter q. This interpolation effectively bounds the loss, mitigating the unbounded penalties of CE that can lead to overfitting noisy labels [51]. By combining GCE's bounded loss properties with the abstention mechanism, GAC provides a dual layer

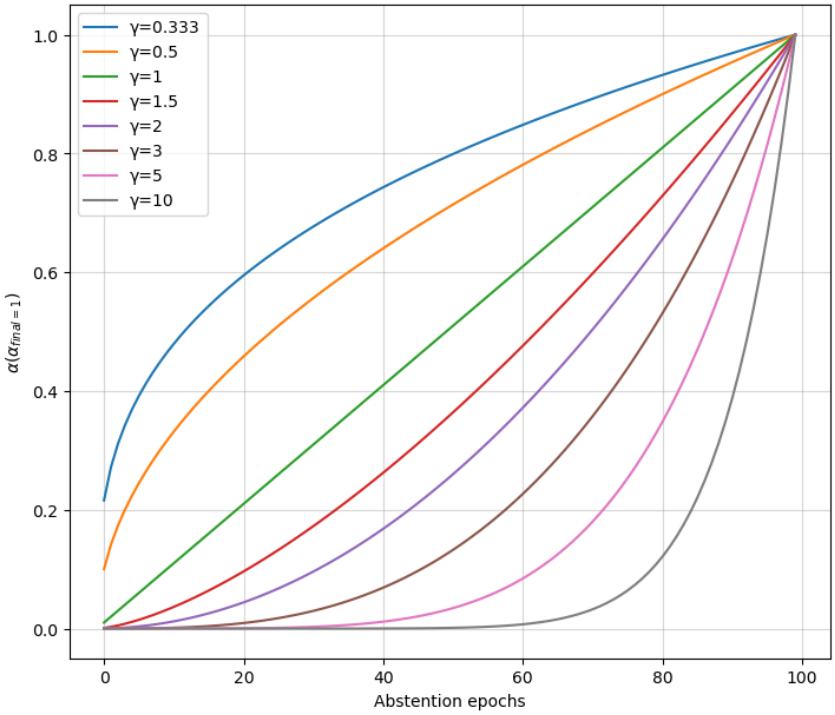


Figure 4.2: The effect of different values of γ on the growth of α with $\alpha_{final} = 1$.

of defence against label noise. The abstention component allows the model to explicitly *opt out* of learning from potentially corrupted samples, while GCE's formulation ensures that even for samples it attempts to classify, the impact of noise is attenuated. This synergy enables GAC to achieve a more resilient training process, particularly beneficial in medical imaging where label noise can be subtle yet impactful, affecting the precise delineation of anatomical structures or pathologies.

Symmetric Abstaining Classifier (SAC)

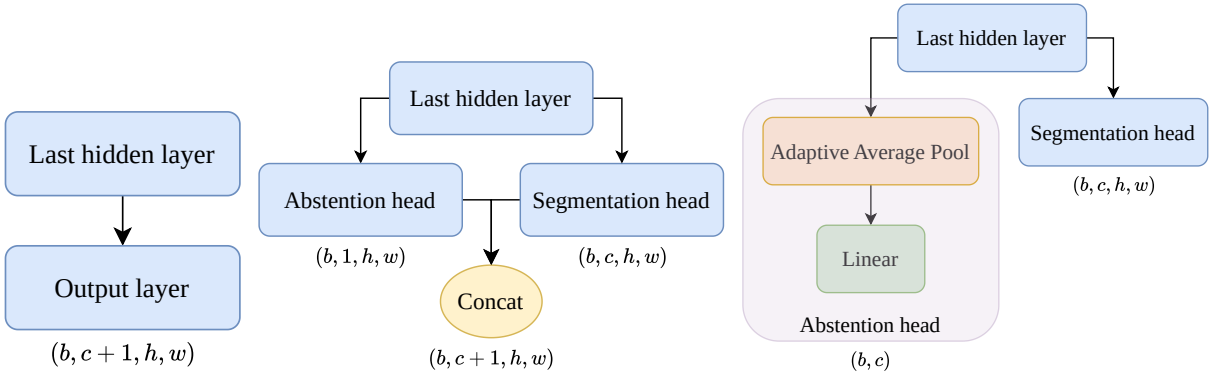
SAC is formed by incorporating the generalized abstention mechanism into the Symmetric Cross Entropy (SCE) loss. SCE is designed to combat label noise by combining standard Cross Entropy with a Reverse Cross Entropy term, creating a symmetrical penalty that addresses both overfitting to noisy labels and under-learning of hard classes [45]. The introduction of abstention into SCE further enhances its balanced learning capabilities. While SCE intrinsically guides the model away from confident incorrect predictions, SAC empowers the model to completely disengage from samples where the label is highly suspect. This means that instead of merely balancing the learning signal, SAC can actively filter out the most egregious noisy examples, allowing the symmetrical CE-RCE components to focus on refining predictions for the more reliable data. This combination is particularly advantageous in scenarios with complex noise patterns, providing a more robust and stable learning trajectory by preventing the model from being misled by highly corrupted labels.

Abstaining Dice Segmenter (ADS)

Adapting the abstention mechanism to Dice Loss for ADS presents unique challenges due to the inherent class-wise nature of Dice Loss, as opposed to the pixel-wise operation of CE-based losses. The standard abstention mechanism, which relies on an extra output

channel representing pixel-wise abstention probability, is not directly compatible with Dice Loss. To overcome this, fundamental changes to the output layer and regularization term are required.

Firstly, to achieve **class-wise abstention** compatible with Dice Loss, the output layer architecture is re-conceptualized. In DAC, the output layer can be viewed as two parallel components whose outputs are concatenated: a standard k-channel segmentation output and a single-channel abstention layer, as illustrated in Fig. 4.3b. For ADS, this single-channel abstention layer is replaced with a module designed to produce class-wise abstention predictions, shown in Fig. 4.3c. This module consists of an Adaptive Average Pooling layer, which reduces the spatial dimensions to a fixed size w, followed by a Linear layer that outputs k values, corresponding to the k classes. Crucially, this module employs a *sigmoid* activation function to obtain the binary class-wise abstention probabilities, in contrast to the *softmax* activation typically used for the non-abstaining output layer. This adaptation ensures that the abstention information is aligned with the class-wise nature of Dice Loss, allowing the model to decide whether to abstain for each specific class within a segmentation task, rather than making a global pixel-wise abstention decision. This fine-grained control is essential for segmentation problems where different classes might exhibit varying levels of noise or ambiguity.



(a) Output layer with abstention channel (b) Output layer with split abstention channel (c) ADS output layer with class-wise abstention

Figure 4.3: Transforming the output layer of an abstaining model from pixel-wise to class-wise abstention. (a) Output layer for pixel-wise abstention [DAC, IDAC, GAC, SAC]. (b) Split segmentation and abstention heads with concatenated outputs. (c) Output layer for class-wise abstention (ADS).

Secondly, the regularization term in ADS is further adapted to the nature of Dice Loss. Instead of a single value $\tilde{\eta}$ representing the estimated noise rate over the entire dataset, ADS benefits from including an estimated noise rate for each class, denoted as $\tilde{\eta}_c$. This class-specific noise estimation allows ADS to guide each class's abstention prediction towards its corresponding noise rate. For instance, if a particular anatomical structure (class) is known to have a higher incidence of labelling errors due to annotation variability or inherent ambiguity, ADS can be specifically encouraged to abstain more frequently on instances of that class. This granular control over abstention, tailored to the noise characteristics of individual classes, significantly increases the flexibility and precision of ADS in combating label noise within complex medical image segmentation tasks, leading to more accurate and reliable segmentations.

4.3.2 Broader Implications and Versatility of Abstention

The systematic expansion of the abstention mechanism to GCE, SCE, and Dice Loss, culminating in GAC, SAC, and ADS, rigorously demonstrates the modularity and profound versatility of this concept. This generalized framework provides a powerful and adaptable tool for developing noise-robust models across a diverse spectrum of segmentation challenges, particularly critical in the medical imaging domain. The benefits for medical image segmentation are multifaceted: it leads to improved reliability of diagnostic systems by actively mitigating the impact of label noise, significantly reduces the propensity for models to overfit to erroneous labels, and consequently enhances generalization performance on unseen, clean data. By empowering models to intelligently manage uncertainty and identify unreliable labels, this contribution moves beyond merely training models on noisy data to actively fostering learning from reliable information. The introduction of GAC, SAC, and ADS serves as concrete evidence of the abstention mechanism's capacity to fundamentally enhance the robustness and trustworthiness of deep learning models, paving the way for more accurate and clinically relevant AI applications in healthcare.

Chapter 5

Experiments

This chapter delineates the empirical validation framework designed to rigorously assess the performance and robustness of the novel loss functions proposed in this thesis. To ensure the generalizability of our findings, our evaluation is conducted across two distinct and challenging, publicly available benchmarks in surgical data science: the Cataract Dataset for Image Segmentation (CaDIS) and the Dresden Surgical Anatomy Dataset (DSAD). The chapter begins by detailing the unique characteristics of each dataset.

A central component of our experimental design is the simulation of realistic annotation errors. We will describe our two-pronged approach to synthetic noise generation, which combines structural perturbations through morphological transformations with semantic label corruption via stochastic flipping. This methodology allows us to create the challenging, noise-corrupted conditions necessary to fairly evaluate the resilience of each loss function.

Subsequently, the chapter outlines the segmentation architectures employed for this evaluation. To substantiate that the efficacy of our methods is fundamentally architecture-agnostic, we utilize two renowned models with distinct design philosophies: the U-Net and DeepLabV3+. We will explain our strategy of leveraging a common ResNet-50 backbone, pre-trained on ImageNet, to provide a powerful feature initialization and facilitate efficient model convergence. Finally, we will specify the training protocols, optimization strategies, and hyperparameter configurations that govern our experiments, paying particular attention to the methods used to ensure a consistent and equitable comparison across all tested functions. Collectively, this chapter establishes a robust and reproducible experimental framework, providing the necessary context for the results and analysis presented in the chapters that follow.

5.1 Datasets

5.1.1 CaDIS

The Cataract Dataset for Image Segmentation (CaDIS) is a comprehensive, high-quality dataset consisting of 4,670 annotated images from cataract surgery procedures, designed to advance the development of computer-assisted interventions [13]. Sourced from 25 videos from the CATARACTS challenge, the dataset contains over 30 classes of anatomical

structures and surgical instruments, providing a robust foundation for training deep learning models. The detailed pixel-wise annotations were created through a rigorous process involving trained artists and validation by medical experts to ensure high fidelity.

