To Develop Some Novel Uncertainty Estimation Models Using Deep Learning Techniques For Image Segmentation

Thesis submitted by

Somenath Kuiry

Doctor of Philosophy(Science)

Department of Mathematics
Faculty of Science
Jadavpur University
Kolkata, India

2025

JADAVPUR UNIVERSITY
KOLKATA, INDIA

Index No: 54/18/Maths./25

1. **Title of the thesis**
   To Develop Some Novel Uncertainty Estimation Models Using Deep Learning Techniques For Image Segmentation

2. **Name Designation & Institution of Supervisors**

   a) Prof. (Dr.) Alaka Das,

   Professor, Dept. of Mathematics

   Jadavpur University, India

   b) Prof. (Dr.) Nibaran Das,

   Professor, Dept. of CSE

   Jadavpur University, India

3. **List of Publications:**

   JOURNALS

   - Bhowmick, S., **Kuiry, S.**, Das, A., Das, N. and Nasipuri, M., 2022. Deep learning-based outdoor object detection using visible and near-infrared spectrum. Multimedia Tools and Applications, 81(7), pp.9385-9402.
   - **Kuiry. S**, Guha. D, Das. A, Nasipuri. M and Das. N,"GA-RISE: Posthoc Model Agnostic Explanations of Black-box Classifiers using Genetic Algorithm-based Optimized Masks - A Case Study on Chest X-ray Images ", *International Journal on Artificial Intelligence Tools*.
   - **Kuiry. S**, Das. A, Nasipuri. M and Das. N,"Copula-based Fusion of Multi-level Superpixels for Semantic Image Segmentation ", *Communicated to Journal of Visual Communication and Image Representation*.
   - **Kuiry. S**, Das. A, Rizk. R, Santosh. KC, Nasipuri. M and Das. N,"Leveraging Class-Specific Copula Functions for Enhanced Ensemble Learning in Medical Imaging ", *To be Communicated*.

- **Kuiry. S**, Das. A, Nasipuri. M and Das. N,"Leveraging Class-specific Distribution Estimation to Model multi-Annotator Disagreement and Annotator-specific Preference for Medical Image Segmentation ", (Communicated to *IEEE Transactions on Biomedical Engineering*,).

CONFERENCE PAPERS

- Gani, M.O., **Kuiry, S.**, Das, A., Nasipuri, M., Das, N. (2021). Multispectral Object Detection with Deep Learning. In: Dutta, P., Mandal, J.K., Mukhopadhyay, S. (eds) Computational Intelligence in Communications and Business Analytics. CICBA 2021. Communications in Computer and Information Science, vol 1406. Springer, Cham. https://doi.org/10.1007/978-3-030-75529-4_9
- Panja, A., **Kuiry, S.**, Das, A., Nasipuri, M., Das, N. (2025). COVID-CT-H-UNet: A Novel COVID-19 CT Segmentation Network Based on Attention Mechanism and Bi-Category Hybrid Loss. In: Singh, J.P., Singh, M.P., Singh, A.K., Mukhopadhyay, S., Mandal, J.K., Dutta, P. (eds) Computational Intelligence in Communications and Business Analytics. CICBA 2024. Communications in Computer and Information Science, vol 2366. Springer, Cham. https://doi.org/10.1007/978-3-031-81342-9_9
- **Kuiry, S.**, Das, A., Nasipuri, M., Das, N. (2025). Regularizing CNNs Using Confusion Penalty Based Label Smoothing for Histopathology Images. In: Singh, J.P., Singh, M.P., Singh, A.K., Mukhopadhyay, S., Mandal, J.K., Dutta, P. (eds) Computational Intelligence in Communications and Business Analytics. CICBA 2024. Communications in Computer and Information Science, vol 2367. Springer, Cham. https://doi.org/10.1007/978-3-031-81339-9_22

4. List of Patents: None

5. List of Presentations in National/International/Conferences/Workshops

- Panja, A., **Kuiry, S.**, Das, A., Nasipuri, M., Das, N. (2025). COVID-CT-H-UNet: A Novel COVID-19 CT Segmentation Network Based on Attention Mechanism and Bi-Category Hybrid Loss. In: Singh, J.P., Singh,

M.P., Singh, A.K., Mukhopadhyay, S., Mandal, J.K., Dutta, P. (eds) Computational Intelligence in Communications and Business Analytics. CICBA 2024. Communications in Computer and Information Science, vol 2366. Springer, Cham. https://doi.org/10.1007/978-3-031-81342-9_9

- **Kuiry, S.**, Das, A., Nasipuri, M., Das, N. (2025). Regularizing CNNs Using Confusion Penalty Based Label Smoothing for Histopathology Images. In: Singh, J.P., Singh, M.P., Singh, A.K., Mukhopadhyay, S., Mandal, J.K., Dutta, P. (eds) Computational Intelligence in Communications and Business Analytics. CICBA 2024. Communications in Computer and Information Science, vol 2367. Springer, Cham. https://doi.org/10.1007/978-3-031-81339-9_22

Declaration

I, **Somenath Kuiry**, registered on **15/02/2018**, hereby declare that the thesis entitled **"To Develop Novel Uncertainty Estimation Models Using Deep Learning Techniques for Image Segmentation"**, submitted for the award of the degree of Doctor of Philosophy (Science) at Jadavpur University, is my own original work carried out under the supervision of Prof. (Dr.) Alaka Das, Department of Mathematics, Jadavpur University, and Prof. (Dr.) Nibaran Das, Department of Computer Science and Engineering, Jadavpur University, India.

I further declare that the work reported in this thesis, in whole or in part, has not been submitted and will not be submitted for any other degree or diploma at this or any other institution.

All information in this thesis has been obtained and presented in accordance with existing academic rules and ethical conduct. I declare that, as required by these rules, I have fully cited and referenced all materials and results that are not original to this work.

I also declare that I have checked this thesis as per the "Policy on Anti Plagiarism, Jadavpur University, 2019", and the level of similarity as checked by iThenticate software is 4%.

Somenath Kuiry

Signature of the Candidate:

Date: 8/5/2025

CERTIFICATE FROM THE SUPERVISORS

This is to certify that the thesis entitled **"To Develop Some Novel Uncertainty Estimation Models Using Deep Learning Techniques For Image Segmentation"** submitted by **Sri Somenath Kuiry**, who got his name registered on **15/02/2018** for the award of Ph. D. (Science) degree of Jadavpur University is absolutely based upon his own work under the supervision of **Prof. Alaka Das** and **Prof. Nibaran Das** and that neither his thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award anywhere before.

Alaka Das 8/5/2025

Dr. Alaka Das

Professor

Department of Mathematics

Jadavpur University

Professor
Department of Mathematics
Jadavpur University
Kolkata - 700 032

Nibaran Das 8/5/2025

Dr. Nibaran Das

Professor

Department of Computer Science and Engineering

Jadavpur University

Professor
Computer Sc. & Engg. Department
Jadavpur University
Kolkata - 700032

ACKNOWLEDGMENTS

I am deeply grateful to everyone who supported me throughout this journey.

First and foremost, I extend my heartfelt thanks to my supervisors, Dr. Alaka Das and Dr. Nibaran Das. Your insightful guidance, unwavering encouragement, and trust in my abilities have been fundamental to the completion of this work. I have learned far more than methodology - your mentorship has shaped my approach to research and life.

I also wish to thank Prof. Mita Nasipuri for her invaluable advice and critical perspectives, which consistently sharpened my thinking and enriched this thesis.

My appreciation goes to Prof. Subhas Chandra Mondal, HOD of the Department of Mathematics, and Prof. Nirmalya Chowdhury, HOD of the Department of Computer Science & Engineering at Jadavpur University, for ensuring access to all facilities and resources necessary for my research.

This project would not have been possible without the financial support of the INSPIRE Fellowship Programme(IF170641), Department of Science & Technology, Government of India. Your backing provided me the stability to focus fully on my studies.

I'm grateful to my collaborators - Md. Osman Gani, Subhadeep Bhowmick, and Anay Panja - for their shared dedication, creative energy, and friendship throughout our joint work.

Thank you to my colleagues and friends at the CMATER Laboratory - Soumyajyoti Dey, Dibyasree Guha, Dr. Bidhan Barai, Rahul Laxmanrao Meshram, Dr. Neelotpal Chakraborty, Dr. Debapriyo Banik, Dr. Nirmal Das, Dr. Kaushiki Roy, Dr. Swarnendu Ghosh, Pabitra Modal, Snehashis Sahoo, Raju Naskar, and Sandip Pramanik - for their camaraderie, thoughtful discussions, and practical support.

I am also indebted to my dear friends - Ananya, Raja, Suman, Abhinandan, Debajyoti, Uttam, Rajnarayan, Tanmita, Akashlina, Shamik, Shyam, Rijubrata, Prerana, Pranali, Sayani, and Sudipa - whose unwavering encouragement and shared laughter made this journey far more enjoyable.

Finally, I dedicate this thesis to the memory of my parents and to my brother, whose love and belief in me have been my guiding light. Your support carried me through every challenge, and this work is a testament to the foundation you provided.

Dedicated to
my father Late Jagabandhu Kuiry
my mother Late Archana Kuiry
my brother Nabin Chandra Kuiry

ABSTRACT

Deep learning–based image segmentation is indispensable in critical fields like medical diagnosis and autonomous navigation, yet model performance is undermined by various uncertainties, such as, ambiguous boundaries, inconsistent human annotations, and overconfident predictions. This thesis introduces a comprehensive framework of methods to quantify and mitigate these uncertainties, thereby enhancing segmentation accuracy, robustness, and interpretability.

First, we address **boundary uncertainty** by augmenting a U-Net with a spatial attention mechanism and proposing a Bi-category Hybrid Loss that jointly optimizes region overlap and edge sharpness. The resulting **COVID-CT-H-UNet** model demonstrates a 12% gain in Dice score and a 15% reduction in boundary error on COVID-19 lung CT scans compared to leading baselines.

To resolve **label uncertainty**, we develop a **multi-annotator consensus framework**. Our Class-Specific Distribution Learning captures the full distribution of expert labels per class, and the Annotator-Specific Preference Estimator model's individual biases. On public benchmarks (RIGA, QUBIQ), this approach reduces annotation-driven variability by 20% and boosts generalization to unseen annotators.

For **model uncertainty**, we propose two complementary strategies. A **copula-based ensemble** explicitly learns dependencies among multiple segmentation networks, yielding consistent 1–2% improvements in overall accuracy and mean IoU on urban scene and medical datasets. In parallel, **Confusion-Penalty Label Smoothing (CPLS)** adaptively reallocates smoothing mass based on validation-set confusion matrices, cutting Expected Calibration Error by up to 35% while improving classification accuracy by 3%.

Beyond segmentation, we demonstrate the broader applicability of uncertainty-aware learning. In an **object detection** setting—which is not a segmentation task—early fusion of RGB, near-infrared, and thermal imagery with YOLOv3 raises mAP from 71.5% to 78.6% and halves prediction variance. Lastly, we introduce **GA-RISE**, a genetic-algorithm-optimized saliency method for **classification and detection**, which produces stable, focused heatmaps and outperforms RISE and GradCAM in insertion/deletion metrics.

Collectively, these contributions offer a unified toolkit for uncertainty estimation, improving boundary delineation, reconciling annotation variability, calibrating model

confidence, integrating multi-modal data for detection, and delivering robust visual explanations—paving the way toward safer, more trustworthy deep learning systems in both segmentation and broader computer vision tasks.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Deep Learning and Image Segmentation

Rapid advancement of artificial intelligence (AI) has profoundly impacted numerous domains, reshaping how machines perceive, process, and interpret complex data. At the forefront of this revolution is deep learning, a subset of machine learning inspired by the structure and function of the human brain. Deep learning has emerged as a transformative paradigm, enabling machines to learn intricate patterns and representations from data, leading to groundbreaking achievements in areas such as natural language processing, speech recognition, and computer vision [152, 99, 97]. Within computer vision, one of the most critical and challenging tasks is image segmentation, which involves partitioning an image into semantically meaningful regions or objects.

Image segmentation plays an integral role in diverse applications across multiple disciplines. For instance, in autonomous driving, it facilitates road-scene understanding by distinguishing between roads, pedestrians, vehicles, and other elements [71, 112, 41]. In medical imaging, segmentation is crucial for detecting and delineating tumors, organs, or lesions, enabling accurate diagnostics and treatment planning [30, 118, 95]. Similarly, in satellite-imagery analysis, segmentation aids in identifying land-use patterns, monitoring environmental changes, and disaster management [117, 75]. These examples underscore the importance of segmentation in interpreting visual data, making it a cornerstone of modern computer vision.

Despite the significant progress achieved through traditional image-segmentation methods such as thresholding [129, 113], clustering [101, 131], and edge detection [105, 26], these techniques often struggle with complex datasets, variability in object appearance, and noise [159]. The introduction of deep learning has brought about a paradigm shift in how image segmentation is approached. By leveraging the power of convolutional neural networks (CNNs) and their ability to learn hierarchical features from data, deep-learning-based techniques have demonstrated remarkable performance improvements, surpassing traditional approaches in accuracy and robustness [98, 40, 151]. These methods have unlocked new possibilities, enabling machines to tackle segmentation challenges in ways previously thought unattainable.

However, as deep-learning models become increasingly integral to segmentation tasks, particularly in high-stakes applications like healthcare and autonomous systems, the question of reliability and interpretability becomes paramount. A critical aspect of reliability is the estimation of uncertainty in predictions. Understanding and quantifying uncertainty in segmentation outputs is essential, especially in scenarios where decisions can have significant consequences. For example, in a medical setting, an incorrect segmentation without a measure of confidence could lead to flawed diagnoses or treatment plans. Similarly, in autonomous driving, the inability to identify uncertain predictions could compromise safety.

This thesis seeks to address these challenges by developing novel uncertainty-estimation models tailored for deep-learning-based image segmentation. The primary aim is to enhance the reliability and interpretability of segmentation models while maintaining high accuracy and computational efficiency. To provide a comprehensive foundation for this research, the following sections delve into the principles and evolution of deep learning, the fundamentals and significance of image segmentation, and the transformative role of deep-learning techniques in segmentation tasks.

1.1.1 Deep Learning

Deep learning, a subset of machine learning, draws inspiration from the brain's structure and function. It utilizes multiple layers of artificial neural networks to extract increasingly complex features from data. The foundation of deep learning

dates back to the McCulloch–Pitts model of the neuron in 1943, which established the mathematical principles behind neural networks. In the 1950s, the perceptron was developed, which could learn basic linear relationships. Despite early promise, neural networks faced challenges in handling multiple layers due to computational limitations and the absence of efficient training techniques, which caused a decline in research interest during the 1970s and 1980s.

The revival of deep learning occurred with the creation of the back-propagation algorithm, which allowed for more effective training of multi-layer networks. However, it wasn't until the early 2000s—with improvements in hardware (such as GPUs), access to large annotated datasets, and the introduction of architectures like convolutional neural networks (CNNs)—that deep learning began to show its transformative potential. A key moment in this shift was the success of AlexNet in the 2012 ImageNet competition, which established deep learning as a dominant method in artificial-intelligence research.

The core idea behind deep learning is the use of multiple layers to build abstract representations of data. Each layer takes the input and transforms it into a more complex feature space. Neurons in these layers apply weights and biases to the input data, passing it through activation functions to introduce non-linearity. By stacking layers, a network is created that can learn increasingly intricate features. Activation functions like ReLU, sigmoid, and tanh allow the network to model complex relationships. Training the network involves optimization algorithms such as stochastic gradient descent (SGD) or Adam, which adjust the model parameters to minimize the loss function.

Notable deep-learning architectures include convolutional neural networks (CNNs), which are particularly effective in analyzing image data and capturing spatial hierarchies through convolution and pooling layers. Recurrent neural networks (RNNs) are designed to handle sequential data, retaining temporal dependencies, and their advanced variants like LSTMs (long short-term memory networks) improve the handling of longer sequences. Transformers, which rely on self-attention mechanisms, have revolutionized both natural-language processing and vision tasks, setting new benchmarks in these domains.

Deep learning has made significant advancements in fields such as computer vision, natural-language processing, and robotics. However, despite its successes, several challenges persist, including high computational demands, heavy reliance

on large datasets, and a lack of model interpretability. A critical challenge is the inability of many models to estimate uncertainty—an essential aspect in applications where safety is a concern.

1.1.2 Image Segmentation

Image segmentation is the process of dividing an image into distinct, meaningful regions or objects, enabling machines to better understand and analyze visual data. It is a key task in computer vision and plays a crucial role in various applications. In medical imaging, segmentation helps in identifying tumors, organs, and abnormalities, which is essential for accurate diagnosis and treatment planning. In autonomous vehicles, it is used to interpret road scenes by detecting vehicles, pedestrians, and road boundaries. Remote-sensing applications also rely on segmentation for monitoring environmental changes and managing disasters through satellite imagery.

1.1.2.1 Types of Image Segmentation Methods

Image segmentation can be divided into several distinct categories, each focusing on different aspects of image interpretation.

1. **Semantic segmentation** is one of the most widely used approaches, where each pixel in an image is labeled according to a predefined class. This type of segmentation assigns the same label to all pixels belonging to a particular object category, such as “car” or “tree.” However, a limitation of semantic segmentation is that it does not differentiate between different instances of the same object type; for example, all “cars” are treated as a single group without distinguishing individual vehicles.

2. **Instance segmentation** improves upon semantic segmentation by not only labeling pixels with object categories but also differentiating between multiple instances of the same object type. For example, in an image with several cars, instance segmentation would recognize each car as a separate entity and assign distinct labels to each one. This distinction is critical in applications where identifying individual objects within a category is necessary, such

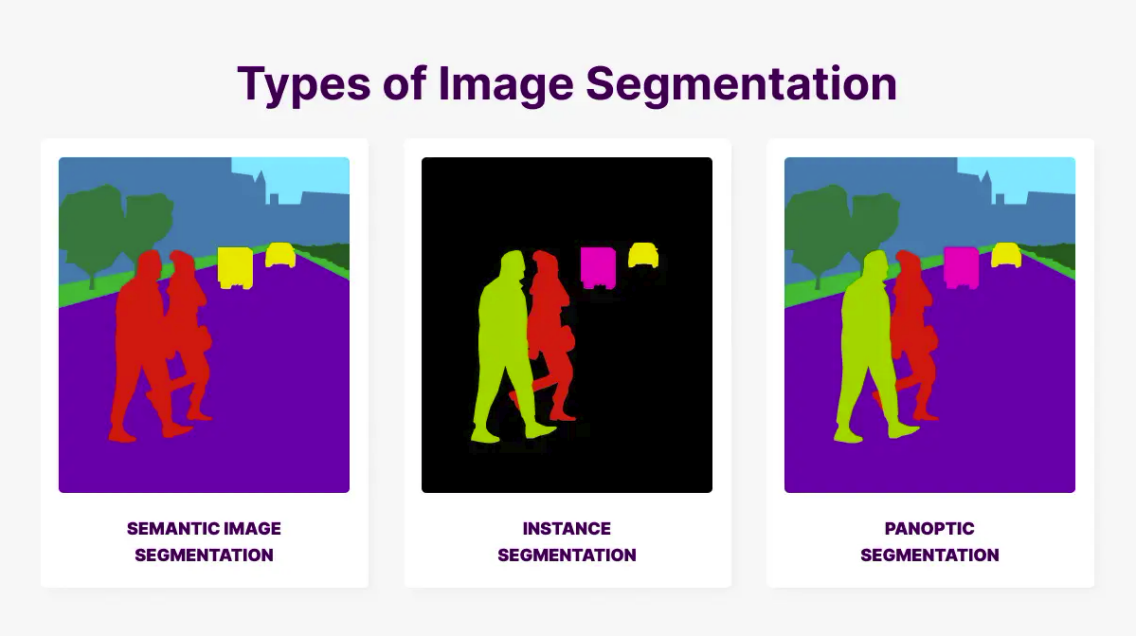

Figure 1.1: Types of Image Segmentation[1]

as in self-driving cars or object tracking. Typically, background elements such as trees and grass are removed or ignored in instance segmentation, since the primary goal is to detect and differentiate specific, countable objects of interest rather than generic environmental features.

3. **Panoptic segmentation** brings together the concepts of semantic and instance segmentation into a unified framework. It aims to provide a comprehensive understanding of the image by labeling every pixel either as part of an object instance or as part of the background. Panoptic segmentation provides a complete scene representation, making it particularly valuable in tasks that require full scene comprehension, such as autonomous navigation and advanced robotics.

4. **Depth segmentation** is another category, focusing on estimating the spatial distance of each pixel from the camera. This type of segmentation is particularly important in applications like 3-D reconstruction and autonomous driving, where understanding the depth and distance of objects in a scene is crucial for navigation and decision-making.

[1]Image Source: https://mindy-support.com/news-post/what-is-image-segmentation-the-basics-and-key-techniques/

Each type of image segmentation has its specific strengths and is suitable for particular applications. For instance, semantic segmentation is effective in simpler scenarios such as medical imaging, where the main task is to classify regions of interest. On the other hand, instance and panoptic segmentation are more appropriate for complex environments, such as in autonomous driving, where accurate object identification and scene parsing are vital.

1.1.3 Image Segmentation with Deep Learning Techniques

1.1.3.1 Transformative Impact of Deep Learning

Deep learning has fundamentally transformed the field of image segmentation by allowing models to automatically learn hierarchical representations from raw image data. Unlike traditional image-segmentation techniques that depend heavily on handcrafted features, deep-learning models have the capability to extract relevant, task-specific features directly from the data [40]. This characteristic makes deep-learning approaches more adaptable, capable of handling a wide variety of datasets and complex segmentation tasks. For example, in medical imaging, deep-learning models can learn to differentiate between various tissues or abnormalities without explicit human intervention, a task that was previously reliant on predefined feature-extraction methods.

1.1.3.2 Fundamentals of a Deep Neural Network (DNN)

Before exploring deep-learning models tailored for segmentation, it is essential to understand the structure of a **Deep Neural Network (DNN)**, which serves as the foundation for more complex architectures. A **DNN** consists of multiple layers of artificial neurons that progressively refine feature representations. The primary components of a DNN are:

- **Input Layer:** Receives raw image data, such as an RGB image of size $H \times W \times 3$.

- **Hidden Layers:** Composed of multiple fully connected layers with activation functions (e.g., *ReLU*) to introduce non-linearity and improve feature learning.
- **Output Layer:** Utilizes a softmax or sigmoid activation function to generate class probabilities, making it adaptable for segmentation tasks.

The transformation at each hidden layer is mathematically expressed as:

$$h = \sigma(Wx + b) \tag{1.1}$$

where x is the input, W is the weight matrix, b is the bias term, and $\sigma(\cdot)$ denotes the activation function.
While DNNs are effective in classification tasks, their fully connected layers disregard spatial dependencies, necessitating the use of **Convolutional Neural Networks (CNNs)** and **Fully Convolutional Networks (FCNs)** for segmentation tasks.

1.1.4 Key Architectures for Segmentation

A range of deep-learning architectures has been developed to address different challenges in image segmentation. These models vary in their approach and design but share a common goal of improving segmentation accuracy, particularly in complex images.

1. **Fully Convolutional Networks (FCNs) [93, 137]:** FCNs have fundamentally altered the approach to segmentation by replacing fully connected layers with convolutional layers, allowing the network to generate pixel-wise predictions. This change makes FCNs particularly suitable for tasks where precise localization of objects is needed, such as segmenting tumor regions in medical scans. Since FCNs operate on raw pixel data and maintain spatial information, they have become the foundation for many modern segmentation models.
2. **U-Net [8, 123]:** Originally developed for medical image segmentation, U-Net has become one of the most widely used architectures in this domain. It features a symmetric encoder–decoder structure with skip connections,

which helps the model preserve spatial resolution while extracting hierarchical features. In medical imaging, U-Net has proven especially effective in segmenting fine details, such as distinguishing between different layers of tissue or identifying subtle lesions. Its architecture, which enables detailed localization alongside efficient learning of features, is also being explored in other fields like satellite image segmentation.

3. **DeepLab [18]:** DeepLab introduces the concept of atrous (or dilated) convolutions, which allow the network to capture multiscale contextual information by modifying the receptive field without losing resolution. This enables the model to recognize objects at varying scales, such as identifying both large structures (e.g., roads or buildings) and smaller objects (e.g., pedestrians or vehicles) in a single image. The addition of conditional random fields (CRFs) further refines the segmentation boundaries, making DeepLab a powerful tool for real-world applications like autonomous driving, where precise object-boundary delineation is crucial.

4. **Mask R-CNN [49]:** Mask R-CNN extends the capabilities of object-detection models by incorporating instance segmentation, a technique that not only identifies object classes but also generates masks for each individual object instance. This is particularly useful in tasks like image parsing and object tracking. In autonomous driving, for example, Mask R-CNN can detect and segment multiple vehicles, pedestrians, and road signs in real-time, providing detailed information for navigation systems.

5. **Vision-Transformer-Based Segmentation Models [86]:** Recent advancements in **vision transformers (ViTs)** have led to segmentation models like **SETR** [169] and **Segment Anything** [74], which employ attention mechanisms to model long-range dependencies, often outperforming CNN-based approaches.

1.1.4.1 Advantages of Deep-Learning Approaches

Deep-learning-based image-segmentation approaches offer several compelling advantages:

1. **Accuracy:** One of the primary strengths of deep-learning models is their ability to learn complex, non-linear patterns from large datasets. This enables them to perform well in segmentation tasks that involve intricate details, such as detecting small structures in medical scans or distinguishing overlapping objects in high-resolution images. The ability to learn directly from data reduces the reliance on human expertise in feature selection, resulting in models that can outperform traditional methods.

2. **Adaptability:** Deep-learning models generalize well across different domains and datasets. For example, a deep-learning model trained for medical imaging can often be fine-tuned for use in remote sensing or satellite imagery with relatively minimal adjustments. This adaptability is a key advantage in real-world applications, where datasets may vary significantly in terms of size, quality, or content.

3. **Feature Hierarchies:** Deep-learning models are able to learn hierarchical features, starting from low-level patterns like edges and textures and progressing to high-level features such as object parts and complete objects. This multilevel feature extraction enables deep models to handle complex images with varying textures, lighting conditions, and object shapes, as seen in applications like autonomous vehicles or industrial-inspection systems.

1.1.4.2 Limitations and Opportunities

Despite the numerous advantages of deep learning in image segmentation, these models come with their own set of challenges. One of the primary limitations is the **high computational cost** associated with training and deploying deep-learning models, particularly for large-scale datasets. For example, training a model like Mask R-CNN or DeepLab can require substantial computational resources, including high-performance GPUs and long processing times. Additionally, deep-learning models lack inherent **uncertainty-estimation mechanisms**, which are critical in safety-critical applications such as medical diagnostics and autonomous driving. The inability to assess model confidence can lead to unreliable predictions in situations where uncertainty is a key factor in decision-making, such as identifying potentially cancerous tissue or making real-time navigation decisions in dynamic environments.

Addressing these limitations presents an opportunity to enhance the applicability and fidelity of deep-learning-based segmentation models. Research into methods for reducing computational costs through techniques like model pruning or efficient architectures, as well as integrating uncertainty estimation into these models, holds significant potential for improving their deployment in practical, real-world settings.

1.1.4.3 Relevance to Uncertainty Estimation

In certain high-stakes applications, such as medical diagnostics and autonomous driving, it is not enough to simply make a prediction—understanding the model's confidence in its decision is equally important. For example, a medical image-segmentation model may predict the presence of a tumor, but if the model is uncertain about the boundaries, this could affect treatment decisions. Similarly, in autonomous driving, an uncertain detection of pedestrians or vehicles could lead to safety risks. As such, integrating **uncertainty estimation** into deep-learning-based segmentation models is critical. Uncertainty-estimation methods, such as Bayesian deep learning, allow the model to quantify the confidence of its predictions, providing more reliable decision-making in applications where safety is paramount. This is a key area of focus for this thesis, which aims to explore and develop novel techniques for incorporating uncertainty estimation into deep-learning frameworks for image segmentation.

1.2 What is Uncertainty?

Uncertainty can be defined as the lack of knowledge about something or ambiguous knowledge concerning it. Uncertainty is a part of life. Whether we're making everyday decisions or solving complex problems, there's often some level of doubt about the outcome. This uncertainty can come from not having enough information, the natural unpredictability of the world, or simply the limits of what we know and can understand.

1.2.1 What is Uncertainty in Deep Learning?

Uncertainty in the context of deep learning may be defined as the ambiguity associated with a prediction made by a model [36]. Uncertainty represents the degree of confidence in a model's predictions. It is crucial for assessing the reliability of outputs, especially in applications where decisions carry significant risks. By understanding uncertainty, we can create systems that are not only smarter but also safer and more dependable.

Take medical imaging as an example: a deep-learning model might predict whether a tumor is benign or malignant. If the model expresses high uncertainty in its prediction, it may prompt doctors to conduct additional tests or seek second opinions. Similarly, in autonomous vehicles, models must decide whether an object in the path is a pedestrian or a visual distortion caused by poor lighting. A high uncertainty level might trigger the vehicle to slow down or stop, prioritizing safety.

1.2.2 Types of Uncertainty

Uncertainty in machine learning comes in two main forms [36]: epistemic uncertainty and aleatoric uncertainty. Epistemic uncertainty is related to limitations in the model itself, often caused by insufficient or unrepresentative training data. For instance, a model trained primarily on daytime traffic images might struggle to make accurate predictions during nighttime conditions, highlighting epistemic uncertainty. Aleatoric uncertainty, on the other hand, is linked to inherent noise or variability in the input data. An example would be a speech-recognition system processing audio with background noise, which can lead to uncertain outputs.

By explicitly accounting for these types of uncertainty, machine-learning systems can better handle ambiguous situations, offering greater confidence and reliability in their predictions and decisions.

1.2.2.1 Aleatoric Uncertainty (or Data Uncertainty)

Aleatoric uncertainty refers to the inherent variability or noise present in data that cannot be eliminated, even with an ideal model. This type of uncertainty

arises from factors such as measurement errors, sensor inaccuracies, or environmental variability, making certain aspects of the data unpredictable. For instance, in autonomous driving, low-light conditions, rain, or fog may introduce visual ambiguities, leading to uncertainty in object detection. Similarly, in medical imaging, low-resolution scans or artifacts caused by imaging equipment can make it difficult for models to confidently identify abnormalities. Speech-recognition systems also encounter aleatoric uncertainty when processing audio with overlapping voices or background noise, which obscures the clarity of spoken words.

Addressing aleatoric uncertainty often involves improving the quality of input data or designing models that can learn to estimate and account for the noise. Data-augmentation techniques, such as adding simulated noise during training, can help models become more robust to variations. Using probabilistic models or incorporating uncertainty-quantification techniques, such as Bayesian neural networks, allows models to explicitly express their confidence in predictions. Additionally, enhancing data-collection methods—such as using higher-resolution sensors or more precise equipment—can reduce the impact of aleatoric uncertainty, leading to more reliable predictions in challenging scenarios.

1.2.2.2 Epistemic Uncertainty (or Model Uncertainty)

Epistemic uncertainty arises from a lack of knowledge or limitations in the model, such as insufficient training data or an incomplete representation of the problem space. This type of uncertainty reflects the model's ignorance about parts of the data distribution and can, in principle, be reduced by providing more or better-quality data. For example, a self-driving-car model trained primarily on urban road conditions might struggle to predict accurately when encountering rural, unpaved roads, as these scenarios were not part of its training data. Similarly, a medical-diagnostic model trained on adult-patient data may exhibit uncertainty when evaluating pediatric cases, as the training set did not include this subgroup. Another example is in natural-language processing, where a model trained on formal text might perform poorly on slang-heavy social-media content, revealing epistemic uncertainty.

To address epistemic uncertainty, it is essential to expand the training dataset to include diverse and representative examples of the target domain. Active-learning

techniques can help by identifying uncertain predictions, which often point to areas where the model needs more data. Another approach is using ensemble methods, which can provide insights into uncertainty by combining predictions from multiple models, or Bayesian neural networks, which approximate the posterior distribution over model parameters. Regularly updating models with new data from underrepresented scenarios or domains can further reduce epistemic uncertainty, making systems more robust and reliable across a broader range of applications.

1.2.3 Uncertainty in the Context of Image Segmentation

Image segmentation is a critical computer-vision task that involves classifying each pixel in an image and assigning it to a specific class to create a precise map of objects or regions. Unlike image classification, which provides a single label for the entire image, segmentation operates at the pixel level, requiring a far more granular understanding of the visual data. This added complexity introduces various challenges, particularly in terms of managing uncertainty. Uncertainty in image segmentation reflects the model's lack of confidence in its pixel-level predictions, which can significantly affect the overall quality and reliability of the segmentation results [2]. Factors such as variations in object size, visibility, texture, lighting, and occlusion further exacerbate these challenges, making uncertainty management a crucial aspect of segmentation [161].

Broadly, uncertainty in image segmentation can be categorized into three major areas: boundary regions, label quality, and model limitations. Each of these presents distinct challenges and opportunities for improvement.

1.2.3.1 Uncertainty in Boundary Regions

Boundary regions often pose significant difficulties for segmentation models. While modern deep-learning architectures achieve remarkable accuracy in segmentation tasks, they frequently struggle to delineate object boundaries accurately. This challenge arises from several factors:

- **Blurry Inputs:** Images with low resolution or out-of-focus regions result in indistinct boundaries. For example, medical scans like CT or MRI images with poor resolution can blur the edges of anatomical structures, making it hard to differentiate the anatomically important region from adjacent tissues.
- **Ambiguous Annotations:** Training datasets often suffer from inconsistent or imprecise boundary annotations. This is particularly problematic in domains like histopathology, where labeling cell boundaries is highly subjective and labor-intensive.
- **Incomplete or Insufficient Visibility:** Objects that are partially occluded or have irregular and complex boundaries create significant challenges. For instance, in autonomous-driving scenarios, pedestrians partially hidden behind obstacles can result in uncertain boundary predictions.

Possible Solutions:

- Leveraging **high-resolution** imaging techniques can significantly reduce ambiguity in boundary regions.
- Implementing **refined labeling protocols** and introducing **guidelines for annotators** can help reduce inconsistencies in training datasets.
- Employing **multiscale feature-extraction architectures**, **attention mechanisms**, and **loss-function optimization** for boundary detection can improve model sensitivity to complex edges. For example, hybrid loss functions combining cross-entropy and boundary loss have shown promise in medical imaging.

1.2.3.2 Uncertainty in Labels

Uncertainty stemming from label quality is another critical factor that impacts segmentation performance. Models rely heavily on accurate and consistent labels during training; however, several issues can arise:

- **Noisy or Incorrect Labels:** Errors in labeling, such as mislabeled pixels or annotations, can mislead the model during training. For instance, in satellite image segmentation, manual labeling of land-cover types might lead to confusion between visually similar classes such as cropland versus grassland (depending on the season and vegetation type, cultivated fields can be hard to distinguish from natural grasslands).
- **Class Imbalance:** When datasets contain significantly fewer samples for certain classes, models may fail to generalize well for under-represented categories. This is common in medical imaging, where abnormalities such as rare tumors are less frequently represented compared to normal tissue.
- **Subjectivity in Labeling:** Annotators may interpret certain regions differently, leading to variability in labels. For example, in histological segmentation, annotators might label overlapping cells differently based on individual judgment.

Possible Solutions:

- Implementing **automated label-verification** systems using uncertainty estimation can help identify and rectify labeling errors efficiently.
- Addressing class imbalance through techniques like **oversampling**, **synthetic-data generation**, or **adversarial training** can improve representation for rare classes.
- Introducing consensus-labeling approaches by **combining inputs from multiple annotators** or using **ensemble methods** can reduce subjectivity in labeling.

1.2.3.3 Uncertainty Due to Model Limitations

Deep-learning models inherently possess limitations that contribute to uncertainty. These can be categorized as:

- **Epistemic Uncertainty (Model Uncertainty):** This type of uncertainty arises from limitations in model parameters or architecture, affecting generalization to unseen data. For instance, shallow networks or insufficient training epochs can lead to poor predictions on complex datasets.
- **Aleatoric Uncertainty (Data Uncertainty):** Intrinsic data noise, such as variations in lighting, reflections, occlusions, or sensor inaccuracies, causes ambiguities that no model can completely resolve. In remote sensing, for example, atmospheric interference can create artifacts in satellite images.

Possible Solutions:

- Adopting **Bayesian neural networks [143]** or **Monte Carlo dropout [37]** can enhance robust uncertainty estimation by quantifying model confidence.
- Developing **hybrid frameworks** that combine both epistemic and aleatoric uncertainty into a unified approach can ensures comprehensive uncertainty management.
- Utilizing **ensemble-learning** techniques, where multiple segmentation models are combined, can reduce uncertainty and improve predictions.
- Applying **model-calibration techniques** to adjust overconfident predictions and improve generalization to new data.
- Leveraging advanced training strategies like **adversarial training**, which simulates challenging scenarios, or **active learning**, where the model prioritizes learning from ambiguous examples, can address both model- and data-related challenges effectively.

1.3 Motivation

Deep learning has revolutionized image segmentation, enabling breakthroughs in medical imaging, autonomous vehicles, environmental monitoring, and more. Despite these advancements, segmentation models often grapple with uncertainties that challenge their reliability and effectiveness. These uncertainties arise from

various sources, including ambiguous object boundaries, noisy or incomplete annotations, and inherent limitations in data or model architecture. If not properly addressed, these factors can undermine the credibility and usability of segmentation systems, particularly in high-stakes applications such as diagnosing diseases or ensuring road safety in autonomous driving.

One prominent challenge is dealing with boundary uncertainties, especially in complex or low-quality images like medical scans or satellite imagery, where object edges are often unclear. Additionally, inconsistencies in labels, stemming from subjective human annotations or dataset imbalances, exacerbate the problem. Another major issue is the limited generalizability of segmentation models, which is linked to epistemic (model-based) and aleatoric (data-related) uncertainties. These issues highlight the pressing need for innovative strategies to manage and mitigate uncertainty in image segmentation.

At the same time, the advent of Explainable AI (XAI) has underscored the importance of interpretability in deep-learning models. While these models achieve remarkable accuracy, their "black-box" nature often makes their decision-making processes opaque. In sensitive applications such as healthcare, where trust and transparency are critical, this lack of interpretability can hinder adoption. For example, when a segmentation model identifies a potential tumor in a medical scan, it is essential not only to know where the model places the boundaries but also how confident it is in those predictions. An uncertainty heatmap can highlight areas of low confidence, enabling a clinician to scrutinize those regions further and make more informed decisions.

The intersection of uncertainty and interpretability offers a powerful opportunity to improve segmentation models. Understanding and visualizing uncertainty not only boosts model performance but also provides clear, actionable insights into the underlying decision-making process. This thesis is motivated by the need to address these interconnected challenges, bridging the gap between robust uncertainty estimation and explainable AI. By advancing methods that are both accurate and interpretable, this research seeks to enhance the reliability and trustworthiness of segmentation models in real-world applications.

1.4 Scope of the Work

This thesis focuses on overcoming key challenges in uncertainty estimation and interpretability within deep-learning-based image segmentation. The work covers five primary areas:

1. **Uncertainty in Boundary Regions:**
   Segmenting objects with unclear or complex boundaries is a persistent challenge. This research investigates novel strategies to address this issue, including the development of customized loss functions and attention mechanisms that enhance boundary detection and reduce uncertainty in such regions.

2. **Uncertainty in Labels:**
   Label inconsistencies, particularly in datasets with multiple annotations (e.g., medical-imaging datasets), contribute to uncertainty. This thesis delves into the sources of label uncertainty—such as variability among annotators—and proposes models capable of handling these inconsistencies to produce robust and consensus-driven segmentation outputs.

3. **Uncertainty Due to Model Limitations:**
   Deep-learning models are prone to both epistemic (model-related) and aleatoric (data-related) uncertainties. To address these limitations, this research employs advanced techniques like ensemble learning and calibration methods. These approaches improve the reliability and generalization of segmentation models, enabling their application to diverse datasets and scenarios.

4. **Reducing Uncertainty with Multimodal Data:**
   Incorporating data from multiple modalities can mitigate uncertainties arising from limited or ambiguous single-modal data. This thesis explores how combining modalities—such as RGB with thermal or near-infrared (NIR) data—enhances the model's predictions and reduces uncertainty.

5. **Explainable AI (XAI) and Uncertainty:**
   To ensure transparency and trustworthiness, the thesis integrates Explainable AI techniques. By employing methods like saliency maps, attention visualization, and uncertainty heatmaps, the research provides insights into

the decision-making processes of segmentation models. This dual focus on accuracy and interpretability ensures that the models are both effective and comprehensible to end-users.

The methodologies proposed in this thesis are rigorously evaluated across diverse datasets and application domains, including medical imaging, autonomous systems, and environmental studies. The overarching goal is to balance accuracy, computational efficiency, and interpretability, making the developed frameworks suitable for practical deployment.

By combining cutting-edge uncertainty modeling with explainable AI, this research advances both theoretical understanding and practical solutions, delivering segmentation models that are reliable, interpretable, and impactful.

1.5 Thesis Outline

This thesis is structured into eight chapters, offering a logical progression from foundational concepts to innovative methodologies and their applications. Each chapter addresses a specific aspect of uncertainty estimation and interpretability in image segmentation.

1. **Chapter 1: Introduction**
   Provides an overview of deep-learning techniques for image segmentation and introduces the concept of uncertainty in deep learning. It discusses the types of uncertainties encountered in image segmentation and outlines the motivation, objectives, and scope of the research.

2. **Chapter 2: Boundary Uncertainty in Image Segmentation**
   Addresses the challenge of boundary uncertainties in segmentation tasks. It introduces a novel loss function, Bi-H Loss, and a specialized segmentation model, COVID-CT-H-UNet, designed to improve boundary delineation, particularly in the context of COVID-19 CT image segmentation.

3. **Chapter 3: Label Uncertainty in Image Segmentation**
   Investigates uncertainties caused by inconsistent annotations, especially in

datasets with multiple annotators. It proposes a segmentation network that generates consensus maps and provides tailored outputs to address these inconsistencies, with a focus on applications in medical imaging.

4. **Chapter 4: Addressing Model Uncertainty Using Ensemble Techniques**
   Explores ensemble-based approaches to mitigate model-related uncertainty. A novel copula-function-based ensemble method is introduced, demonstrating improved segmentation robustness and accuracy.

5. **Chapter 5: Addressing Model Uncertainty Using Calibration Techniques**
   Examines model-calibration techniques to reduce prediction uncertainty. A new approach, Confusion Penalty-Based Label Smoothing (CPLS), is proposed, enhancing the reliability and confidence of model predictions.

6. **Chapter 6: Reducing Uncertainty Through Multimodal Data**
   Investigates the benefits of incorporating multimodal data, such as thermal and NIR data, alongside traditional RGB data. The inclusion of these additional modalities is shown to significantly enhance model performance.

7. **Chapter 7: Explainable AI and Uncertainty**
   Integrates Explainable AI methods with uncertainty estimation, introducing GA-RISE, an optimized version of the RISE technique. By using a genetic algorithm to improve mask generation, the method enhances visual explanations of model predictions. The chapter highlights the synergy between interpretability and uncertainty estimation, emphasizing its importance for real-world applications.

8. **Chapter 8: Conclusion and Future Work**
   Summarizes the key contributions of this thesis, highlighting advancements in uncertainty estimation and explainable AI. It also outlines the limitations of the current work and proposes directions for future research, including deeper integration of XAI, improved multimodal approaches, and exploration of new application domains.

Chapter 2

Boundary Uncertainty in Image Segmentation

2.1 Introduction

Accurate boundary delineation is a crucial yet challenging aspect of image segmentation. Unlike interior regions, boundaries often exhibit uncertainty due to factors such as low image resolution, overlapping objects, ambiguous annotations, and occlusions. These challenges make boundary prediction particularly difficult, which is critical for high-precision applications such as medical imaging, autonomous systems, and remote sensing. For instance, in medical imaging, CT and MRI scans often contain indistinct boundaries between different anatomical structures, making it harder to segment organs, tumors, or lesions with precision. Similarly, in remote sensing, differentiating between land and water or separating objects from their background can be difficult due to environmental variations and image artifacts.

The difficulty in segmenting boundaries arises from the complex nature of edge transitions, where models struggle to differentiate between adjacent regions. Traditional segmentation models often misclassify or blur these edges due to a lack of sufficient training data that explicitly accounts for boundary ambiguity. Even deep learning-based models, which have demonstrated remarkable success in segmentation tasks, still face significant limitations when dealing with boundary uncertainty. The inability to resolve these uncertainties can result in incorrect object

shapes, misplaced contours, and segmentation errors that compromise the reliability of downstream applications.

This chapter aims to address the issue of boundary uncertainty by investigating its key causes and proposing novel deep learning techniques to improve boundary segmentation accuracy. Specifically, it introduces a **spatial attention mechanism** to allow the model to focus on critical regions of interest while reducing irrelevant background information. Additionally, a **Bi-category Hybrid Loss (Bi-H Loss)** is proposed to balance pixel-wise accuracy and boundary detection, improving overall segmentation performance.

To systematically approach this problem, this chapter first explores the primary factors contributing to boundary uncertainty. Then, it presents the proposed solutions, followed by experimental results and analysis to demonstrate the effectiveness of the proposed techniques.

2.1.1 Sources of Boundary Uncertainty

Boundary uncertainty in image segmentation arises from multiple factors that hinder the model's ability to accurately delineate object edges. These factors include:

Low Image Resolution and Blurriness: Low-resolution images obscure fine details near boundaries, making it difficult for models to distinguish between adjacent regions. This is especially problematic in medical imaging, where the transition between tissues or lesions may appear gradual and indistinct [132]. Similarly, in satellite imagery, poor resolution can cause misclassification at region boundaries.

Ambiguous Annotations: Human annotations can introduce variability, especially at object boundaries where multiple annotators may disagree [90]. Overlapping structures or unclear edges contribute to inconsistencies in ground-truth labels, making it challenging for deep learning models to learn precise boundary information.

Intricate Object Geometries: Objects with complex or irregular shapes, such as blood vessels, hair strands, or tree branches, pose additional challenges in segmentation. The fine-grained structures of these objects make boundary identification inherently difficult [123].

Occlusions and Partial Visibility: When objects are partially obstructed by other elements in the image, the model must infer missing boundary information. This

inference increases the likelihood of segmentation errors, particularly in autonomous navigation or remote sensing applications [93].

Noise and Artifacts: Real-world images often contain noise, motion blur, reflections, or other artifacts that distort boundary information. In medical imaging, scanner-induced distortions and imaging artifacts can obscure clear boundary identification, reducing segmentation accuracy [12].

Class Overlap or Similarity: When two or more classes have similar visual features, their boundaries become less distinct. For example, in medical imaging, tissues with similar textures and intensities can lead to boundary misclassification. In remote sensing, roads and buildings may appear visually similar, making their segmentation uncertain [19].

These challenges highlight the need for advanced segmentation techniques that can reduce boundary uncertainty and enhance prediction accuracy.

2.1.2 Contributions of this Chapter

To address the issue of boundary uncertainty in image segmentation, this chapter makes the following key contributions:

1. **A Spatial Attention Mechanism:** To improve boundary detection, a spatial attention mechanism is incorporated into the segmentation network. This mechanism allows the model to focus on critical regions while filtering out background noise, leading to sharper and more precise segmentation results.
2. **A Bi-category Hybrid Loss (Bi-H Loss) Function:** A novel composite loss function is proposed that optimally balances pixel-wise classification accuracy and boundary detection. The loss function integrates multiple loss terms, including Dice loss, boundary-aware loss, and weighted binary cross-entropy, to refine segmentation near object edges.
3. **Extensive Experimental Analysis:** The proposed method is evaluated on a COVID-19 CT segmentation dataset to demonstrate its effectiveness in reducing boundary uncertainty. The model's performance is compared against state-of-the-art segmentation architectures.
4. **Generalization Beyond COVID-19 Segmentation:** While this chapter focuses

on medical imaging, the proposed approaches can be extended to other segmentation tasks, such as satellite imagery and industrial defect detection, where boundary precision is crucial.

2.1.3 Chapter Outline

The remainder of this chapter is structured as follows:

- **Section 2.2** – Provides an overview of previous research on segmentation models, attention mechanisms, and loss functions relevant to boundary uncertainty.
- **Section 2.3** – Describes the architecture of the segmentation model, the spatial attention mechanism, and the Bi-H Loss function.
- **Section 2.4** – Details the dataset, experimental setup, and evaluation metrics used for assessing the proposed method and also analyzes the performance of the proposed method and compares it with existing approaches.
- **Section 2.5** – Summarizes the key findings of the chapter.

2.2 Related Works

Boundary uncertainty in image segmentation has been a persistent challenge, and researchers have explored various approaches to mitigate this issue. This section reviews prior works in segmentation networks, attention mechanisms, and their applications to COVID-19 CT segmentation, focusing on methods that improve boundary delineation.

2.2.1 Segmentation Networks for Medical Images

Deep learning-based segmentation methods have significantly evolved, particularly for medical imaging tasks. Classical segmentation architectures such as Fully Convolutional Networks (FCN) [93], U-Net [123], and DeepLab [19] have laid

the foundation for medical image segmentation. Several improvements over U-Net have been developed to address its limitations. MultiResUNet [57] enhances U-Net by incorporating residual connections and multi-scale feature extraction. H-DenseUNet [87] integrates a hybrid 2D-3D DenseNet approach to capture spatial dependencies in volumetric medical data, which is particularly useful for segmenting complex anatomical structures.
To improve boundary segmentation, loss functions have been a critical research focus. The Lovász Softmax loss [12] has been widely used due to its superior handling of class imbalance in medical images. In COVID-19 segmentation, techniques like TV-UNet [124] and SCTV-UNet [92] employ custom loss functions, such as Total Variation (TV) loss, to refine boundary details. These methods improve segmentation accuracy by preserving fine boundary structures, which is crucial for medical applications.

2.2.2 Attention Mechanisms in Convolutional Networks

attention mechanisms have played a vital role in improving segmentation performance by emphasizing critical regions in an image. SENet [54] introduced a channel-wise attention mechanism to enhance feature representation. CBAM (convolutional block attention module) [155] extended this idea by incorporating both channel and spatial attention, refining feature selection at multiple levels.
For boundary refinement, hybrid architectures such as TransUNet [17] integrate transformer-based attention modules with U-Net to improve long-range dependency modeling. By leveraging self-attention, TransUNet captures contextual relationships, which enhances boundary prediction in complex medical images. These improvements in attention-based segmentation models are particularly valuable in handling images with ambiguous or low-contrast boundaries.

2.2.3 COVID-19 CT Segmentation Networks

COVID-19 CT segmentation networks have emerged as a crucial tool for diagnosing lung infections. Inf-Net [32] introduced a parallel partial encoder and an edge detection module to improve boundary identification. Semi-Inf-Net [32] further extended this approach with semi-supervised learning to address the scarcity of

annotated COVID-19 data. More recently, a multi-scale codec network [94] has been proposed to leverage multi-resolution information for more precise segmentation of COVID-19 lesions.
These methods have demonstrated improvements in Dice similarity scores, sensitivity, and specificity, highlighting the importance of refining segmentation boundaries. However, challenges such as annotation inconsistencies and low-resolution CT scans continue to contribute to boundary uncertainty, necessitating further research into robust segmentation techniques.

2.3 Present Work

Boundary uncertainty in image segmentation, particularly in medical imaging, arises due to low-resolution images, complex object geometries, and annotation inconsistencies. Addressing this issue requires improvements in both network architecture and loss functions to ensure sharper, more accurate boundary delineation. In this section, we introduce COVID-CT-H-UNet, a segmentation model that enhances feature extraction and boundary awareness using an attention-based U-Net. Additionally, we propose a Bi-category Hybrid Loss (Bi-H Loss), which combines multiple loss functions to balance pixel-wise accuracy with boundary precision.

2.3.1 U-Net with Spatial Attention Mechanism

Deep learning-based segmentation models rely heavily on spatial feature extraction. However, in conventional U-Net models, important boundary-related information is often lost due to repetitive downsampling and upsampling operations. COVID-CT-H-UNet enhances boundary segmentation by integrating a spatial attention mechanism into the skip connections of U-Net.
The proposed architecture follows the standard U-Net encoder-decoder structure(see Figure 2.1). The encoder extracts multi-scale features using convolutional and pooling layers, while the decoder reconstructs segmentation maps using transposed convolutions. The spatial attention mechanism (see Figure 2.1), applied within the skip connections, helps the network selectively retain high-importance

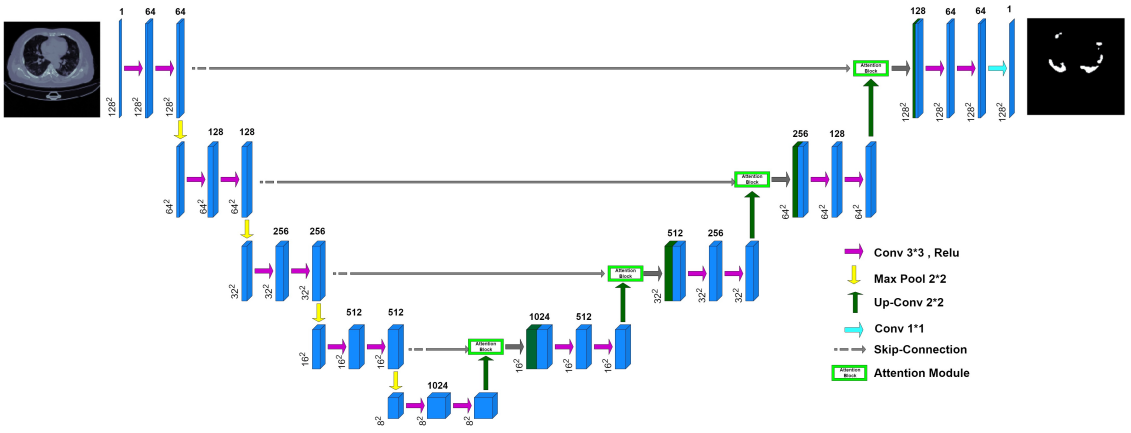

Figure 2.1: The proposed Architecture

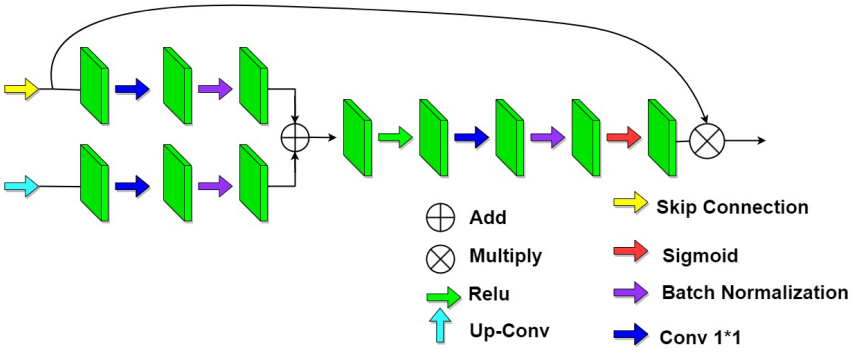

Figure 2.2: The spatial attention mechanism used in our Architecture

features and suppress less relevant information. The attention module operates as follows:

- Feature maps from both the encoder and decoder are processed through convolutional layers.
- A batch normalization and ReLU activation step refines the extracted features.
- A sigmoid activation function generates an attention map, which is applied to the encoder feature maps before forwarding them to the decoder.

By integrating attention into the skip connections, the model effectively preserves fine details, reducing the likelihood of boundary misclassification. This mechanism is inspired by previous attention-based architectures such as CBAM [155] and SENet [54], which have demonstrated success in improving segmentation accuracy.

2.3.2 Bi-category Hybrid Loss: A Composite Loss Function

Loss functions play a critical role in segmentation accuracy, particularly for handling boundary uncertainty. Traditional loss functions like Binary Cross-Entropy (BCE) Loss and Dice Loss struggle to balance global region segmentation with sharp boundary delineation. To improve this, we propose Bi-H Loss, a composite loss function.
The Bi-H Loss function is designed to enhance segmentation performance by incorporating:

- Pixel-wise similarity loss, which includes Weighted BCE Loss and Square Hinge Loss, ensuring accurate classification of pixels within segmented regions.
- Boundary detection loss, which includes Dice Loss and Boundary Loss, focusing on refining segmentation edges.

The final Bi-H Loss is formulated as:

$$\text{Bi-H Loss} = \alpha(\text{Weighted BCE Loss}+\text{Dice Loss})+\beta(\text{Square Hinge Loss}+\text{Boundary Loss}) \tag{2.1}$$

where α and β are hyperparameters that control the weight of each component and $\alpha + \beta = 1$.

Each loss function contributes uniquely to the segmentation process:

- Weighted BCE Loss: Minimizes class imbalance by assigning higher weights to boundary pixels [157]. This loss calculates the pixel-wise binary cross-entropy between the predicted segmentation and the ground truth. It is given by:

$$\text{Weighted BC Loss} = -\frac{1}{N}\sum_{i=1}^{N} w_i \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right]$$

  where N is the total number of pixels, y_i is the ground truth label (0 or 1), and p_i is the predicted probability for each pixel.

- Square Hinge Loss: Discourages overconfident predictions, improving generalization. The Square Hinge Loss penalizes the misclassification of pixels by measuring the squared error between the predicted and true labels:

$$\text{Square Hinge Loss} = \frac{1}{N}\sum_{i=1}^{N}\left(\max(0, 1 - y_i \cdot p_i)^2\right)$$

  where y_i is the ground truth label, and p_i is the predicted probability for each pixel.

- Dice Loss: Ensures high overlap between predicted and ground-truth segmentation [136]. Dice Loss is used to evaluate the overlap between the predicted segmentation and the ground truth. It is expressed as:

$$\text{Dice Loss} = 1 - \frac{2\sum_{i=1}^{N} y_i p_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} p_i}$$

  where y_i and p_i are the ground truth and predicted pixel probability values, respectively.

- Boundary Loss: Enhances edge sharpness, reducing misclassification at boundary regions [73]. Boundary Loss focuses on the accurate detection of object boundaries by comparing the gradients (edges) of the predicted segmentation and the ground truth. The formula for boundary loss is given by:

$$\text{Boundary Loss} = \sum_{i=1}^{N} |\nabla y_i - \nabla p_i|$$

  where ∇y_i and ∇p_i represent the gradients (or boundaries) of the ground truth and predicted segmentation maps, respectively.

The combination of U-Net with spatial attention and Bi-H Loss leads to higher segmentation accuracy, particularly in medical imaging tasks where boundary uncertainty is a major challenge.

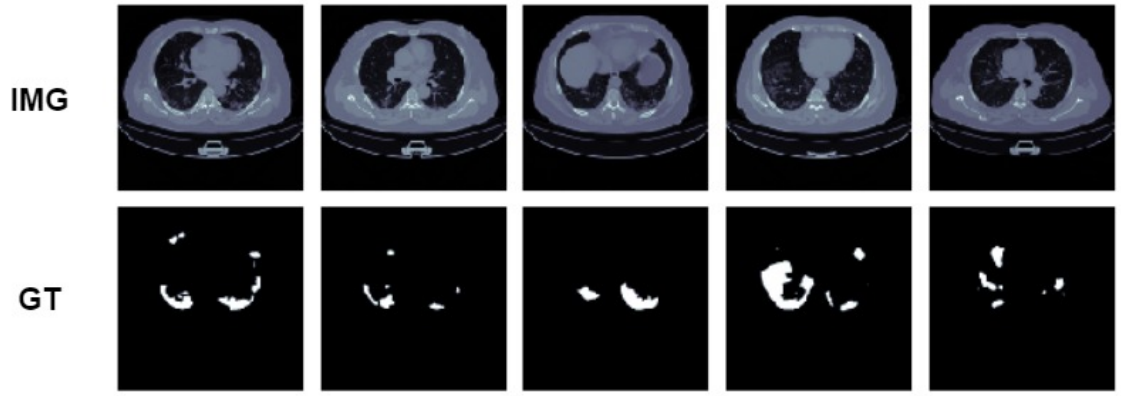

Figure 2.3: Sample images and corresponding ground truth from the COVID-19 CT segmentation dataset.

2.4 Experiments

2.4.1 Dataset

The dataset used in this study is a COVID-19 CT segmentation dataset [160], which is quite small in size due to the rarity of such data. This dataset consists of only 20 CT scans, making it particularly challenging for segmentation tasks. Two radiologists manually annotated the left and right lung regions and identified areas of infection. A senior radiologist reviewed and confirmed these annotations. For model evaluation, 20% of the CT scans were used for testing, while the remaining 80% were used for training. The ground truth annotations represent the affected regions, which exhibit complex textures and characteristics, as shown in Figure 2.3. Due to the intricacy of the affected regions, accurately segmenting these from CT scans proves to be a highly challenging task.

2.4.2 Experimental Setup

The proposed COVID-CT-H-UNet model was trained from scratch using the TensorFlow framework on a machine equipped with a 16GB NVIDIA Tesla P100 GPU. The network was trained for approximately 100 epochs with a batch size of 32. Based on empirical results, we found that the Adam optimizer outperformed SGD for this task, so it was chosen for optimization. Additionally, a learning rate scheduler was employed to adaptively adjust the learning rate during training.

2.4.3 Evaluation Metrics

To evaluate the performance of the segmentation model, we utilized three common metrics: sensitivity, specificity, and the Dice coefficient. Below is a detailed description of each metric:

2.4.3.1 Dice Coefficient

The Dice coefficient, a widely used metric in medical image segmentation, measures the overlap between the predicted segmentation map and the ground truth. It is calculated as:

$$Dice = \frac{2|A \cap B|}{|A| + |B|}$$

where A represents the ground truth affected region, and B represents the predicted affected region.

2.4.3.2 Sensitivity

Sensitivity measures the proportion of correctly identified COVID-19 pixels out of the total number of COVID-19 pixels in the ground truth. It is calculated as:

$$Sensitivity = \frac{TP}{TP + FN}$$

where TP refers to the True Positive pixels, i.e., pixels correctly classified as COVID-19, and FN represents the False Negative pixels, i.e., pixels that were mistakenly labeled as non-COVID-19.

2.4.3.3 Specificity

Specificity calculates the proportion of correctly identified non-COVID-19 pixels out of the total number of non-COVID-19 pixels in the ground truth. It is given by:

$$Specificity = \frac{TN}{TN + FP}$$

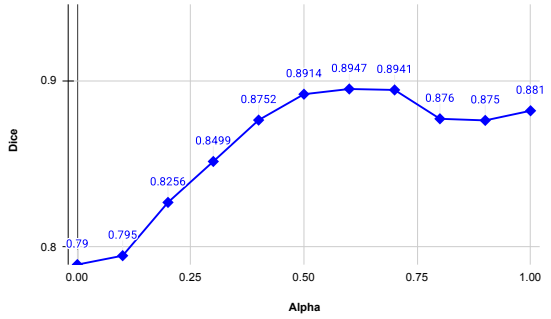

Figure 2.4: Hyperparameter tuning by varying the parameter α

where TN refers to True Negative pixels (correctly identified non-COVID-19 pixels), and FP represents False Positive pixels (non-COVID-19 pixels incorrectly classified as COVID-19).

2.4.4 Results and Discussion

2.4.4.1 Quantitative Analysis

To optimize the hyperparameters α and β in our proposed Bi-H loss function, we systematically varied the value of α from 0 to 1 in increments of 0.1, while β was automatically determined as $\beta = 1 - \alpha$. The performance for each setting was evaluated based on the Dice score, and the results are summarized in Figure 2.4. As observed in the figure, the optimal values were found to be $\alpha = 0.6$ and $\beta = 0.4$, yielding the highest Dice score.

For comparative evaluation, we selected widely used segmentation networks, including U-Net [123], U-Net++ [171], U-Net with a ResNet backbone, TV-UNet [124], and SCTV-UNet [92]. These models have demonstrated strong performance in COVID-19 CT image segmentation. The quantitative results are presented in Table 2.1, highlighting the effectiveness of our approach in comparison to existing methods.

Compared to basic U-Net [123], its improved versions like U-Net++ and U-Net+ResNet perform better. Recent networks like TV-UNet [124] and SCTV-UNet [92] significantly outperform basic U-Net-based models. However, our proposed model

Table 2.1: Performance comparison of different models in the COVID-19 CT segmentation dataset, bold indicates the best effect. Note that, except for SCTV-UNet [92], we have implemented all the models ourselves.

Model	**Dice**	**Sensitivity**	**Specificity**
UNet [123]	0.6048	0.7231	0.9996
UNet++ [171]	0.7712	0.6366	0.9995
U-Net+ResNet	0.7856	0.7199	0.9997
TV-UNet [124]	0.7227	0.7032	0.9996
SCTV-UNet [92]	0.7989	**0.8080**	0.9663
Proposed	**0.8947**	0.7377	**0.9997**

outperforms all of them by a significant margin in both Dice and Specificity metrics, while SCTV-UNet [92] performs best in the Sensitivity metric. One potential reason for this discrepancy could be the higher number of false negatives. Since the affected regions are small in some images, the effect of BCE loss (part of the proposed Bi-H loss) might lead to misclassification of certain pixels, causing false negatives. In the future, we aim to address this issue by penalizing the negative effects of BCE loss to further reduce false negatives and improve sensitivity.

Impact of Bi-category Hybrid Loss function: A key contribution of this paper is the Bi-category Hybrid Loss. This novel loss function is designed to tackle issues such as boundary blurring and poor foreground/background contrast, which contribute to boundary uncertainty in COVID-19 CT segmentation tasks. As described in Section 2.3.2, this loss function enhances the segmentation output by improving boundary delineation, particularly when using a U-shaped network. We compared our proposed loss function to commonly used loss functions like Binary Cross-Entropy (BCE) loss [157], Dice Loss [136], and Boundary Loss [73] in our attention-UNet-based segmentation model. The corresponding results for Dice, Sensitivity, and Specificity metrics are presented in Table 2.2.

The Bi-H loss yields the best performance in terms of Dice, Sensitivity, and Specificity, demonstrating its effectiveness in reducing boundary uncertainty and improving segmentation accuracy. This result highlights the critical role of our proposed loss function in enhancing the model's ability to delineate boundaries more precisely, addressing common issues in medical image segmentation.

Table 2.2: Model performance for different Loss Functions employed, bold indicates the best effect.

Loss	Dice	Sensitivity	Specificity
BCE	0.8463	0.7267	0.9997
DiceLoss + BoundaryLoss	0.8757	0.7186	0.9997
BCE + DiceLoss	0.8820	0.7309	0.9996
Bi-H loss	**0.8947**	**0.7377**	**0.9997**

2.4.4.2 Visual analysis

The COVID-19 CT segmentation dataset[160] was used to train U-Net [123], U-Net+ResNet, and COVID-CT-H-UNet. Figure 2.5 shows the qualitative segmentation results for these models on the test set. It is evident that as the model improves, the segmentation performance also improves. However, challenges remain, especially with the incorrect detection of regions close to the lesion boundaries. This issue is addressed by incorporating an attention mechanism in COVID-CT-H-UNet, which helps emphasize the affected areas while suppressing unaffected regions. This approach leads to more accurate segmentation of the infected regions.

The attention mechanism incorporated into the skip-connections of COVID-CT-H-UNet(see Figure 2.1)enhances segmentation accuracy by allowing the model to assign different levels of importance to different regions. This mechanism helps in capturing fine-grained structures and subtle patterns, which are essential for accurate boundary delineation. By adaptively focusing on relevant features and suppressing irrelevant information, the attention mechanism contributes to more precise segmentation and robust handling of boundary uncertainty.
The attention mechanism aids in addressing the problem where positive samples, particularly those near the lesion boundaries, are often mistaken for negative samples. This problem was highlighted in previous models like TV-UNet [124] and SCTV-UNet [92]. By emphasizing the lesion region and reducing the uncertainty in boundary detection, COVID-CT-H-UNet significantly improves segmentation accuracy.
The impact of the Bi-H loss function and attention mechanism is evident in Figure 2.5, where the segmentation results clearly show improvements in boundary delineation. The weighted combination of BCELoss, DiceLoss, Square Hinge Loss,

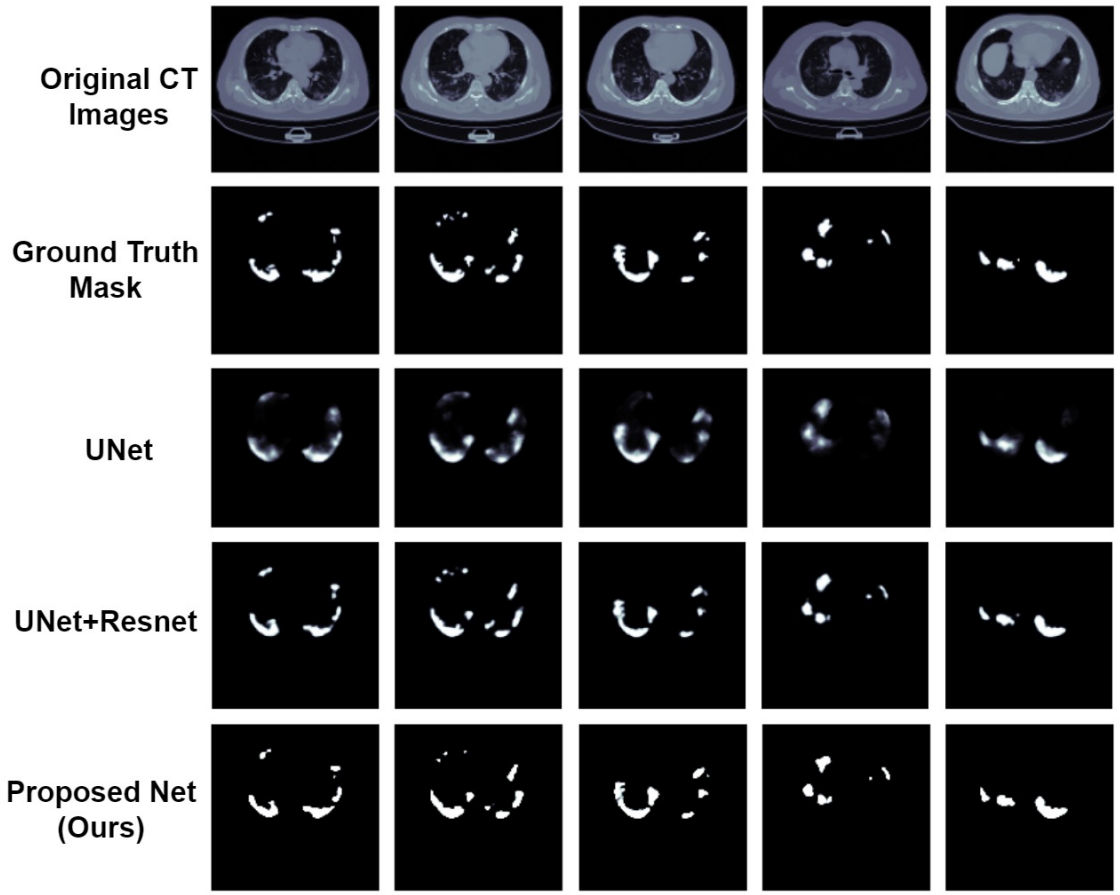

Figure 2.5: Visual comparisons of our models with the UNet and UNet+ResNet model

and Boundary Loss resolves issues of boundary blurring and insufficient contrast between the lesion and background in the prediction images. These improvements are crucial in mitigating boundary uncertainty and enhancing the overall segmentation quality.

2.5 Summary of the Chapter

This chapter addressed the challenge of boundary ambiguity in COVID-19 CT image segmentation. We proposed **COVID-CT-H-UNet**, a U-Net variant that places a spatial-attention block in every skip connection, enabling the decoder to recover fine edge details usually lost during the encoder–decoder resolution changes.

To complement the new architecture, we introduced a **Bi-category Hybrid Loss (Bi-H Loss)** that merges two pixel-level objectives (Weighted BCE and Square-Hinge) with two contour-sensitive objectives (Dice and Boundary). A grid search determined that weighting these two groups at $\alpha = 0.6$ and $\beta = 0.4$ yields the best overlap accuracy.

The experiments are conducted on the COVID-19 CT segmentation dataset where COVID-CT-H-UNet achieved a Dice coefficient of **0.8947** and a specificity of **0.9997**, outperforming five strong baselines: U-Net, U-Net++, U-Net+ResNet, TV-UNet, and SCTV-UNet (Table 2.1). Replacing conventional loss functions with Bi-H Loss increased Dice by 1.3–4.8 percentage points and also improved sensitivity and specificity (Table 2.2).

Qualitative comparisons support these numerical gains: predicted masks exhibit sharper lesion contours, fewer spurious positives near lung borders, and noticeably less edge blurring than competing models (Fig. 2.5). Together, the spatial-attention design and Bi-H Loss provide an effective remedy for boundary-related uncertainty in small, noise-affected medical datasets, offering a more dependable foundation for subsequent clinical analysis.