The CaDIS dataset is structured to support different experimental goals by offering three class groupings. For the scope of our work, we use the first variant of the dataset, which limits the number of classes to just 8. This configuration simplifies the segmentation problem by grouping all surgical tools into a single "instrument" class, allowing our model to focus on the fundamental task of differentiating instruments from key anatomical structures. In our implementation, we preprocess the data by resizing the images to 480 x 256 and normalizing them during the training phase. Fig. 5.1 illustrates a sample image and its corresponding annotation with all classes included.

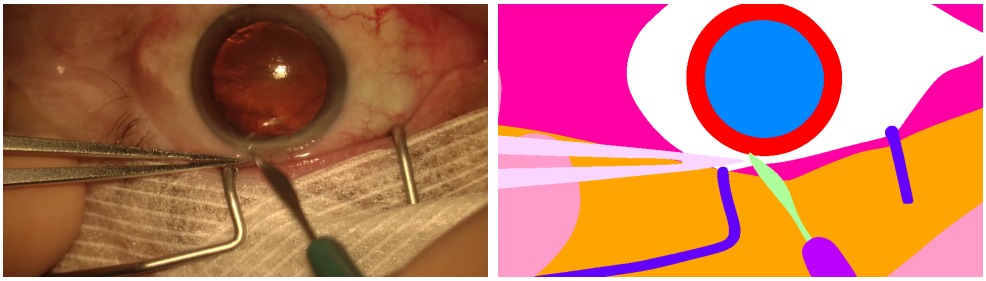


Figure 5.1: Example image frame (left) and semantic segmentation labels (right) from the CaDIS Dataset [13].

5.1.2 DSAD

The Dresden Surgical Anatomy Dataset (DSAD) is a significant, publicly available benchmark in the field of surgical data science, created to address the critical scarcity of densely annotated laparoscopic imaging data [3]. It comprises 13,195 high-resolution (1920×1080) images extracted from video recordings of 32 distinct robot-assisted rectal resections. A key contribution of DSAD lies in its provision of meticulous, pixel-wise semantic segmentations for 11 crucial abdominal anatomical structures. These include eight organs—the colon, liver, pancreas, small intestine, spleen, stomach, ureter, and vesicular glands—as well as the abdominal wall and two vital vessel structures, the inferior mesenteric artery and intestinal veins. This level of detail distinguishes it from earlier datasets that were often limited to weak, image-level annotations, thereby enabling more sophisticated research into context-aware computer vision models for surgical applications [3].

The dataset's high-fidelity annotations are the product of a rigorous, multi-stage validation process, illustrated in Fig. 5.2. For each image, three independent annotators first generated segmentation masks. These were subsequently fused into a single consensus mask using the STAPLE (Simultaneous Truth and Performance Level Estimation) algorithm. Finally, each merged segmentation underwent a conclusive review and refinement by an experienced surgeon to ensure maximum anatomical accuracy. Furthermore, the dataset intentionally preserves the inherent challenges of the intraoperative environment by including images with varying levels of motion blur, smoke, inhomogeneous lighting, and partial occlusion of organs by instruments or other tissues. This realism makes DSAD a robust resource for developing and validating algorithms intended for clinical translation [3].

DSAD is structured to support diverse research objectives, offering both single-structure datasets and a dedicated multi-organ subset. For our research purposes, we specifically utilize the *multilabel* portion of the dataset. This subset consists of 1430 images, that feature concurrent visibility and annotation of seven distinct anatomical structures within the same frame. For our experimental pipeline, these images and their corresponding masks are normalized and resized to a uniform resolution of 480×384.

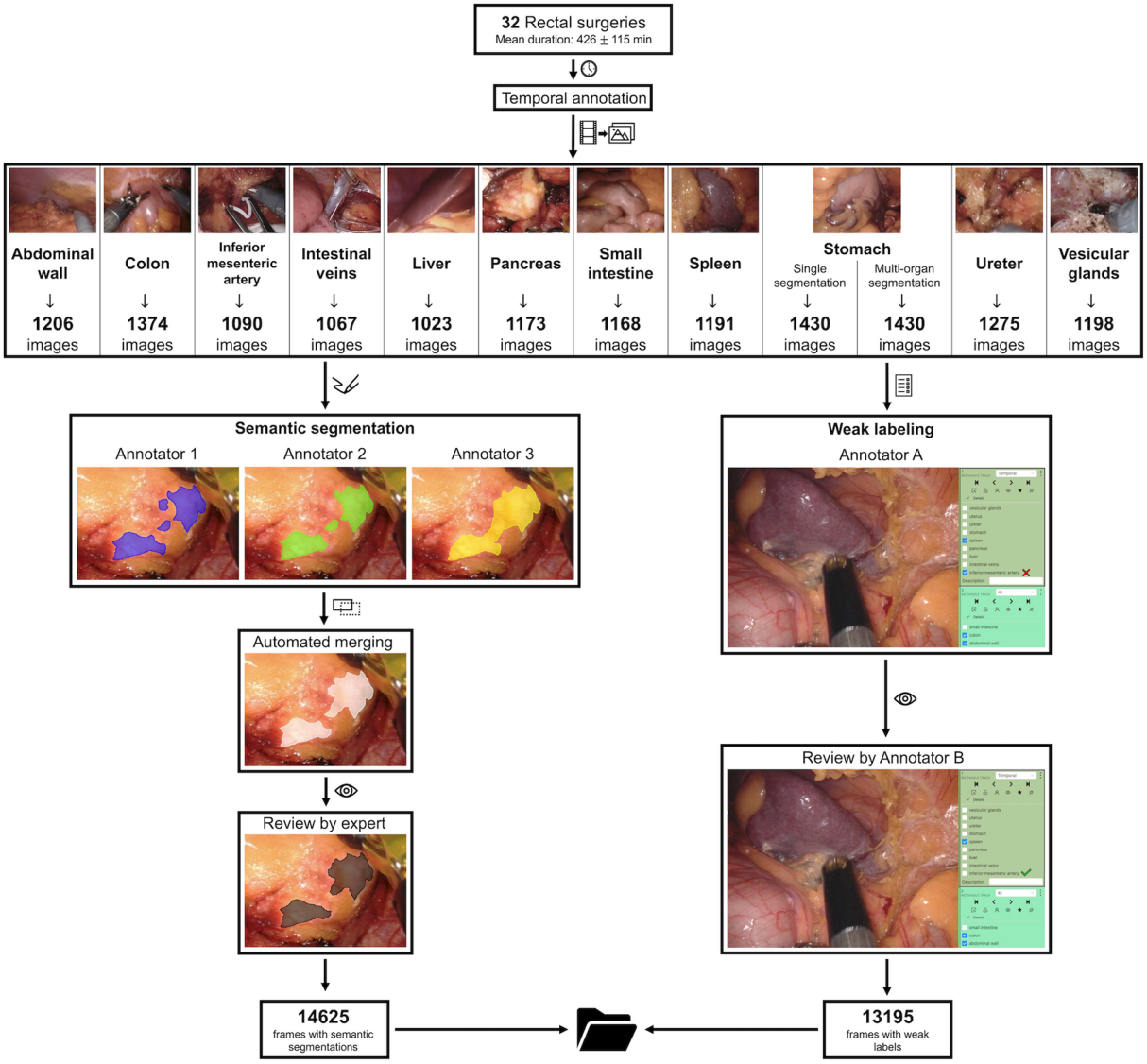
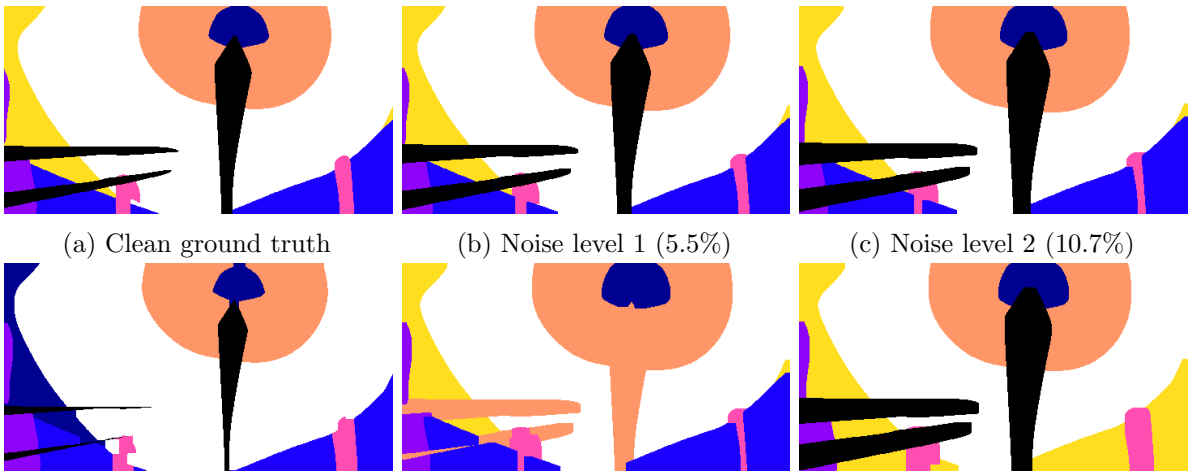


Figure 5.2: Overview of the data acquisition and validation process of DSAD [3].

5.1.3 Noise Simulation

Figure 5.3: Two examples of Erosion and Dilation. Correct segmentation boundaries in red [54].