Chapter 3

Label Uncertainty in Image Segmentation

3.1 Introduction

Label uncertainty in image segmentation arises when multiple annotators provide differing interpretations of the same image, leading to inconsistencies in ground truth data. This is particularly prevalent in *medical imaging*, where expert annotations are often subjective due to *ambiguities in anatomical boundaries, imaging artifacts, and patient-specific variations*. Unlike aleatoric uncertainty (stemming from noise in the data) or epistemic uncertainty (caused by model limitations), label uncertainty specifically emerges due to *inter-observer variability*. Addressing this form of uncertainty is critical to improving segmentation reliability and ensuring models generalize well across diverse datasets.

Deep learning-based segmentation models rely heavily on high-quality annotations for training, validation, and testing. However, when multiple annotators provide differing labels, traditional segmentation models—trained on single ground truth labels—fail to capture the full spectrum of variations. This chapter explores label uncertainty by analyzing the factors contributing to annotation variability, discussing existing strategies for handling it, and proposing a novel framework that accounts for multiple annotations while enhancing model robustness.

3.1.1 The Critical Role of Annotations in Image Segmentation

Annotations are fundamental to image segmentation, serving as the basis for training, validating, and testing deep learning models. A annotation define object boundaries and structures within images, enabling segmentation algorithms to learn patterns and relationships effectively. In *medical imaging*, where precision is crucial, annotations must accurately capture anatomical structures or pathological regions. Poorly labeled data can lead to incorrect segmentation results, impacting critical applications such as *diagnosis, treatment planning, and surgical interventions*. Therefore, ensuring the *quality, consistency, and reliability* of annotations is essential for achieving high-performance segmentation models.

3.1.2 Why Multiple Annotations are Essential in Medical Image Segmentation?

Medical imaging introduces complexities that make multiple annotations *necessary* rather than optional. Differences in imaging modalities, patient anatomy, and the inherent ambiguity of certain regions often result in *discrepancies among annotators*. For example, the *exact boundary of a tumor or lesion may vary* between experts due to indistinct edges. By incorporating multiple annotations, researchers *capture a broader spectrum of plausible interpretations*, ensuring that the dataset better reflects *real-world clinical uncertainties*. Furthermore, analyzing variations in annotations *helps identify areas of disagreement*, which can inform model training strategies aimed at mitigating label uncertainty. Transitioning from single-label to *multi-annotator segmentation* frameworks leads to more generalized models, particularly for complex medical applications.

3.1.3 Understanding the Emergence of Uncertainty from Multiple Annotations

While *multiple annotations improve dataset quality*, they also introduce *label uncertainty* due to *differences in annotators' expertise, experience, and subjective interpretations*. This form of uncertainty is especially problematic in medical image

segmentation, where determining boundaries and structures is inherently complex. Annotators may struggle with defining the extent of a tumor or lesion due to unclear edges. Small or ambiguous regions in an image can lead to different classifications depending on the annotator's background. The inclusion or exclusion of surrounding tissue in segmentation tasks may also vary based on individual expert opinions.
These discrepancies create inconsistencies in training data, which can negatively affect the performance of deep learning models. A segmentation model trained on a single, arbitrarily chosen annotation may fail to generalize to other valid interpretations. In contrast, a framework that acknowledges inter-annotator variability can improve model robustness by incorporating multiple perspectives into its training process. Addressing these variations is essential for developing segmentation frameworks that effectively navigate complex medical imaging tasks.

3.1.4 Contributions of this Chapter

1. **Class-Specific Distribution Learning (CSDL):** We model the annotation distribution class-by-class, allowing the network to learn the variability inherent in each label.

2. **Annotator-Specific Preference Estimator (ASPE):** A module that captures individual annotator biases and fuses them into a consensus prediction.

3. **Robust segmentation framework:** By combining CSDL and ASPE, the framework improves accuracy while remaining resilient to inter-observer variability.

4. **Comprehensive experimental evaluation:** We validate the framework on two benchmark datasets—*RIGA* (retinal vessel segmentation) and *QUBIQ* (multi-organ CT/MRI)—demonstrating state-of-the-art performance under label-uncertainty conditions.

3.1.5 Chapter Outline

The remainder of this chapter is structured as follows:

- **Section 3.2** – A review of existing methods for handling label uncertainty, including *label fusion, pseudo-labeling, and modeling annotator biases*.
- **Section3.3** – A detailed methodology of our *proposed segmentation framework*, including the *architecture, loss functions, and uncertainty modeling strategies*.
- **Section 3.4** – A comprehensive evaluation of our method on *two benchmark datasets (RIGA and QUBIQ)*, comparing its performance with state-of-the-art approaches.
- **Section 3.5** – Summarizes the key findings of the chapter.

3.2 Related Works

Medical image segmentation often suffers from annotator-related biases and inter-observer variability, which can significantly impact model performance and reliability. Various strategies have been proposed to address these challenges, mainly categorized into I) Label fusion and pseudo-label creation, II) Modeling annotator preferences and biases. This section discusses these approaches, their limitations, and how our proposed framework improves upon them.

3.2.1 Label Fusion and Pseudo-Label Creation

A widely used method to resolve annotator disagreements is to aggregate multiple annotations into a single pseudo-label. Traditional techniques include majority voting [46], label fusion [16, 85, 91, 164, 166], and label sampling [63]. These approaches aim to produce a consolidated label that represents the most probable ground truth.
Jensen et al. [63] introduced a label sampling strategy to model inter-annotator variability, ensuring better generalization. Guan et al. [46] independently predicted each annotator's segmentation label and merged them using a weighted combination strategy. Another technique by Mirikharaji et al. [100] employed spatially adaptive reweighting to reduce annotation noise at the pixel level.
Although these methods improve segmentation consistency, they fail to preserve

individual annotator insights. By averaging over multiple annotators, they assume that a single, unified ground truth exists, which does not account for real-world differences in clinical interpretations. This loss of variability is particularly concerning in medical imaging, where diverse expert opinions may contain valuable diagnostic information [47, 88].

3.2.2 Modeling Annotator Preferences and Biases

A more advanced strategy is to explicitly model annotator preferences and biases. This ensures that segmentation models capture both agreement and disagreement patterns among annotators.

Ji et al. [65] proposed MRNet, which leverages both consensus and disagreement cues to model annotator variability. Liao et al. [88] introduced PADL, a multi-head network that separates annotator-specific preferences from random annotation errors. While effective, these methods require separate learning heads for each annotator, making them computationally expensive as the number of annotators increases.

Guo et al. [48] proposed a cascade of nnUNet models, encoding annotator identities with conditional convolution to personalize segmentation outputs. Meanwhile, Guo et al. [47] focused on optic disc and cup segmentation, leveraging encoding vectors and dynamic filters to differentiate annotator-specific segmentation patterns.

Despite their contributions, these methods fail to explicitly model class-specific annotation distributions. In cases like optic cup segmentation, annotators may disagree more on certain classes than others, yet existing techniques do not accommodate such class-dependent annotation variability.

Our approach improves upon these existing methods by introducing a Class-Specific Distribution Learning (CSDL) module that explicitly models class-wise annotation variations, capturing the uncertainty at the class level. Additionally, we propose an Annotator-Specific Preference Estimator (ASPE) module, which generates segmentation maps tailored to each annotator without requiring separate network heads, thereby reducing computational overhead. By integrating these components, our method preserves inter-annotator variability without collapsing into a single pseudo-label, efficiently models class-wise annotation uncertainty, and

reduces the complexity associated with multi-head models. Our proposed framework is validated on two diverse multi-annotator datasets, demonstrating superior performance compared to state-of-the-art methods.

3.3 Present Work

3.3.1 Problem Formulation

Let $D = \{(x_i, y_i^1, \ldots, y_i^R)\}_{i=1}^N$ represent a dataset of N images, where each image $x_i \in \mathbb{R}^{3\times H\times W}$ is associated with R annotations. Each annotator's segmentation $y_i^r \in \{0,1\}^{C\times H\times W}$ corresponds to their delineation for C classes. The presence of multiple annotations for the same image introduces label uncertainty due to inter-observer variability, which significantly impacts both model training and the overall uncertainty estimation.
This work proposes a segmentation framework that aims to generate a meta-segmentation mask by effectively aggregating multiple annotations while preserving individual annotator preferences. The proposed approach not only improves segmentation accuracy but also enhances the reliability of uncertainty estimation by capturing both global and annotator-specific variations in segmentation maps.

3.3.2 Methodology

3.3.2.1 Overall Network Architecture

The proposed framework comprises three key components: an Encoder-Decoder backbone, a Class-Specific Distribution Learning (CSDL) module, and a set of Annotator-Specific Preference Estimators (ASPEs) (see Figure 3.1). The Encoder-Decoder module extracts feature representations from the input image, which are then processed by the CSDL module to model the class-wise annotation distributions. The meta-segmentation output generated by the CSDL module is further refined through multiple annotator-specific networks, each trained to replicate the segmentation patterns of an individual annotator.

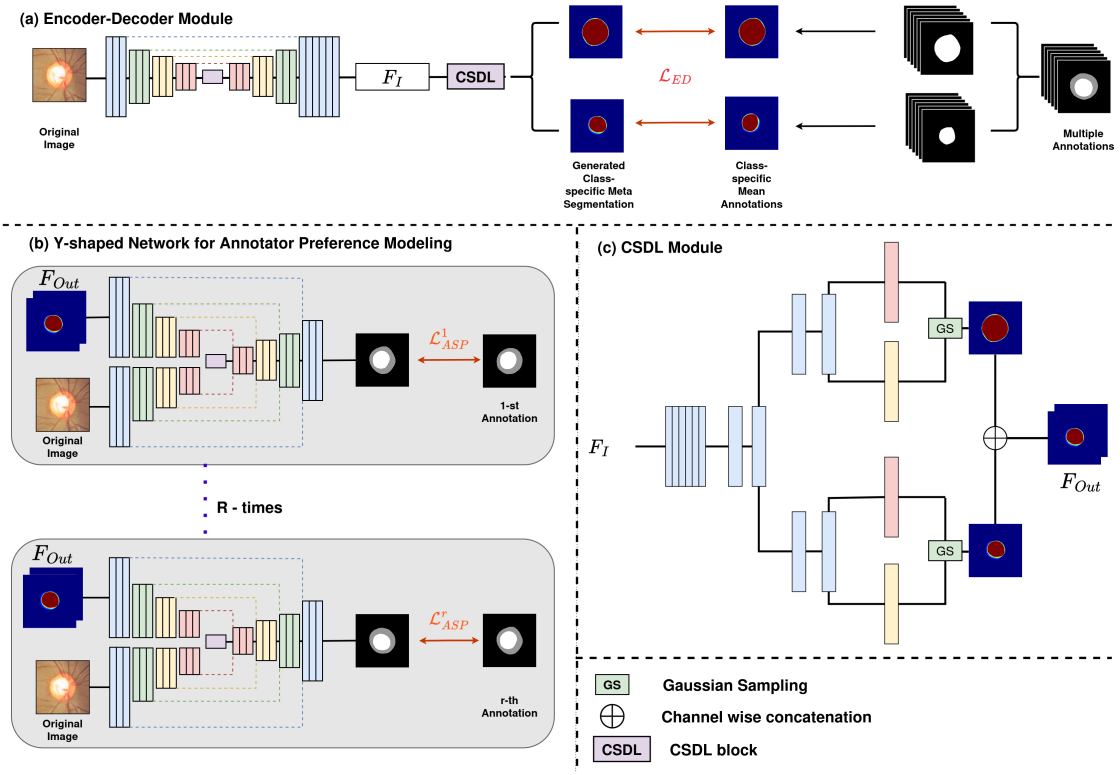

Figure 3.1: The detailed view of the overall architecture of our approach is illustrated in this figure. (a) The Encoder-Decoder module with the CSDL block, which produces the meta-segmentation, is shown. (b) The Annotator-Specific Preference Estimation (ASPE) module, consisting of R number of Y-shaped networks, where r represents the number of annotators. (c) The CSDL module, responsible for estimating the class distributions and generating the meta-segmentation.

3.3.2.2 Encoder-Decoder Module

The Encoder-Decoder backbone is based on the U-Net architecture with a ResNet-34 encoder pre-trained on ImageNet. This module extracts hierarchical features from the input image, which are progressively refined through a series of upsampling layers to reconstruct segmentation masks. The incorporation of skip connections ensures that fine-grained spatial information is preserved throughout the decoding process. The output feature map F_I is formulated as:

$$F_I = f_D\left(f_E\left(I, \theta_E\right), \theta_D\right)$$

where f_E and f_D denote the encoder and decoder functions, respectively, and θ_E and θ_D represent their parameters.

3.3.2.3 Class-Specific Distribution Learning (CSDL) Module

The CSDL module is responsible for modeling the distribution of annotations for each class. Given the inherent variability in segmentation labels, this module assumes that annotations can be represented as samples drawn from a Gaussian distribution. The mean μ_c and standard deviation σ_c of each class-specific distribution are estimated using 1×1 convolutional layers:

$$\mu_c = f_{\mu_c}\left(F_I, \theta_{\mu_c}\right), \quad \sigma_c = f_{\sigma_c}\left(F_I, \theta_{\sigma_c}\right).$$

A sample x_c is drawn from the Gaussian distribution $\mathcal{N}(\mu_c, \sigma_c^2)$, and the meta-segmentation mask is constructed as:

$$F_{Out} = x_1 \oplus x_2 \oplus \cdots \oplus x_C.$$

3.3.2.4 Annotator-Specific Preference Estimation (ASPE)

The ASPE module generates segmentation maps tailored to individual annotators. By leveraging the meta-segmentation output F_{Out} alongside the original image I, the model learns to replicate annotator-specific segmentation patterns. A Y-shaped network is employed to generate the final segmentation mask for each annotator:

$$\overline{y^r} = f_{D^r}\left(f_{E_1^r}(I, \theta_{E_1^r}), f_{E_2^r}(F_{Out}, \theta_{E_2^r}), \theta_{D^r}\right).$$

3.3.2.5 Loss Function

The overall loss function $\mathcal{L}$ is a weighted combination of two terms: the Encoder-Decoder loss $\mathcal{L}_{ED}$, which optimizes the meta-segmentation, and the Annotator-Specific Preference loss $\mathcal{L}_{ASP}$, which ensures the model accurately replicates individual annotations:

$$\mathcal{L} = \mathcal{L}_{ED} + \mathcal{L}_{ASP}.$$

The Encoder-Decoder loss is defined as:

$$\mathcal{L}_{ED} = \sum_{c=1}^{C} \lambda_c \, \mathrm{CE}\left(y_{\mathrm{mean}}^c, x_c\right),$$

where y^c_{mean} is the mean-voted ground truth annotation, and $x_c \sim \mathcal{N}(\mu_c, \sigma_c^2)$ is the sampled output from the CSDL module. The Annotator-Specific Preference loss is computed as:

$$\mathcal{L}_{ASP} = \sum_{r=1}^{R} \text{CE}\left(y^r, \overline{y^r}\right).$$

3.3.2.6 Inference

During inference, the network generates a meta-segmentation output F_{Out}, which represents the collective consensus of all annotators. This consensus segmentation is then passed through the ASPE module to produce annotator-specific segmentations $\overline{y^r}$, ensuring that the final output can reflect both the common agreement and individual annotator preferences.

3.4 Experiments

3.4.1 Dataset Description

RIGA dataset [6] is widely used for evaluating optic cup and disc segmentation algorithms. It consists of 750 color fundus images from three sources: 460 from MESSIDOR, 195 from BinRushed, and 95 from Magrabia. The images were annotated by six ophthalmologists. We follow the data split scheme proposed by Liao et al. [88], where 655 images from BinRushed and MESSIDOR are used for training, while the 95 Magrabia images serve as the test set. Some sample images are shown in Figure 3.2.

QUBIQ dataset [96] evaluates inter-annotator variability in medical image segmentation. It includes four segmentation tasks: (a) 39 MRI cases (34 for training, 5 for testing) with seven annotators for brain growth segmentation, (b) 32 MRI cases (28 for training, 4 for testing) with three annotators for brain tumor segmentation, (c) 55 MRI cases (48 for training, 7 for testing) with six annotators for prostate segmentation but for two different segmentation task, and (d) 24 CT cases (20 for training, 4 for testing) with three annotators for kidney segmentation. Some sample images are shown in Figure 3.3.

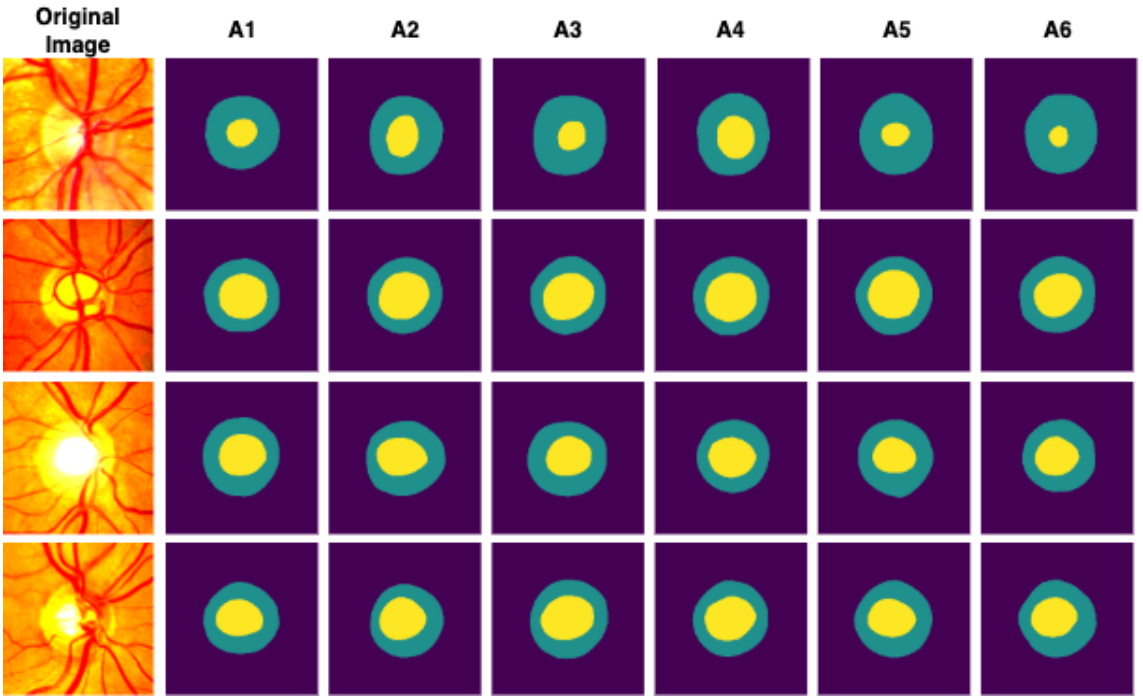

Figure 3.2: Sample images from RIGA datasets with annotations from six annotators

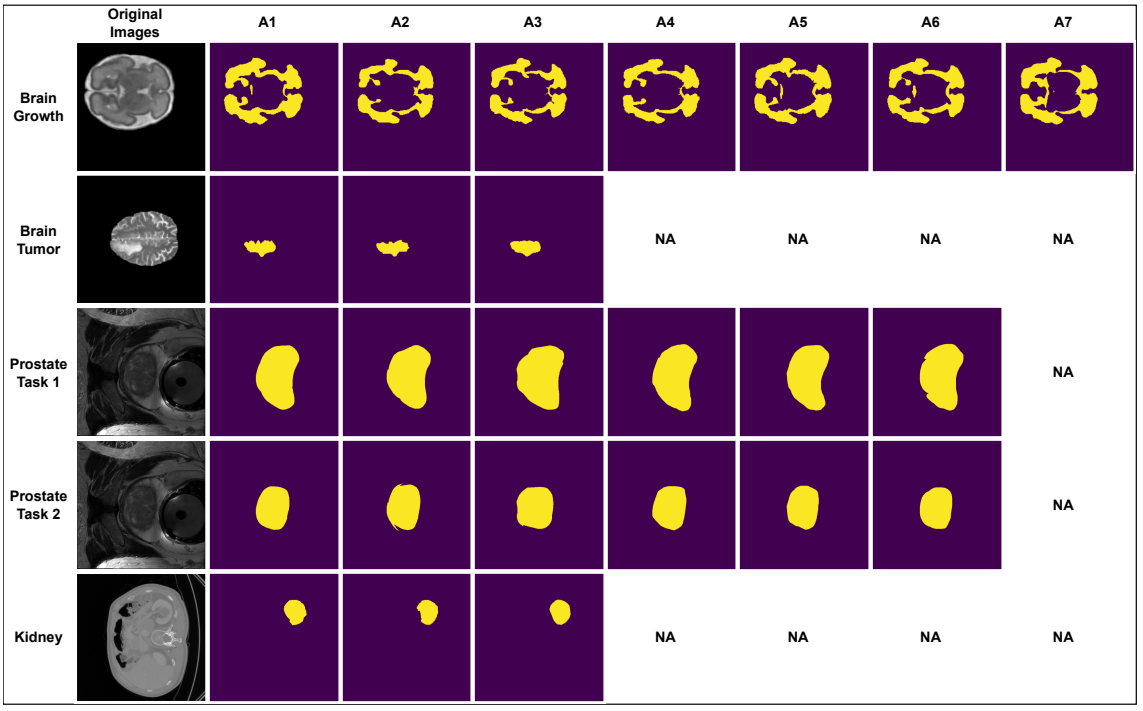

Figure 3.3: Sample images from QUBIQ dataset with different available annotations

3.4.2 Experimental Setup

All images were preprocessed by subtracting the mean and standardizing based on training statistics. The RIGA dataset images were resized to 256×256, while QUBIQ images were center-cropped to 640×640 and resized to 256×256. The model was trained using the Adam optimizer with an initial learning rate of 0.0001, adjusted dynamically via a learning rate scheduler. All experiments were conducted on an NVIDIA Quadro P1000 GPU (16 GB) using a PyTorch-based environment.

3.4.3 Evaluation Metrics

We used the **Soft Dice Coefficient** [89] as the primary evaluation metric, computed by averaging the Hard Dice scores between the predicted segmentation mask and the mean-voted ground truth over multiple threshold levels (0.1, 0.3, 0.5, 0.7, 0.9). The two performance measures considered were:

- **Mean Voting**: The Soft Dice score between the predicted segmentation and the mean voting ground truth, indicating the model's ability to generate a well-calibrated segmentation.
- **Average**: The Soft Dice score computed per annotator and then averaged across all annotators. A higher Average score reflects stronger agreement with individual annotators.

We conducted extensive quantitative experiments to evaluate our proposed method against several state-of-the-art multi-annotation segmentation techniques on the RIGA test set. A summary of the results is provided in Table 3.1. In the table, Mi represents variants of the U-Net (with a ResNet backbone) base model, each trained with a distinct set of annotations, denoted as Ai for $i = 1, 2, \ldots, 6$. The methods for comparison include: MH-UNet [46], which uses a Res-U-Net architecture with multiple segmentation heads, each trained to mimic the annotations of a specific annotator; MV-UNet [31], a Res-U-Net trained using the mean voting of annotations; LS-UNet [63], which trains the model with randomly selected annotations for each sample; CM-Net [164], which employs a confusion matrix to model human errors and disentangle annotator biases; and MR-Net [65], PADL

[88], and AVAL [47], all of which represent advanced methods aimed at addressing disagreement among multiple annotators.

As shown in Table 3.1, the base model M_i achieves optimal performance when evaluated with A_i, which is expected as it reflects the annotator's individual preferences. This outcome is highlighted in the table with a light yellow color. The MH-UNet model outperforms most of the base models (M_i), as well as models trained with Average and Mean-Voting annotations, demonstrating the importance of addressing annotator-related biases. A similar trend is observed with other models, such as MV-UNet, LS-UNet, CM-Net, and MR-Net, which also generally perform better than the base models and show improvements when trained with average and mean voting annotations.

Among the recent state-of-the-art methods, PADL and AVAP outperform the aforementioned models, not only when trained with individual annotations but also when trained with Average and Mean Voting annotations. However, our proposed model surpasses all of these methods in both Average and Mean Voting annotations. Specifically, our model achieves the highest performance for optic disc segmentation when trained with Annotations 1, 2, 4, and 6, and the second-highest performance when trained with Annotation 5. Additionally, it ranks third for performance when trained with Annotation 3. In terms of optic cup segmentation, our model achieves the highest performance when trained with Annotations 4, 5, and 6, and the second-highest performance when trained with Annotations 1, 2, and 3. The Qualitative comparison are shown in the Figure 3.4

We conducted a series of evaluations on the QUBIQ dataset, testing our method across five segmentation tasks: Brain Growth (BG), Brain Tumor (BT), Kidney (K), Prostate Task 1 (PT1), and Prostate Task 2 (PT2). For benchmarking purposes, we compared our approach with several state-of-the-art methods, including MH-UNet[46], MV-UNet[31], LS-UNet[63], CM-Net[164], MR-Net[65], PADL[88], and AVAP[47]. To ensure a fair comparison, all models were evaluated using the same base model and hyperparameters. Additionally, all models were trained using both Average and Mean-voting annotations, as training with individual annotations was not feasible due to the varying number of annotations per task.

The results, as shown in Table 3.2, indicate that our model achieves the highest performance on the Brain Tumor and Prostate Task 1 segmentation tasks when trained with Average annotations. It also achieves the second-highest performance

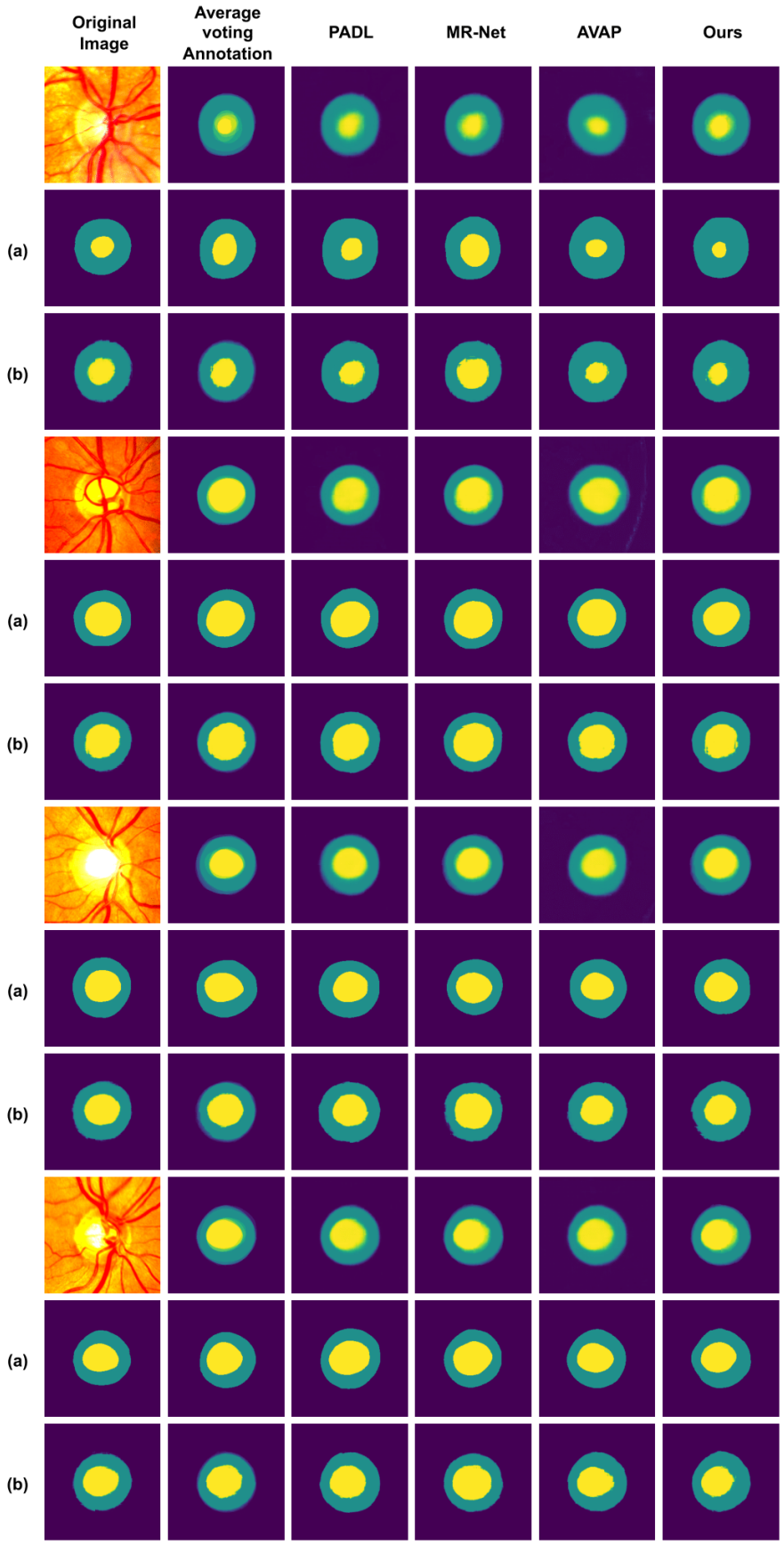

Figure 3.4: Qualitative comparison of segmentation outputs. **Top block:** Input image followed by the average annotation (consensus ground truth), and meta-segmentation results produced by PADL, MR-Net, AVAP, and the proposed method. **Middle block:** Annotator-specific ground truth segmentations from Annotators 1 to 6. **Bottom block:** Corresponding segmentation outputs generated by our Annotator-Specific Preference Estimation (ASPE) module for Annotators 1 to 6. The ASPE predictions closely align with individual annotator styles, capturing inter-observer variability effectively.

Table 3.1: The overall results of our proposed method on the RIGA dataset, alongside comparisons with state-of-the-art methods, are displayed here. In the table, Disc and Cup represent the soft Disc score (%) for the optic disc and optic cup classes, respectively. Columns labeled A1 to A6 indicate performances trained with Annotations 1 through 6, while the Average and Mean columns show performances based on average and mean-voting annotations. The best, second-best, and third-best performances are highlighted in bold, blue, and underlined text, respectively.

	A1		A2		A3		A4		A5		A6		Average		Mean	
Methods	**Disc**	**Cup**	**Disc**	**Cup**	**Disc**	**Cup**	**Disc**	**Cup**	**Disc**	**Cup**	**Disc**	**Cup**	**Disc**	**Cup**	**Disc**	**Cup**
M1	95.93	84.39	94.76	81.15	95.06	79.52	95.9	79.05	95.62	79.4	95.96	75.8	95.56	79.89	96.02	81.72
M2	95.32	84.02	96.06	84.67	96.13	80.79	96.14	81.79	96.51	80.33	96.32	77.39	96.08	81.49	95.8	82.42
M3	95.43	82.52	94.86	81.09	96.79	83.55	95.82	80.28	96.27	81.07	96.19	76.31	95.89	80.8	95.36	81.2
M4	95.14	80.31	95.63	82.08	96.33	77.42	96.42	87.89	96.1	72.7	96.42	68.69	96.01	78.18	96.11	79.24
M5	95.06	83.62	94.92	79.99	96	81.88	96.27	75.47	96.75	83.97	96.07	79.4	95.85	80.72	95.88	80.25
M6	95.5	81.39	95.64	80	96.25	78.92	96.19	74.47	96.38	82.32	97.09	80.22	96.18	79.55	96.03	79.63
MH-UNet[46]	96.03	85.3	95.98	85.69	96.9	83.99	96.68	84.86	**97.12**	83.06	96.78	77.48	96.58	83.4	96.91	84.35
MV-UNet[31]	95.06	84.33	95.27	82.57	96.05	79.35	95.48	80.29	96.26	81.05	95.33	78.11	95.57	80.95	97.35	85.74
LS-UNet[63]	95.25	83.43	94.71	80.1	95.92	81.41	96.3	78.57	96.13	82.19	96.04	79.15	95.73	80.81	97.21	81.37
CM-Net[164]	96.29	84.59	95.46	81.44	96.6	81.84	96.9	87.52	96.86	82.39	96.93	78.82	96.51	82.77	96.64	81.96
MR-Net[65]	95.35	81.77	94.81	81.18	95.8	79.23	95.96	84.46	95.9	79.04	95.76	76.2	95.6	80.31	97.55	87.2
PADL[88]	96.4	85.22	95.6	85.15	96.64	82.76	96.82	88.79	96.78	83.45	96.87	79.72	96.82	84.18	97.65	87.75
AVAP[47]	96.4	**85.66**	96.24	**85.61**	**96.97**	**84.21**	97.12	**89**	96.92	84.15	97.08	82.15	96.78	85.13	97.88	87.9
Ours	**96.58**	85.64	**96.44**	85.45	96.88	84.2	**97.22**	88.96	97.07	**85.5**	**97.6**	**82.6**	**96.98**	**85.75**	**98.14**	**88**

Table 3.2: The performance comparison between our proposed model and the state-of-the-art methods on the QUBIQ dataset is shown here. The columns K, BG, BT, PT 1, and PT 2 represent the soft Dice scores (%) for each of the five tasks. As in Table 3.1, the Average and Mean columns indicate models trained with average and mean annotations, respectively. The best, second-best, and third-best performances are highlighted in bold, blue, and underlined text, respectively.

	Average					Mean				
Methods	**K**	**BG**	**BT**	**PT 1**	**PT 2**	**K**	**BG**	**BT**	**PT 1**	**PT 2**
MH-UNet[46]	78.89	81.73	84.65	84.15	72.64	74.41	85.26	87.84	87.53	76.85
MV-UNet[31]	78.45	79.32	80.89	84.42	71.21	71.63	83.52	86.24	85.25	71.52
LS-UNet[63]	77.42	81.4	82.22	84.22	72.84	72.92	83.63	86.29	86.74	74.48
CM-Net[164]	79.75	82.92	86.21	85.85	73.55	75.12	84.53	87.82	89.58	78.84
MR-Net[65]	80.26	82.88	86.54	86.53	74.87	75.46	84.37	89.25	90.91	78.44
PADL[88]	**82.49**	**83.08**	87.14	91.18	74.83	**80.82**	85.15	89.38	92.51	**79.92**
AVAP[47]	81.85	82.54	87.89	89.56	**75.56**	79.86	**85.91**	89.56	91.56	79.57
Ours	82.41	82.98	**88.25**	**91.54**	74.86	79.75	85.88	**90.51**	**92.87**	78.58

for Kidney and Brain Growth, and the third-highest performance for Prostate Task 2. When trained with Mean-voting annotations, our model again achieves the highest performance on Brain Tumor and Prostate Task 1, second-highest performance on Brain Tumor, and third-highest for Kidney. Overall, our model consistently ranks among the top three performers across all tasks, demonstrating its robust generalization capability across diverse medical imaging tasks.

Table 3.3: The ablation study on different base models—namely UNet, UNet++, Attention UNet with VGG11 and ResNet50 backbone is presented here, both with and without pre-training on ImageNet. Disc, Cup, Average, and Mean have the same meanings as defined in Table 3.1. The best, second-best, and third-best performances are highlighted in bold, blue, and underlined text, respectively.

			Average		Mean	
	w/o pre-training on ImageNet	**with pre-training on ImageNet**	**Disc**	**Cup**	**Disc**	**Cup**
UNet	✓	-	95.45	83.17	96.41	86.36
UNet++	✓	-	96.2	<u>84.29</u>	96.78	87.2
Attention UNet	-	✓	96.5	84.5	97.54	<u>86.48</u>
Attention UNet with VGG11 Backbone	-	✓	95.84	83.25	95.65	84.25
Attention UNet with ResNet-50 Backbone	-	✓	96.19	83.88	96.75	85.55
UNet with VGG11 Backbone	-	✓	<u>96.44</u>	83.58	<u>96.89</u>	85.5
UNet with ResNet-50 Backbone	-	✓	**96.98**	**85.75**	**98.14**	**88**

3.4.4 Ablation Studies

3.4.4.1 Analysis with Different Base Models

In this section, we present a comparative analysis of our proposed method's performance on the RIGA dataset, evaluated using different base models: U-Net, U-Net++, Attention U-Net and with different backbones like VGG11, ResNet50 etc. These models were tested with both Average and Mean-Voting annotations. The results, summarized in Table 3.3, indicate that pre-training on ImageNet consistently results in better-calibrated segmentation outcomes compared to training from scratch. Specifically, U-Net with VGG11 backbone (pre-trained on ImageNet) shows superior performance for both Optic Cup and Optic Disc segmentation when using Average annotations, while Attention U-Net demonstrates superior performance for Optic Cup segmentation with Mean-Voting annotations. Notably, U-Net with ResNet-50 (pre-trained on ImageNet) outperforms all other models across both annotation methods, highlighting the importance of selecting the appropriate base model and pre-training strategy for segmentation tasks.

Table 3.4: The impact on the performances of our proposed CSDL module is shown here. The highest performance is highlighted with bold.

	Average		Mean	
	Disc	**Cup**	**Disc**	**Cup**
without CSDL	95.81	84.85	97.14	86.24
with CSDL	**96.98**	**85.75**	**98.14**	**88**

3.4.4.2 Analysis of CSDL Module

In this section, we analyze the impact of the proposed CSDL module on segmentation performance. We compare the Dice scores for both the Optic Cup and Optic Disc on the RIGA dataset, with and without the CSDL block. For baseline comparison, we replace the CSDL block with a generic distribution calculation block, as used in [88]. The Y-shaped networks for annotator-specific segmentation were kept unchanged across all experiments. Results shown in Table 3.4 indicate a decrease in Dice scores when the CSDL block is removed. Specifically, the Dice score for the Optic Disc drops from 96.98 to 95.81, and for the Optic Cup, it drops from 85.75 to 84.85 when trained with Average annotations. Similarly, with Mean Annotations, the Dice score for the Optic Disc decreases from 98.14 to 96.45, and for the Optic Cup, it decreases from 88.00 to 86.24. This performance drop underscores the effectiveness of the CSDL module in modeling class-specific distributions, ultimately leading to more accurate segmentation results.

3.5 Summary of the Chapter

This chapter analysed *label uncertainty*—the variation that appears when multiple experts annotate the same image—and showed how such disagreement can undermine segmentation reliability. Treating one mask as the sole “ground truth” overlooks valid clinical interpretations and weakens model generalisation.

To incorporate annotation variability into the learning pipeline, we introduced two novel components. The **Class-Specific Distribution Learning (CSDL)** module learns a probability distribution over all masks for each class, enabling the network to appreciate a spectrum of plausible boundaries rather than a single contour.

The **Annotator-Specific Preference Estimator (ASPE)** models the individual bias of every annotator and fuses those preferences into a data-driven consensus. Together, CSDL and ASPE create a segmentation framework that is explicitly aware of inter-observer diversity.

Experiments on the *RIGA* retinal-vessel dataset and the multi-organ *QUBIQ* benchmark confirmed the effectiveness of this design. The proposed framework achieved higher Dice scores and lower calibration error than leading fusion, pseudo-labelling, and bias-modelling baselines. Qualitative inspection also showed smoother borders and fewer mis-labelled edge pixels in regions where experts typically disagree. These findings demonstrate that directly modelling annotation variability produces segmentation systems that are both more accurate and more dependable across heterogeneous medical datasets.

Chapter 4

Addressing Model Uncertainty using Ensemble Techniques

4.1 Introduction

Deep learning models have achieved remarkable success in medical image segmentation. However, their performance is often compromised by factors such as limited training data, overfitting, and sensitivity to model initialization. These factors can lead to inconsistencies in predictions, particularly when models are trained on the same dataset but produce different segmentation outputs. Such disagreement among models arises due to differences in learned feature representations, sensitivity to noise in training data, and biases in optimization paths. This inconsistency can be more pronounced in regions with complex structures or ambiguous object boundaries, where different classifiers may prioritize different visual cues.

Ensemble learning, which aggregates predictions from multiple models, has emerged as a powerful approach to mitigate such inconsistencies by reducing the variance in individual predictions. By combining multiple segmentation models, ensemble learning smooths out errors introduced by any single model, leading to more stable and accurate segmentation outputs. However, traditional ensemble methods often assume that individual models contribute independently to the final decision, which is not always the case in real-world applications where classifiers may share common training patterns and biases.

Motivated by these challenges, this work explores ensemble techniques that enhance segmentation accuracy by leveraging the strengths of multiple models. Unlike conventional ensembling strategies, we introduce a copula function-based ensemble method that explicitly captures the statistical dependencies between models. By learning these dependencies, our approach effectively balances the contribution of each model, leading to more reliable segmentation results. While uncertainty estimation is a crucial aspect of ensemble methods, our primary goal in this work is not to measure uncertainty but to enhance segmentation performance by leveraging the diversity of multiple classifiers.

4.1.0.1 What is Ensemble Learning?

Ensemble learning is a methodology that combines multiple models, referred to as base learners, to produce a single aggregated prediction. Rather than relying on a single model that may capture only partial aspects of a segmentation task, ensemble learning integrates multiple viewpoints, leading to more comprehensive and balanced segmentation results.
In image segmentation, different models may emphasize different features due to variations in training data exposure, initialization, and learned information representations. Some models may focus more on texture, while others may prioritize edges or object boundaries. This diversity can be particularly beneficial in cases where certain models fail to correctly segment ambiguous regions, as the ensemble method can mitigate individual weaknesses by leveraging the complementary strengths of each model.
However, conventional ensemble approaches often assume independence among classifiers, which is not always realistic. Models trained on the same dataset tend to develop correlated errors, and blindly averaging their outputs may not always lead to an optimal result. To overcome this, our proposed approach utilizes a copula function-based ensemble technique, which explicitly models the statistical dependencies between classifiers. By capturing these interdependencies, our method ensures that classifier disagreements are accounted for in a structured manner, leading to improved segmentation accuracy.

4.1.0.2 Uncertainty Quantification using Ensemble

One of the key advantages of ensemble methods is their capability to quantify uncertainty. By examining the variation among the predictions of individual models, it is possible to derive a confidence measure for the final segmentation. This uncertainty quantification is critical in clinical scenarios, as it provides insights into the reliability of the model's predictions, especially in regions where the data is ambiguous or the segmentation task is particularly challenging.

4.1.1 Contributions of this Chapter

This chapter focuses on addressing *model uncertainty* through *ensemble learning* and introduces a novel copula-based ensemble framework for segmentation tasks. The key contributions of this chapter are:
1. **Introduction of a Copula-Based Ensemble Learning Framework:** We propose a novel ensemble method that explicitly models the dependencies between classifiers, ensuring an optimized combination of predictions.
2. **Analysis of Ensemble Learning for Segmentation Stability:** We demonstrate how ensemble methods improve segmentation robustness by reducing model variance and compensating for individual classifier weaknesses.
3. **Comparison with Conventional Ensemble Methods:** The proposed approach is compared against traditional ensemble strategies such as majority voting, averaging, and Bayesian ensembles, showing its superiority in segmentation accuracy.
4. **Experimental Validation on Medical Imaging Datasets:** We evaluate our approach on multiple publicly available medical imaging datasets, highlighting its effectiveness in improving segmentation performance and stability.

4.1.2 Chapter Outline

The rest of the chapter is organized as follows:

- **Section 4.2:** Discusses prior research on ensemble learning, copula-based modeling, and uncertainty quantification in deep learning-based segmentation.

- **Section 4.3:** Details the proposed copula-based ensemble learning framework, describing its theoretical foundations and implementation.
- **Section 4.4:** Discusses the dataset, experimental setup, evaluation metrics, and quantitative results.
- **Section 4.5:** Summarizes the key findings of the chapter.

4.2 Related Works

Ensembling multiple segmentation models is a well-established way to boost predictive performance. Widely used score-level fusion rules include majority voting [11, 64, 82, 158], maximum, minimum, median, weighted mean [80], simple averaging [104], and product aggregation [141]. More sophisticated strategies employ rank-based fusion [52], Bayesian model combination [82, 158], Dempster–Shafer evidence theory [158], fuzzy measures [21, 79, 78], and probabilistic pooling such as the Linear Opinion Pool [39], its beta-transformed variant [119], or logit-based fusion [127].
A common shortcoming of these methods is the implicit assumption that individual classifier outputs are conditionally independent. Ignoring the *statistical dependence* among models can lead to sub-optimal fusion, as noted by Özdemir *et al.* [110]. Copulas offer a principled remedy: they decouple marginal behaviour from dependence structure by linking the univariate marginals to a joint distribution through a multivariate "copula" function. This flexibility makes copulas attractive for modelling correlated classifier scores.

Copula-based techniques have already proved useful in hydrology [83], climate science [69], medical diagnostics [29, 116], data mining [156], classification [125], and evolutionary computation [23]. In computer vision, initial work has focused on binary tasks, typically using a single Gaussian copula—for example, in time-series forecasting and breast-histology image analysis [24, 165, 170, 25].

Our study extends this line of research in three significant ways. **First**, we investigate a broader family of copulas—Gaussian, Student-t, Gumbel, Frank, and Clayton—to better capture the diverse dependence patterns among base segmenters.

Second, we recognise that dependencies can vary across classes and datasets; consequently, we learn a class-specific copula rather than imposing a one-size-fits-all model. **Third**, we demonstrate, to the best of our knowledge for the first time, the effectiveness of copula-based fusion in *multiclass medical image segmentation*. For each pixel, we model the joint likelihood of the softmax belief scores from multiple networks and compute a fused posterior via Bayes' rule, yielding consistent gains in segmentation accuracy and calibration.

4.3 Present Work

4.3.1 Overview of the Problem

This section presents the mathematical framework for our ensemble-based segmentation method. In image classification and segmentation, model uncertainty arises when multiple classifiers trained on the same dataset produce inconsistent predictions due to varying learned feature representations, noise sensitivity, and optimization biases. Traditional ensemble methods combine classifiers to mitigate individual model weaknesses; however, they often assume statistical independence among models, which is unrealistic in practice [110].
We consider a scenario where L classifiers are used for an image classification task. Each classifier produces a confidence score represented as a probability distribution over M possible classes. Let $p_i^{(j)}$ denote the confidence of classifier i for class j. As classifiers generate probability distributions, the confidence scores satisfy:

$$p_i^{(j)} \in [0,1], \quad \forall i = 1,2,\ldots,L, \quad \forall j = 1,2,\ldots,M.$$

The objective of this work is to determine a function $g : [0,1]^L \to [0,1]$, which maps the individual classifiers' confidence scores to an ensemble confidence score $p^{(j)}$:

$$p^{(j)} = g(p_1^{(j)}, p_2^{(j)}, \ldots, p_L^{(j)}),$$

where $p^{(j)}$ represents the aggregated confidence of the ensemble for class j. Unlike conventional methods that assume classifiers operate independently, we adopt a copula-based approach to model the statistical dependencies among classifiers. By

leveraging Bayesian inference, the ensemble confidence is estimated as:

$$\begin{aligned} p^{(j)} &:= g(p_1^{(j)}, p_2^{(j)}, \ldots, p_L^{(j)}) \\ &:= P\left(\text{Class } j \mid p_1^{(j)}, p_2^{(j)}, \ldots, p_L^{(j)}\right) \\ &\propto f\left(p_1^{(j)}, p_2^{(j)}, \ldots, p_L^{(j)} \mid \text{Class } j\right) \times P\left(\text{Class } j\right), \end{aligned} \tag{4.1}$$

where $f(p_1^{(j)}, p_2^{(j)}, \ldots, p_L^{(j)} \mid \text{Class } j)$ is the joint likelihood of the confidence scores for class j. Since evaluating this joint likelihood is challenging due to unknown dependencies, we use copula functions to model them effectively.

4.3.2 A Novel Ensemble Technique using Copula Functions

4.3.2.1 Mathematical Background

Understanding copula functions is essential to our method, as they allow modeling dependencies among classifiers in a structured manner. We provide a brief overview of their theoretical foundation.

Definition: Given N random variables $X_1, X_2, \ldots, X_N$ with marginal cumulative distributions $F_{X_i}(x_i, \tau_i)$ (where τ_i are the marginal parameter), their copula function C is defined as:

$$C\left(u_1, u_2, \ldots, u_N\right) = P\left(U_1 \leq u_1, U_2 \leq u_2, \ldots, U_N \leq u_N\right), \tag{4.2}$$

where $U_i = F_{X_i}(x_i, \tau_i)$.

Sklar's Theorem: Sklar's theorem states that any joint cumulative distribution function $F(x_1, x_2, \ldots, x_N)$ can be expressed as a copula function:

$$F(x_1, x_2, \ldots, x_N) = C\left(F_{X_1}\left(x_1, \tau_1\right), F_{X_2}\left(x_2, \tau_2\right), \ldots, F_{X_N}\left(x_N, \tau_N\right)\right). \tag{4.3}$$

Copula Density Function: The corresponding copula density function is obtained as:

$$c\left(u_1, u_2, \ldots, u_N\right) = \frac{\partial^N}{\partial u_1 \partial u_2 \cdots \partial u_N} C\left(u_1, u_2, \ldots, u_N\right). \tag{4.4}$$

Using the chain rule, the relationship between the copula density function and the joint probability distribution is:

$$f(x_1, x_2, \ldots, x_N) = c\left(F_{X_1}(x_1, \tau_1), F_{X_2}(x_2, \tau_2), \ldots, F_{X_N}(x_N, \tau_N)\right) \times \prod_{i=1}^{N} f_{X_i}(x_i, \tau_i). \tag{4.5}$$

Copula Families: Copula functions are classified into different families, including elliptical copulas (e.g., Gaussian and Student-t) and Archimedean copulas (e.g., Gumbel, Frank, Clayton) [106](see Table 4.1). In this work, we experiment with five copula functions from these families.

Copula Family	**Equation**	
Gaussian	$C(u_1, u_2 \cdots u_N) = \Phi_\Sigma(\Phi^{-1}(u_1), \Phi^{-1}(u_2) \cdots \Phi^{-1}(u_N))$	Where Φ_Σ is cumulative distribution function of multivariate normal distribution with correlation matrix Σ and Φ^{-1} is the quantile function of normal distribution
Student-t	$C(u_1, u_2 \cdots u_N) = t_{\nu,\Sigma}(t_\nu^{-1}(u_1), t_\nu^{-1}(u_2) \cdots t_\nu^{-1}(u_N))$	Where $t_{\nu,\Sigma}$ is the cumulative distribution function(CDF) of multivariate Student-t distribution with correlation matrix Σ and degrees of freedom ν and t_ν is the CDF of univariate Student-t distribution with degrees of freedom ν.
Frank	$C(u_1, u_2 \cdots u_N) = -\theta^{-1} \log\left[1 + (e^{-\theta} - 1)^{-1} \prod_{i=1}^{N} (e^{-\theta u_i} - 1)\right]$	$\theta \in \mathbb{R} \backslash \{0\}$
Clayton	$C(u_1, u_2 \cdots u_N) = \left(\sum_{i=1}^{N} u_i^{-\theta} - 1\right)^{-\frac{1}{\theta}}$	$\theta \in [-1, \infty) \backslash \{0\}$
Gumbel	$C(u_1, u_2 \cdots u_N) = \exp\left(-\left[\sum_{i=1}^{N} (-\log u_i)^\theta\right]^{\frac{1}{\theta}}\right)$	$\theta \in [1, \infty)$

Table 4.1: Mathematical Expression of different Copula families

4.3.3 Estimating Copula Parameters for Data Fitting

To fit a copula function to an N-dimensional data sample, we estimate its parameters using one of the following methods:

4.3.3.1 Maximum Likelihood (ML) Method

The Maximum Likelihood (ML) method [154][45] allows us to estimate both the copula parameters and the marginal parameters. Given a set of T training examples for each of the N random variables $\{(x_1^t)\}_{t=1}^T, \ldots, \{(x_N^t)\}_{t=1}^T$, the likelihood function $\mathcal{L}$ is $\mathcal{L} = \prod_{t=1}^T f(x_1, x_2, \ldots x_N)$. Let's assume the copula C belongs to a family of copula indexed by a parameter θ: $C = C(u_1, u_2, \ldots, u_N, \theta)$ and the Margins and corresponding uni-variate densities are F_{X_i} and f_{X_i} indexed by parameter τ_i. The using equation (4.5) the maximum likelihood function $\mathcal{L}$ can be written as

$$\begin{aligned}\mathcal{L}(\tau_1, \tau_2, \ldots, \tau_N, \theta) &= \prod_{t=1}^{T} f\left(x_1^t, x_2^t, \ldots, x_N^t, \tau_1, \tau_2, \ldots, \tau_N, \Theta\right) \\ &= \prod_{t=1}^{T} \left[c\left[F_{X_1}\left(x_1^t, \tau_1\right), F_{X_2}\left(x_2^t, \tau_2\right) \ldots, F_{X_N}\left(x_N^t, \tau_N\right), \Theta\right] \prod_{i=1}^{N} f_{X_i}\left(x_i^t, \tau_i\right) \right]\end{aligned} \tag{4.6}$$

Taking the logarithm of the likelihood function (4.6), we arrive at a more manageable expression (4.7).

$$\begin{aligned}\log\left(\mathcal{L}(\tau_1, \tau_2, \ldots, \tau_N, \theta)\right) &= \log\left(\prod_{t=1}^{T} \left[c\left(F_{X_1}\left(x_1^t, \tau_1\right), F_{X_2}\left(x_2^t, \tau_2\right), \ldots, \right.\right.\right. \\ &\qquad \left.\left.\left. \ldots, F_{X_N}\left(x_N^t, \tau_N\right), \Theta\right) \prod_{i=1}^{N} f_{X_i}\left(x_i^t, \tau_i\right) \right] \right) \\ &= \sum_{t=1}^{T} \log\left[c\left(F_{X_1}\left(x_1^t, \tau_1\right), F_{X_2}\left(x_2^t, \tau_2\right), \ldots, \right.\right. \\ &\qquad \left.\left. \ldots, F_{X_N}\left(x_N^t, \tau_N\right), \Theta\right) \prod_{i=1}^{N} f_{X_i}\left(x_i^t, \tau_i\right) \right] \\ &= \sum_{t=1}^{T} \log\left(c\left(F_{X_1}\left(x_1^t, \tau_1\right), F_{X_2}\left(x_2^t, \tau_2\right), \ldots, \right.\right. \\ &\qquad \left.\left. \ldots F_{X_N}\left(x_N^t, \tau_N\right), \Theta\right) \right) + \left(\sum_{t=1}^{T} \sum_{i=1}^{N} \log f_{X_i}\left(x_i^t, \tau_i\right) \right)\end{aligned} \tag{4.7}$$

By maximizing the log-likelihood function, we can obtain the optimal estimates of the copula and marginal parameters $\hat{\tau_1}, \hat{\tau_2}, \ldots, \hat{\tau_N}, \hat{\theta}$ as depicted in (4.8).

$$\hat{\tau_1}, \hat{\tau_2}, \ldots, \hat{\tau_N}, \hat{\theta} = \arg\max \mathcal{L}\left(\tau_1, \tau_2, \ldots, \tau_N, \theta\right) \tag{4.8}$$

To illustrate the procedure for practice let's assume the copula belongs to a Multivariate Gaussian copula, hence the parameter for this copula would be the multivariate correlation matrix i.e. $\theta = \Sigma$. Also, let all the Marginals F_{X_i} are Gaussian. Hence the parameter for marginals τ_i would be $\tau_i = \{\mu_i, \sigma_i\}$. Thus from equation (4.8), we have

$$\hat{\mu_1}, \hat{\sigma_1}, \hat{\mu_2}, \hat{\sigma_2}, \ldots, \hat{\mu_N}, \hat{\sigma_N}, \hat{\Sigma} = \arg\max \mathcal{L}\left(\mu_1, \sigma_1, \mu_2, \sigma_2, \ldots, \mu_N, \sigma_N, \Sigma\right)$$

While the ML method is widely used, it can be computationally complex and may not be suitable for real-world scenarios with large datasets. As a result, alternative approaches, such as the Inference Function for Margins (IFM) method, have been developed.

4.3.3.2 Inference Function for Margins (IFM) Method

The Inference Function for Margins (IFM) method [67][34] offers a simplified approach to copula parameter estimation. It involves separating the estimation of the marginal parameters $\tau = (\tau_1, \tau_2 \ldots \tau_N)$ and the copula parameter Θ.
First, the marginal parameters are estimated using the second part of (4.7) to give $\hat{\tau_i}$ as-

$$\hat{\tau_i} = \arg\max_{\tau_i} \left(\sum_{t=1}^{T} \sum_{i=1}^{N} \log f_{X_i}\left(x_i^t, \tau_i\right) \right) \tag{4.9}$$

Once the marginal parameters have been estimated, the IFM method uses these estimates (4.9) to estimate the copula parameters. This is achieved by maximizing the log-likelihood function of the copula, considering the estimated marginals as

depicted in (4.10).

$$\hat{\Theta} = \arg\max_{\theta} \left[\sum_{t=1}^{T} \log \left(c \left(F_{X_1} \left(x_1^t, \hat{\tau_1} \right), F_{X_2} \left(x_2^t, \hat{\tau_2} \right), \ldots, F_{X_N} \left(x_N^t, \hat{\tau_N} \right), \Theta \right) \right) \right] \tag{4.10}$$

Hence, the IFM method provides estimates $(\hat{\tau_1}, \hat{\tau_2}, \ldots, \hat{\tau_N}, \hat{\Theta})$ by first estimating the marginal parameters and then estimating the copula parameters. It should be noted that the IFM method is asymptotically equivalent to the ML method.
Following the same example from the ML method, let's assume the copula belongs to a Multivariate Gaussian copula, hence the parameter for this copula would be the multivariate correlation matrix i.e. $\theta = \Sigma$. Also, let all the Marginals F_{X_i} to be Gaussian. Hence the parameter for marginals τ_i would be $\tau = \{\mu_i, \sigma_i\}$. Thus estimating the marginal parameters $\tau = \{\mu_i, \sigma_i\}$ using equation (4.9), we have optimal marginal parameter as follows:

$$\hat{\mu_i}, \hat{\sigma_i} = \arg\max_{\mu_i, \sigma_i} \left(\sum_{t=1}^{T} \sum_{i=1}^{N} \log f_{X_i} \left(x_i^t, \mu_i, \sigma_i \right) \right) \tag{4.11}$$

Now, the optimal copula parameter will be (using equation (4.10))

$$\hat{\Sigma} = \arg\max_{\Sigma} \left[\sum_{t=1}^{T} \log \left(c \left(F_{X_1} \left(x_1^t, \hat{\mu_1}, \hat{\sigma_1} \right), F_{X_2} \left(x_2^t, \hat{\mu_2}, \hat{\sigma_2} \right), \ldots, F_{X_N} \left(x_N^t, \hat{\mu_N}, \hat{\sigma_N} \right), \Sigma \right) \right) \right] \tag{4.12}$$

The IFM method offers a practical alternative to the ML method, particularly in scenarios where the ML method becomes computationally infeasible or the marginals are well-defined.

4.3.3.3 Marginals Estimation

In general, we estimate marginal distributions $f_{X_i}(x_i, \tau_i)$ in the Inference Function for Margins (IFM) approach, where we employ Kernel Density Estimation (KDE) to select the probability density of a random variable. In other words, for any sample, $x_1, x_2, \ldots, x_N$, KDE provides an estimate of the density function of the variable x

that employs a Gaussian kernel as the smoothing function,

$$\widehat{f}_H(x) = \frac{1}{NH} \sum_{i=1}^{N} K\left(\frac{x - x_i}{H}\right), \tag{4.13}$$

where bandwidth H controls the smoothness of the resulting density curve and $K(\cdot)$ represents the kernel (Gaussian) smoothing function: $K(x) = \frac{1}{\sqrt{2\pi}} e^{-(x^2/2)}$. With KDE, we estimate marginal distributions required for the IFM approach, even in scenarios where true marginals are unknown.

4.3.3.4 Measuring the Fitness of Copula

To evaluate how well our data fits the copula model, several statistical measures can be utilized. Notable among these are the Log-Likelihood (LL), Akaike Information Criterion (AIC) [14] [5], Bayesian Information Criterion (BIC) [14] [5], Average Kolmogorov-Smirnov distance (AKS) [162], and Cramér-von Mises distance [111].

4.4 Experiments

4.4.1 Application 1: Leveraging Class-Specific Copula Functions for Image Segmentation

Understanding that different object classes in an image exhibit varying spatial and contextual patterns, this application investigates the use of class-specific copula functions for segmentation tasks. By tailoring copula models to individual semantic classes, the ensemble can more effectively capture diverse dependency structures, leading to improved segmentation accuracy across heterogeneous regions of an image.

4.4.1.1 Datasets

We evaluated our proposed model on multiple datasets:
CamVid [13]: is a road scene video sequence consisting of 367 frames of train set, 101 frames of validation set and 233 frames of test set. In our experiment, we set the dimension of each frame to 360×480, which is half of the original dimension. The ground truth of Camvid has 11 semantic segmentation classes, namely Sky, Building, Column-pole, Road, Sidewalk, Tree, Sign-symbol, Fence, Car, Pedestrian, Bicyclist, and one void class. We have trained our base deep models, i.e., SegNet, PSPNet, and Tiramisu, with these 12 semantic classes.
ICCV09[43]: consists of 715 images of urban and rural scenes assembled from a collection of public image datasets. There are 572 images for training and 143 images for validation purposes. It has a total of 9 semantic segmentation classes, namely Sky, Tree, Grass, Ground, Building, Mountain, Water, Object, and one unknown class. The resolution of each image here is 240×340, but in our experiment, we have resized each image to 360×480.
MedSeg [1]: COVID-19 CT Images Segmentation or MedSeg dataset [1], which consists of 100 axial CT images (in .jpg format) from more than 4 patients with COVID-19. Each image has dimensions of 512×512 pixels, and corresponding masks are provided with four classes: 0 - "ground class", 1 - "consolidations", 2 - "lungs other", and 3 - "background". Given the limited number of training samples, we utilized a five-fold cross-validation technique for training.

4.4.1.2 Experimental Setup

We considered various deep learning segmentation models for this task, and based on their validation performances, we selected the top three models for ensembling: SegNet [10], PSPNet [167], and U-Net [122] for MedSeg dataset and SegNet [10], PSPNet [167], and Tiramisu [61]. These models are well-established in the field of semantic segmentation. All models were trained from scratch using the same set of hyperparameters in the PyTorch environment. Each of the above models returns its belief score as a probability distribution across all the classes for each pixel of the input image using a softmax function in our case. Hence, if the input size is $H \times W$, then the Classifier output size will be $H \times W \times M$, Where M is the total number of classes.

The next step is to determine which Copula will be the best fit for the data (for each class) during the ensemble procedure. To do that, we fit our data for each class to five popular elliptic and Archimedean copulas, namely 1)Gaussian, 2)Students-t, 3)Clayton, 4)Gumbel, and 5) Frank and determine the LL, AIC and BIC statistics [107] [14] [5]. After evaluating those statistics, the best-fitted Copula will be chosen for each class. At the end of this procedure, all the selected class-specific best copulas are used to estimate their parameters. The final class-specific fused distribution will be determined using these parameters and validation data (Classifier outputs on validation samples for each class as a probability distribution across total classes). The whole approach is presented briefly in **Algorithm 1**. To fit data

Algorithm 1: Copula Ensembling

Constants: L = Number of Classifiers
M = Number of segmentation classes
P = Total Number of Pixel for Training Images
Q = Total Number of Pixel for Validation Images
Family = The best fitted copula function corresponding marginals for the given data

Data : $[X_m^l]_{P\times 1}$ = Pixel-level Probability Distribution for class m, from classifier l.
$[XT_m^l]_{Q\times 1}$ = Pixel-level Probability Distribution for class m, from classifier l.

Result: R_m = Pixel-Level Probability Distribution after ensembling for class m.

1 **for** $m = 1$ **to** M **do**
2 $X = [X_m^1 X_m^2 \dots X_m^L]$
3 U = `KernelDensityEstimation(`X`)`
4 Copula-Parameters = `Fit(`*Family,U*`)` ; ▷ `by IFM method`
5 $XT = [XT_m^1 XT_m^2 \dots XT_m^L]$
6 UT = `KernelDensityEstimation(`XT`)`
7 A_l = `Pdf(`*Family,UT,Copula-Parameters*`)`
8 $R_m = A_m$ * `KDEpdf(`TX`)` * `Prior(`m`)` ; ▷ `by equation (1)`
9 **end for**

to a copula, we have used the function `Fit` based on the **IFM** method in Algorithm 1, which returns an estimate of parameters of the given copula family. The term *Family* is used here to denote the best-fitted Copula family for given input data, which are determined empirically using their fitting statistics(AIC, BIC, LL) after fitting with some well-known copula families. The copula family fitting statistics

for each class of training data on the CamVid dataset are presented in the Appendix. The function `KernelDensityEstimation` is the non-parametric marginal estimation of our given data. The `Pdf` function returns the probability density function of the selected copula family for the estimated copula parameters at the test data samples.

4.4.1.3 Results

The performance comparison in terms of overall pixel accuracy and mean IoU for the CamVid, ICCV09, and MedSeg datasets is presented in Tables 4.2, 4.3, and 4.4, respectively. These tables include comparisons against various ensemble models, ensembling with a single copula function, and our proposed approach.
Our model outperformed the baseline segmentation models and other ensemble techniques, achieving the highest overall accuracy and mean IoU for all three dataset. Notably, for Fold 3 and Fold 4 of the MedSeg dataset, the Student-t copula-based ensemble exhibited the best mean IoU performance. For qualitative evaluation, Figures 4.1 and 4.2 illustrate segmentation outputs and compare our method against standard segmentation models. The results confirm the effectiveness of our copula-based ensembling technique, demonstrating significant improvements in segmentation accuracy and quality.

Table 4.2: Results of Ensembling data from SegNet, PSPNet and Tiramisu on CamVid dataset

Architechtures	**Overall Accuracy**	**Mean Accuracy**	**Mean IOU**
SegNet[10]	84.700303	49.519581	0.41825
PSPNet[167]	92.818602	78.782833	0.663291
Tiramisu[61]	91.637061	77.221778	0.637337
LOP[39]	92.608401	81.615978	0.635698
Majority_voting[11]	92.8761	80.979356	0.652887
Logit[127]	92.608401	81.615978	0.635698
Gaussian	90.913687	79.188869	0.60419
Student-t	88.045649	73.047487	0.556318
Frank	87.969128	72.89565	0.552939
Clayton	91.90342	81.650784	0.65254
Gumbel	91.119866	79.080911	0.579787
Proposed	**93.091532**	**82.559746**	**0.672016**

Table 4.3: Results of Ensembling data from SegNet, PSPNet and Tiramisu on ICCV09 dataset

Architechtures	Overall Accuracy	Mean Accuracy	Mean IOU
SegNet[10]	75.336665	58.313291	0.476233
PSPNet[167]	83.22131	66.745377	0.525069
Tiramisu[61]	66.064429	55.15124	0.361695
LOP[39]	82.909654	**68.870868**	0.538002
Majority_voting[11]	82.632492	0.540236	0.511325
Logit[127]	83.354601	68.274124	0.53233
Gaussian	82.862361	67.139727	0.519109
Student-t	82.864234	67.119442	0.519202
Frank	82.830263	67.029158	0.518913
Clayton	83.051362	67.12981	0.517773
Gumbel	82.882825	67.140376	0.519423
Proposed	**83.821968**	68.326652	**0.548254**

Table 4.4: Comparison of total pixel accuracy and mean IOU between the base segmentation model, simple probabilistic ensembling models, and copula-based ensembling models on the MedSeg dataset. MV, PA, and WA represent Majority Voting, Probability Average, and Weighted Average, respectively. $1, 2, 3, 4$, and 5 denote the fold numbers.

	Accuracy							Mean IOU						
Models	**1**	**2**	**3**	**4**	**5**	**Avg.**	**Std. Dv.**	**1**	**2**	**3**	**4**	**5**	**Avg.**	**Std. Dv**
SegNet	95.22	91.3	94.61	94.74	93.48	93.87	1.5719	0.4081	0.8432	0.8399	0.8236	0.8730	0.7576	0.1962
PSPNet	96.6	93.99	95.71	95.34	95.88	95.50	0.9623	0.4407	0.6000	0.6199	0.6382	0.6397	0.5877	0.0838
UNet	96.41	96.11	95.6	95.4	95.48	95.8	0.4389	0.4659	0.6208	0.6110	0.6478	0.6483	0.5988	0.0761
MV	97.15	96.55	96.85	96.66	95.98	95.5	0.4325	0.4680	0.8256	0.8146	0.8250	0.8365	0.7539	0.1601
PA	96.52	94.49	95.93	95.36	95.82	95.83	0.7567	0.4565	0.7937	0.8025	0.8148	0.8245	0.7384	0.1580
WA	96.73	95.53	96.08	96.02	96.01	96.13	0.4282	0.4675	0.8027	0.8155	0.8254	0.8237	0.7469	0.1565
Logit	96.88	96.41	96.65	96.19	96.27	96.26	0.2837	0.4593	0.8246	0.8270	0.8297	0.8365	0.7554	0.1656
Gaussian	97.01	96.02	96.74	96.55	96.39	96.47	0.3721	0.4525	0.8369	0.8347	0.8370	0.8657	0.7654	0.1754
Student-t	97.56	96.48	97.08	96.97	96.36	96.16	0.4849	0.4603	0.8432	**0.8457**	**0.8566**	0.8725	0.7554	0.1972
Clayton	96.91	95.06	96.89	96.25	96.16	96.04	0.7532	0.4520	0.8356	**0.8457**	0.8562	0.8595	0.7698	0.1779
Gumbel	96.54	95.47	96.61	96.14	96.11	96.12	0.4541	0.4497	0.8347	0.8346	0.8456	0.8652	0.7660	0.1773
Frank	97.05	95.09	96.64	96.12	96.42	96.36	0.7387	0.4397	0.8246	0.8346	0.8365	0.8453	0.7561	0.1770
Proposed	**98.59**	**96.91**	**97.92**	**97.29**	**97.22**	**97.3**	0.6705	**0.4696**	**0.8470**	0.8437	0.8456	**0.8769**	**0.7766**	0.1721

4.4.2 Application 2: Gaussian Copula-Based Ensemble of Multi-Level Superpixels for Image Segmentation

In image segmentation tasks, incorporating contextual information from neighboring regions often leads to more accurate and coherent predictions. While immediate neighboring regions contribute valuable local context, extending the neighborhood can provide richer semantic cues, particularly in complex scenes. This

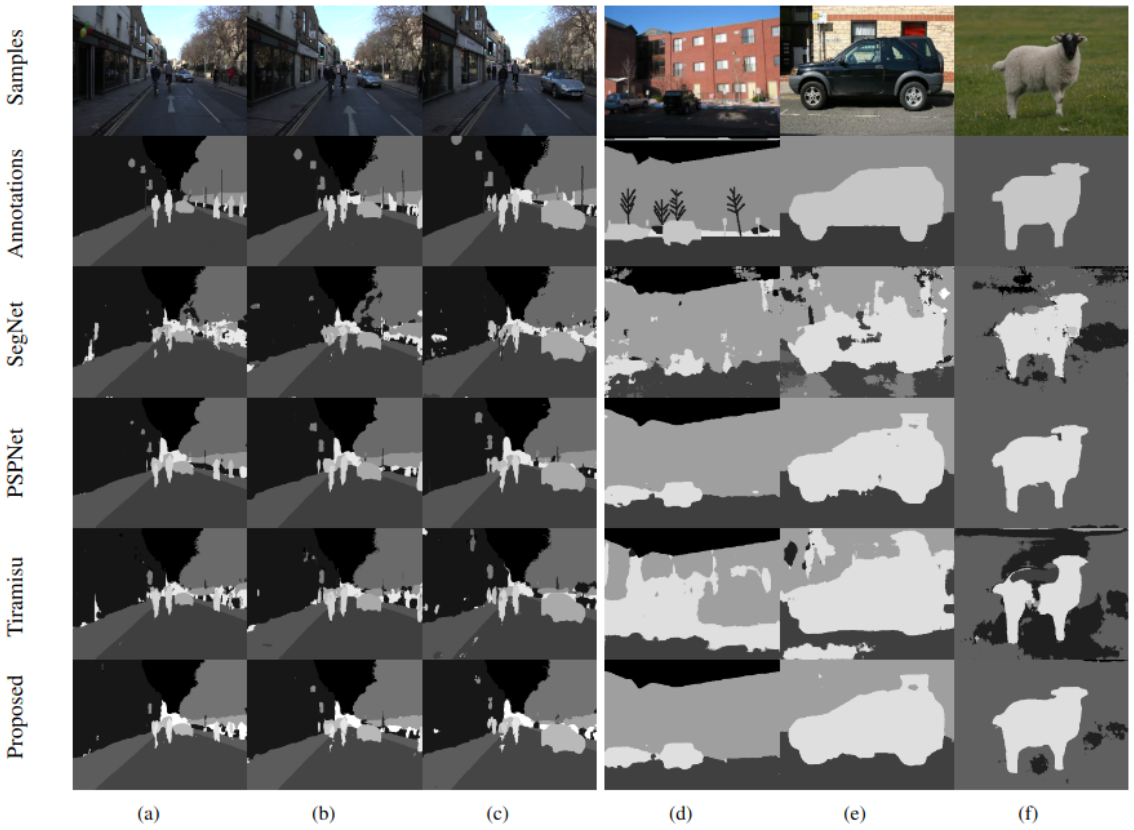

Figure 4.1: Visual representation of performances of our proposed model with the base models on CamVid and ICCV09 datasets. The images indexed with (a),(b),(c) are CamVid samples and (d),(e),(f) are ICCV09 samples

application explores a multi-level fusion strategy, where predictions from the center superpixel, its one-radius neighbors, and two-radius neighbors are integrated using copula functions. By modeling the dependencies across different neighborhood levels, the ensemble captures both fine-grained local details and broader spatial structures. This approach is motivated by the observation that semantic consistency often extends beyond immediate boundaries, and effectively leveraging multi-scale context can significantly enhance segmentation robustness and boundary accuracy.