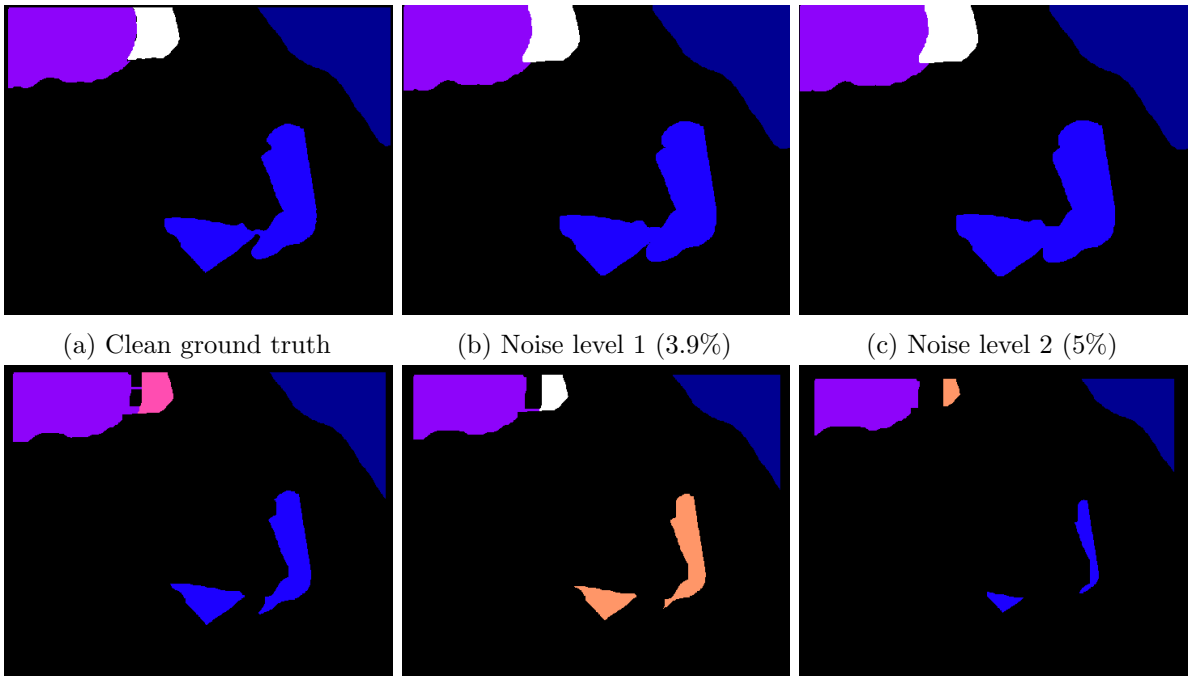
Following [18, 50, 28, 21], we implement a two-pronged approach for synthetic label noise generation, combining structural perturbations and semantic label corruption. Structural noise is introduced through morphological transformations: iterative erosion (shrinking object boundaries) and dilation (expanding boundaries) to simulate segmentation errors arising from ambiguous anatomical edges, as illustrated in Fig. 5.3. Concurrently, semantic noise is injected via stochastic label flipping, wherein a uniformly sampled subset of pixels undergoes class reassignment, mimicking systematic annotation biases. We evaluate five noise levels per dataset, calibrated to balance realism and severity: 3–15% pixel-wise corruption for DSAD (increments of 3%) and 5–25% for CaDIS (increments of 5%). Each percentage here refers to the overall average noise rate in the training datasets, however the actual noise rates of individual masks under the same noise setup will vary, as shown in Figs. 5.4 and 5.5. While it is possible to simulate higher noise levels in both datasets, the resulting noisy ground truth masks at high noise rates tend to look severely unrealistic, potentially rendering our results irrelevant to the real-world application.



(a) Clean ground truth (b) Noise level 1 (5.5%) (c) Noise level 2 (10.7%)

(d) Noise level 3 (15.3%) (e) Noise level 4 (20%) (f) Noise level 5 (23.3%)

Figure 5.4: A sample annotation from the CaDIS dataset under different noise setups. The number in the parenthesis represents the actual noise rate in each noisy mask, not the overall average noise rate of the training dataset.



(a) Clean ground truth (b) Noise level 1 (3.9%) (c) Noise level 2 (5%)

(d) Noise level 3 (8.2%) (e) Noise level 5 (13%) (f) Noise level 5 (15.2%)

Figure 5.5: A sample annotation from the DSAD dataset under different noise setups. The number in the parenthesis represents the actual noise rate in each noisy mask, not the overall average noise rate of the training dataset.

5.2 Models

A pivotal development in medical image analysis was the introduction of the U-Net architecture by Ronneberger, Fischer, and Brox [33]. Designed specifically to address the challenges of biomedical image segmentation, U-Net presented an elegant and powerful solution that has since become a foundational component of the field. Prior to its introduction, methods often relied on patch-based classification using a sliding window, a process that was computationally intensive and often struggled to effectively balance local detail with broader image context. U-Net was conceived to overcome these limitations by processing entire images in an end-to-end fashion [33].

The architecture is named for its distinctive, symmetric U-shape, which consists of two main pathways: a contracting path (the encoder) and an expansive path (the decoder), as illustrated in Fig. 5.6.

- The Contracting Path (Encoder): This path follows the typical structure of a convolutional neural network. It is composed of sequential blocks of 3x3 convolutions and Rectified Linear Unit (ReLU) activations, followed by max pooling operations. This process progressively downsamples the spatial dimensions of the input while simultaneously increasing the number of feature channels. The purpose of the encoder is to capture the hierarchical contextual features of the image—learning 'what' is present at various scales.

- The Expansive Path (Decoder): This path is responsible for upsampling the condensed feature maps to reconstruct a full-resolution segmentation map. It systematically increases the spatial dimensions using up-convolutions (transposed convolutions)

while reducing the feature channel depth. The most critical innovation of the U-Net lies in the use of skip connections. These connections bridge the two paths by concatenating feature maps from the encoder directly with the corresponding layers in the decoder. By merging the high-resolution, fine-grained spatial information from the contracting path with the rich, contextual information from the expansive path, these skip connections enable the network to achieve highly precise localization, which is crucial for accurately delineating anatomical boundaries.

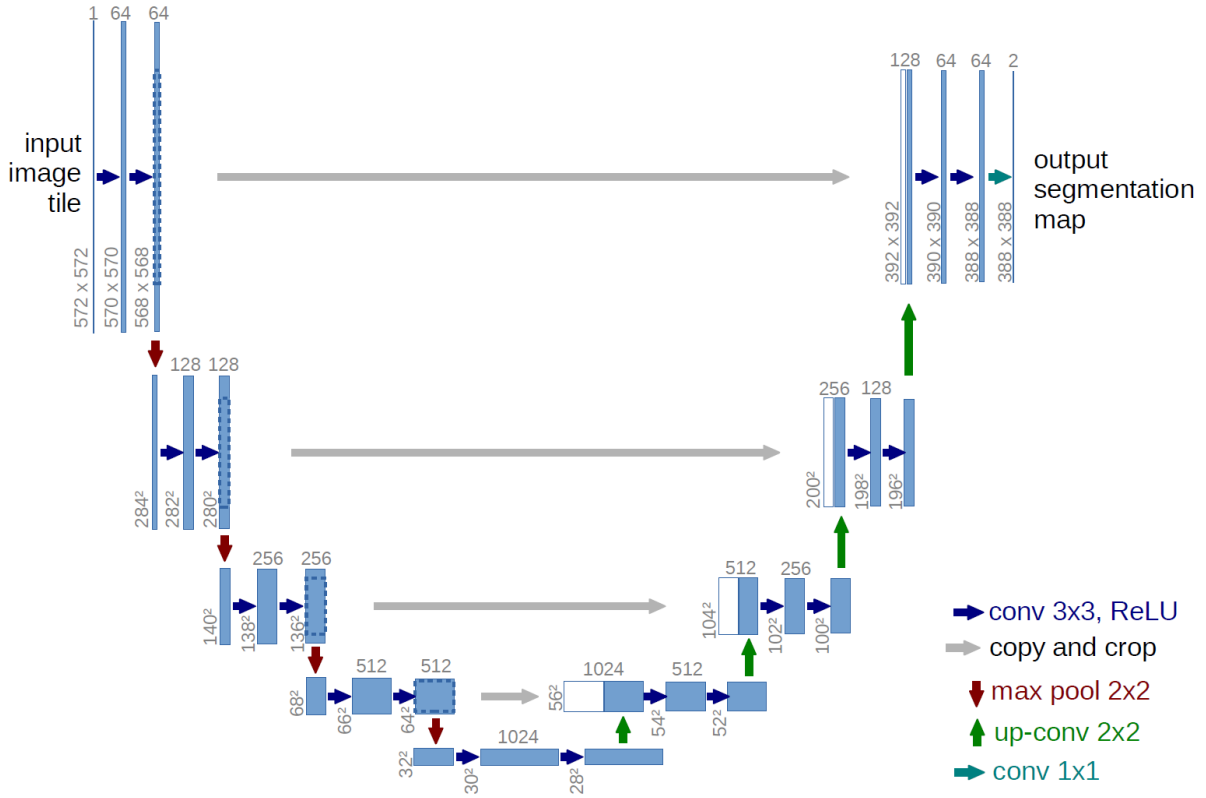


Figure 5.6: The U-Net architecture. The contracting path on the left captures context, while the expansive path on the right enables precise localization. Grey arrows represent the skip connections that fuse high-resolution features from the encoder with the upsampled features in the decoder [33].

The contributions of U-Net to the field of medical image segmentation are substantial. One of its most celebrated strengths, is its ability to be trained effectively on very small datasets [33], a common scenario in medical imaging where annotated data is scarce. Its architectural design proved so effective and robust that it rapidly became the de facto standard and a primary benchmark for a vast range of medical segmentation tasks [23]. Indeed, its success has spawned an entire family of 'U-Net-like' architectures, and it remains one of the most widely used and cited models in the medical image analysis literature, serving as the backbone for countless applications and further research [40]. For these reasons, the U-Net is employed as the base segmentation architecture for the experiments conducted in this thesis.

While the U-Net architecture forms the cornerstone of our experiments, it is crucial to demonstrate that the contributions of this thesis are not merely a product of this specific architectural choice. To affirm the generalizability of the loss functions discussed in this thesis, we conduct a validation experiment utilizing the DeepLabV3+ [4] architecture,

which is another renowned segmentation architecture that is fundamentally different from U-Net. By showing the successful application of our approach on a model with a distinct design philosophy, characterized by atrous convolutions and spatial pyramid pooling [4], we substantiate the robustness of our solution and assert that its efficacy is fundamentally architecture-agnostic. In our experimental framework, we leverage the principles of transfer learning to facilitate efficient and effective model training. Specifically, both the U-Net and DeepLabV3+ segmentation architectures are built upon a common feature extractor, or backbone: a ResNet-50 [16] network previously trained on the large-scale ImageNet [19] dataset. This transfer learning approach provides a powerful initialization, significantly reducing the computational cost required to achieve high performance.

5.3 Experimental Setup

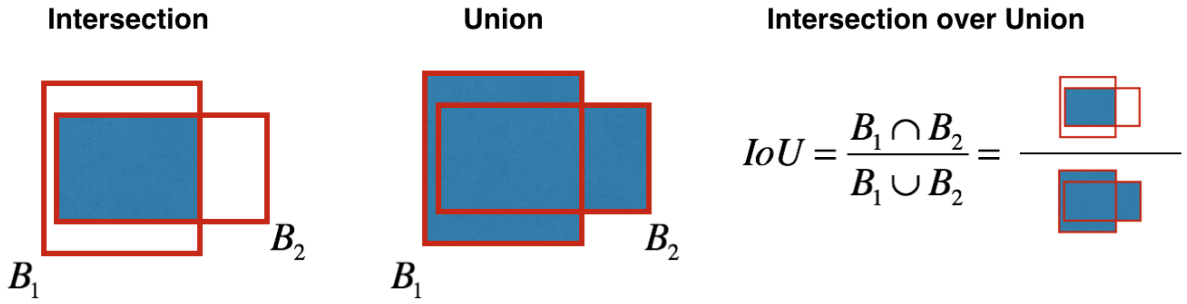


Figure 5.7: Visual explanation of the Intersection over Union (IoU) metric, calculated as the ratio of the overlapping area (Intersection) to the total combined area (Union) of the predicted and ground truth regions.