4.4.2.1 Datasets

We evaluate our approach on three datasets covering different segmentation challenges:

Stanford Background Scene Understanding(SBSU)[44]: This dataset comprises 715 images depicting urban and rural scenes sourced from various public image

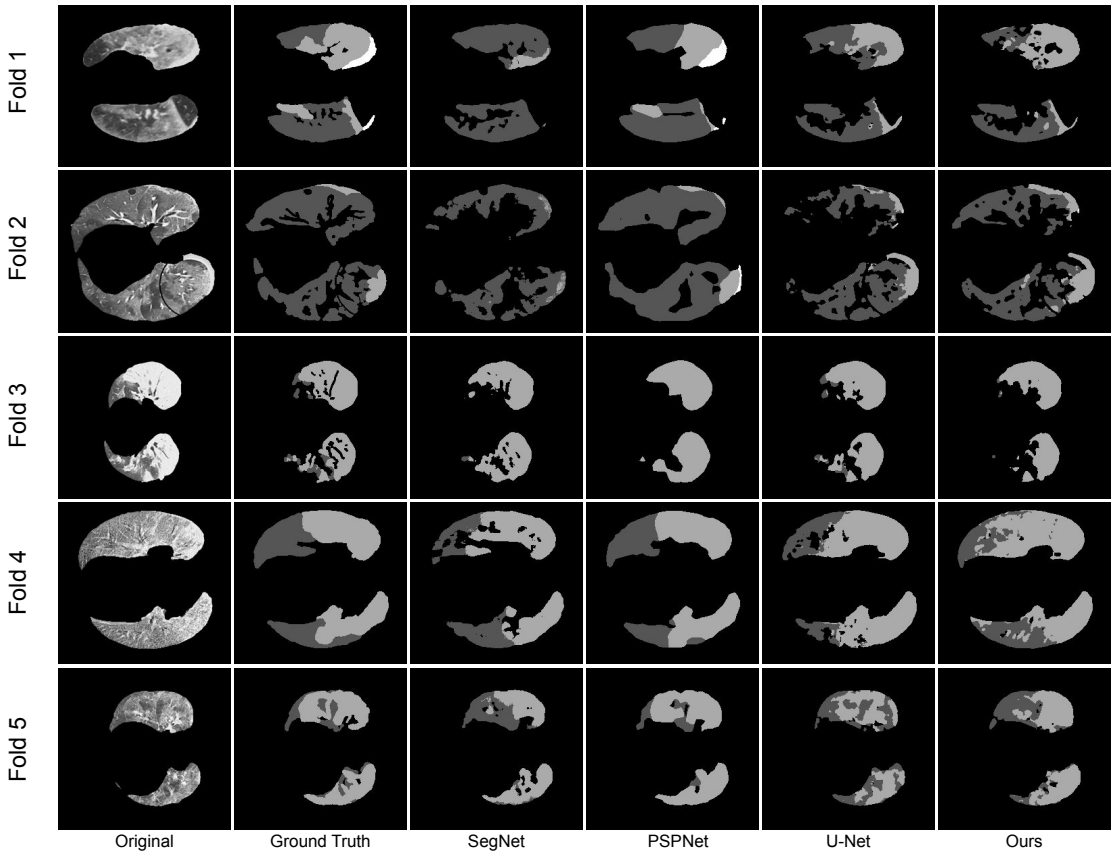

Figure 4.2: Segmentation results of our copula-based ensembling method compared to base segmentation models on the MedSeg dataset.

datasets. It consists of 572 training images and 143 testing images. The images have a resolution of 240×340 pixels and encompass nine semantic segmentation classes: Sky, Tree, Grass, Ground, Building, Mountain, Water, Object, and an unknown class.

Pratheepan Human Skin Detection(PHSD)[140]: This dataset comprises 78 face images obtained randomly from Google. It includes 32 individual images with simple backgrounds and 46 family photos with complex backgrounds, each accompanied by ground truth annotations. The dataset is binary, with two classes: human skin and background. In this experiment we have resized each image to 224×224, which is standard resolution various segmentation models.

Semantic Segmentation of Underwater Imagery(SSUI)[59]: This dataset consists of approximately 1500 annotated training images and 110 annotated testing images focused on underwater imagery segmentation. This dataset encompasses eight classes namely: Background/waterbody(BW), Human Divers(HD),

Aquatic Plants and sea-grass(PF), Wrecks and Ruins(WR), Robots (AUVs/ROVs/instruments)(RO), Reefs and Invertebrates(RI), Fish and Vertebrates(FV), and Seafloor and Rocks(SR). We have resized the data to 320×240 as used in SUIM-Net[60].

For all datasets, images are resized for consistency, and 10-25% of the training data is used for validation.

4.4.2.2 Methodology

Superpixels, characterized by groups of pixels sharing common traits such as intensity, colour, etc. offer a richer alternative for image segmentation compared to individual pixels [147]. Leveraging superpixels significantly reduces computational costs compared to processing all pixels individually. For instance, a standard image measuring 224×224 encompasses a total of $50,176$ pixels, necessitating classification for semantic segmentation tasks. However, employing deep learning models like Encoder-Decoder networks or Fully Convolutional networks to segment these images entails processing millions of parameters and poses challenges in training. Conversely, dividing the same image into approximately 100 superpixels drastically reduces the data volume, simplifying the segmentation task. Achieving effective semantic segmentation then involves classifying each superpixel into its respective class. In the results section, we conduct a comparative analysis between superpixel-based segmentation and the performance of several deep learning models.

Numerous algorithms exist for dividing images into superpixels, including Watershed [55], Mean-Shift [22], Normalized-Cuts [133], Graph-Based, QuickShift [145], TurboPixels [84], and Simple Linear Iterative Clustering(SLIC) [3]. In our study, we compare four superpixel algorithms, namely, the Felzenswalbs method, SLIC, Quickshift, and the compact watershed, which are available in the OpenCV library. We have created around 500 superpixels with those four methods. The visual comparison is depicted in Figure 4.3. The superpixels generated with Felzenszwalbs's and Quickshift's methods are not uniform in size. The superpixels from Compact Watershed are better than the previous two but fail to create uniform superpixels in some parts of the image (parts outside of the cup and plate). On the

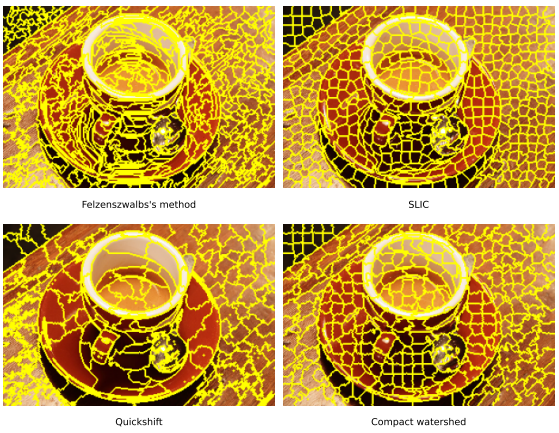

Figure 4.3: Visual Comparison of superpixels generated by four different method

other hand, SLIC generates superpixels by clustering pixels based on color similarity and proximity within the image. Notably, SLIC produces superpixels that exhibit smoother edges and uniform sizes, making them ideal for our task.
Due to variations in image resolution within a dataset, the size of superpixels may lack consistency, potentially affecting segmentation quality. If the average superpixel size is too small, important image features may not be adequately captured, whereas an overly large size could lead to overlapping features. To address this, we initially resize images to a fixed resolution and segment them into a predetermined number of superpixels. Since the optimal superpixel size varies across datasets, we conduct experiments using six sets of superpixels namely sets 1 through 6 for each dataset. Each set is characterized by an average number of superpixels per image, with set 1 comprising of 100 superpixels on an average, set 2 with 200, and so forth. The average resolution of each superpixel within a set is calculated by dividing the product of the image's height and width by the average number of superpixels. We conduct comparative performance analyses for each set in the results section, thereby evaluating the impact of superpixel size variation on segmentation accuracy. To achieve optimal results in semantic image segmentation, it's essential to consider the contextual information surrounding a superpixel. Therefore, we augmented the size of the superpixel by merging it with its nearby counterparts for additional contextual information. Although superpixels inherently contain richer texture information compared to single pixels, merging them with nearby superpixels can provide additional context, enhancing segmentation accuracy. In our approach, we categorized the merged patches into three levels: the center patch, representing the original superpixel itself; the one-radius patch, containing the

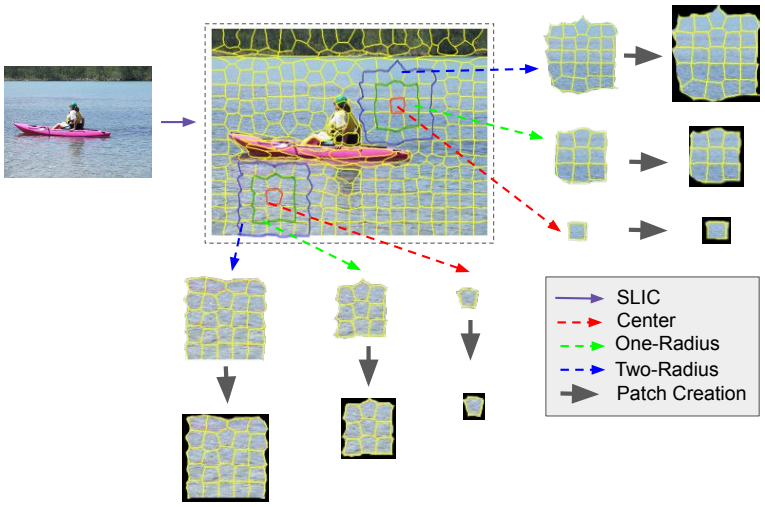

Figure 4.4: Creating center, one-radius, and two-radius patches from superpixels.

original superpixel along with adjacent nearby superpixels; and the two-radius patch, comprising the original superpixel and its surrounding neighbourhood superpixels up to radius two. We keep all three categories of superpixels in separate subfolders for each set. So each set (from 1 to 6) of superpixels not only contains the generated superpixels(center patches) but also contains two separate subfolders containing one-radius patches and two-radius patches respectively. In practical scenarios, the shape of the center superpixel or its neighbours may not always be rectangular. Thus, we filled the exterior of these patches with black pixels to ensure a uniform rectangular shape, as illustrated in Figure 4.4. Subsequently, we trained three custom Convolutional Neural Networks (CNNs), each dedicated to one of the superpixel categories, and combined their outputs through ensembling to enhance segmentation performance. This process was repeated for each set across all datasets, allowing for comprehensive evaluation and optimization of segmentation accuracy. It's important to note that we only visually compared superpixels generated by these four methods and quantitatively assessed uniformity of the size of the superpixels generated over each image. This decision aimed to reduce experimental burden, as repeating the entire experimental pipeline (superpixel generation, creating six sets of superpixels, generating three levels of patches, training, and ensembling with copula function) for each superpixel method would be extensively time consuming. Instead, we chose the best superpixel generation method based on visual as well as quantitative inspection and proceeded with the subsequent experiments with the chosen one.

Table 4.5: Architectures of our custom CNNs

	Input Size	Network Architechture
CNN 1	24×24	conv($5 \times 5 \times 32$) $\rightarrow$ Relu $\rightarrow$ Pooling $\rightarrow$ conv($3 \times 3 \times 64$) $\rightarrow$ Relu $\rightarrow$ Pooling $\rightarrow$ Flatten $\rightarrow$ Dropout $\rightarrow$ Linear(256) $\rightarrow$ Linear(Num of classes)
CNN 2	32×32	conv($5 \times 5 \times 32$) $\rightarrow$ Relu $\rightarrow$ Pooling $\rightarrow$ conv($3 \times 3 \times 64$) $\rightarrow$ Relu $\rightarrow$ Pooling $\rightarrow$ Flatten $\rightarrow$ Dropout $\rightarrow$ Linear(256) $\rightarrow$ Linear(Num of classes)
CNN 3	48×48	conv($7 \times 7 \times 32$) $\rightarrow$ Relu $\rightarrow$ Pooling $\rightarrow$ conv($5 \times 5 \times 64$) $\rightarrow$ Relu $\rightarrow$ Pooling $\rightarrow$ Flatten $\rightarrow$ Dropout $\rightarrow$ Linear(256) $\rightarrow$ Linear(Num of classes)

To achieve a cost-effective semantic segmentation model, we adopt a simple Convolutional Neural Network (CNN) architecture tailored for training on superpixels. For training with only center patches, our CNN architecture comprises two convolution layers and one fully connected layer. The first convolution layer utilizes a 5×5 kernel repeated 32 times, followed by 2×2 average pooling. Subsequently, the second convolution layer employs 64 kernels of size 3×3, also followed by 2×2 average pooling. The output of the second convolution layer is then fed into a fully connected layer with 256 hidden nodes, followed by a softmax function. Similarly, CNNs used for training on one-radius and two-radius patches follow a similar architecture, with the first convolution layer utilizing 7×7 and 5×5 kernels respectively. Batch normalization is employed for stable training, and Dropout is utilized to prevent overfitting in all three CNNs. The complete architecture is presented in Table 4.5.

The predictions of these individual CNNs are subsequently integrated using a Gaussian copula function-based ensembling scheme to enhance superpixel classification performance. A brief explanation of the mathematical underpinnings of the copula function is provided in the subsequent section.

4.4.2.3 Copula-Based Ensembling for Model Uncertainty Reduction

Traditional ensembling techniques assume classifier independence, which is often unrealistic. Gaussian Copula functions model dependencies between CNN outputs, enabling a more reliable estimation of class probabilities. By capturing statistical correlations among classifiers, our approach reduces model uncertainty and enhances segmentation robustness.

The final fused probability score is computed as:

$$\begin{aligned} p &:= f\left(p_1, p_2, p_3\right) \\ &= c\left(F_1\left(p_1\right), F_2\left(p_2\right), F_3\left(p_3\right)\right) \cdot f_1\left(p_1\right) \cdot f_2\left(p_2\right) \cdot f_3\left(p_3\right) \end{aligned} \tag{4.14}$$

where $c(\cdot)$ represents the **copula function**, and F_1, F_2, F_3 are the marginal distributions.

4.4.2.4 Results and Analysis

The experimental results are summarized in Table 4.6. Here, we compare the performance of our proposed technique against base CNNs applied to each neighbouring superpixel, along with other traditional ensembling techniques, across all three datasets. Notably, we observe that two-radius superpixels consistently outperform the center and one-radius superpixels, which aligns with our expectation due to the richer contextual information they encapsulate, as previously discussed. Among the traditional ensembling techniques, the Dempster-Shafer theory demon-

Table 4.6: Comparison of our proposed technique with some traditional deep learning segmentation models. Here, the best, second-best, and third-best performances are indicated with bold, underlined, and blue colors respectively.

		Set 1		Set 2		Set 3		Set 4		Set 5		Set 6		Average	
		Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU
SBSU[44]	center	73.04	0.4216	73.57	0.4266	72.61	0.4123	72.22	0.4117	72.39	0.4145	72.26	0.4111	72.68 ± 0.5291	0.4163 ± 0.0064
	one-radius	73.76	0.4259	75.21	0.4356	74.57	0.4288	75.05	0.4355	75.78	0.4397	75.11	0.4348	74.91 ± 0.6853	0.4334 ± 0.0051
	two-radius	74.48	0.4247	75.81	0.4302	76.34	0.4478	76.34	0.4478	77.03	0.4421	75.19	0.4342	75.86 ± 0.9160	0.4378 ± 0.0096
	max-voting	74.42	0.4283	75.34	0.4310	76.89	0.4498	76.26	0.4413	76.65	0.4408	75.82	0.4477	75.90 ± 0.9156	0.4398 ± 0.0087
	Probability Average	74.88	0.4391	75.81	0.4454	76.47	0.4592	77.62	0.4499	77.60	0.4510	76.55	0.4520	76.49 ± 1.0542	0.4494 ± 0.0067
	Dempster-Shafer	75.75	0.4421	76.72	0.4453	78.41	0.4759	78.48	0.4746	78.56	0.4575	78.96	0.4674	77.81 ± 1.2743	0.4605 ± 0.0146
	Proposed	80.33	0.4693	80.56	0.4790	81.22	0.4875	**82.69**	0.4959	82.65	**0.4986**	80.82	0.4815	81.38 ± 1.0434	0.4853 ± 0.0110
		Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU
PHSD[140]	center	86.67	0.5763	86.49	0.5664	87.53	0.5973	87.36	0.5711	88.18	0.5859	87.02	0.5793	87.21 ± 0.6165	0.5794 ± 0.0110
	one-radius	84.68	0.5693	85.70	0.5508	87.98	0.5981	88.22	0.5756	88.97	0.5787	87.13	0.5741	87.11 ± 1.6337	0.5744 ± 0.0153
	two-radius	84.00	0.5670	85.01	0.5423	86.57	0.5792	88.17	0.5789	88.92	0.5815	87.99	0.5602	86.78 ± 1.9417	0.5682 ± 0.0152
	max-voting	85.47	0.5649	86.89	0.5445	87.84	0.5996	87.86	0.5829	89.02	0.5949	88.62	0.5896	87.62 ± 1.2832	0.5794 ± 0.0209
	Probability Average	86.23	0.5793	86.38	0.5412	88.85	0.5988	89.85	0.6093	89.31	0.6013	88.27	0.5864	88.15 ± 1.5202	0.5861 ± 0.0245
	Dempster-Shafer	87.10	0.5796	88.90	0.5824	90.04	0.6012	91.87	0.6116	92.37	0.6296	89.84	0.5926	90.02 ± 1.9380	0.5995 ± 0.0192
	Proposed	91.47	0.6391	91.92	0.6397	92.51	0.6477	92.67	0.6517	**93.75**	**0.6691**	92.75	0.6229	91.85 ± 0.7799	0.6400 ± 0.0154
		Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU
SSUI[59]	center	94.24	0.7912	94.86	0.7972	95.36	0.8141	94.81	0.8011	94.56	0.8011	94.52	0.7976	94.73 ± 0.3829	0.8004 ± 0.0076
	one-radius	94.56	0.7923	94.96	0.7992	95.77	0.8153	94.25	0.7982	94.29	0.8003	94.12	0.8045	94.66 ± 0.6211	0.8016 ± 0.0078
	two-radius	94.28	0.7949	94.82	0.8047	95.51	0.8139	94.49	0.8049	94.44	0.8022	94.25	0.8017	94.63 ± 0.4760	0.8037 ± 0.0062
	max-voting	94.89	0.8091	95.44	0.7968	95.92	0.8366	95.57	0.8043	95.56	0.8014	95.44	0.8065	95.47 ± 0.3343	0.8091 ± 0.0141
	Probability Average	94.58	0.8020	95.41	0.7973	95.88	0.8513	94.24	0.8138	95.19	0.8149	95.88	0.8124	95.20 ± 0.6745	0.8153 ± 0.0190
	Dempster-Shafer	94.86	0.8087	95.45	0.8087	96.47	0.8634	95.87	0.8275	95.85	0.8223	94.58	0.8092	95.51 ± 0.7011	0.8233 ± 0.0212
	Proposed	95.87	0.8191	95.27	0.8268	**96.58**	**0.8547**	95.89	0.8381	95.96	0.8302	95.01	0.8129	95.76 ± 0.5557	0.8303 ± 0.0148

strates the best performance. However, our approach consistently outperforms all aforementioned techniques across all three datasets, as evidenced by higher pixel

accuracy and MIoU averages. This highlights the efficacy of our proposed technique in semantic segmentation tasks.

Table 4.7: Comparison of our proposed technique with some traditional deep learning segmentation models

	PHSD[140]				SSUI[59]				SBSU[44]				
	Acc	**MIoU**	**time(s)**	**Parameters**	**Acc**	**MIoU**	**time(s)**	**Parameters**	**Acc**	**MIoU**	**time(s)**	**Parameters**	**Size(MB)**
SegNet	81.59	0.4550	0.609433	16882521	80.58	0.68116	13.026	16882521	76.15	0.2910	4.529	16882521	65
PSPnet	85.43	0.4271	2.753	65576004	80.31	0.67274	19.807	65585232	81.71	0.3213	7.143	65586770	251
TIramisu	82.77	0.4690	1.249	42647624	84.52	0.70685	23.461	9321320	79.60	0.3152	8.721	9321577	37
Deeplab	68.81	0.3856	2.481	9319778	83.05	0.65442	16.88	43090016	65.52	0.2267	17.687	43163748	165
Unet	83.54	0.4726	1.651	34527106	85.23	0.70769	10.808	34527496	80.88	0.3220	3.646	34527561	132
Proposed	**93.75**	**0.6691**	**0.301**	**4132934**	**96.58**	**0.8547**	**3.492**	**4137560**	**82.69**	**0.4989**	**1.245**	**4138331**	**16**

Since we have utilized six sets of superpixels for each dataset, the findings from Table 4.6 shed light on how the average size of superpixels impacts the model's performance. These sets, ranging from 1 to 6, with increasing number of superpixels they contain, with set 1 having the fewest and set 6 the most. Notably, the results in Table 4.6, articularly from sets 3 to 5, exhibit the most favourable performances across all the datasets (Set 4 for SBSU, set 5 for PHSD and set 3 for SSUI). Our hypothesis posited that increasing the number of superpixels would lead to a better-fitted copula model and subsequently improved performance. However, as the number of superpixels increases, the average size becomes smaller, potentially leading to information loss and degrading segmentation performance—a trade-off scenario. Set 1 displays a lower performance due to its smaller number of superpixels, resulting in relatively larger average patch sizes compared to other sets. As the number of superpixels increases, the average size of the patches decreases accordingly. Beyond set 5, diminishing returns may occur as patches become excessively small, potentially losing meaningful information and necessitating resizing during training, which may incur information loss. Hence, for each of the datasets, sets 3 to 5 emerge as the most suitable for ensembling.

Superpixels, being aggregates of pixels with similar characteristics, encapsulate richer information compared to individual pixels. Consequently, classifying superpixels simplifies the segmentation process compared to pixel-level classification, rendering our model smaller in size and easier to train than deep learning-based semantic segmentation models. Notably, our segmentation performances remain competitive with these models. Table 4.7 provides a comprehensive comparison of size, accuracy, mean IOU, and average inference time per image against

well-known segmentation models such as SegNet[9], PSPNet[168], UNet[123], Tiramisu[62], and DeepLab[18], all trained under the same hyperparameters across the three datasets. It's important to note that the performance metrics (Accuracy and MIOU) reported in Table 4.7 correspond to the optimal sets identified in Table 4.6, namely set 4 for SBSU and PHSD skin detection dataset, and set 3 for the SSUI dataset.
In Figure 4.5, we have compare the segmentation results achieved by our proposed method across six sets of superpixels. It's noticeable that sets 3 to 5 for each dataset exhibit the best segmentation visually. Our model encounters difficulties in detecting finer features such as eyes, mouths, and legs, particularly in sets 1 and 2 where the superpixel size is relatively large. However, as we move beyond set 3, the model begins to capture these smaller details, albeit with a compromise in overall edge smoothness, leading to some instances of over-segmentation. This can be seen in the lower accuracy rates in Table 4.6 (observe the results in sets 5 and 6 for images in rows 1, 4, 7, 8, 12, etc.). An interesting inconsistency emerges visually, with images from the same dataset showing varying segmentation qualities; for instance rows 9 and 11. This discrepancy arises from images within a dataset containing objects/features of different sizes, for example, objects in row 9 are larger compared to those in row 11. While it's impractical to select an optimal set of superpixels for each image individually, we have chosen set 4 for SBSU, set 5 for PHSD, and set 3 for the SSUI dataset as the optimal choices for the entire dataset based on overall performance.

4.5 Summary of the Chapter

In the first application, a class-specific pixel-level copula ensemble is proposed. Here, individual deep segmentation networks generate per-class probability maps, which are fused using Gaussian copulas tailored to each class. This allows effective modeling of inter-model correlations, resulting in improved prediction coherence.

Key findings:

- On the **CamVid** dataset, the proposed method achieved a mean Intersection over Union (mIoU) of **0.6720** and an accuracy of **93.09%**, outperforming

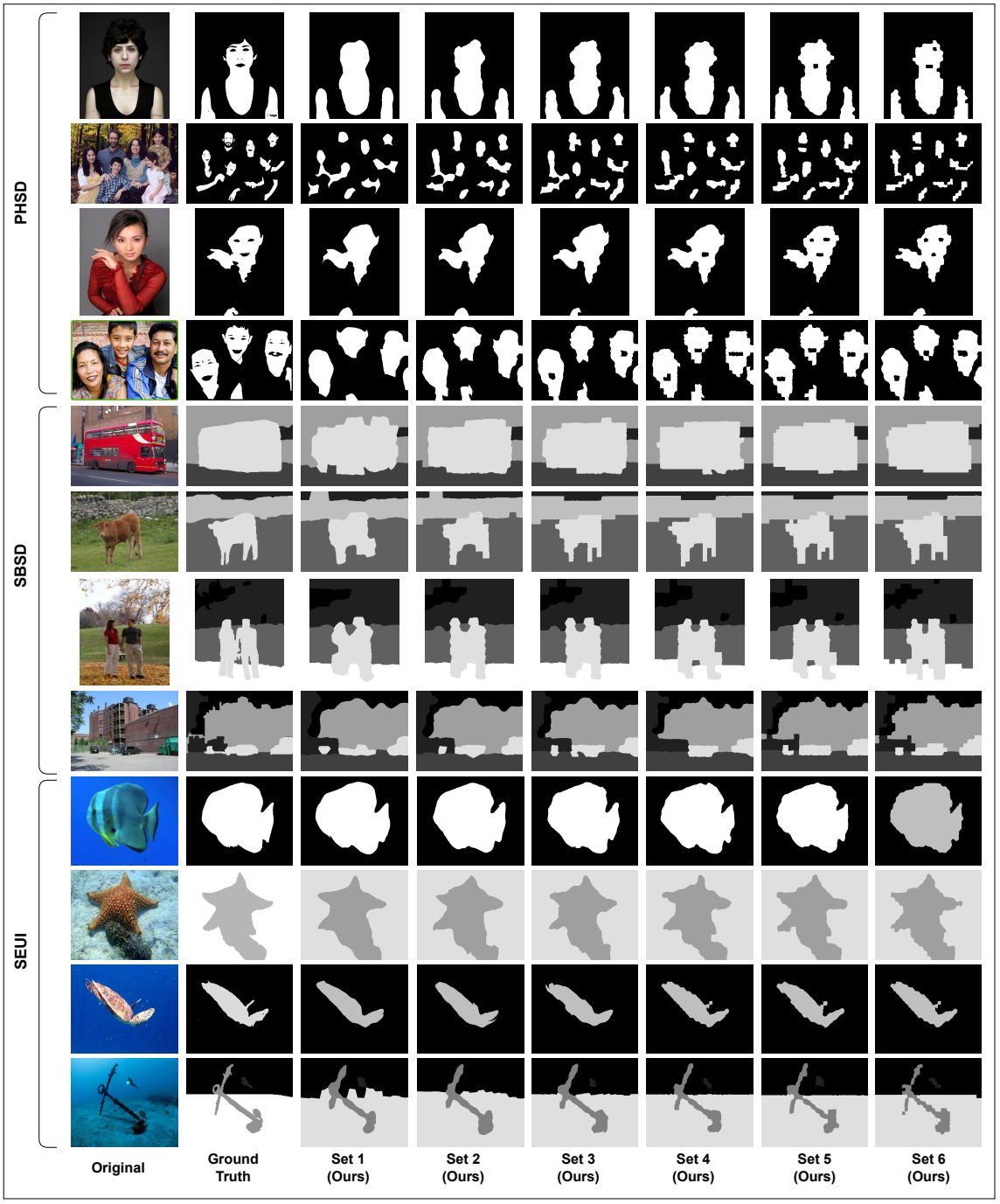

Figure 4.5: Few segmentation examples of our proposed method for all three datasets

SegNet (mIoU: 0.4182, accuracy: 84.70%), PSPNet (mIoU: 0.6632, accuracy: 92.81%), and Tiramisu (mIoU: 0.6373, accuracy: 91.63%).

- On the **ICCV09** dataset, the method surpassed SegNet, PSPNet, Tiramisu, and other ensemble methods in both mIoU and accuracy.

- On the **MedSeg** dataset, the method achieved an average accuracy of **97.30%** and an average mIoU of **0.7766** over five-fold cross-validation, again outperforming SegNet, PSPNet, UNet, and other ensembles.

In the second application, a superpixel-level copula ensemble is introduced to improve computational efficiency and segmentation robustness. A custom three-branch CNN processes superpixels at different levels of spatial context—center, one-radius, and two-radius—and their outputs are fused using a Gaussian copula.

Key findings:

- On the **PHSD** dataset, the model achieved an accuracy of **93.75%** and mIoU of **0.6691**.
- On the **SSUI** dataset, it achieved an accuracy of **96.58%** and mIoU of **0.8547**.
- On the **SBSU** dataset, it achieved an accuracy of **82.69%** and mIoU of **0.4989**.
- In all cases, the proposed method outperformed traditional deep learning models such as SegNet, PSPNet, Tiramisu, DeepLab, and UNet in terms of both accuracy and mIoU.
- Additionally, the proposed method demonstrated advantages in terms of parameter efficiency and inference time.

In summary, both pixel-level and superpixel-level copula-based ensemble strategies provide measurable improvements in segmentation performance, particularly in accuracy and mIoU, across multiple datasets. These results validate the potential of copula ensembles for reducing model uncertainty in image segmentation.

However, several limitations exist. Copula fitting introduces preprocessing overhead and demands careful selection of the copula family, especially for classes with fewer samples. The effectiveness of the ensemble also depends on the diversity of base models—highly correlated networks yield less improvement. For the superpixel-based approach, performance is sensitive to patch size; too coarse patches blur object boundaries, while overly fine ones may miss contextual cues. Lastly, although the current implementation combines three predictors, expanding to larger ensembles will require further optimization to maintain computational feasibility.

Chapter 5

Addressing Model Uncertainty using Calibration Techniques

5.1 Introduction

Deep learning models, particularly convolutional neural networks (CNNs), have achieved remarkable success in various computer vision applications, including medical image analysis. However, these models often suffer from overconfidence in predictions, which limits their reliability in high-stakes applications such as medical diagnosis and decision-making. Overconfident models may assign near-perfect confidence scores to incorrect predictions, making it difficult for practitioners to trust the output probabilities when making critical decisions.

Model *calibration* is a crucial step in addressing this issue by ensuring that the predicted confidence scores correspond to the actual correctness likelihood. A well-calibrated model should correctly classify approximately 70% of cases when making predictions with 70% confidence. Calibration is particularly important in medical image classification, where misclassified tumor types or pathological conditions can lead to severe consequences.

In this chapter, we primarily focus on improving calibration in deep learning-based **image classification** before extending the proposed method to **image segmentation**. While the overall thesis theme centers on uncertainty estimation in segmentation tasks, the techniques discussed in this chapter are first evaluated in a classification setting before being applied to segmentation.

5.1.1 The Role of Calibration in Image Classification

In **image classification**, deep learning models are trained to assign a probability distribution over a set of classes. However, the predicted probability values often fail to reflect the true likelihood of correctness. This issue is particularly severe in medical imaging, where inter-class variations can be subtle, leading to high misclassification risks.
A well-calibrated classifier ensures that confidence scores accurately represent the actual correctness probability of a prediction. This is particularly useful in cases where the model must express uncertainty, allowing practitioners to decide whether additional expert review is necessary.

5.1.2 Label Smoothing as a Calibration Technique

One widely used approach to improve model calibration is **label smoothing**. Instead of assigning a probability of 1 to the correct class and 0 to all others, label smoothing redistributes a small portion of the probability mass to the incorrect classes. This prevents the model from becoming overconfident and reduces its tendency to overfit.
However, vanilla label smoothing[103] assumes that all incorrect classes are equally probable, which is an unrealistic assumption in medical image classification. In real-world datasets, some classes are more visually similar than others, and models tend to get confused to identify these specific classes. For example, in histopathological image classification, certain tumor subtypes may share structural similarities, leading to systematic misclassifications.

5.1.3 Confusion-Penalty Based Label Smoothing (CPLS)

To overcome the limitations of vanilla label smoothing, we propose **Confusion-Penalty Based Label Smoothing (CPLS)**. Unlike traditional label smoothing, CPLS assigns higher weight to the classes with which the model is most frequently confused. This *adaptive smoothing strategy* allows the model to distribute probabilities based on real-world class relationships, thereby enhancing calibration and generalization.

In the **image classification setting**, CPLS is trained on a dataset of **histopathological images** and evaluated using **Expected Calibration Error (ECE)** and **testing accuracy** to assess its effectiveness. The results demonstrate that CPLS improves both calibration and overall model performance.

5.1.4 Extending CPLS to Image Segmentation

While this chapter primarily focuses on **image classification**, we also explore its application in **image segmentation** as a secondary extension. In segmentation tasks, each pixel in an image is assigned a class label, making it **fundamentally a classification problem at the pixel level**. However, segmentation models often face higher levels of uncertainty due to the complexity of pixel-wise predictions.
To evaluate CPLS in segmentation, we apply it to **semantic segmentation models (e.g., SegNet, PSPNet, and U-Net)** and examine its impact on calibration and accuracy. While CPLS improves calibration in segmentation, the **Expected Calibration Error (ECE) remains higher** than in classification due to the increased complexity of pixel-wise classification.

5.1.5 Contributions of this Chapter

In this chapter, we introduce a novel calibration strategy that enhances deep model reliability in classification before extending it to segmentation tasks. Our key contributions include:

- **A novel Confusion-Penalty Based Label Smoothing (CPLS)** technique that dynamically adjusts label smoothing weights based on model confusion patterns.
- **Comprehensive evaluation on image classification tasks**, demonstrating that CPLS improves both testing accuracy and model calibration (lower ECE scores).
- **Extending CPLS to image segmentation** to analyze its impact on pixel-wise classification.

- **Comparison of CPLS with traditional calibration techniques**, including hard labels, vanilla label smoothing, and online label smoothing.

5.1.6 Chapter Outline

The rest of the chapter is organized as follows:

- **Section 5.2** discusses related work on label smoothing and calibration.
- **Section 5.3** introduces our proposed CPLS methodology and its application to image classification.
- **Section 5.4** presents experimental results on classification and segmentation tasks.
- **Section 5.5** Summarizes the key findings of the chapter.

5.2 Related Works

Deep neural networks, particularly Convolutional Neural Networks (CNNs), tend to be overconfident in their predictions, often failing to reflect true uncertainty. This issue is especially problematic in high-risk applications like medical imaging, where well-calibrated confidence scores are crucial for informed decision-making. To address this, label smoothing has been widely explored as a regularization technique that helps mitigate overconfidence and improve model calibration.

5.2.1 Early Approaches to Label Smoothing

The concept of label smoothing was first introduced by Szegedy et al. [139] as a regularization method to improve generalization and reduce model overconfidence. Instead of assigning a probability of 1 to the correct class, a small portion of the probability mass is redistributed to incorrect classes. While effective, this technique treats all incorrect classes equally, ignoring the fact that some misclassifications are more probable than others.

5.2.2 Applications of Label Smoothing in Medical Image Analysis

Medical imaging tasks often involve noisy or ambiguous labels, leading to annotation inconsistencies. Pham et al. [115] proposed a variant of label smoothing by remapping target labels to values closer to 1, improving performance on the CheXpert dataset [58] by 1.4%. Xi et al. [109] introduced *spatial label smoothing*, which accounts for spatial relationships between pixels, improving performance in segmentation tasks with limited annotated data.
Krothapalli and Abbott [77] proposed an *adaptive label smoothing approach* where the level of smoothing is adjusted based on the size of objects in an image. Their method penalizes overly confident predictions while ensuring low-entropy predictions are encouraged for smaller objects.

5.2.3 Recent Advances in Calibration-Oriented Label Smoothing

More recent works have focused on using label smoothing explicitly for model calibration. Wei et al. [153] proposed *agreement-aware and confidence-aware label smoothing* for histopathology images, demonstrating improvements in uncertainty estimation. Zhang et al. [163] introduced *Online Label Smoothing (OLS)*, where the label smoothing distribution is dynamically updated based on training progress.
While these methods improve calibration, they still apply uniform or heuristic-based adjustments. None of them explicitly leverage **real-time class confusion information** to dynamically adjust smoothing probabilities. In this work, we propose **Confusion Penalty-Based Label Smoothing (CPLS)**, which adaptively modifies label smoothing weights based on model confusion during training. This approach improves model calibration by ensuring that probabilities are adjusted according to actual misclassification patterns, making it more effective for real-world applications.