We evaluate our experiments with the mean Intersection over Union (mIoU) metric, which is the benchmark standard for quantifying performance in semantic segmentation tasks. The metric assesses the spatial correspondence between the predicted segmentation and the ground truth by measuring the overlap of their respective regions.

For a single class, the Intersection over Union (IoU), depicted in Fig. 5.7, is defined as the ratio of the area of their intersection to the area of their union:

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|\text{Prediction} \cap \text{Ground Truth}|}{|\text{Prediction} \cup \text{Ground Truth}|}$$

This can also be expressed in terms of True Positives (TP), False Positives (FP), and False Negatives (FN) as:

$$IoU = \frac{TP}{TP + FP + FN}$$

The mean Intersection over Union (mIoU) is simply the average of the IoU scores computed for each individual class present in the dataset. This averaging makes it a holistic measure for multi-class segmentation problems.

A standardized training and evaluation protocol was established to ensure a rigorous and fair comparison across all experiments. All networks were trained for a total of 50 epochs, utilizing the AdamW optimizer [27]. We employed a step-wise learning rate schedule, with an initial rate of 0.003 that was subjected to a decay factor of 5 every

10 epochs. This schedule allows for rapid initial learning while facilitating fine-grained convergence in the later stages of training. Reflecting the differing memory footprints of the datasets, we used a batch size of 128 for CaDIS and 50 for DSAD. To account for the inherent stochasticity of the deep learning training process and to ensure the statistical reliability of our findings, each experimental run was conducted five times using distinct random seeds. The final performance metrics, presented in the subsequent chapter, are therefore reported as the mean mIoU and standard deviation across these independent trials, providing a robust measure of both central tendency and result stability. For the sake of complete reproducibility, all experiments were conducted on an NVIDIA A100 80GB GPU within a consistent software environment comprising Python 3.12, CUDA Toolkit 11.8, PyTorch 2.6, and Torchvision 0.21, orchestrated using the PyTorch Lightning 2.5.1 framework.

5.4 Hyperparameter Analysis

We performed hyperparameter optimization to determine a reasonably ideal set of parameters that yield the highest validation mIoU scores for each loss function. These optimizations were conducted on the highest noise level of each dataset to insure maximal resistance to noise. The resulting parameters are detailed in Table 5.1.

A critical methodological point must be addressed regarding the comparison between the baseline losses (GCE, SCE) and our novel abstaining variants (GAC, SAC). These pairs share underlying hyperparameters: q for GCE/GAC, α and β for SCE/SAC. The introduction of the abstention mechanism, however, can alter the optimal value for these shared parameters. To ensure a methodologically consistent and equitable comparison, we first determined the optimal hyperparameters for the baseline GCE and SCE functions. We then deliberately retained these exact settings for GAC and SAC. While this approach is likely suboptimal for our novel functions, it rigorously isolates the impact of the abstention mechanism itself, ensuring that any performance gains are attributable solely to abstention rather than to confounding factors from differential hyperparameter tuning. Thus, the results for GAC and SAC—detailed in he following chapter—likely represent a conservative estimate of their full potential.

Dataset	DAC	IDAC	GCE	GAC	SCE	SAC	ADS
CaDIS	$\alpha_{final} = 1$ $L = 10$	$\alpha = 1$ $L = 10$	q=0.5	$\alpha_{final} = 3$ $L = 10$ $\gamma = 3$	$\alpha = 1$ $\beta = 1$	$\alpha_{final} = 1$ $L = 10$ $\gamma = 1.5$	$\alpha_{final} = 1$ $L = 10$ $\gamma = 3$ $w = 16$
DSAD	$\alpha_{final} = 2$ $L = 18$	$\alpha = 1$ $L = 10$	q=0.1	$\alpha_{final} = 2$ $L = 15$ $\gamma = 2$	$\alpha = 0.5$ $\beta = 1$	$\alpha_{final} = 1$ $L = 20$ $\gamma = 3$	$\alpha_{final} = 4$ $L = 10$ $\gamma = 1.5$ $w = 16$

Table 5.1: Hyperparameter configurations used in our experiments. q controls GCE's sensitivity to noise, while α and β are weights for the terms in SCE. For the abstaining losses, L is the number of warm-up epochs, α is IDAC's fixed abstention penalty, and α_{final} is the target penalty for DAC, GAC, SAC, and ADS. γ is the growth factor for our enhanced α auto-tuning algorithm , and w is the pooling output size for the class-wise abstention module in ADS.

Chapter 6

Evaluations and Analysis

This chapter presents a comprehensive analysis of the experimental results. Table 6.1 provides the average test mean Intersection over Union (mIoU) percentages and their corresponding standard deviations, computed across five independent experimental runs. The evaluation was conducted on CaDIS and DSAD under a spectrum of synthetically introduced label noise rates (η). The primary objective of this analysis is to rigorously compare the performance of five abstaining loss functions against their respective non-abstaining baseline counterparts under various noise intensities. The overarching narrative of this investigation is to demonstrate the profound effectiveness of the abstention mechanism in mitigating label noise and to establish its remarkable versatility as an extension applicable to loss functions possessing diverse complexities and characteristics.

The observed disparity in mIoU results, where performance on the DSAD dataset consistently lags significantly behind that on CaDIS across all evaluated loss functions and noise levels, can be attributed to fundamental differences in dataset characteristics and inherent image complexity. Firstly, the annotation density varies considerably between the two datasets; the CaDIS dataset is characterized by dense annotations where every pixel is assigned to a meaningful class representing an anatomical structure or a surgical instrument, and the specific variant utilized in this study contains no explicit background or 'ignore' class. In contrast, the DSAD dataset is sparsely annotated, with approximately 82% of the pixels in the training split designated as background. This substantial class imbalance in DSAD inherently presents a greater challenge for segmentation models. Secondly, the intrinsic complexity of the images themselves contributes to the performance differential. DSAD images are typically more intricate to analyse, featuring internal organs that exhibit high visual similarity, and some acquisitions were performed during laparoscopic camera movement, resulting in motion blur that further obscures organ boundaries and makes precise distinction exceptionally difficult. Collectively, these factors render DSAD an inherently more challenging dataset for medical image segmentation compared to CaDIS.

6.1 Performance Analysis and Trends

At a noise rate of $\eta = 0\%$, representing a scenario with perfectly clean labels, all models generally achieve their peak performance, serving as a benchmark for their inherent capabilities without the confounding factor of noise. On the CaDIS dataset, the Dice Loss and our proposed Abstaining Dice Segmenter (ADS) exhibit the strongest initial

Dataset	Noise rate η (%)	Loss function								
		CE	DAC	IDAC	GCE	GAC	SCE	SAC	Dice	ADS
CaDIS	0	**76.02±0.70**	75.29±0.79	75.36±0.73	73.49±3.27	**73.76±2.80**	75.38±0.75	**75.83±0.62**	76.52±0.47	**77.04±0.37**
	5	**73.67±1.03**	73.14±0.46	72.89±0.41	**72.83±1.11**	71.73±2.79	73.41±0.71	**73.51±1.59**	73.48±0.28	**75.22±0.85**
	10	66.39±0.17	**67.43±0.49**	66.92±0.49	**64.82±0.86**	64.16±2.57	65.92±0.91	**67.29±1.65**	66.51±0.61	**71.12±0.55**
	15	64.15±2.47	**65.85±1.05**	64.87±0.91	**64.81±0.46**	64.44±2.70	62.16±1.99	**65.48±2.11**	67.31±0.73	**70.80±1.08**
	20	59.56±1.21	**63.42±0.87**	60.54±2.27	60.73±1.41	**60.91±1.64**	57.62±4.22	**62.70±0.31**	63.64±0.82	**68.88±0.49**
	25	52.27±1.70	**60.63±2.73**	58.19±4.77	55.71±1.30	**59.46±0.76**	55.08±0.93	**61.27±1.22**	61.04±1.41	**66.39±0.67**
DSAD	0	**34.25±2.50**	34.01±0.96	33.60±0.72	**35.14±1.65**	32.26±0.53	32.78±1.19	**33.86±1.83**	**31.28±0.87**	30.09±1.10
	3	**33.69±1.85**	33.67±2.01	32.76±2.03	**33.84±2.56**	32.94±2.23	**32.11±1.09**	30.90±2.76	**30.83±4.78**	28.64±2.76
	6	**30.70±2.47**	29.47±1.97	29.11±2.10	29.69±1.96	**29.78±4.27**	30.51±2.16	**31.55±2.43**	28.56±1.00	**30.48±3.61**
	9	**24.65±2.90**	24.58±2.61	23.47±2.48	22.95±2.93	**28.84±4.17**	28.02±2.37	**28.55±1.29**	19.04±1.92	**26.23±2.05**
	12	21.00±3.15	**22.59±4.35**	20.94±1.86	19.84±2.89	**25.00±4.13**	21.57±0.67	**23.73±0.68**	16.15±1.49	**22.63±0.51**
	15	14.41±2.59	**17.69±3.97**	16.24±1.45	14.12±2.91	**20.01±2.56**	15.31±0.75	**15.91±3.53**	14.65±1.50	**18.05±1.63**

Table 6.1: Average test mIoU (%) and standard deviation (5 runs) of a U-Net model trained on CaDIS and DSAD datasets with various rate of label noise, comparing five abstaining loss functions [DAC, IDAC, GAC, SAC, ADS] against their non-abstaining baselines [CE, GCE, SCE, Dice]. Best results in each bracket are in **bold**.

performance, registering 76.52% and 77.04% mIoU, respectively. These figures marginally surpass the standard Cross Entropy (CE) baseline (76.02%) and other non-abstaining counterparts, suggesting that for clean segmentation tasks, Dice-based metrics inherently provide a robust foundation, likely due to their focus on region overlap. Conversely, on the DSAD dataset, Generalized Cross Entropy (GCE) achieves the highest mIoU at 35.14% under clean conditions, indicating its baseline effectiveness for this particular dataset.

As the label noise rate η progressively increases from 0% to 25% for CaDIS and from 0% to 15% for DSAD, a consistent and anticipated degradation in the performance of all loss functions is observed. This decline underscores the detrimental impact of label corruption on model training. However, a critical distinction emerges in the rate and extent of this performance degradation when comparing the non-abstaining baselines against their abstaining extensions. This differential response forms the crux of our investigation into the noise-robustness conferred by the abstention mechanism.

Deep Abstaining Classifier (DAC) and Informed Deep Abstaining Classifier (IDAC): As established methods, DAC and IDAC serve as crucial points of comparison. On the CaDIS dataset, at the highest noise rate of η=25%, the standard CE loss experiences a substantial drop to 52.27% mIoU. In stark contrast, DAC maintains a significantly higher mIoU of 60.63%, demonstrating a clear advantage of approximately 8.36%. This improvement highlights DAC's ability to prevent overfitting to noisy labels by allowing the model to abstain. IDAC, an extension of DAC, shows a mIoU of 58.19% at η=25%, indicating its continued robustness, though slightly lower than DAC at this specific point. On the DSAD dataset, at η=15%, CE performance plummets to 14.41% mIoU, whereas DAC achieves 17.69%, showcasing an improvement of approximately 3.28%s. IDAC, at 16.24% mIoU, also outperforms CE, further solidifying the foundational benefits of abstention for segmentation tasks.

Generalized Abstaining Classifier (GAC): Our proposed GAC, which integrates the generalized abstention mechanism with GCE, consistently outperforms its GCE baseline across all noise levels on both datasets. On CaDIS, at η=25%, GAC achieves 59.46%

mIoU, a notable lead of 3.75%s over GCE's 55.71%. This demonstrates that combining GCE's inherent robustness (derived from its interpolation between CE and MAE) with the explicit noise-filtering capability of abstention creates a more resilient learning objective. The ability of GAC to selectively disengage from noisy samples complements GCE's bounded loss properties, leading to a more stable optimization trajectory in the presence of label corruption. On the DSAD dataset, GAC's superiority is even more pronounced. At η=15%, GAC achieves 20.01% mIoU, a substantial 5.89% lead over GCE's 14.12%. This robust performance on DSAD underscores GAC's effectiveness even in more challenging segmentation contexts.

Symmetric Abstaining Classifier (SAC): Similarly, our proposed SAC, which extends Symmetric Cross Entropy (SCE) with the abstention mechanism, consistently demonstrates superior performance over its SCE baseline. On CaDIS, at η=25%, SAC achieves 61.27% mIoU, outperforming SCE's 55.08%. This indicates that the active filtering of noisy labels through abstention further enhances SCE's symmetrical penalty, which is designed to address both overfitting to noise and under-learning of hard classes. By allowing the model to completely ignore highly suspect labels, SAC enables the core SCE mechanism to focus on refining predictions for more reliable data. On the DSAD dataset, SAC maintains a superior performance over SCE, particularly at higher noise rates, achieving 15.91% mIoU compared to SCE's 15.31% at η=15%. While the absolute improvement might appear smaller on DSAD, the consistent trend across both datasets highlights the added value of abstention for SCE.

Abstaining Dice Segmenter (ADS): ADS represents a critical contribution, demonstrating the adaptability of the abstention mechanism to a fundamentally different class of loss functions—those based on Dice Similarity. The results for ADS are particularly compelling, showcasing its exceptional robustness against label noise in segmentation tasks. On the CaDIS dataset, ADS consistently emerges as the top performer at higher noise rates. At η=25%, ADS maintains an impressive mIoU of 66.39%, significantly outperforming its Dice baseline (61.04%) by 5.35% and surpassing all other loss functions, including DAC, IDAC, GAC, and SAC, at this high noise level. This remarkable resilience is attributed to the class-wise nature of ADS's abstention mechanism, which is specifically tailored to the Dice Loss. By allowing the model to abstain on a per-class basis, guided by class-specific noise estimations $\tilde{\eta}_c$, ADS can more precisely mitigate the impact of noise on individual anatomical structures or pathologies. This is crucial in medical image segmentation where different classes tend to have varying levels of annotation noise or inherent ambiguity. The trends are equally strong on the DSAD dataset. At the highest noise rate of η=15%, ADS secures 18.05% mIoU, outperforming its Dice baseline (14.65%) by 3.4%. While the overall mIoU values on DSAD are lower, ADS's consistent lead over its non-abstaining counterpart and other abstaining methods at high noise levels underscores its superior ability to handle label noise in challenging segmentation scenarios. The results for ADS provide strong empirical evidence that the abstention mechanism, when meticulously adapted to the unique characteristics of Dice Loss, yields a highly resilient and effective solution for noisy medical image segmentation.

A graphical analysis of the performance trends, presented in Figs. 6.1 and 6.2, provides several key insights into the behaviour of the evaluated loss functions under increasing label noise. For the CaDIS dataset, Fig. 6.1 visually confirms that while all methods degrade

with more noise, the *rate* of this degradation varies substantially. The trajectory for ADS is visibly flatter than its counterparts, graphically representing its superior noise resilience at high noise levels. In contrast, Fig. 6.2 illustrates the more pronounced performance decline across all functions on the more challenging DSAD dataset. Despite the steeper overall degradation, the graphical data still accentuates the advantage of the abstaining methods at the highest level, with DAC, GAC, and ADS showing significantly higher mIoU scores compared to their respective baselines. Collectively, Figs. 6.1 and 6.2 illustrate that the introduction of an abstention mechanism consistently results in a more graceful degradation of performance, an effect that is particularly evident at high noise contamination levels across both datasets.

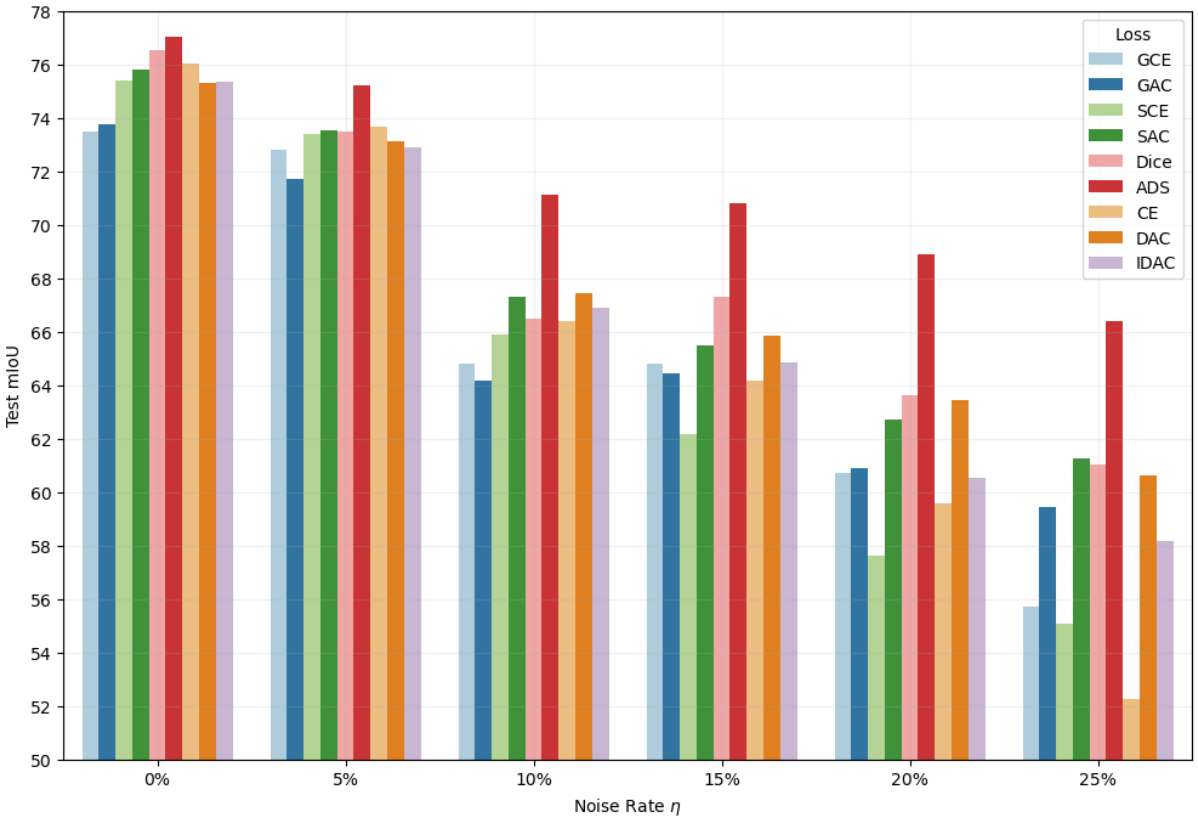


Figure 6.1: Performance comparison of all evaluated loss functions on the CaDIS dataset across increasing levels of label noise η. The y-axis represents the average test mIoU (%), illustrating the degradation trend for each loss function as the noise rate increases.

6.2 Architectural Robustness: DeepLabV3+

To substantiate the architectural agnosticism of the abstention mechanism and demonstrate its inherent robustness, we evaluated the loss functions were using a second segmentation architecture that is distinctively different from U-Net. Further analysis of model performance, as detailed in Table 6.2, reveals nuanced insights into the behaviour of abstaining loss functions when applied to a DeepLabV3+ [4] architecture. Specifically, at 25% label noise on the CaDIS dataset, DAC (57.02% mIoU) and IDAC (56.29% mIoU) exhibit only marginal improvements over the standard CE baseline (56.02% mIoU). More notably, on the inherently more challenging DSAD dataset with 15% label noise, DAC

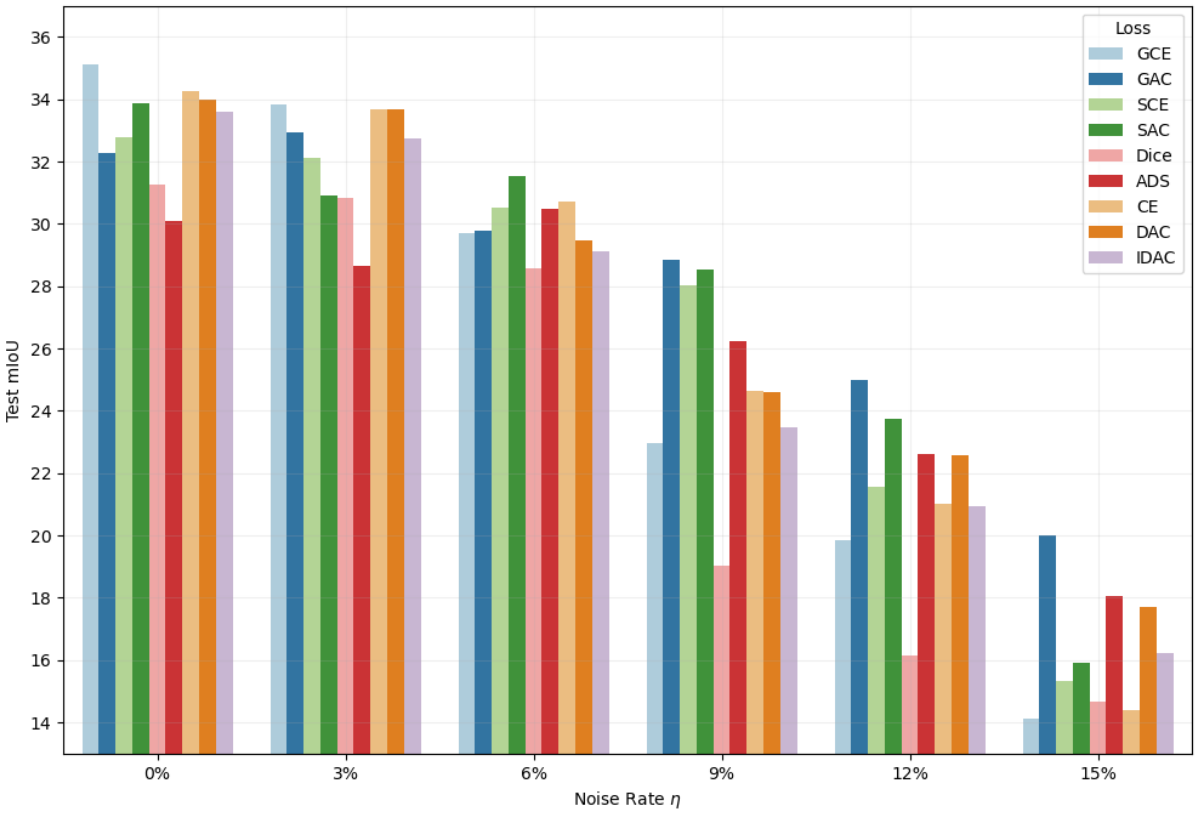


Figure 6.2: Performance comparison of all evaluated loss functions on the more challenging DSAD dataset across its corresponding noise levels η. The graph visually represents the steeper performance degradation inherent to this sparsely annotated and complex dataset.