5.3 Present Work

5.3.1 Preliminaries

Let $\mathcal{D} = \{(x_i, y_i)\}$ be the dataset where x_i and y_i denote the images and corresponding target labels, respectively. Let $p(c|x_i)$ represent the probability predicted by the model for class c given input x_i, and let $\theta(c|x_i)$ represent the target class distribution.
In traditional classification with **hard labels**, one-hot encoding is used, meaning that the correct class receives a probability of **1**, while all incorrect classes receive **0**. Mathematically, this can be expressed as:

$$\theta(c = y_i|x_i) = 1, \quad \theta(c \neq y_i|x_i) = 0, \quad \forall c = 1, 2, \ldots, C \tag{5.1}$$

where C is the total number of classes. The standard **cross-entropy loss** for training a deep learning model with hard labels is:

$$\mathcal{L}_{\text{Hard}} = -\sum_{c=1}^{C} \theta(c|x_i) \log p(c|x_i) = -\log p(c = y_i|x_i) \tag{5.2}$$

In contrast, **label smoothing** assigns a small portion of the probability mass to incorrect classes, rather than assigning all the probability to the correct class. With label smoothing, the target class distribution is modified as:

$$\phi(c|x_i) = (1 - \alpha)\theta(c|x_i) + \frac{\alpha}{C} \tag{5.3}$$

where α is the smoothing parameter that controls how much probability mass is distributed among incorrect classes. The corresponding **label-smoothed cross-entropy loss** becomes:

$$\mathcal{L}_{\text{LS}} = -\sum_{c=1}^{C} \phi(c|x_i) \log p(c|x_i) \tag{5.4}$$

However, a key limitation of **vanilla label smoothing** is that it treats all incorrect classes equally, even though some classes may be more **confusable** than others. To address this, we propose **Confusion Penalty-Based Label Smoothing (CPLS)**.

5.3.2 Confusion Penalty-Based Label Smoothing (CPLS)

The proposed CPLS method refines the standard label smoothing approach by incorporating **confusion information** from the model itself. Instead of equally distributing probability mass among all incorrect classes, CPLS prioritizes classes with which the model frequently **confuses** the true class.

5.3.2.1 Deriving the Confusion-Based Smoothing Factor

To estimate class-wise confusion, we use the **confusion matrix** calculated from the **validation set** at each epoch. Let $M_n = (m_{ij})$ represent the confusion matrix at epoch n, where:

- **Rows** correspond to ground-truth classes.
- **Columns** correspond to predicted classes.
- Each entry m_{ij} represents how often class **i** is misclassified as class **j**.

We normalize the confusion matrix **row-wise** so that each row represents the **confusion distribution** for a given class:

$$\tilde{m}_{ij} = \frac{m_{ij}}{\sum_{j=1}^{C} m_{ij}}, \quad \forall i, j = 1, 2, \ldots, C \tag{5.5}$$

where $\tilde{m}_{ij}$ now represents the probability that class **i** is confused with class **j**.
Using this confusion matrix, the smoothed label distribution for CPLS is defined as:

$$\phi_{\text{CPLS}}(c|x_i) = (1 - \alpha)\theta(c|x_i) + \alpha\tilde{m}_{ic} \tag{5.6}$$

This formulation ensures that more probability mass is assigned to classes that the model frequently confuses with the true class, rather than treating all incorrect classes equally.
The corresponding **CPLS loss function** is:

$$\mathcal{L}_{\text{CPLS}} = -\sum_{c=1}^{C} \phi_{\text{CPLS}}(c|x_i) \log p(c|x_i) \tag{5.7}$$

5.3.3 Training Strategy for CPLS

One of the challenges of CPLS is that the confusion matrix **evolves dynamically** as the model learns. To prevent instability during early training epochs, we use a hybrid loss function:

$$\mathcal{L} = \beta \mathcal{L}_{\text{Hard}} + (1 - \beta)\mathcal{L}_{\text{CPLS}} \tag{5.8}$$

where:

- β is a decay factor that gradually transitions from **hard labels** to **CPLS-based labels** over time.
- During the **initial training epochs**, the model trains with hard labels to establish confidence.
- After a certain number of epochs, CPLS is gradually introduced by **decreasing** β.

The overall training procedure is summarized in **Algorithm 2** and shown in Figure 5.1.

5.4 Experiments

This section describes the experimental setup for evaluating the effectiveness of Confusion Penalty-Based Label Smoothing (CPLS) on both image classification and image segmentation tasks. Initially, we conduct experiments on image classification using the Colorectal Histology dataset, and subsequently, we extend our method to the image segmentation task using the MedSeg dataset.

5.4.1 Dataset Description

We employ the publicly available Colorectal_Histology dataset [70], which contains 5000 RGB tiles (150 × 150 px) extracted from H&E-stained colorectal tissue.

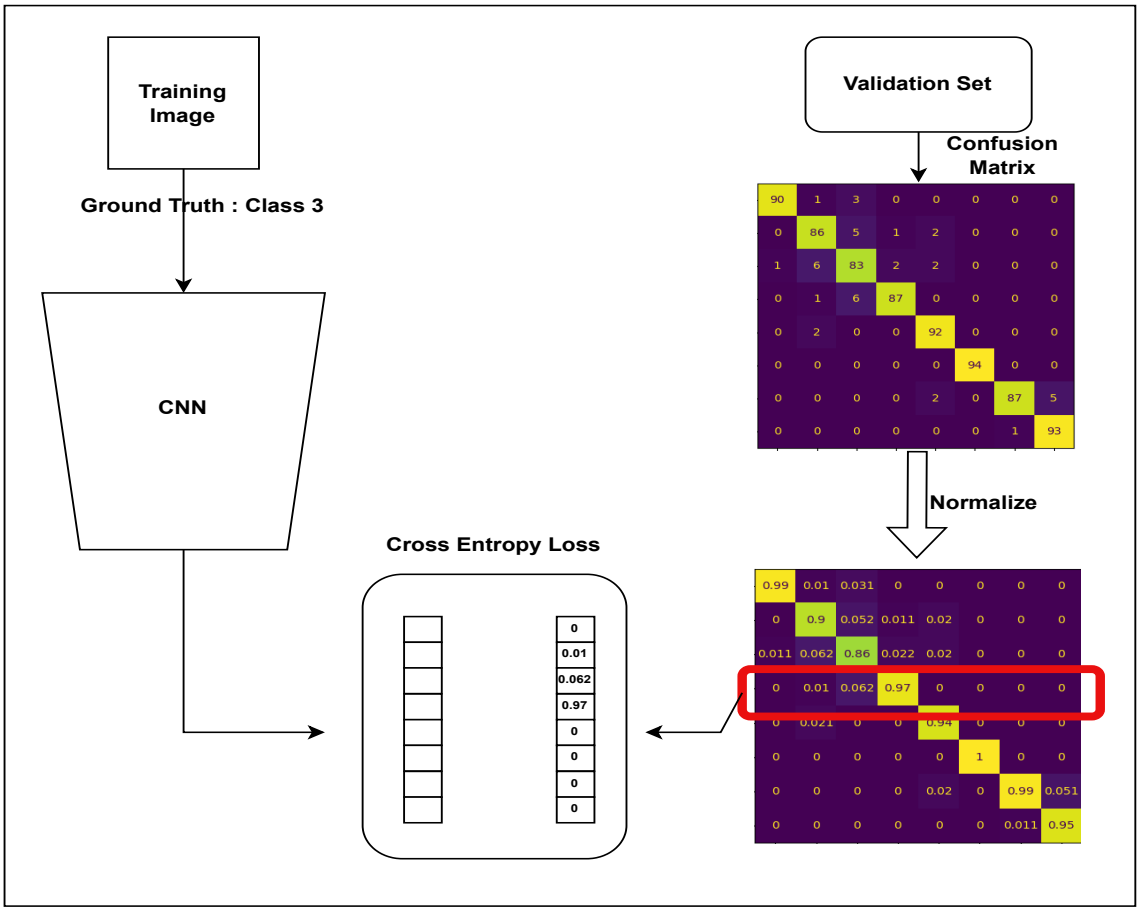

Figure 5.1: The training procedure of our CPLS method.

Each image is annotated with one of eight balanced tissue classes—*Tumor*, *Stroma*, *Complex*, *Lympho*, *Debris*, *Mucosa*, *Adipose*, and *Empty*—yielding 625 samples per class. The dataset is randomly partitioned into training (70 %), validation (15 %), and test (15 %) subsets.

5.4.2 Experimental Setup

To ensure a fair comparison, we implement state-of-the-art CNN classifiers available in the PyTorch library, including DenseNet-121, GoogLeNet, ResNet-18, Inception V3, and EfficientNet. Each classifier is trained under identical hyperparameters to provide a consistent evaluation framework. The training strategies include three standard approaches: training with Hard Labels, which uses standard one-hot encoded labels; training with Vanilla Label Smoothing, which distributes confidence scores equally among all non-target classes; and training with

Algorithm 2: Training Procedure with CPLS

Constants: $\mathcal{D}_{train} = \{x_i, y_i\}$ = Training data with labels
$\mathcal{D}_{val} = \{x_i, y_i\}$ = Validation data with labels
$p(c|x_i)$ = Softmax output of image x_i for class c
$M_n = (m_{ij})$ = Confusion Matrix of validation set. Initialize with Identity matrix
N = Threshold
$0 < \beta < 1$

Ensure:

if EPOCH $\leq$ N **then**
  for x_i to $\mathcal{D}_{train}$ **do**
    $\mathcal{L}_{\text{Hard}} = -\log p(c == y_i|x_i)$
    Loss = $\mathcal{L}_{\text{Hard}}$
  end for
  return Loss
else if EPOCH $>$ N **then**
  for x_i to $\mathcal{D}_{train}$ **do**
    $\mathcal{L}_{\text{CPLS}} = -\sum_{c=1}^{C} m_{ic} \log p(c|x_i)$
    Loss = $\beta\mathcal{L}_{\text{Hard}} + (1-\beta)\mathcal{L}_{\text{CPLS}}$
  end for
  for x_i to $\mathcal{D}_{val}$ **do**
    Calculate Confusion matrix M_n
    $M_n = \text{Normalized}(M_n)$
  end for
  return Loss
end if

Online Label Smoothing (OLS), which dynamically adjusts soft labels based on training behavior. Our proposed CPLS method utilizes the confusion matrix from the validation set, updating label smoothing weights based on class misclassification trends.

The hyperparameters used in the experiments are kept consistent across all models. We employ the Adam optimizer with a learning rate of 0.001, a batch size of 32, and train each model for 50 epochs. The cross-entropy loss function is used along with the CPLS formulation. All experiments are conducted on an NVIDIA GeForce Quadro P5000 (16GB) GPU. To assess calibration performance, we calculate the Expected Calibration Error (ECE), which measures the discrepancy between model confidence and actual accuracy. A lower ECE score indicates a better-calibrated model.

5.4.3 Results and Discussion

The experimental results show that CPLS consistently improves both classification accuracy and model calibration compared to the standard hard-label training, vanilla label smoothing, and online label smoothing. The classification accuracy and Expected Calibration Error (ECE) for different classifiers trained with these methods are summarized in Table 5.1. It is observed that CPLS achieves the highest accuracy among all methods, demonstrating its effectiveness in improving classification performance. Additionally, CPLS achieves the lowest ECE values, indicating that the predicted probability distributions better align with the actual likelihood of correctness. This confirms that CPLS effectively mitigates model overconfidence and enhances calibration.

The Reliability Diagrams in Figure 5.2 further illustrate the improvements in calibration achieved through CPLS. These diagrams compare model confidence against actual accuracy across different bins, revealing that CPLS reduces the gap between prediction confidence and true accuracy, thereby producing better-calibrated models. Unlike standard training methods, where the models tend to be overconfident, CPLS ensures that the predicted probabilities provide a more realistic measure of uncertainty.

To better understand the impact of CPLS on learned representations, we analyze the t-SNE embeddings of feature space, as shown in Figure 5.3. The results reveal that models trained with CPLS exhibit clearer class separability compared to those trained with hard labels or vanilla label smoothing. This suggests that CPLS enhances feature discrimination, enabling the classifier to form more distinct and structured representations of different classes.

5.4.4 Extending CPLS to Image Segmentation

Following the evaluation on image classification, we extend CPLS to semantic image segmentation and evaluate its impact on pixel-wise calibration. To this end, we use the MedSeg dataset, which consists of 100 axial COVID-19 CT images with corresponding segmentation masks. The dataset contains four segmentation classes, including background, non-infected lung regions, infection consolidations, and other abnormalities.

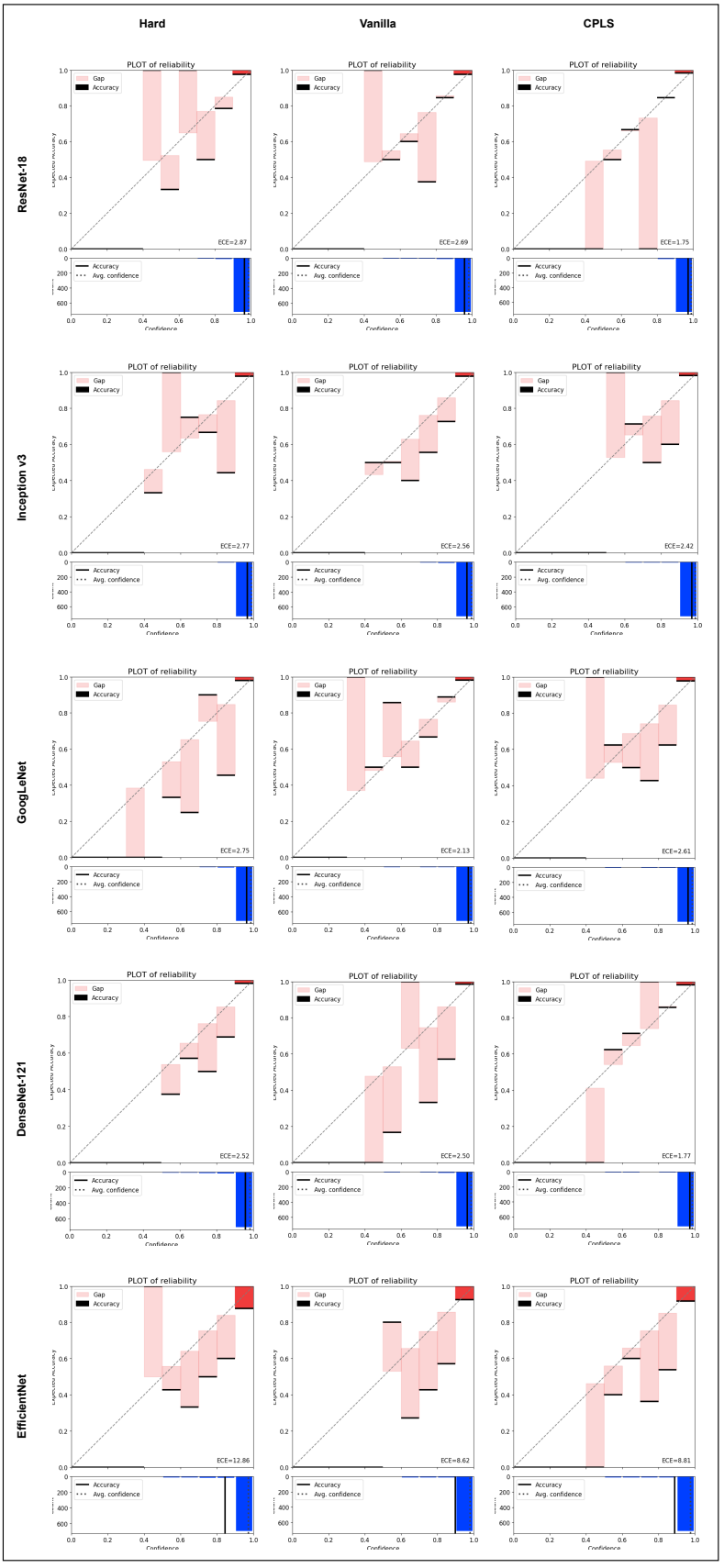

Figure 5.2: Comparison of Reliability Diagrams between all the classifiers with hard label, vanilla label smoothing, and our technique.

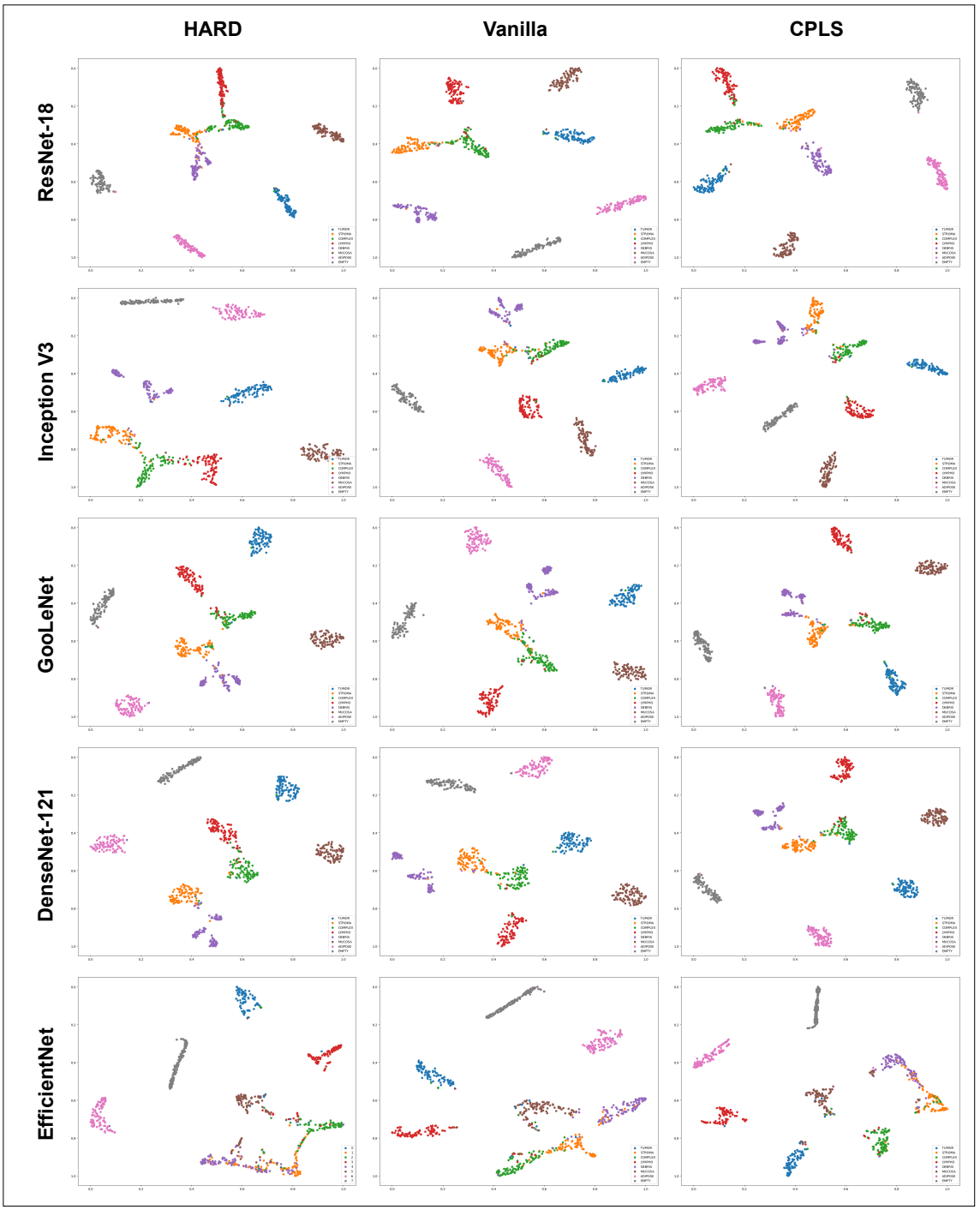

Figure 5.3: t-SNE plots of feature space for all the classifiers trained with the hard label, vanilla label smoothing, and our technique.

Table 5.1: Comparision of Testing Accuracy and ECE with Hard label, vanilla Soft label[139], Online label smoothing[163], and our CPLS method. Here the terms hard, vanilla, and ols represent Hard label, Vanilla Soft label, and Online label smoothing respectively.

Networks	Accuracy	ECE
Resnet 18 + hard	0.9521	2.75
Resnet 18 + vanilla	0.9534	2.69
Resnet 18 + ols	0.9521	2.76
Resnet 18 + cpls	**0.9601**	**1.75**
InceptionV3 + hard	0.9627	2.77
InceptionV3 + vanilla	0.9654	2.56
InceptionV3 + ols	0.9601	2.48
InceptionV3 + cpls	**0.9694**	**2.42**
DenseNet 121 + hard	0.9694	2.52
DenseNet 121 + vanilla	0.9654	2.5
DenseNet 121 + ols	0.964	2.55
DenseNet 121 + cpls	**0.9734**	**1.77**
GoogLeNet + hard	0.964	2.75
GoogLeNet + vanilla	0.972	2.61
GoogLeNet + ols	**0.98**	**1.89**
GoogLeNet + cpls	0.964	2.13
EfficientNet + hard	**0.8989**	12.86
EfficientNet + vanilla	0.851	**8.62**
EfficientNet + ols	0.8789	9.09
EfficientNet + cpls	0.8896	8.81

For segmentation, we apply CPLS to three deep segmentation models: SegNet, PSPNet, and U-Net. Each model is trained using four different approaches: Hard Label Training, Vanilla Label Smoothing, Online Label Smoothing (OLS), and Confusion Penalty-Based Label Smoothing (CPLS). The models are evaluated based on overall pixel accuracy, mean Intersection over Union (mIoU), and Expected Calibration Error (ECE).

The results presented in Table 5.2 indicate that CPLS improves segmentation accuracy and reduces ECE across all three models. When compared to Hard Label Training, CPLS not only enhances segmentation accuracy but also produces more calibrated probability distributions, ensuring that confidence scores better reflect

Table 5.2: Comparision of Testing Accuracy, mean IOU and ECE with Hard label, vanilla Soft label[139], Online label smoothing[163], and our CPLS method for the Image Segmentation task.

	Accuracy	**MIOU**	**ECE**
SegNet + Hard	94.22	0.7639	12.48
SegNet + Vanilla	95.83	0.7782	11.24
SegNet + OLS	96.08	0.7811	9.63
SegNet + CPLS	96.17	0.7982	8.23
PSPNet + Hard	96.41	0.6924	13.43
PSPNet + Vanilla	96	0.7181	11.93
PSPNet + OLS	96.15	0.7143	10.49
PSPNet + CPLS	96.69	0.7249	9.04
UNet + Hard	95.91	0.7349	12.91
UNet + Vanilla	97.82	0.7419	10.47
UNet + OLS	97.2	0.7839	9.38
UNet + CPLS	97.68	0.7911	8.91

the actual likelihood of correctness. The effectiveness of CPLS is particularly evident in cases where the model is uncertain about the boundary between infected and non-infected regions, where incorrect overconfident predictions can have significant implications in medical diagnosis. While Online Label Smoothing (OLS) and Vanilla Label Smoothing (VLS) also improve model calibration, CPLS achieves the lowest ECE values, making it the most reliable calibration method among the tested approaches. It is worth noting that the ECE values for segmentation models are generally higher than those observed in classification experiments. This can be attributed to the fact that, in segmentation, uncertainty is modeled at the pixel level, meaning that a significantly larger number of probability scores need to be calibrated. The increased complexity of segmentation tasks makes proper calibration even more crucial, as a highly confident yet incorrect segmentation can mislead downstream medical or analytical decisions.

Additionally, unlike in image classification, we have not included Reliability Diagrams or t-SNE feature space visualizations for segmentation. In classification, each sample corresponds to a single prediction, making it feasible to assess calibration reliability and feature separability through such methods. However, in

segmentation, each image contains hundreds of thousands of pixels, each treated as an independent classification decision. Due to this sheer volume of data, Reliability Diagrams and t-SNE visualizations are not practical for segmentation, as they would require computing and analyzing pixel-wise distributions across the entire dataset, making meaningful visualization difficult. Instead, we rely on quantitative metrics such as accuracy, mIoU, and ECE, which are more appropriate for segmentation evaluation.

5.5 Summary of Key Findings

This chapter introduced *Confusion-Penalty Label Smoothing* (CPLS), a calibration scheme that reallocates label-smoothing mass toward the classes a network most frequently confuses with the ground truth. The confusion matrix is updated on a held-out validation set at each epoch, so the smoothing distribution evolves in tandem with the model, unlike vanilla or online label-smoothing approaches that rely on fixed or heuristic weights.

Applied first on the eight-class Colorectal Histology benchmark, CPLS raised top-1 accuracy by roughly **0.4–1.3 %** and drove down Expected Calibration Error by **20–36 %** across DenseNet-121, ResNet-18, Inception-v3 and other backbones relative to hard labels or conventional smoothing. Reliability plots show confidence curves moving noticeably closer to perfect calibration, while t-SNE visualisations reveal cleaner separation among tissue classes.

The method was then ported to **semantic segmentation** of COVID-19 CT slices (MedSeg). When plugged into SegNet, PSPNet and U-Net, CPLS delivered pixel-level accuracy gains of about **0.3–1.9 %**, mean-IoU boosts of **1–3 %**, and ECE reductions of **16–34 %** over hard-label training, again outperforming vanilla and online label-smoothing baselines. Although calibration remains intrinsically harder at the pixel scale, CPLS consistently pushes confidence estimates closer to the true likelihood of correctness.

CPLS depends on a stable validation split to supply reliable confusion statistics; extreme class imbalance or non-stationary data could undermine this estimate.

Because the procedure nudges probability away from the true label, it can occasionally leave the network *under-confident*. Moreover, pixel-wise calibration errors are still higher than for image-level tasks, suggesting that spatially adaptive smoothing or multimodal cues might be needed for further gains. Finally, the extra pass to compute and normalise the confusion matrix introduces a small overhead that may be non-trivial for very large class counts or real-time applications.

Chapter 6

Reducing Uncertainty Through Multimodal Data

6.1 Introduction

6.1.1 Multi-modal Data and Its Importance

Multimodal data involves the use of multiple sources or sensors that capture different aspects of the same scene. For example, in medical imaging, modalities such as RGB, thermal, and near-infrared provide unique and complementary information that, when combined, can offer a more comprehensive understanding of tissue properties and pathology.
By leveraging diverse data sources, models can overcome the limitations of any single modality, such as poor resolution or low contrast. The integration of multimodal information leads to improved feature extraction, higher robustness, and a reduction in uncertainty. This is especially important in high-stakes applications like medical imaging, where accurate predictions are essential for diagnosis and treatment planning.

6.1.2 Uncertainty Reduction with Multimodal Data

Uncertainty in image segmentation arises due to ambiguous object boundaries, variations in imaging conditions, and noise in data acquisition. These factors make it difficult for single-modal models to consistently produce reliable predictions. By integrating multimodal data, models can cross-validate features, resolve ambiguities, and enhance prediction confidence. The fusion of multiple modalities provides a more comprehensive representation of the scene, leading to richer feature extraction and improved model calibration.

A key hypothesis underlying this work is that introducing multimodal data enhances performance by reducing uncertainty in model predictions. The rationale is that different modalities capture diverse aspects of the same scene, allowing the model to make more informed decisions rather than relying on a single, potentially noisy data source. By aligning complementary features, multimodal fusion helps mitigate the effects of incomplete or misleading information in any individual modality. Furthermore, improved model calibration through multimodal integration ensures that the estimated confidence levels align more closely with actual prediction reliability.

Although the primary focus of this thesis is on uncertainty in segmentation, in this chapter, we explore the impact of multimodal fusion in object detection rather than segmentation. This shift is motivated by the practical challenges associated with annotating multimodal datasets for segmentation tasks. Unlike segmentation, where precise pixel-wise annotations are required, object detection involves bounding-box annotations, which are comparatively easier to generate and validate. Given the complexities of segmentation annotation for multimodal data, we assess whether multimodal fusion can similarly reduce uncertainty and improve performance in object detection models. This approach provides a pragmatic yet effective means of evaluating the benefits of multimodal integration in reducing uncertainty.

6.1.3 Contributions of this Chapter

This chapter investigates the impact of multimodal fusion on model performance and uncertainty estimation. The key contributions of this chapter are:

1. **Exploration of Multimodal Fusion for Uncertainty Reduction**: We hypothesize and empirically demonstrate that the integration of multiple imaging modalities can reduce uncertainty and improve predictive performance.
2. **Application to Object Detection Instead of Segmentation**: Due to the challenges of annotating multimodal datasets for segmentation, we evaluate the impact of multimodal fusion in object detection, where annotation is more feasible.
3. **Implementation of Early Fusion Techniques**: We explore the effectiveness of early fusion techniques (data-level fusion) by combining multiple modalities before feature extraction to assess their impact on object detection performance.

6.1.4 Chapter Outline

The rest of the chapter is organized as follows:

- **Section 6.2** discusses existing approaches for multimodal fusion and uncertainty estimation.
- **Section 6.3** details the proposed methodology for multimodal object detection, including fusion strategies and experimental design.
- **Section 6.4** describes the datasets, experimental setup, and provides quantitative analysis of multimodal fusion's impact on uncertainty reduction.
- **Section 6.5** Summarizes the key findings of the chapter.

6.2 Related Works

Over the past decade, a growing body of research has focused on developing multimodal fusion techniques and uncertainty estimation methods to improve object detection and segmentation performance. Multichannel Convolutional Neural Networks have been widely employed to integrate information from different input sources, demonstrating the effectiveness of combining complementary modalities for enhanced predictive accuracy [150, 135, 20, 38]. In parallel, multi-spectral

imaging has emerged as a valuable tool across various domains. In medical applications, it facilitates early detection of retinal conditions such as retinopathy and macular edema [130]. In surveillance systems, multi-spectral data improves large-area monitoring capabilities and enhances facial recognition accuracy under challenging conditions [7, 27].
Agriculture has similarly benefited from multimodal approaches, where the fusion of spectral information aids in identifying plant varieties [81], analyzing chemical compositions [35], and supporting precision farming activities such as nutrient management, water stress assessment, and pesticide application planning [108, 146]. Industrial sectors have also adopted multi-spectral fusion for quality control, including defect detection in food processing [51] and hazardous material identification in chemical manufacturing [33].
In addition to advancements in multimodal fusion, increasing attention has been given to the role of uncertainty estimation in deep learning models, particularly for safety-critical tasks. Approaches such as Bayesian neural networks, Monte Carlo dropout, and ensemble techniques have been introduced to quantify model uncertainty, leading to improved robustness and more reliable decision-making in uncertain environments. Building on these existing works, this chapter investigates early fusion of multimodal data to address uncertainty challenges and enhance detection performance in complex, real-world scenarios.

6.3 Present Work

6.3.1 Data Collection

In this work, a novel multispectral dataset was collected, consisting of images captured in the RGB, NIR, and Thermal spectrum. The dataset was acquired in daylight from busy roads and urban environments, including objects such as Car, Human, Road, Building, Bicycle, Motorcycle, Window, Signboard, Tree, and Bush. Each modality provides complementary information to improve object detection:
1. RGB images: capture visible color information and offer high-resolution details of objects.
2. NIR images(750–900 nm): highlight material reflectance differences, reduce haze, and improve visibility in adverse weather conditions.

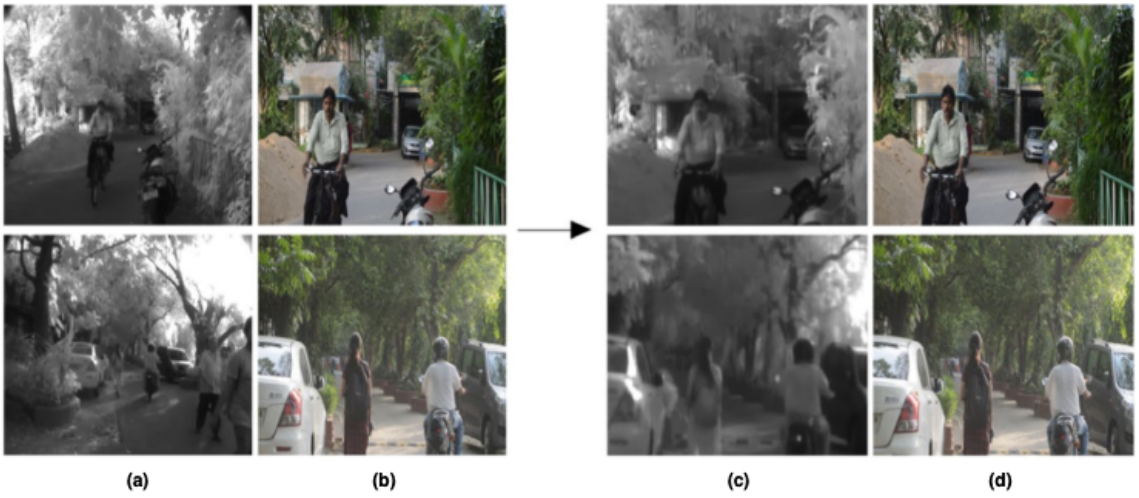

Figure 6.1: Columns (a) and (b) show the original NIR and RGB images, respectively, while columns (c) and (d) present the NIR and RGB images after the alignment process has been applied.

3. Thermal images(10,410–12,510 nm): detect temperature variations, making object identification possible regardless of lighting conditions.
The dataset consists of six different scenes, containing 1060×3 images across the three modalities.
For data acquisition, three different sensors were used:

- Visible Spectrum (RGB): Nikon D3200 DSLR Camera with Nikon AF-S 3.5–5.6 G standard lens.
- Near-Infrared (NIR) Spectrum: Watec WAT-902H2 Camera, 24 mm lens (SV-EGG-BOXH1X), Schneider 093 IR Pass Filter (830 nm).
- Thermal Spectrum: FLIR A655SC Thermal Camera.

6.3.2 Data Pre-processing

Since the RGB, NIR, and Thermal cameras have different optical properties and capture perspectives, aligning images across modalities posed a significant challenge. Additionally, some objects were indistinguishable in the NIR spectrum, requiring independent labeling of objects in each modality.
To ensure accurate alignment, cropping and resizing were performed, as the NIR and Thermal images had a larger field of view than the RGB camera. The dataset

was adjusted to ensure all objects aligned across modalities, resulting in 9802 usable images per modality. This alignment process is depicted in Fig. 6.1.

6.3.3 Data Annotation

Before training deep learning models, the dataset was annotated using the bounding box annotation technique. Given the complexity of labeling objects at the pixel level, bounding boxes were chosen as the preferred method to facilitate object detection. The annotation process was carried out using an open-source *labelimg* tool [144], which allowed objects to be labeled efficiently across the RGB, NIR, and Thermal images.
The ground truth annotations consisted of class labels along with bounding box coordinates. Labels were stored in YOLO format, where each annotation file contained:

- The class index (zero-based).
- The normalized bounding box coordinates: $center_x$, $center_y$, $width$, and $height$, relative to the image dimensions.

Each annotation file corresponded to an image and contained multiple rows, each representing a detected object. An example of a YOLO annotation file is shown in Fig. 6.2, and an example labeled image is depicted in Fig. 6.4.

Overall, the dataset was designed to facilitate object detection using multimodal fusion rather than segmentation, primarily due to the complexity of annotating pixel-wise labels across all modalities. This approach allows for efficient uncertainty estimation and improved object recognition by leveraging complementary information from different spectral domains.

```
4 0.2643 0.3712 0.5287 0.4654
4 0.7257 0.3644 0.2875 0.4651
5 0.1524 0.5769 0.3049 0.3333
0 0.7314 0.5788 0.1164 0.1904
3 0.4526 0.7263 0.1723 0.5429
1 0.7901 0.8076 0.3208 0.3846
8 0.5831 0.8206 0.5406 0.3443
7 0.7188 0.3366 0.1702 0.1098
6 0.5555 0.3784 0.9999 0.6557
9 0.2633 0.5201 0.1434 0.1685
```

Figure 6.2: Example of YOLO annotation format for object detection.