(15.90% mIoU) and IDAC (16.20% mIoU) actually demonstrate a slight performance decrement compared to CE (16.73% mIoU). This unexpected outcome for DAC and IDAC, particularly on DSAD, can be primarily attributed to the fact that the hyperparameters for these loss functions, as detailed in Table 5.1, were originally optimized for the U-Net architecture. It is highly probable that re-optimizing these parameters specifically for the DeepLabV3+ architecture would yield improved performance for DAC and IDAC. However, performing a second hyperparameter sweep of the same scope as our primary U-Net experiments was not feasible due to significant time and hardware constraints. Additionally, in the case of DAC, these results further underscore the potential benefits of the two enhancements introduced in our generalized abstention mechanism in Eqs. (4.3) and (4.4): the incorporation of the expected noise rate $\tilde{\eta}$ into the regularization term, and the more flexible power-law-based α auto-tuning algorithm. These enhancements, which are integral to our proposed GAC, SAC, and ADS loss functions, are designed to provide a more informed and adaptable abstention behaviour, suggesting that DAC could significantly benefit from their integration.

In contrast to DAC and IDAC, our newly proposed abstaining loss functions—GAC, SAC, and ADS—consistently demonstrate superior performance over their respective non-abstaining baselines even when evaluated with the DeepLabV3+ architecture. On the CaDIS dataset (25% noise), GAC achieves 58.08% mIoU, surpassing GCE (55.56%) by 2.52%. Similarly, SAC records 59.77% mIoU, outperforming SCE (58.37%) by 1.4%. Most notably, ADS achieves 61.84% mIoU, significantly exceeding the Dice baseline (59.55%)

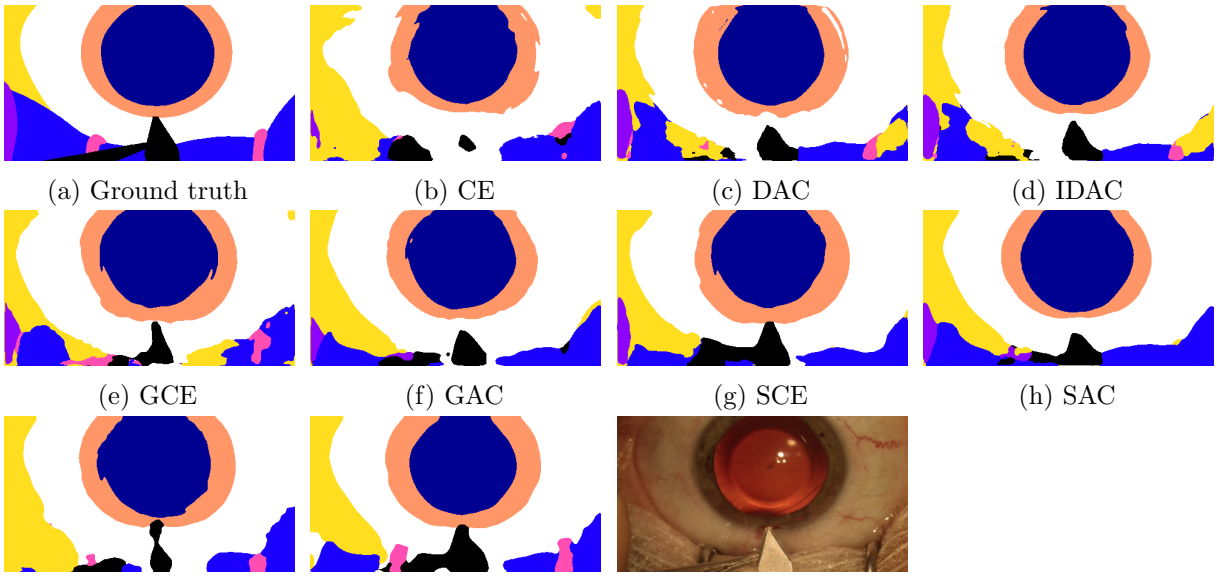
Dataset	Loss function								
	CE	DAC	IDAC	GCE	GAC	SCE	SAC	Dice	ADS
CaDIS	56.02±1.30	**57.02±0.81**	56.29±1.05	55.56±2.08	**58.08±1.43**	58.37±0.53	**59.77±1.17**	59.55±1.66	**61.84±2.23**
DSAD	**16.73±2.34**	15.90±3.19	16.20±1.37	16.26±1.37	**19.01±1.69**	12.74±2.03	**14.03±3.53**	12.46±0.86	**17.16±2.02**

Table 6.2: Average test mIoU (%) and standard deviation (5 runs) of a DeepLabV3+ model trained on CaDIS and DSAD datasets at 25% and 15% label noise, respectively. Best results in each bracket are in **bold**.

by 2.29% and emerging as the top performer among all evaluated losses on CaDIS for DeepLabV3+. On the more challenging DSAD dataset (15% noise), this trend of superiority persists. GAC (19.01% mIoU) shows a substantial improvement over GCE (16.26%) by 2.75%. SAC (14.03% mIoU) also outperforms SCE (12.74%) by 1.29%. Furthermore, ADS (17.16% mIoU) maintains its lead over the Dice baseline (12.46%) by 4.7% and other abstaining losses on DSAD. These consistent gains across different underlying loss functions and on both datasets, despite the architectural shift, provide strong empirical evidence for the robustness and generalizability of our enhanced abstention mechanism. The results suggest that the improved regularization and α auto-tuning effectively guide the model to leverage abstention for noise mitigation, irrespective of the trained architecture.

6.3 Visual Analysis

6.3.1 CaDIS



(a) Ground truth (b) CE (c) DAC (d) IDAC

(e) GCE (f) GAC (g) SCE (h) SAC

(i) Dice (j) ADS (k) Input

Figure 6.3: Visualisation of a sample test frame (k) and clean ground truth (a) from CaDIS. (b) to (j) show the segmentation predictions of a U-Net model trained with each loss function at 25% noise.

In this section, we provide a qualitative analysis of the segmentation predictions generated by a U-Net model, trained with each loss function under 25% label noise on the CaDIS dataset, as depicted in Fig. 6.3. The visual comparison of these masks against the

ground truth highlights the impact of different loss functions, particularly emphasizing the benefits of the abstention mechanism in mitigating the effects of label noise.

Observing Figs. 6.3b to 6.3d, it is evident that both DAC and IDAC produce visually superior segmentation masks compared to the traditional CE loss. The CE prediction exhibits noticeable inaccuracies, particularly around the boundaries of iris and in the surrounding anatomical regions. There are instances of misclassified pixels and less precise delineation. In contrast, DAC and IDAC show cleaner and more accurate segmentations, with sharper boundaries and fewer erroneous pixels. IDAC, in particular, appears to offer a slight edge over DAC, suggesting that its informed abstention mechanism leads to marginally better boundary adherence and reduced noise in the predictions. This qualitative improvement aligns with the expectation that abstaining losses are better equipped to handle noisy labels by avoiding overfitting to corrupted information.

When comparing GCE with GAC, a similar pattern of improvement is discernible. The segmentation produced by GCE in Fig. 6.3e shows some irregularities and less precise contours, particularly in the lower left and right regions of the image. GAC, which incorporates the abstention mechanism, yields a smoother and more accurate segmentation mask, showin in Fig. 6.3f. The boundaries appear more consistent with the ground truth, and the overall segmentation quality is enhanced, indicating that GAC's ability to abstain on noisy samples helps in learning more robust features and producing better-defined segmentations than its non-abstaining counterpart.

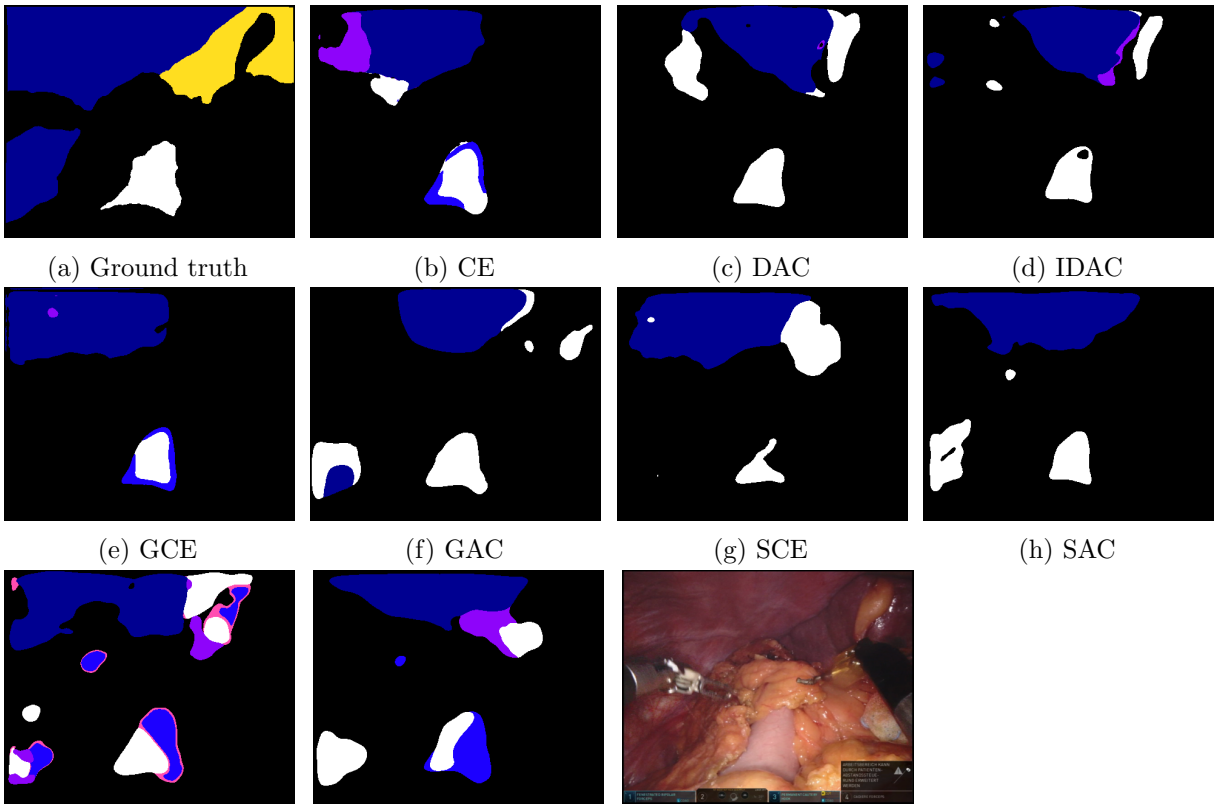
The visual comparison between SCE and SAC further underscores the benefits of abstention. SCE exhibits some fragmentation and less accurate delineation of the anatomical structures, especially in the peripheral regions, as see in Fig. 6.3g. SAC, with its abstention capability, produces a more coherent and accurate segmentation in Fig. 6.3h. The contours are better defined, and the presence of misclassified pixels appears reduced. This suggests that SAC's mechanism for handling noisy labels, by selectively abstaining, allows for the training of a more reliable segmentation model compared to SCE.

Finally, comparing Dice and ADS (Figs. 6.3i and 6.3j), the superiority of ADS is quite pronounced. While Dice provides a reasonable segmentation, ADS clearly offers the most accurate and visually clean mask among all evaluated loss functions. The boundaries are exceptionally sharp, and the segmentation closely mirrors the ground truth in Fig. 6.3a, with minimal noise or misclassifications. This strong performance of ADS highlights the synergistic effect of combining the inherent class-wise nature of Dice loss, which is robust to class imbalance, with the noise-filtering capabilities of the abstention mechanism. ADS appears to be highly effective at preserving fine details and producing high-fidelity segmentations even in the presence of significant label noise.

6.3.2 DSAD

Similar to the previous section's analysis of the CaDIS dataset, Fig. 6.4 illustrates a qualitative examination of segmentation outputs from a U-Net model, trained with each loss function under 15% label noise on the intricate DSAD dataset. The visual comparison against the ground truth Fig. 6.4a highlights the differential impact of distinct loss functions, acknowledging that the inherent complexity and sparse annotations of DSAD generally lead to lower overall segmentation quality compared to CaDIS.

The conventional CE prediction, shown in Fig. 6.4b exhibits fragmented and inaccurate



(a) Ground truth (b) CE (c) DAC (d) IDAC

(e) GCE (f) GAC (g) SCE (h) SAC

(i) Dice (j) ADS (k) Input

Figure 6.4: Visualisation of a sample test frame (k) and clean ground truth (a) from DSAD. (b) to (j) show the segmentation predictions of a U-Net model trained with each loss function at 15% noise.

segmentations on DSAD, often struggling with boundary definition and producing spurious activations. DAC and IDAC (Figs. 6.4c and 6.4d) show some improvement over CE, yielding slightly more coherent masks with fewer extraneous pixels and a somewhat better delineation of the main structures, though the overall quality remains challenging given the dataset's characteristics.

Comparing GCE with GAC Figs. 6.4e and 6.4f, GCE's output appears noisy and less defined. GAC, with its generalized abstention, generally produces a somewhat more defined segmentation with improved coherence and reduced small-scale noise, indicating a marginal but noticeable benefit in handling the dataset's complexities.

SCE (Fig. 6.4g) often yields disjointed and noisy segmentations, particularly in regions with fine details or complex boundaries. SAC (Fig. 6.4h), leveraging abstention, tends to generate a slightly more unified segmentation with reduced misclassified pixels and a somewhat cleaner appearance, suggesting some advantage in handling noise and improving overall mask integrity.

While Dice (Fig. 6.4i) provides a more creative yet flawed segmentation, ADS (Fig. 6.4j) typically offers a more accurate and visually cleaner mask for DSAD. ADS demonstrates a consistent, albeit modest, improvement in boundary definition and overall coherence over its Dice baseline, indicating its effectiveness even on this difficult dataset by producing segmentations that are generally more faithful to the ground truth.

Chapter 7

Conclusions

While the critical importance of noise-robust loss functions in deep learning is widely acknowledged, there remains limited dedicated research investigating their specific utility and effectiveness within the domain of image segmentation. The abstention mechanism, as presented by the Deep Abstaining Classifier (DAC) and its informed extension IDAC, has demonstrably proven its effectiveness and capability in mitigating label noise primarily within classification tasks. In this thesis, we initially addressed this research gap by adapting DAC and IDAC to the segmentation domain, thereby demonstrating that the effectiveness of the abstention mechanism indeed extends to image segmentation, offering a viable strategy against label noise in this context.

Building upon this foundational adaptation, this thesis takes a significant step further by rigorously demonstrating the profound versatility and modularity of the abstention mechanism. Our core contribution lies in the systematic extension of this powerful concept through the integration of an enhanced and universal abstention framework into three other distinct and widely utilized loss functions: Generalized Cross Entropy (GCE), Symmetric Cross Entropy (SCE), and Dice Loss. This endeavour culminated in the introduction of three novel noise-robust loss functions: the Generalized Abstaining Classifier (GAC), the Symmetric Abstaining Classifier (SAC), and the Abstaining Dice Segmenter (ADS). The generalized abstention mechanism, as formulated in Section 4.3, incorporates an informed regularization term that guides abstention based on estimated noise rates and employs a flexible power-law-based α auto-tuning algorithm, offering superior control over the abstention behaviour compared to prior approaches. Furthermore, ADS introduces specific architectural adaptations for class-wise abstention, demonstrating the mechanism's adaptability to the unique characteristics of Dice Loss for segmentation tasks, thereby proving its broad applicability and robustness across different loss function paradigms.

The empirical evaluations presented in this work unequivocally validate the efficacy of the proposed abstaining loss functions. As detailed in Table 6.1, experiments conducted with a U-Net architecture consistently demonstrated that GAC, SAC, and ADS achieved superior performance compared to their respective non-abstaining baselines (GCE, SCE, and Dice Loss) across varying levels of label noise. This consistent performance uplift underscores the fundamental advantage conferred by the abstention mechanism: by intelligently identifying and disengaging from potentially corrupted training samples, the models can avoid internalizing erroneous information. This selective learning process, guided by the informed regularization and dynamic α tuning, enables the networks to focus

their learning capacity on reliable data, thereby fostering the development of more robust feature representations and decision boundaries. Qualitatively, as observed in Fig. 6.3, the abstaining variants generally produced visually cleaner, sharper, and more accurate segmentations on the CaDIS dataset, with ADS showing particularly strong performance.

Crucially, the architectural robustness of the abstention mechanism was further substantiated through evaluations with a distinct DeepLabV3+ architecture, as summarized in Table 6.2. While DAC and IDAC exhibited only marginal improvements or even slight performance decrements compared to CE on this architecture (likely attributable to hyperparameters not being specifically optimized for DeepLabV3+), our newly proposed GAC, SAC, and ADS consistently maintained their superior performance over their respective non-abstaining baselines. This provides strong empirical evidence for the architectural agnosticism and generalizability of our enhanced abstention mechanism. Even on the inherently more challenging DSAD dataset, characterized by sparse annotations and complex anatomical features, the abstaining variants generally produced segmentations that were marginally cleaner and more coherent compared to their non-abstaining baselines, as illustrated in Fig. 6.4. These findings collectively underscore the critical role of abstention in mitigating the detrimental effects of label noise and improving the generalization capabilities of deep learning models in challenging real-world scenarios.

The successful adaptation of the abstention mechanism to diverse loss functions—ranging from the probabilistic nature of Cross Entropy variants (GCE, SCE) to the overlap-centric Dice Loss—rigorously substantiates its universal applicability. This work establishes that abstention is not merely a specialized technique for specific loss functions but rather a powerful, generalizable extension that can fundamentally enhance the noise resistance of a broad spectrum of deep learning models. This modularity simplifies the development of robust AI systems, as existing high-performing architectures can be readily augmented with abstention capabilities without requiring extensive redesign.

7.1 Future Work

The present thesis has rigorously demonstrated the versatility and effectiveness of the generalized abstention mechanism in enhancing the noise robustness of deep learning models for medical image segmentation. Building upon this foundation, several promising directions for future research emerge.

Firstly, the current framework relies on a fixed, prior estimation of the noise rate ($\tilde{\eta}$ or $\tilde{\eta}_c$). A significant advancement would be to investigate adaptive, data-driven methods for noise level estimation. Instead of relying on a pre-defined hyperparameter, future research could explore techniques for dynamically learning or inferring the noise probability for each sample or region during training. This could involve integrating meta-learning approaches, uncertainty quantification methods, or even small, auxiliary networks trained to predict label quality. Such an 'end-to-end informed' abstention mechanism would make the models even more robust and less dependent on external, potentially inaccurate, noise estimations, thereby enhancing their applicability to real-world datasets with unknown or complex noise characteristics.

Secondly, while this thesis focused on synthetically introduced label noise, a crucial next step is to validate the proposed abstaining loss functions on real-world medical

image datasets afflicted with naturally occurring label noise. Real-world noise is often more complex, heterogeneous, and correlated with specific image features or annotation processes. Testing GAC, SAC, and ADS on such datasets would provide invaluable insights into their practical robustness and generalizability beyond controlled experimental settings. This could also involve exploring how the abstention mechanism interacts with other real-world challenges, such as inter-observer variability in annotations, which inherently introduces a form of label uncertainty.

Finally, given the demonstrated versatility of the abstention mechanism across different loss function characteristics, future work could explore its application to other challenging computer vision tasks beyond semantic segmentation. This includes extending the framework to instance segmentation, object detection, or even tasks in other domains where label noise or inherent ambiguity is prevalent. Furthermore, the abstention probability itself could be leveraged as a direct measure of model uncertainty, opening avenues for human-in-the-loop systems where highly uncertain or abstained predictions are flagged for expert review, thereby optimizing annotation efforts and enhancing the trustworthiness of AI-driven diagnostic tools. This would transform the abstention mechanism from merely a training technique into a comprehensive framework for robust and interpretable AI systems.

List of Figures

List of Tables

Bibliography

[1] Christopher Bishop. "Pattern Recognition and Machine Learning". In: vol. 16. Jan. 2006, pp. 140–155. DOI: 10.1117/1.2819119.