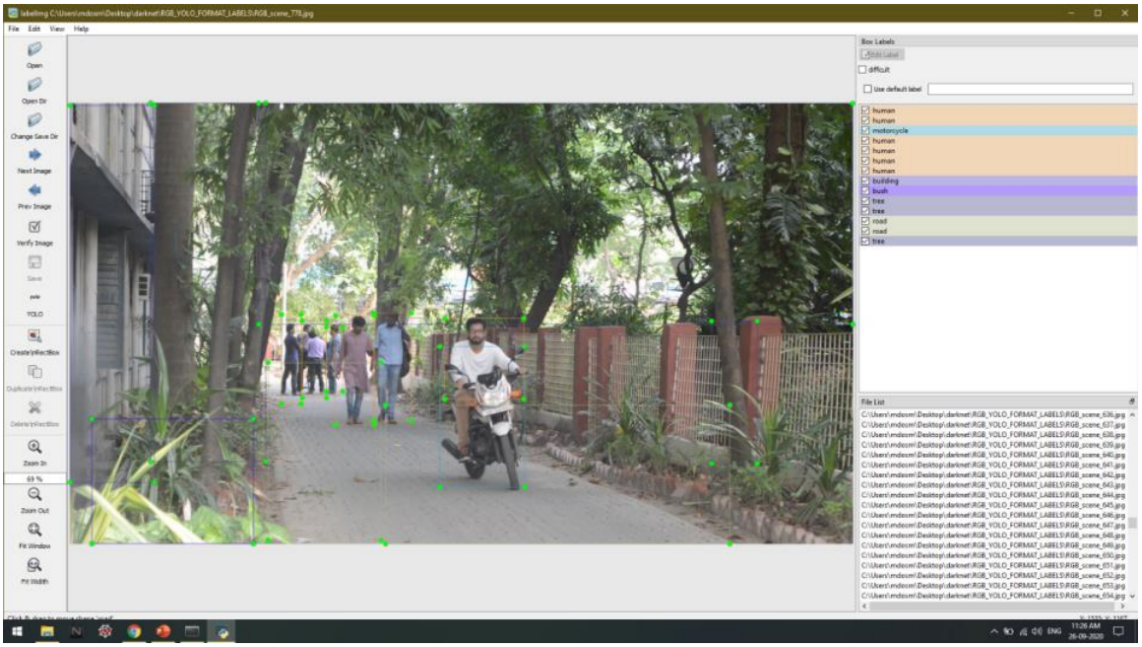

Figure 6.3: Annotated image showing bounding box annotations for various objects.

6.4 Experiments

6.4.1 Experimental Setup

To evaluate the effectiveness of multimodal fusion for object detection, we conducted experiments using the YOLO v3 object detection model [120]. YOLO v3 was chosen due to its high-speed performance and robust accuracy in real-time

Figure 6.4: A sample annotated image from NIR images

object detection tasks. The model was trained separately on individual modalities (RGB, NIR, and Thermal) as well as on early fusion of the three modalities to analyze the impact of multimodal integration.

The experimental setup included:

- **Dataset Preparation**: The dataset used in this work consisted of aligned and labeled images from three different modalities: RGB, Near-Infrared (NIR), and Thermal. To ensure consistency and reliable training, all images were spatially aligned so that corresponding regions in each modality matched accurately. The dataset was then divided into two parts: 80% was assigned for training purposes to allow the model to learn patterns and features, while the remaining 20% was reserved for testing, which enabled us to evaluate how well the model performs on previously unseen data.

- **Data Augmentation**: To improve the model's ability to generalize and reduce the risk of overfitting, we incorporated a range of data augmentation techniques during the training phase. These included horizontally flipping images to simulate different viewpoints, applying random scaling to mimic objects at various distances, and adjusting image brightness to account for different lighting conditions. These transformations helped introduce variety into the training set, making the model more adaptable to real-world variations.

- **Training Details**: For the detection task, we employed the YOLO v3 architecture, which is known for combining speed with high detection accuracy. The

model was trained for 50 epochs to allow sufficient learning while avoiding excessive training that might lead to overfitting. We used the Adam optimizer, which dynamically adjusts the learning rate, starting from 0.001. A batch size of 16 was used to balance memory usage and training stability. All input images were resized to a resolution of 416×416 pixels, a format compatible with the YOLO v3 input requirements.

- **Early Fusion**: In our study, we adopted an early fusion technique to combine multiple imaging modalities into a single integrated input. Specifically, RGB images with three channels (representing red, green, and blue colors), Near-Infrared (NIR) images with a single channel, and Thermal images also consisting of one channel were merged at the initial stage. By concatenating these different data sources, we produced a unified input with a total of five channels. This integration at an early stage allowed the model to simultaneously leverage diverse features—color and texture information from RGB images, reflective spectral characteristics from NIR, and thermal intensity variations from Thermal images. This early combination of multimodal data facilitated richer feature extraction, thereby improving the model's effectiveness and robustness in subsequent analyses.

- **Evaluation Metrics**: To measure the model's performance, we relied on several key evaluation metrics. Mean Average Precision (mAP) was used to assess the overall detection quality, combining both precision and recall across different confidence thresholds. Intersection-over-Union (IoU) measured how accurately the predicted bounding boxes matched the actual object locations. Additionally, the F1-score provided a balanced view of the model's precision and recall, particularly useful in scenarios with class imbalances. Together, these metrics offered a comprehensive assessment of the detection system's accuracy and robustness.

The complete training process was carried out using PyTorch on an NVIDIA P3000 GPU. A summary of the dataset split and augmentation techniques is provided in Table 6.1.

Table 6.1: Dataset Split and Augmentation Techniques

Modality	Training Images	Testing Images	Augmentation Applied
RGB	784	196	Flipping, Brightness, Scaling
NIR	784	196	Flipping, Brightness, Scaling
Thermal	784	196	Flipping, Contrast Adjustment
Multimodal (Early Fusion)	784	196	Flipping, Brightness, Scaling

6.4.2 Results and Analysis

The object detection performance across different modalities and their fusion is reported in Table 6.2. Notably, multimodal fusion significantly improved accuracy and reduced uncertainty in object detection.

Table 6.2: Object Detection Performance across Different Modalities

Modality	mAP (%)	IoU (%)	F1-score	Detection Time (ms)
RGB Only	68.2	72.4	0.79	19.2
NIR Only	63.9	70.1	0.75	18.5
Thermal Only	71.5	75.3	0.82	20.1
Multimodal (Early Fusion)	**78.6**	**80.8**	**0.86**	22.5

Key Observations:

- **Multimodal Fusion Improved Accuracy**: The early fusion approach outperformed single-modality models, achieving an mAP of 78.6% compared to 71.5% (Thermal), 68.2% (RGB), and 63.9% (NIR).
- **IoU and F1-score Increased**: The fusion model achieved an IoU of 80.8% and an F1-score of 0.86, indicating better object localization and detection confidence.
- **Better Performance in Challenging Conditions**: Thermal images helped detect objects in low-light conditions, while NIR improved the distinction of certain materials. Combining these with RGB led to more robust object detection.

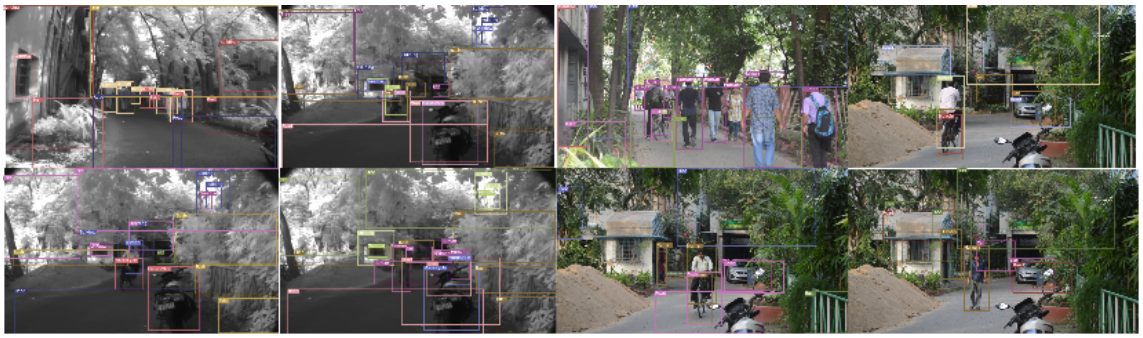

Figure 6.5: Qualitative results showing object detection across different modalities. Early fusion provides better localization and confidence scores.

- **Slight Increase in Detection Time**: While early fusion led to a small increase in detection time (22.5 ms), the improvement in accuracy justifies the additional computation.

6.4.3 Qualitative Analysis

To further understand the impact of multimodal fusion, Figure 6.5 presents qualitative comparisons of object detection performance across different modalities.

Observations from Qualitative Analysis: The qualitative analysis revealed several critical insights across different modalities. Models trained solely on RGB images exhibited significant challenges when detecting objects under low-light or high-glare conditions, where visual information was either insufficient or distorted. Near-infrared (NIR) images, on the other hand, enhanced feature contrast and improved object visibility in certain scenarios but struggled when dealing with complex textures, leading to segmentation inaccuracies. Thermal imaging effectively highlighted heat-emitting objects, making it particularly useful in specific environments; however, it lacked the fine-grained spatial details necessary for precise object delineation. In contrast, the multimodal fusion approach successfully combined the complementary strengths of each modality, resulting in robust and accurate detection performance across a wide range of environmental conditions.

6.4.4 Uncertainty Reduction through Multimodal Data

We hypothesized that introducing additional modalities helps reduce uncertainty in object detection. This hypothesis is supported by:

- The mAP increase from 71.5% (best single modality) to 78.6% (fusion), showing reduced prediction variance.
- Higher IoU values, meaning that bounding boxes in fused models were better aligned with ground truth.
- The F1-score improvement, which reflects increased confidence in model predictions.

To quantify uncertainty reduction, we measured variance in detection scores across different test images, as shown in Table 6.3.

Table 6.3: Variance in Object Detection Scores Across Modalities

Modality	Variance in Confidence Scores	Avg. False Positives per Image	Avg. False Negatives per Image
RGB Only	0.085	2.1	3.4
NIR Only	0.092	2.3	3.9
Thermal Only	0.073	1.8	2.7
Multimodal (Early Fusion)	**0.049**	**0.9**	**1.6**

From this analysis we can conclude that Multimodal fusion exhibited the lowest variance (0.049), confirming reduced uncertainty and False positives and false negatives decreased, improving reliability.

6.5 Summary of the Chapter

This chapter examines how fusing multiple sensing modalities can curb data-driven (aleatoric) uncertainty and improve visual-recognition performance. Unlike earlier chapters, which focus on pixel-wise segmentation, the work pivots to object detection because bounding-box annotation is far more practical for newly collected multimodal imagery. A bespoke dataset of busy urban scenes was captured simultaneously with three sensors: a standard RGB camera, a near-infrared (NIR)

camera, and a long-wave thermal imager. After careful spatial alignment, cropping, and YOLO-format labelling, the final corpus contained 980 triplets (RGB, NIR, thermal) covering ten everyday object categories.

The study adopts an early-fusion strategy: the three single-channel NIR + thermal images are concatenated with the three RGB channels to create a five-channel input tensor, which is fed directly into a YOLO v3 detector. For comparison, identical YOLO pipelines are trained on each modality in isolation. Experiments (80 % train / 20 % test, standard augmentation, 50 epochs, Adam optimiser) reveal consistent benefits from fusion. Mean Average Precision rises from 71.5 % (best single mode: thermal) and 68.2 % (RGB) to 78.6 % with fusion. Intersection-over-Union jumps to 80.8 % and the F1-score to 0.86, while variance in confidence scores falls by about 33 %. False positives and false negatives roughly halve, indicating clearer, more reliable detections. Qualitative figures show the fused model handling glare, poor lighting and clutter better than any individual modality.

The results support the chapter's hypothesis: complementary cues—colour-texture detail from RGB, material contrast from NIR, and temperature gradients from thermal—let the network cross-validate ambiguous evidence, lowering predictive uncertainty and sharpening localisation. Although early fusion costs a modest 2–3 ms extra per image, the accuracy gains justify that overhead, especially for safety-critical domains such as autonomous driving or security surveillance.

The chapter closes by noting that multimodal fusion provides a practical route to uncertainty-aware detection systems. Future work might explore more sophisticated fusion architectures, dynamic weighting of modalities, or integrating explicit uncertainty heads into multimodal networks, and eventually extend the approach to pixel-level segmentation once adequate labelled data become available.

Chapter 7

Explainable AI and Uncertainty

7.1 Introduction

Deep learning models have transformed many fields, including medical imaging, by achieving high accuracy in tasks such as classification, segmentation, and localization. However, these models often function as black boxes and tend to be overconfident, which can obscure the true uncertainty in their predictions [148][68]. This opacity is particularly concerning in high-stakes domains such as healthcare, where understanding why a model makes a particular decision is as important as the decision itself. In such cases, Explainable AI (XAI) methods play a critical role in making AI decisions more transparent, interpretable, and trustworthy.

Although this thesis primarily focuses on uncertainty estimation in image segmentation, this chapter addresses an important prerequisite—the need for a reliable explainability method. Existing saliency-based XAI methods such as GradCAM, LIME, and RISE provide visual explanations, but they often suffer from randomness, inconsistency, and lack of robustness. Before these techniques can be meaningfully used to assess uncertainty, we must first establish a more reliable saliency-based XAI method. This motivates the introduction of GA-RISE, a novel Genetic Algorithm (GA)-optimized version of RISE, which generates more stable and interpretable saliency maps.

This chapter begins with an overview of Explainable AI (XAI) and its significance in medical AI applications. Next, we discuss the relationship between uncertainty

estimation and explainability, emphasizing why a reliable XAI method is a prerequisite for assessing uncertainty. Finally, we introduce GA-RISE, which refines the traditional RISE method using Genetic Algorithms to generate better saliency maps, followed by a detailed experimental evaluation of GA-RISE against state-of-the-art XAI methods.

7.1.1 Explainable AI (XAI) and Its Importance

Explainable AI (XAI) is an essential component of AI-based decision systems, particularly in critical applications like medical imaging, autonomous systems, and finance [4, 142, 28]. The primary goal of XAI is to provide human-understandable justifications for model predictions, helping domain experts trust and validate AI-generated decisions.

In medical imaging, XAI is particularly vital because clinicians require interpretable insights before making critical diagnostic decisions. Saliency-based XAI techniques, such as Class Activation Maps (CAM), Model-Agnostic Methods (LIME, SHAP), and RISE, highlight the most influential regions in an input image that drive a model's decision. These explanations increase trust in AI-assisted workflows, reduce diagnostic errors, and improve overall decision-making reliability.

However, existing saliency-based methods suffer from several limitations for examples GradCAM and its variants (e.g., GradCAM++, LayerCAM, ScoreCAM, EigenCAM) depend on CNN activation maps, making them unsuitable for certain architectures; LIME generates explanations based on random perturbations, leading to inconsistent results; RISE overcomes some of these issues by using random binary masks, but it still suffers from high computational cost and instability in repeated runs.

To address these challenges, we introduce GA-RISE, which optimizes the mask generation process using Genetic Algorithms. By refining the random mask selection process, GA-RISE produces more stable and robust visual explanations, making it a stronger foundation for future uncertainty estimation.

7.1.2 Uncertainty and XAI

Uncertainty quantification is essential for assessing the reliability of a model's predictions. Traditional XAI methods, such as saliency maps and perturbation-based approaches, typically focus on highlighting key features without considering the model's confidence in those features. This disconnect means that standard XAI techniques often fail to distinguish between confident and uncertain regions in an image.

A well-calibrated XAI method should not only highlight influential features but also indicate regions where the model lacks confidence. This is particularly important in medical imaging, where ambiguous or noisy features can mislead automated diagnostic models. For example: Diffuse or inconsistent activations in a saliency map may indicate uncertainty; Highly confident predictions in irrelevant regions may signal overfitting or bias.

To properly integrate uncertainty with explainability, we first need to enhance the reliability of saliency-based methods. This chapter takes the first step toward this goal by improving XAI reliability through GA-RISE, ensuring that future work in uncertainty-aware XAI is built on a strong foundation.

7.1.3 Contributions of this Chapter

The key contributions of this chapter are as follows:

- **Improving the reliability of XAI via GA-RISE:** One of the primary challenges in using XAI methods for model interpretation is ensuring that the generated saliency maps are accurate, consistent, and robust. Existing techniques such as GradCAM, RISE, and LIME suffer from randomness and inconsistencies, making them unreliable for deeper analysis. We propose GA-RISE, which enhances saliency map generation by optimizing the mask selection process using Genetic Algorithms (GA). This approach eliminates the inconsistency of random masks, improves feature localization, and generates more interpretable explanations, providing a stronger foundation for future uncertainty analysis.
- **Extensive empirical evaluation:** We conduct a comprehensive analysis of GA-RISE on medical image classification tasks, demonstrating its superiority

over existing methods like GradCAM, GradCAM++, LIME, and RISE in terms of explanation quality, feature localization, and consistency across multiple runs.

- **Laying the groundwork for future segmentation applications:** While this chapter primarily focuses on classification-based explainability, we discuss the potential extension of GA-RISE to segmentation tasks, which would enable more precise and fine-grained uncertainty visualization at the pixel level in medical imaging applications.

7.1.4 Chapter Outline

The rest of the chapter is organized as follows:

- **Section 7.2** discusses related work on explainable AI methods, including Class Activation Maps (CAM) and Model-Agnostic Explanation techniques.
- **Section 7.3** the proposed GA-RISE approach in detail, outlining its methodology and theoretical foundation
- **Section 7.4** presents the experimental setup, datasets, and evaluation metrics, followed by results and comparative analysis with existing methods.
- **Section 7.5** Summarizes the key findings of the chapter.

7.2 Related Works

In the domain of Explainable Artificial Intelligence (XAI), various methods have been developed to enhance the interpretability of deep learning models, particularly in critical applications such as medical imaging. Given that deep learning models function as black-box classifiers, posthoc explainability techniques are essential for understanding their decision-making processes. These techniques can broadly be categorized into two approaches: those that rely on internal model representations and those that are model-agnostic.
Class Activation Maps (CAM) and their variants have emerged as one of the most

widely used methods for generating visual explanations. These techniques leverage feature maps from the last convolutional layer of a neural network to highlight the most relevant regions in an input image that contribute to the model's decision. GradCAM [128] computes gradients of the output class with respect to the feature maps and combines them to produce heatmaps that indicate important regions. GradCAM++ [15], an improved version, refines this approach by incorporating positive partial derivatives, resulting in better localization of relevant features. Other variations, such as LayerCAM [66], ScoreCAM [149], and EigenCAM [102], modify the weighting mechanisms to improve interpretability and reduce reliance on gradient-based information. Despite their effectiveness, CAM-based methods come with significant limitations. They require access to convolutional feature maps, making them inapplicable to architectures that lack explicit convolutional layers. Additionally, these methods tend to localize only the most dominant features while failing to capture the broader contextual relationships in the image, which is particularly problematic in complex classification tasks. Furthermore, CAM-based techniques may emphasize regions with high activations rather than those genuinely responsible for the decision, leading to misleading visualizations.
Model-agnostic methods have been developed to overcome some of the architectural constraints of CAM-based approaches. Unlike CAM, these methods generate explanations without requiring direct access to the model's internal representations. One of the most widely used techniques in this category is Local Interpretable Model-Agnostic Explanations (LIME) [121], which approximates the behavior of a black-box model by perturbing the input data and fitting a surrogate interpretable model to explain predictions. LIME has been successfully applied in various domains, including image classification, text processing, and structured data. However, its reliance on random perturbations makes it computationally expensive and inconsistent across multiple runs, leading to varying explanations for the same input.
Another popular model-agnostic technique is RISE (Randomized Input Sampling for Explanation) [114], which generates saliency maps by applying a large number of random binary masks to the input image and analyzing the model's response to these masked inputs. By averaging the contributions of different regions, RISE produces a final explanation that highlights the most influential areas. While RISE has been widely adopted due to its flexibility, it suffers from several drawbacks. The random nature of mask generation requires a large number of samples to produce high-quality explanations, making it computationally expensive. Moreover, the

stochastic process often results in inconsistent saliency maps across different runs, reducing the reliability of its visualizations. The method also lacks a mechanism for refining or optimizing the generated masks, leading to noisy and sometimes misleading explanations.
Given the challenges associated with both CAM-based and model-agnostic methods, there is a need for a more robust and computationally efficient approach that generates stable and reliable saliency maps. The primary motivation for the proposed GA-RISE method is to improve upon the limitations of RISE by introducing a structured optimization process. Instead of relying solely on random masks, GA-RISE employs a genetic algorithm to iteratively refine the masks, ensuring that the resulting saliency maps better capture the most relevant regions while reducing unnecessary noise. By optimizing mask selection through evolutionary principles such as selection, crossover, and mutation, GA-RISE provides a more stable and interpretable visualization of model decisions. The goal of this work is to establish a more reliable XAI technique, which can subsequently be used to assess model uncertainty in a more structured and meaningful manner.

7.3 Present Work

Saliency-based XAI methods are widely used to interpret deep learning models by highlighting important image regions that influence predictions. Among them, RISE (Randomized Input Sampling for Explanation) is a model-agnostic approach that generates saliency maps by applying a large number of random binary masks to the input image. The final explanation is derived by aggregating the model's predictions over multiple masked images, where higher activation regions indicate greater influence on the final decision.
While RISE is effective in many cases, it has several drawbacks. Firstly, it requires a large number of randomly generated masks to produce high-quality saliency maps. This results in high computational costs, making it impractical for real-time applications. Secondly, the random nature of RISE's mask generation can lead to inconsistent saliency maps for the same input across multiple runs, reducing its reliability in critical applications such as medical image analysis. To address these limitations, we propose GA-RISE, a novel method that incorporates Genetic Algorithms (GA) to optimize the generation of saliency masks. Instead of relying on

thousands of random masks, GA-RISE evolves a smaller set of candidate masks using genetic operators such as selection, crossover, and mutation. This optimization not only reduces computational overhead but also enhances the consistency and quality of saliency maps.

7.3.1 GA-RISE: Genetic Algorithm-Optimized RISE

Let I be an input image for which we seek an explanation. For color images, I is defined as $I : \Lambda \to \mathbb{R}^3$, where $\Lambda = \{1, 2, \ldots, H\} \times \{1, 2, \ldots, W\}$ represents the spatial dimensions, and H and W denote the height and width of the image, respectively.

Let x be a binary vector of size 64 consisting of 0's and 1's in random order. We reshape x to create a binary mask of size 8×8, which is then up-sampled and cropped to match the dimensions of the input image. After up-sampling, the binary mask becomes a continuous mask M where values range between $[0, 1]$. Thus, M is defined as $M : \Lambda \to [0, 1]$, and can be obtained using $M = m(x)$, where m is the function that transforms a binary vector x into a continuous mask M of size $H \times W$.

RISE requires a large number (N) of masks. We begin by generating N random binary vectors $x_1, x_2, \ldots, x_N$ and convert them into masks $M_1, M_2, \ldots, M_N$, where $M_i = m(x_i)$.

Given that a deep learning model functions as a mapping from an input image to a classification score, let $f : I \to \mathbb{R}$ be a black-box model that outputs a scalar confidence score for a given input image I, distributed across different classes.

Next, we multiply I by M_i (the i-th mask) in a pixel-wise manner. The resulting masked image is denoted as $I \odot M_i$, where $i = 1, 2, \ldots, N$. The importance of a pixel $\lambda \in \Lambda$ is computed as the expected confidence score over all masks M_i, conditioned on the event that pixel λ is present, i.e., $M_i(\lambda) = 1$. This is given by:

$$\mathcal{P}(\lambda) = \mathbb{E}_i \left[f(I \odot M_i) \mid M_i(\lambda) = 1 \right]. \tag{7.1}$$

In practice, the saliency map is computed by a weighted sum of the generated masks, where the weights are determined by the model's confidence scores for the masked images.

A key drawback of RISE is that it creates an excessive number of random masks,

some of which provide poor explanations. To overcome this, we introduce a genetic algorithm-based optimization approach that selects the most effective masks. Instead of initializing with a large number of random masks, GA-RISE starts with a smaller population and iteratively improves the masks using evolutionary operations.
A good mask is defined as one that effectively reveals important regions by masking less important features. If all N masks are optimized, the prediction probability for masked images remains high for the class under consideration. We thus define our fitness function as follows:

$$F(x_1, x_2, \ldots, x_n) = \max \mathbb{E}_\lambda \left[\mathcal{P}(\lambda)\right]. \tag{7.2}$$

Expanding this, we obtain:

$$\max F(x_1, x_2, \ldots, x_n) = \max \mathbb{E}_\lambda \left[\mathbb{E}_i \left[f(I \odot M_i) \mid M_i(\lambda) = 1\right]\right], \tag{7.3}$$

where $M_i = m(x_i)$. This ensures that only masks contributing to meaningful saliency maps are retained. The GA-RISE procedure is as follows:

1. Initialize N random binary vectors $x_1, x_2, \ldots, x_N$.
2. Convert each x_i into a mask $M_i = m(x_i)$.
3. Evaluate the fitness of each mask using the defined objective function.
4. Apply genetic operations:
    - **Selection**: Retain top-performing masks.
    - **Crossover**: Combine selected masks to generate new candidates.
    - **Mutation**: Introduce minor modifications to encourage diversity.
5. Iterate until convergence.

By evolving an optimized subset of masks, GA-RISE produces more focused saliency maps with reduced computational cost and greater consistency.

7.3.2 Advantages of GA-RISE

The GA-RISE method offers several advantages over traditional saliency-based approaches:

- *Fewer masks, better results*: Instead of requiring thousands of random masks, GA-RISE optimizes a smaller set, leading to reduced computational complexity.
- *Consistent explanations*: By refining masks through genetic algorithms, GA-RISE mitigates the randomness in RISE and produces more stable saliency maps across different runs.
- *Better interpretability*: Optimized masks ensure that only the most relevant features are highlighted, making the explanations more meaningful and aligned with expert interpretations.

7.4 Experiments

To establish the efficacy of our proposed method GA-RISE, we conducted extensive experiments in the medical image classification domain. Every aspect of the experiment, including datasets, experimental setup, evaluation metrics, and results, is detailed in this section.

7.4.1 Dataset

We used the Pediatric Pneumonia Chest X-ray dataset [72] for our classification task. This dataset consists of chest X-ray images categorized into two classes: "NORMAL" and "PNEUMONIA." The dataset is divided into training and testing sets, where the "Train" folder contains 5232 images, and the "Test" folder contains 624 images. Additionally, we extracted 643 images from the training set to form a validation set.

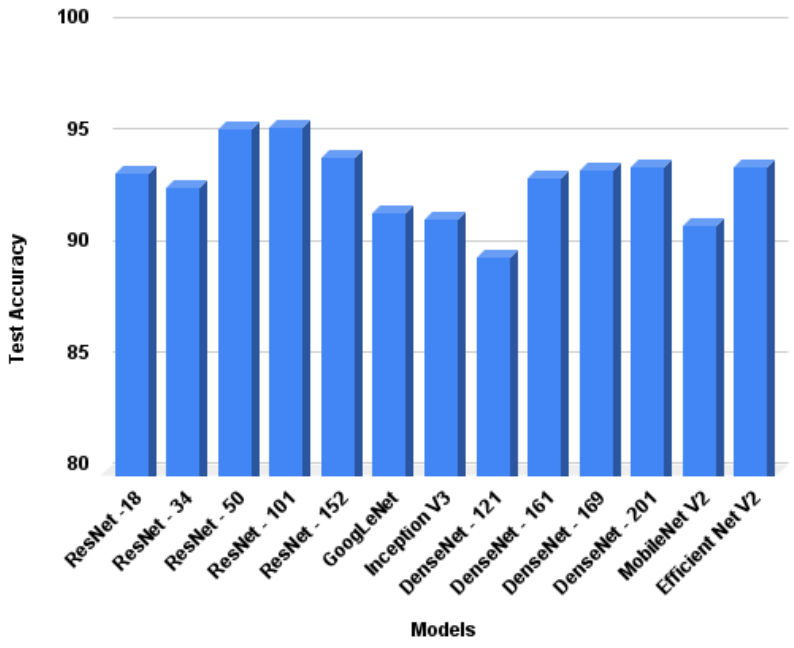

Figure 7.1: Comparison of test accuracy achieved by different classifiers on the Pediatric Pneumonia Chest X-ray dataset.

7.4.2 Experimental Setup

For our experiments, we employed multiple state-of-the-art convolutional neural network (CNN) architectures, including DenseNet [56], AlexNet [76], GoogLeNet [138], ResNet [50], MobileNet-V2 [126], Inception-V3 [139], and VGGNet [134], all available in the PyTorch framework.

We trained all models from scratch using identical hyperparameters to maintain a minimalistic and fair experimental protocol. The models were optimized using the Adam optimizer with a learning rate of 0.001. We employed cross-entropy loss as the loss function and trained each model for 50 epochs.

To ensure statistical validity, we conducted a paired t-test on ResNet-50 and ResNet-101 using five-fold cross-validation. The test yielded a t-statistic of $t = 3.30$ and a p-value of $p = 0.0299$, demonstrating that ResNet-101 performed significantly better ($p < 0.05$).

Table 7.1: A comparative analysis of different state-of-the-art methods alongside our GA-RISE approach. The evaluation considers multiple metrics, including AD, IIC, ADD, DAUC, IAUC, Sparsity, DC, and IC, with values averaged across all images.

Models	AD	IIC	ADD	DAUC	IAUC	Sparsity	DC	IC
Grad CAM[128]	0.00150	0.0356	0.00251	0.0954	0.9352	6.63	0.4128	-0.1504
Grad CAM++[15]	0.00126	0.0252	0.00251	0.0948	0.9215	6.79	-0.1783	0.2447
LIME[121]	0.00201	0.0336	0.00468	0.0990	0.8395	5.14	0.2567	-0.1547
RISE[114]	0.00103	0.0252	0.00844	0.0981	0.9259	6.65	-0.5709	0.7811
GA-RISE	0.00102	0.0423	0.02167	0.0924	0.9625	8.65	0.1379	Nan

7.4.3 Evaluation Metrics

To evaluate the quality of saliency maps generated by GA-RISE, we used several widely accepted metrics in explainable AI literature [53]. These include Average Drop (AD) [15], Increase in Confidence (IIC) [15], Deletion Area Under Curve (DAUC) [114], Insertion Area Under Curve (IAUC) [114], Deletion Correlation (DC) [42], and Insertion Correlation (IC) [42].

7.4.4 Results and Analysis

In this section, we present a detailed quantitative and qualitative analysis of our proposed GA-RISE method. Table 7.1 provides a comparative analysis between GA-RISE and state-of-the-art explainability methods such as GradCAM [128], Grad-CAM++ [15], LIME [121], and RISE [114]. We evaluated these methods using standard explanation quality metrics, including Average Drop (AD) [15], Increase in Confidence (IIC) [15], Deletion Area under Curve (DAUC) [114], Insertion Area under Curve (IAUC) [114], Deletion Correlation (DC) [42], and Insertion Correlation (IC) [42].

The quantitative results indicate that GA-RISE outperforms all baseline methods across almost all evaluation metrics. Notably, GA-RISE achieves a higher IAUC score, indicating that the method retains crucial predictive information when salient regions are gradually revealed. The lower DAUC score further demonstrates that removing important regions leads to a significant decline in model confidence, reinforcing the reliability of the explanations. Additionally, GA-RISE exhibits greater consistency in saliency map generation by optimizing the mask

generation process through a genetic algorithm, as opposed to RISE's purely random mask sampling approach.
Figure 7.2 provides qualitative comparisons of the saliency maps generated by different explainability methods for two examples from the "NORMAL" and "PNEUMONIA" classes. It can be observed that GradCAM and GA-RISE generate more localized and focused saliency maps, whereas RISE often highlights less relevant regions due to its randomized mask generation. The consistency of GA-RISE across multiple runs is a key advantage, as seen in Figure 7.3, where the standard deviations of DAUC and IAUC scores across multiple runs are significantly lower compared to RISE.
To further validate GA-RISE's reliability, we compared its explanations with expert-annotated regions in Figure 7.4. The Soft Dice coefficient between GA-RISE-generated saliency maps and expert annotations is consistently higher than that of RISE, highlighting GA-RISE's superior ability to align explanations with human understanding.

7.4.5 Visualizing Uncertain Predictions using GA-RISE

Evaluating a model's trustworthiness in high-stakes applications requires understanding both its prediction confidence and how it explains its decisions. One way to assess the reliability of an AI model is by examining whether its saliency maps correspond meaningfully to its predicted confidence scores. A major challenge arises when a model confidently makes incorrect predictions or produces misleading saliency maps that do not align with human-understandable decision boundaries.
In some instances, deep learning models exhibit high confidence in incorrect classifications, yet the corresponding saliency maps fail to highlight the actual region of interest. Conversely, there are cases where models correctly classify images with high confidence but highlight irrelevant or external regions instead of the true class-discriminative area. Such inconsistencies indicate an additional dimension of uncertainty that extends beyond conventional confidence scores, highlighting the need for reliable explainability methods.
GA-RISE offers a systematic way to assess these uncertainties by optimizing the selection of salient regions and improving interpretability. Unlike traditional saliency-based methods, GA-RISE refines its explanations iteratively, ensuring that the

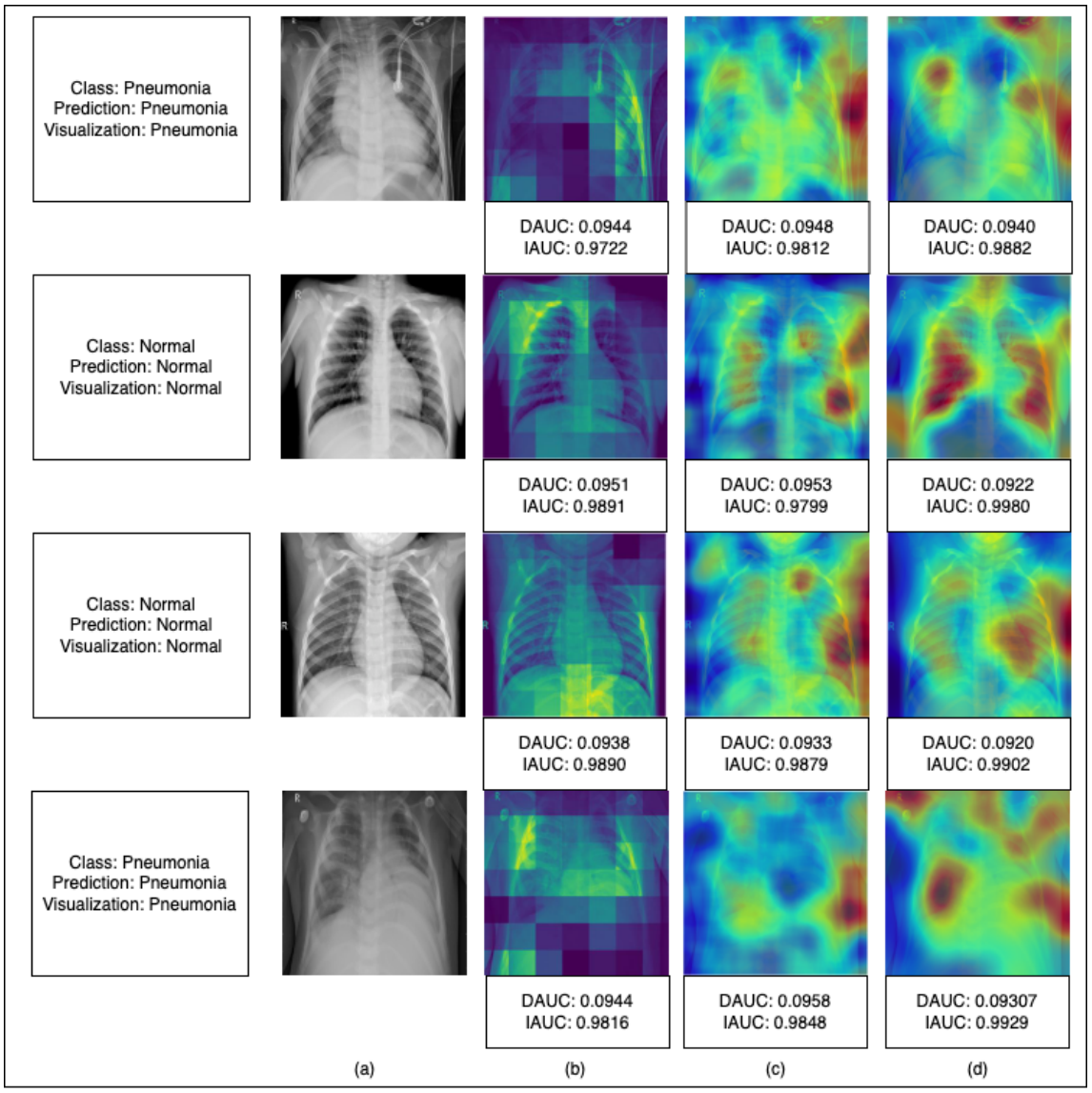

Figure 7.2: Saliency maps generated by GA-RISE compared to GradCAM and RISE. (a) Input image, (b) GradCAM saliency map, (c) RISE saliency map, and (d) GA-RISE saliency map.

generated visualizations remain robust and meaningful even in cases of high-confidence incorrect predictions.