[2] Léon Bottou. "Large-Scale Machine Learning with Stochastic Gradient Descent". In: *Statistical Learning and Data Science*. Ed. by Mireille Gettler Summa et al. 0th ed. Chapman and Hall/CRC, Dec. 2011, pp. 33–42. ISBN: 978-0-429-10768-9. DOI: 10.1201/b11429-6.

[3] Matthias Carstens et al. "The Dresden Surgical Anatomy Dataset for Abdominal Organ Segmentation in Surgical Data Science". In: *Sci Data* 10.1 (Jan. 2023), p. 3. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01719-2.

[4] Liang-Chieh Chen et al. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Vol. 11211. Cham: Springer International Publishing, 2018, pp. 833–851. DOI: 10.1007/978-3-030-01234-2_49.

[5] Kehui Ding et al. "Improve Noise Tolerance of Robust Loss via Noise-Awareness". In: *IEEE Transactions on Neural Networks and Learning Systems* (2024), pp. 1–15. DOI: 10.1109/TNNLS.2024.3457029.

[6] Omar Elharrouss et al. *Loss Functions in Deep Learning: A Comprehensive Review*. Apr. 2025. DOI: 10.48550/arXiv.2504.04242. arXiv: 2504.04242 [cs].

[7] Boyan Gao, Henry Gouk, and Timothy M. Hospedales. "Searching for Robustness: Loss Learning for Noisy Classification Tasks". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 6650–6659. ISBN: 978-1-6654-2812-5. DOI: 10.1109/ICCV48922.2021.00660.

[8] Alberto Garcia-Garcia et al. *A Review on Deep Learning Techniques Applied to Semantic Segmentation*. Apr. 2017. DOI: 10.48550/arXiv.1704.06857. arXiv: 1704.06857 [cs].

[9] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, May 2010, pp. 249–256. URL: https://proceedings.mlr.press/v9/glorot10a.html.

[10] Alvaro Gonzalez-Jimenez et al. *Robust T-Loss for Medical Image Segmentation*. June 2023. DOI: 10.48550/arXiv.2306.00753. arXiv: 2306.00753 [cs].

[11] César González-Santoyo, Diego Renza, and Ernesto Moya-Albor. "Identifying and Mitigating Label Noise in Deep Learning for Image Classification". In: *Technologies* 13.4 (Apr. 2025), p. 132. ISSN: 2227-7080. DOI: 10.3390/technologies13040132.

[12] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[13] Maria Grammatikopoulou et al. *CaDIS: Cataract Dataset for Image Segmentation*. Feb. 2022. DOI: 10.48550/arXiv.1906.11586. arXiv: 1906.11586 [cs].

[14] Erjian Guo et al. *Imbalanced Medical Image Segmentation with Pixel-dependent Noisy Labels*. Jan. 2025. DOI: 10.48550/arXiv.2501.06678. arXiv: 2501.06678 [cs].

[15] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. Feb. 2009. ISBN: 0387848576.

[16] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, June 2016, pp. 770–778. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.90.

[17] Mohammad Hesam Hesamian et al. "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges". In: *J Digit Imaging* 32.4 (Aug. 2019), pp. 582–596. ISSN: 1618-727X. DOI: 10.1007/s10278-019-00227-x.

[18] Davood Karimi et al. "Deep Learning with Noisy Labels: Exploring Techniques and Remedies in Medical Image Analysis". In: *Medical Image Analysis* 65 (Oct. 2020), p. 101759. ISSN: 13618415. DOI: 10.1016/j.media.2020.101759.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Commun. ACM* 60.6 (May 2017), pp. 84–90. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3065386.

[20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning". In: *Nature* 521.7553 (May 2015), pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539.

[21] Peixia Li et al. "Semi-Supervised Semantic Segmentation under Label Noise via Diverse Learning Groups". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, pp. 1229–1238. ISBN: 979-8-3503-0718-4. DOI: 10.1109/ICCV51070.2023.00119.

[22] Julian Lienen and Eyke Hüllermeier. "Mitigating Label Noise through Data Ambiguation". In: *AAAI Conference on Artificial Intelligence 2024*. 12. 2024, pp. 13799–13807. URL: http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-118343-5.

[23] Geert Litjens et al. "A Survey on Deep Learning in Medical Image Analysis". In: *Medical Image Analysis* 42 (Dec. 2017), pp. 60–88. ISSN: 13618415. DOI: 10.1016/j.media.2017.07.005. arXiv: 1702.05747 [cs].

[24] Chenying Liu et al. *CromSS: Cross-modal Pre-Training with Noisy Labels for Remote Sensing Image Segmentation*. Mar. 2025. DOI: 10.48550/arXiv.2405.01217. arXiv: 2405.01217 [cs].

[25] Sheng Liu et al. "Adaptive Early-Learning Correction for Segmentation from Noisy Annotations". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, June 2022, pp. 2596–2606. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.00263.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully Convolutional Networks for Semantic Segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.

[27] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. Jan. 2019. DOI: 10.48550/arXiv.1711.05101. arXiv: 1711.05101 [cs].

[28] Michał Marcinkiewicz and Grzegorz Mrukwa. "Quantitative Impact of Label Noise on the Quality of Segmentation of Brain Tumors on MRI Scans". In: *2019 Federated*

Conference on Computer Science and Information Systems. Sept. 2019, pp. 61–65. DOI: 10.15439/2019F273.

[29] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *2016 Fourth International Conference on 3D Vision (3DV)*. Oct. 2016, pp. 565–571. DOI: 10.1109/3DV.2016.79.

[30] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning Series. Cambridge, MA: MIT Press, 2012. ISBN: 978-0-262-01825-8.

[31] Chengxuan Qian et al. *Adaptive Label Correction for Robust Medical Image Segmentation with Noisy Labels*. Mar. 2025. DOI: 10.48550/arXiv.2503.12218. arXiv: 2503.12218 [cs].

[32] Shenghai Rong et al. "Boundary-Enhanced Co-training for Weakly Supervised Semantic Segmentation". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, June 2023, pp. 19574–19584. ISBN: 979-8-3503-0129-8. DOI: 10.1109/CVPR52729.2023.01875.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. May 2015. DOI: 10.48550/arXiv.1505.04597. arXiv: 1505.04597 [cs].

[34] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning Representations by Back-Propagating Errors". In: *Nature* 323.6088 (Oct. 1986), pp. 533–536. ISSN: 1476-4687. DOI: 10.1038/323533a0.

[35] Jürgen Schmidhuber. "Deep Learning in Neural Networks: An Overview". In: *Neural Networks* 61 (2015), pp. 85–117. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2014.09.003.

[36] Helen Schneider et al. *Informed Deep Abstaining Classifier: Investigating Noise-Robust Training for Diagnostic Decision Support Systems*. Oct. 2024. DOI: 10.48550/arXiv.2410.21014. arXiv: 2410.21014 [cs].

[37] C. E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[38] Jialin Shi and Ji Wu. "Distilling Effective Supervision for Robust Medical Image Segmentation with Noisy Labels". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Ed. by Marleen de Bruijne et al. Cham: Springer International Publishing, 2021, pp. 668–677. ISBN: 978-3-030-87193-2. DOI: 10.1007/978-3-030-87193-2_63.

[39] Yucheng Shu, Xiao Wu, and Weisheng Li. "LVC-Net: Medical Image Segmentation with Noisy Label Based on Local Visual Cues". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen et al. Cham: Springer International Publishing, 2019, pp. 558–566. ISBN: 978-3-030-32226-7. DOI: 10.1007/978-3-030-32226-7_62.

[40] Nahian Siddique et al. "U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications". In: *IEEE Access* 9 (2021), pp. 82031–82057. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3086020.

[41] Max Staats, Matthias Thamm, and Bernd Rosenow. "Enhancing Noise-Robust Losses for Large-Scale Noisy Data Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 39.7 (Apr. 2025), pp. 7006–7014. DOI: 10.1609/aaai.v39i7.32752.

[42] Sunil Thulasidasan et al. *Combating Label Noise in Deep Learning Using Abstention*. Aug. 2019. DOI: 10.48550/arXiv.1905.10964. arXiv: 1905.10964 [stat].

[43] William Toner and Amos Storkey. "Label Noise: Correcting a Correction Loss". In: *The Second Workshop on New Frontiers in Adversarial Machine Learning*. 2023. URL: https://openreview.net/forum?id=FenYb7HXSy.

[44] Petar Veličković et al. *Graph Attention Networks*. Feb. 2018. DOI: 10.48550/arXiv.1710.10903. arXiv: 1710.10903 [stat].

[45] Yisen Wang et al. "Symmetric Cross Entropy for Robust Learning With Noisy Labels". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 322–330. ISBN: 978-1-7281-4803-8. DOI: 10.1109/ICCV.2019.00041.

[46] Yan Xu et al. "Advances in Medical Image Segmentation: A Comprehensive Review of Traditional, Deep Learning and Hybrid Approaches". In: *Bioengineering* 11.10 (Oct. 2024), p. 1034. ISSN: 2306-5354. DOI: 10.3390/bioengineering11101034.

[47] Jiachen Yao et al. *Learning to Segment from Noisy Annotations: A Spatial Correction Approach*. July 2023. DOI: 10.48550/arXiv.2308.02498. arXiv: 2308.02498 [eess].

[48] Xichen Ye et al. *Active Negative Loss: A Robust Framework for Learning with Noisy Labels*. Dec. 2024. DOI: 10.48550/arXiv.2412.02373. arXiv: 2412.02373 [cs].

[49] Rumeng Yi et al. "Learning from Pixel-Level Label Noise: A New Perspective for Semi-Supervised Semantic Segmentation". In: *IEEE Trans. on Image Process.* 31 (2022), pp. 623–635. ISSN: 1057-7149, 1941-0042. DOI: 10.1109/TIP.2021.3134142. arXiv: 2103.14242 [cs].

[50] Le Zhang et al. "Disentangling Human Error from Ground Truth in Segmentation of Medical Images". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 15750–15762.

[51] Zhilu Zhang and Mert Sabuncu. "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018.

[52] Bolei Zhou et al. "Learning Deep Features for Discriminative Localization". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, June 2016, pp. 2921–2929. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.319.

[53] Yuyin Zhou et al. "L2B: Learning to Bootstrap Robust Models for Combating Label Noise". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, June 2024, pp. 23523–23533. ISBN: 979-8-3503-5300-6. DOI: 10.1109/CVPR52733.2024.02220.

[54] Haidong Zhu, Jialin Shi, and Ji Wu. *Pick-and-Learn: Automatic Quality Evaluation for Noisy-Labeled Image Segmentation*. July 2019. DOI: 10.48550/arXiv.1907.11835. arXiv: 1907.11835 [cs].