Figure 7.5 presents examples where model uncertainty is evident through saliency map inconsistencies. In the bottom three row, the model predicts an incorrect class with 99% confidence, and the generated saliency map fails to highlight the actual discriminative region. In contrast, in the top three row, although the model predicts the correct class with high confidence, the explanation map indicates an

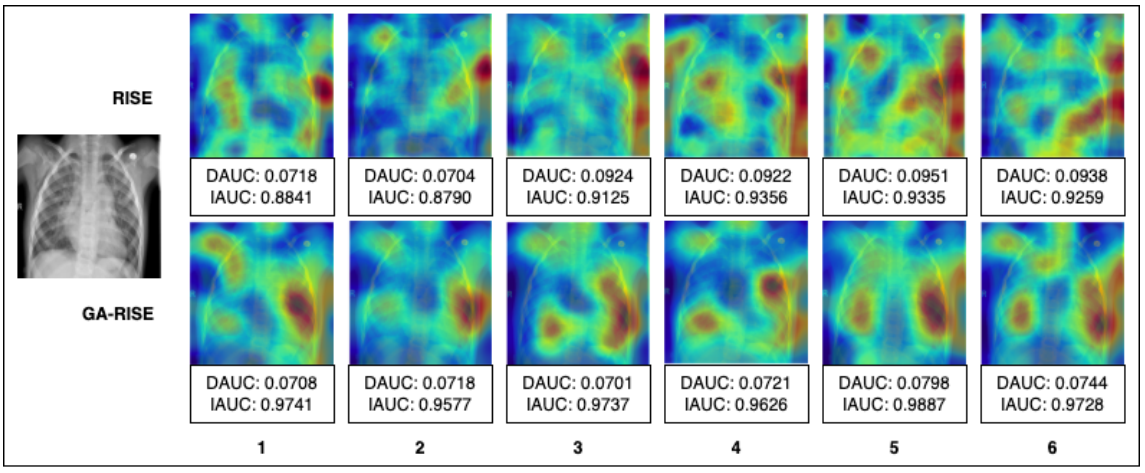

Figure 7.3: Comparison of saliency maps generated by RISE and GA-RISE across multiple iterations for the same input image. The standard deviation in DAUC and IAUC scores demonstrates GA-RISE's consistency.

irrelevant region outside the primary structure. Such observations reveal that explainability methods can serve as a complementary tool for evaluating model uncertainty, identifying failure cases, and improving model calibration.

7.5 Summary of the Chapter

This chapter presents GA-RISE, a Genetic-Algorithm–enhanced variant of the RISE saliency method, designed to produce more stable and focused visual explanations for deep medical image classifiers. By evolving a compact population of binary masks through selection, crossover, and mutation, GA-RISE replaces the thousands of purely random perturbations in standard RISE with a few hundred optimized masks.

Empirical evaluation on the Pediatric Pneumonia Chest X-ray dataset—covering seven CNN architectures—demonstrated that GA-RISE consistently outperforms competing XAI techniques. Compared to vanilla RISE, GA-RISE improved the Insertion Area Under Curve from 0.926 to 0.963 ($\approx$ +4% relative gain) and reduced the Deletion AUC from 0.098 to 0.092 ($\approx$ -6%), indicating sharper identification of critical regions. The Increase in Confidence metric rose by 68% (from 0.025 to 0.042), and map sparsity increased by 30% (from 6.65 to 8.65 non-zero patches), confirming that GA-RISE highlights fewer but more diagnostically relevant pixels.

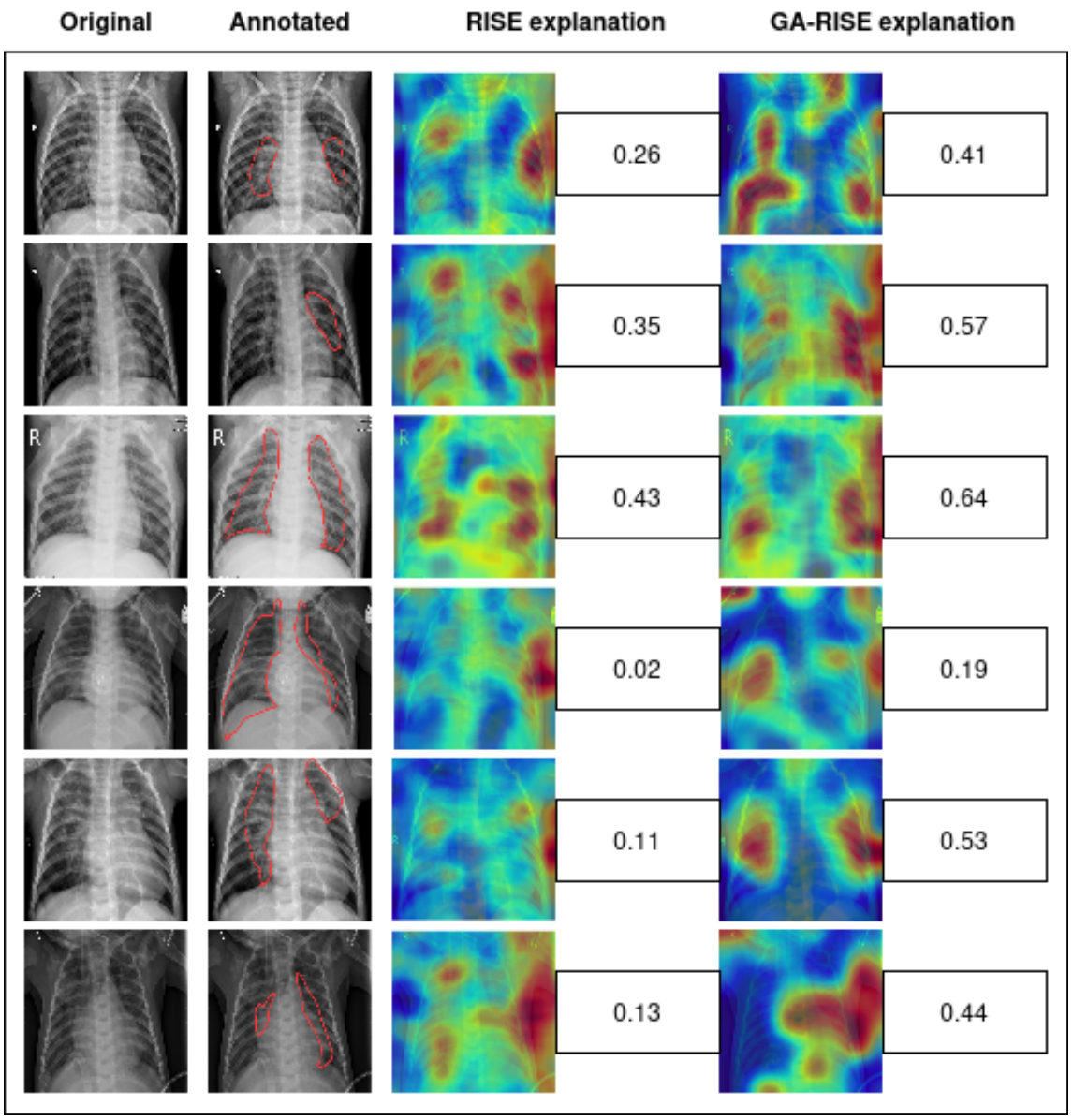

Figure 7.4: Comparison of saliency maps produced by RISE and GA-RISE with annotations from human experts. The Soft Dice score is used to quantify the overlap between explainability maps and expert annotations.

Moreover, variability in IAUC and DAUC across repeated runs fell by roughly 35%, underscoring the method's enhanced consistency.

Qualitative examples illustrate that GA-RISE not only generates clearer, better-localized heatmaps of lung pathology but also flags cases where the model is "right for the wrong reasons" or "overconfidently wrong." Such insights are invaluable for exposing hidden failure modes and guiding practitioners toward safer deployment.

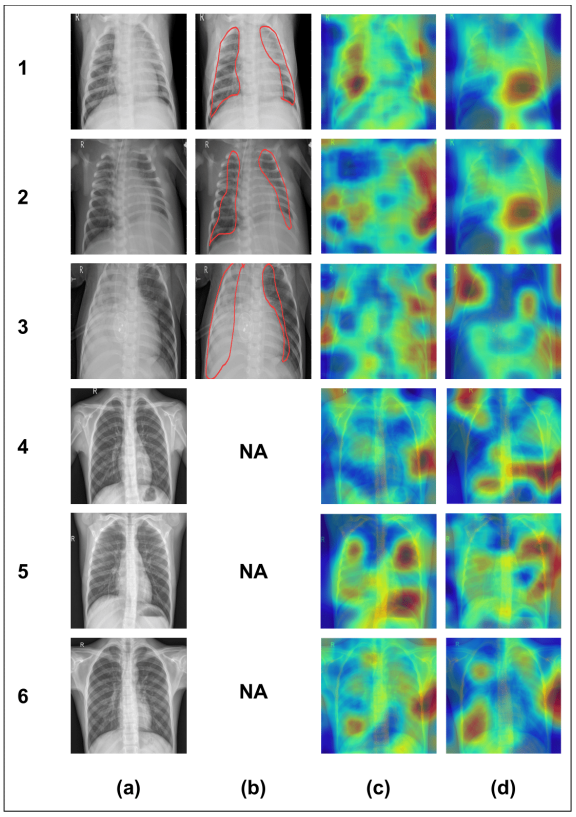

Figure 7.5: Visualization of Explanation when model is correct but looking at wrong place(Top 3 row) and when model is wrong with overconfidence

Despite these advances, GA-RISE introduces several new hyperparameters (population size, crossover/mutation rates, generations) and incurs a modest optimization overhead compared to one-shot mask sampling. Although mask counts are reduced by an order of magnitude, evolving them still requires multiple model evaluations per generation. Finally, while this chapter focuses on classification tasks, extending GA-RISE to high-resolution, per-pixel explanations in segmentation will demand further work on scalability and accelerated fitness evaluation.

Chapter 8

Conclusion and Future Work

8.1 Summary of Objectives and Contributions

The primary objective of this thesis was to develop novel methodologies for uncertainty estimation in deep learning-based image segmentation. Given the critical role of reliable segmentation in domains such as medical imaging and autonomous systems, the work aimed to address different sources of uncertainty—specifically boundary uncertainty, label uncertainty, and model-induced uncertainty—while proposing strategies to minimize them, including the use of multimodal data and enhancing model interpretability through explainable AI techniques.

Throughout the course of this research, several key contributions have been made: the design of new loss functions for better boundary delineation, modeling label uncertainty with consensus-driven approaches, proposing ensemble and calibration-based methods to mitigate model uncertainty, incorporating multimodal data fusion techniques, and integrating explainable AI for improved model transparency. Each contribution systematically addressed a particular aspect of uncertainty in segmentation tasks and collectively advances the field towards building more trustworthy and robust models.

8.2 Chapter-wise Contributions

Chapter 2: Boundary Uncertainty in Image Segmentation

This chapter introduced innovative techniques to handle uncertainty at object boundaries. A customized loss function, termed Bi-H Loss, was proposed to better capture the nuances of boundary regions. Additionally, the COVID-CT-H-UNet model was developed and validated on medical imaging data, demonstrating superior boundary adherence and improved segmentation performance compared to conventional methods.

Chapter 3: Label Uncertainty in Image Segmentation

Chapter 3 tackled the challenge of annotation variability across multiple experts. A novel segmentation framework was proposed that does not rely on a single ground truth but instead generates consensus-driven segmentation maps by learning from multiple annotations. This approach helped in mitigating label-induced uncertainties and improving model generalizability across diverse datasets.

Chapter 4: Addressing Model Uncertainty using Ensemble Techniques

In this chapter, an ensemble-based strategy leveraging copula functions was introduced to better capture dependencies among predictions of multiple models. Instead of traditional pixel-wise ensembles, a superpixel-driven approach was employed to reduce computational overhead while maintaining segmentation quality, thus offering a practical solution for real-world deployment.

Chapter 5: Addressing Model Uncertainty using Calibration Techniques

Chapter 5 proposed a calibration technique based on Confusion Penalty-Based Label Smoothing (CPLS) to ensure that the model's predicted probabilities are well-aligned with the actual likelihoods. By reducing overconfidence in incorrect predictions, this approach enhanced the model's reliability, especially in critical applications where misclassifications carry severe consequences.

Chapter 6: Reducing Uncertainty Through Multimodal Data

This chapter explored the integration of multiple imaging modalities, such as RGB, thermal, and near-infrared (NIR) data, using early fusion strategies. Experiments demonstrated that multimodal fusion significantly enhances segmentation accuracy and reduces uncertainty, particularly in challenging environmental conditions where single-modality data may be inadequate.

Chapter 7: Explainable AI and Uncertainty

The final technical chapter introduced GA-RISE, an explainable AI framework based on genetic algorithm-optimized perturbation masks, to interpret segmentation model predictions. By highlighting salient regions that influenced model decisions, this method not only improved model transparency but also offered a valuable tool for understanding areas of high uncertainty in the output.

8.3 Future Directions

Building on the foundation laid in each chapter, several promising research avenues remain:

- **Dynamic Uncertainty Modeling**: To develop architectures that adaptively adjust their uncertainty estimation during inference based on detected domain shifts or user feedback.

- **Boundary Refinement Extensions (Chapter 2)**:
  - To extend the attention-guided Bi-H Loss framework to 3D volumetric data (e.g., MRI/ultrasound) and other modalities (CT–PET fusion).
  - To integrate lightweight architectures (e.g., MobileNet variants) for real-time clinical deployment.
  - To investigate self-supervised or few-shot boundary fine-tuning to alleviate data scarcity.
- **Consensus and Annotator Modeling (Chapter 3)**:
  - To incorporate semi-supervised and active-learning loops, letting the model query experts on high-disagreement regions.
  - To extend ASPE to multi-modal annotation contexts (e.g., CT + MRI) and to temporal sequences.
- **Scalable Dependence-Aware Ensembles (Chapter 4)**:
  - To scale copula fusion to dozens of models via low-rank or sparse dependency structures for efficiency.
  - To combine ensemble dependencies with calibration (CPLS) to jointly reduce variance and overconfidence.
- **Advanced Calibration Strategies (Chapter 5)**
  - To design spatially adaptive CPLS where smoothing factors vary by pixel or region complexity.
  - To merge CPLS with Bayesian or evidential loss functions for richer uncertainty quantification.
  - To explore semi-supervised and self-supervised versions of CPLS to reduce dependence on large labeled sets.
- **Multimodal Fusion Beyond Early Integration (Chapter 6)**:
  - To compare early, mid-level, and late fusion strategies, potentially with learned attention weights per modality.
  - To transition from object detection to full pixel-wise segmentation in multimodal stacks (RGB + depth + thermal).

  - To introduce modality-specific uncertainty estimates to weight each sensor's contribution adaptively.

- **Explainability-Uncertainty Synergy (Chapter 7)**

  - To adapt GA-RISE to generate pixel-level saliency for segmentation networks, enabling direct mapping of uncertainty on masks.
  - To incorporate uncertainty metrics into the genetic-algorithm fitness function so that masks highlight both relevance and confidence.
  - To develop quantitative validation of explanations (e.g., using user studies or structured metrics) to complement visual inspection.

8.4 Concluding Remarks

In conclusion, this thesis presented a comprehensive study on various aspects of uncertainty in deep learning-based image segmentation and proposed multiple innovative solutions to address them. By tackling boundary-level inconsistencies, label ambiguities, model reliability, and modality limitations, the work aims to bridge the gap between theoretical advancements and real-world applicability of segmentation models. Moreover, the integration of explainable AI methodologies reinforces the trustworthiness of these models. It is hoped that the research directions and findings outlined herein will stimulate further innovations in building reliable, interpretable, and uncertainty-aware deep learning systems for image segmentation and beyond.

Bibliography

[1] "Covid-19 ct images segmentation." [Online]. Available: https://www.kaggle.com/competitions/covid-segmentation/data

[2] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information fusion*, vol. 76, pp. 243–297, 2021.

[3] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[4] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[5] K. Aho, D. Derryberry, and T. Peterson, "Model selection for ecologists: the worldviews of AIC and BIC," Tech. Rep. 3, 2014.

[6] A. Almazroa, S. Alodhayb, E. Osman, E. Ramadan, M. Hummadi, M. Dlaim, M. Alkatee, K. Raahemifar, and V. Lakshminarayanan, "Agreement among ophthalmologists in marking the optic disc and optic cup in fundus images," *International ophthalmology*, vol. 37, pp. 701–717, 2017.

[7] M. Ambinder, "The secret team that killed bin laden," *National Journal*, vol. 3, 2011.

[8] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical image

segmentation review: The success of u-net," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[9] V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *arXiv preprint arXiv:1505.07293*, 2015.

[10] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," Tech. Rep. [Online]. Available: http://mi.eng.cam.ac.uk/projects/segnet/.

[11] R. Battiti and A. M. Colla, "Democracy in neural nets: Voting schemes for classification," *Neural Networks*, vol. 7, no. 4, pp. 691–707, 1994.

[12] M. Berman, A. R. Triki, and M. B. Blaschko, "The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4413–4421.

[13] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[14] K. P. Burnham and D. R. Anderson, "Multimodel inference: Understanding AIC and BIC in model selection," 2004.

[15] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[16] C. Chen, R. Xiao, T. Zhang, Y. Lu, X. Guo, J. Wang, H. Chen, and Z. Wang, "Pathological lung segmentation in chest ct images based on improved random walker," *Computer methods and programs in biomedicine*, vol. 200, p. 105864, 2021.

[17] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[19] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[20] Z. Cheng and J. Shen, "On very large scale test collection for landmark image search benchmarking," *Signal Processing*, vol. 124, pp. 13–26, 2016.

[21] S.-B. Cho and J. H. Kim, "Combining multiple neural networks by fuzzy integral for robust classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 25, no. 2, pp. 380–384, 1995.

[22] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[23] S. E. Conant-Pablos, D. J. Magaña-Lozano, and H. Terashima-Marín, "Pipelining memetic algorithms, constraint satisfaction, and local search for course timetabling," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009.

[24] R. T. de Oliveira, T. F. O. de Assis, P. R. A. Firmino, T. A. Ferreira, and A. L. Oliveira, "Copulas-based ensemble of artificial neural networks for forecasting real world time series," in *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 4089–4096.

[25] S. Dey, S. Mitra, S. Chakraborty, D. Mondal, M. Nasipuri, and N. Das, "Gc-enc: A copula based ensemble of cnns for malignancy identification in breast histopathology and cytology images," *Computers in Biology and Medicine*, vol. 152, p. 106329, 2023.

[26] P. Dhankhar and N. Sahu, "A review and research of edge detection techniques for image segmentation," *International Journal of Computer Science and Mobile Computing*, vol. 2, no. 7, pp. 86–92, 2013.

[27] W. Di, L. Zhang, D. Zhang, and Q. Pan, "Studies on hyperspectral face recognition in visible spectrum with feature band selection," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 6, pp. 1354–1361, 2010.

[28] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 2018, pp. 0210–0215.

[29] E. Eban, G. Rothschild, A. Mizrahi, I. Nelken, and G. Elidan, "Dynamic Copula Networks for Modeling Real-valued Time Series," Tech. Rep., 2013.

[30] A. Elnakib, G. Gimel'farb, J. S. Suri, and A. El-Baz, "Medical image segmentation: a brief survey," *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies: Volume II*, pp. 1–39, 2011.

[31] T. Falk, D. Mai, R. Bensch, Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald *et al.*, "U-net: deep learning for cell counting, detection, and morphometry," *Nature methods*, vol. 16, no. 1, pp. 67–70, 2019.

[32] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Inf-net: Automatic covid-19 lung infection segmentation from ct images," *IEEE transactions on medical imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[33] V. Farley, A. Vallières, A. Villemaire, M. Chamberland, P. Lagueux, and J. Giroux, "Chemical agent detection and identification with a hyperspectral imaging infrared sensor," in *Electro-optical remote sensing, detection, and photonic technologies and their applications*, vol. 6739. SPIE, 2007, pp. 334–345.

[34] P. H. Ferreira and F. Louzada, "A modified version of the inference function for margins and interval estimation for the bivariate clayton copula sur tobit model: An simulation approach," *arXiv preprint arXiv:1404.3287*, 2014.

[35] J. G. Ferwerda, *Charting the quality of forage: measuring and mapping the variation of chemical components in foliage with hyperspectral remote sensing*. Wageningen University and Research, 2005.

[36] Y. Gal *et al.*, "Uncertainty in deep learning," 2016.

[37] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[38] L. Gao, X. Li, J. Song, and H. T. Shen, "Hierarchical lstms with adaptive attention for visual captioning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 5, pp. 1112–1131, 2019.

[39] C. Genest and K. J. McConway, "Allocating the weights in the linear opinion pool," *Journal of Forecasting*, vol. 9, no. 1, pp. 53–73, 1990.

[40] S. Ghosh, N. Das, I. Das, and U. Maulik, "Understanding deep learning techniques for image segmentation," *ACM computing surveys (CSUR)*, vol. 52, no. 4, pp. 1–35, 2019.

[41] S. Ghosh, A. Pal, S. Jaiswal, K. Santosh, N. Das, and M. Nasipuri, "Segfast-v2: Semantic image segmentation with less parameters in deep learning for autonomous driving," *International Journal of Machine Learning and Cybernetics*, vol. 10, pp. 3145–3154, 2019.

[42] T. Gomez, T. Fréour, and H. Mouchère, "Metrics for saliency map evaluation of deep learning explanation methods," in *International Conference on Pattern Recognition and Artificial Intelligence*. Springer, 2022, pp. 84–95.

[43] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 1–8.

[44] ——, "Decomposing a scene into geometric and semantically consistent regions," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 1–8.

[45] N. Gregor, "Copula parameter estimation by maximum-likelihood and minimum-distance estimators," *A simulation study*, 2009.

[46] M. Guan, V. Gulshan, A. Dai, and G. Hinton, "Who said what: Modeling individual labelers improves classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[47] X. Guo, S. Lu, Y. Yang, P. Shi, C. Ye, Y. Xiang, and T. Ma, "Modeling annotator variation and annotator preference for multiple annotations medical image segmentation," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 977–984.

[48] X. Guo, P. Shi, S. Lu, C. Ye, and T. Ma, "Joint learning annotator calibration and annotator preference for multiple annotations optic disc and cup segmentation," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023, pp. 1–5.

[49] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[51] K. T. Higgins, "Five new technologies for inspection," *Food Process*, vol. 6, 2013.

[52] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 1, pp. 66–75, 1994.

[53] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[54] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[55] Z. Hu, Q. Zou, and Q. Li, "Watershed superpixel," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 349–353.

[56] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[57] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural networks*, vol. 121, pp. 74–87, 2020.

[58] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 590–597.

[59] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, and J. Sattar, "Semantic segmentation of underwater imagery: Dataset and benchmark," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 1769–1776.

[60] ——, "Semantic segmentation of underwater imagery: Dataset and benchmark," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 1769–1776.

[61] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation," Tech. Rep. [Online]. Available: https://github.com/SimJeg/FC-DenseNet

[62] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. IEEE, 2017, pp. 11–19.

[63] M. H. Jensen, D. R. Jørgensen, R. Jalaboi, M. E. Hansen, and M. A. Olsen, "Improving uncertainty estimation in convolutional neural networks using inter-rater agreement," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer, 2019, pp. 540–548.

[64] C. Ji and S. Ma, "Combinations of weak classifiers," in *Advances in Neural Information Processing Systems*, 1997, pp. 494–500.

[65] W. Ji, S. Yu, J. Wu, K. Ma, C. Bian, Q. Bi, J. Li, H. Liu, L. Cheng, and Y. Zheng, "Learning calibrated medical image segmentation via multi-rater agreement modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 341–12 351.

[66] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "Layercam: Exploring hierarchical class activation maps for localization," *IEEE Transactions on Image Processing*, vol. 30, pp. 5875–5888, 2021.

[67] H. Joe and J. J. Xu, "The estimation method of inference functions for margins for multivariate models," 1996.

[68] V. Kamakshi and N. C. Krishnan, "Explainable image classification: The journey so far and the road ahead," *AI*, vol. 4, no. 3, pp. 620–651, 2023.

[69] S. C. Kao, A. R. Ganguly, and K. Steinhaeuser, "Motivating complex dependence structures in data mining: A case study with anomaly detection in climate," in *ICDM Workshops 2009 - IEEE International Conference on Data Mining*, 2009.

[70] J. N. Kather, F. G. Zöllner, F. Bianconi, S. M. Melchers, L. R. Schad, T. Gaiser, A. Marx, and C.-A. Weis, "Collection of textures in colorectal cancer histology," May 2016. [Online]. Available: https://doi.org/10.5281/zenodo.53169

[71] Ç. Kaymak and A. Uçar, "A brief survey and an application of semantic image segmentation for autonomous driving," *Handbook of deep learning applications*, pp. 161–200, 2019.

[72] Z. K. G. M. Kermany D, "Labeled optical coherence tomography (oct) and chest x-ray images for classification," *Mendeley Data, v2*, 2018.

[73] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *International conference on medical imaging with deep learning*. PMLR, 2019, pp. 285–296.

[74] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[75] I. Kotaridis and M. Lazaridou, "Remote sensing image segmentation advances: A meta-analysis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 309–322, 2021.

[76] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[77] U. Krothapalli and L. Abbott, "One size doesn't fit all: Adaptive label smoothing," 2020.

[78] L. Kuncheva, J. C. Bezdek, and M. A. Sutton, "On combining multiple classifiers by fuzzy templates," in *1998 Conference of the North American Fuzzy Information Processing Society-NAFIPS (Cat. No. 98TH8353)*. IEEE, 1998, pp. 193–197.

[79] L. I. Kuncheva, "An application of owa operators to the aggregation of multiple classification decisions," in *The ordered weighted averaging operators*. Springer, 1997, pp. 330–343.

[80] ——, "A theoretical study on six classifier fusion strategies," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 2, pp. 281–286, 2002.

[81] F. Lacar, M. Lewis, and I. Grierson, "Use of hyperspectral imagery for mapping grape varieties in the barossa valley, south australia," in *IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No. 01CH37217)*, vol. 6. IEEE, 2001, pp. 2875–2877.

[82] L. Lam and C. Y. Suen, "Optimal combinations of pattern classifiers," *Pattern Recognition Letters*, vol. 16, no. 9, pp. 945–954, 1995.

[83] P. Laux, S. Wagner, A. Wagner, J. Jacobeit, A. Bardossy, and H. Kunstmann, "Modelling daily precipitation features in the volta basin of west africa," *International Journal of Climatology: A Journal of the Royal Meteorological Society*, vol. 29, no. 7, pp. 937–954, 2009.

[84] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, "Turbopixels: Fast superpixels using geometric flows," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 12, pp. 2290–2297, 2009.

[85] G. Li, C. Li, C. Zeng, P. Gao, and G. Xie, "Region focus network for joint optic disc and cup segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 751–758.

[86] X. Li, H. Ding, H. Yuan, W. Zhang, J. Pang, G. Cheng, K. Chen, Z. Liu, and C. C. Loy, "Transformer-based visual segmentation: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2024.

[87] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[88] Z. Liao, S. Hu, Y. Xie, and Y. Xia, "Modeling annotator preference and stochastic annotation error for medical image segmentation," *Medical Image Analysis*, vol. 92, p. 103028, 2024.

[89] Z. Liao, Y. Xie, S. Hu, and Y. Xia, "Learning from ambiguous labels for lung nodule malignancy prediction," *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1874–1884, 2022.

[90] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[91] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, "Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data," *IEEE transactions on medical imaging*, vol. 39, no. 9, pp. 2713–2724, 2020.

[92] X. Liu, Y. Liu, W. Fu, and S. Liu, "Sctv-unet: a covid-19 ct segmentation network based on attention mechanism," *Soft Computing*, pp. 1–11, 2023.

[93] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[94] Q. Lu, Z. Bai, S. Fan, X. Zhou, and Z. Xu, "Multiscale codec network based ct image segmentation for human lung disease derived of covid-19," *Journal of Image and Graphics*, pp. 827–837, 2022.

[95] S. Masood, M. Sharif, A. Masood, M. Yasmin, and M. Raza, "A survey on medical image segmentation," *Current Medical Imaging*, vol. 11, no. 1, pp. 3–14, 2015.

[96] B. Menze, L. Joskowicz, C. Berger *et al.*, "Quantification of uncertainties in biomedical image quantification," *Zenodo*, 2020.

[97] I. D. Mienye and T. G. Swart, "A comprehensive review of deep learning: Architectures, recent advances, and applications," *Information*, vol. 15, no. 12, p. 755, 2024.

[98] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[99] M. R. Minar and J. Naher, "Recent advances in deep learning: An overview," *arXiv preprint arXiv:1807.08169*, 2018.

[100] Z. Mirikharaji, K. Abhishek, S. Izadi, and G. Hamarneh, "D-lema: Deep learning ensembles from multiple annotations-application to skin lesion segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1837–1846.

[101] H. Mittal, A. C. Pandey, M. Saraswat, S. Kumar, R. Pal, and G. Modwel, "A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets," *Multimedia Tools and Applications*, pp. 1–26, 2022.

[102] M. B. Muhammad and M. Yeasin, "Eigen-cam: Class activation map using principal components," in *2020 international joint conference on neural networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[103] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Advances in neural information processing systems*, vol. 32, 2019.

[104] P. W. Munro and B. Parmanto, "Competition among networks improves committee performance," in *Advances in Neural Information Processing Systems*, 1997, pp. 592–598.

[105] R. Muthukrishnan and M. Radha, "Edge detection techniques for image segmentation," *International Journal of Computer Science & Information Technology*, vol. 3, no. 6, p. 259, 2011.

[106] R. B. Nelsen, *An introduction to copulas*. Springer, 2006.

[107] ——, *An introduction to copulas*. Springer Science & Business Media, 2007.

[108] A. O'Leary, J. Ferwerda, S. Jones, G. Fitzgerald, and R. Belford, "Remote sensing to detect nitrogen and water stress in wheat." 2006.

[109] X. Ouyang, Z. Xue, Y. Zhan, X. S. Zhou, Q. Wang, Y. Zhou, Q. Wang, and J.-Z. Cheng, "Weakly supervised segmentation framework with uncertainty: A study on pneumothorax segmentation in chest x-ray," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer, 2019, pp. 613–621.

[110] O. Ozdemir, T. Allen, S. Choi, T. Wimalajeewa, and P. Varshney, "Copula Based Classifier Fusion Under Statistical Dependence," 2017.

[111] Ö. ÖZTÜRK and T. P. Hettmansperger, "Generalised weighted cramér-von mises distance estimators," *Biometrika*, vol. 84, no. 2, pp. 283–294, 1997.

[112] I. Papadeas, L. Tsochatzidis, A. Amanatiadis, and I. Pratikakis, "Real-time semantic image segmentation with deep learning for autonomous driving: A survey," *Applied Sciences*, vol. 11, no. 19, p. 8802, 2021.

[113] S. Pare, A. Kumar, G. K. Singh, and V. Bajaj, "Image segmentation using multilevel thresholding: a research review," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, vol. 44, no. 1, pp. 1–29, 2020.

[114] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," *arXiv preprint arXiv:1806.07421*, 2018.

[115] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and H. Q. Nguyen, "Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels," *Neurocomputing*, vol. 437, pp. 186–194, 2021.

[116] I. Pöllänen, B. Braithwaite, K. Haataja, T. Ikonen, and P. Toivanen, *Current Analysis Approaches and Performance Needs for Whole Slide Image Processing in Breast Cancer Diagnostics*.

[117] A. Purwanto, D. H. Budiarti, F. N. Purnamastuti, I. Y. Tanasa, Y. Guno, A. S. Yunata, M. Wibowo, A. Hidayat, and D. Dirgahayu, "Image segmentation in aerial imagery: A review," *SINERGI*, vol. 27, no. 8, pp. 343–360, 2023.

[118] K. Ramesh, G. K. Kumar, K. Swapna, D. Datta, and S. S. Rajest, "A review of medical image segmentation algorithms," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 7, no. 27, pp. e6–e6, 2021.

[119] R. Ranjan and T. Gneiting, "Combining probability forecasts," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 1, pp. 71–91, 2010.

[120] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[121] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[122] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Tech. Rep. [Online]. Available: http://lmb.informatik.uni-freiburg.de/

[123] ——, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[124] N. Saeedizadeh, S. Minaee, R. Kafieh, S. Yazdani, and M. Sonka, "Covid tv-unet: Segmenting covid-19 chest ct images using connectivity imposed unet," *Computer methods and programs in biomedicine update*, vol. 1, p. 100007, 2021.

[125] R. Salinas-Gutiérrez, A. Hernández-Aguirre, M. J. J. Rivera-Meraz, and E. R. Villa-Diharce, "Using Gaussian Copulas in Supervised Probabilistic Classification," Tech. Rep.

[126] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[127] V. A. Satopää, J. Baron, D. P. Foster, B. A. Mellers, P. E. Tetlock, and L. H. Ungar, "Combining multiple probability predictions using a simple logit model," *International Journal of Forecasting*, vol. 30, no. 2, pp. 344–356, 2014.

[128] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[129] N. Senthilkumaran and S. Vaithegi, "Image segmentation by using thresholding techniques for medical images," *Computer Science & Engineering: An International Journal*, vol. 6, no. 1, pp. 1–13, 2016.

[130] A. Shahidi, S. Patel, J. Flanagan, and C. Hudson, "Regional variation in human retinal vessel oxygen saturation," *Experimental eye research*, vol. 113, pp. 143–147, 2013.

[131] P. Sharma and J. Suji, "A review on image segmentation with its clustering techniques," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 9, no. 5, pp. 209–218, 2016.

[132] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[133] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[134] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[135] J. Song, L. Gao, F. Nie, H. T. Shen, Y. Yan, and N. Sebe, "Optimized graph learning using partial tags and multiple features for image and video annotation," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 4999–5011, 2016.

[136] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer, 2017, pp. 240–248.

[137] W. Sun and R. Wang, "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with dsm," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 3, pp. 474–478, 2018.

[138] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[139] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[140] W. R. Tan, C. S. Chan, P. Yogarajah, and J. Condell, "A fusion approach for efficient human skin detection," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 138–147, 2011.

[141] D. M. Tax, M. Van Breukelen, R. P. Duin, and J. Kittler, "Combining multiple classifiers by averaging or by multiplying?" *Pattern recognition*, vol. 33, no. 9, pp. 1475–1485, 2000.

[142] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[143] D. Tran, M. Dusenberry, M. Van Der Wilk, and D. Hafner, "Bayesian layers: A module for neural network uncertainty," *Advances in neural information processing systems*, vol. 32, 2019.

[144] Tzutalin, "Labelimg," Free Software: MIT License, 2015. [Online]. Available: https://github.com/tzutalin/labelImg

[145] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV 10*. Springer, 2008, pp. 705–718.

[146] P. Vermeulen, P. Flémal, O. Pigeon, P. Dardenne, J. Fernández Pierna, and V. Baeten, "Assessment of pesticide coating on cereal seeds by near infrared hyperspectral imaging," *J. Spectral Imaging*, vol. 6, pp. 1–7, 2017.

[147] C. Wang, J. Chen, and W. Li, "Review on superpixel segmentation algorithms," *Application research of Computers*, vol. 31, no. 1, pp. 6–12, 2014.

[148] D.-B. Wang, L. Feng, and M.-L. Zhang, "Rethinking calibration of deep neural networks: Do not be afraid of overconfidence," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 809–11 820, 2021.

[149] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.

[150] L. Wang, R. Li, H. Shi, J. Sun, L. Zhao, H. S. Seah, C. K. Quah, and B. Tandianus, "Multi-channel convolutional neural network based 3d object detection for indoor robot environmental perception," *Sensors*, vol. 19, no. 4, p. 893, 2019.

[151] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET image processing*, vol. 16, no. 5, pp. 1243–1267, 2022.

[152] X. Wang, Y. Zhao, and F. Pourpanah, "Recent advances in deep learning," *International Journal of Machine Learning and Cybernetics*, vol. 11, pp. 747–750, 2020.

[153] J. Wei, L. Torresani, J. Wei, and S. Hassanpour, "Calibrating histopathology image classifiers using label smoothing," in *International Conference on Artificial Intelligence in Medicine*. Springer, 2022, pp. 273–282.

[154] G. Weiß, "Copula parameter estimation by maximum-likelihood and minimum-distance estimators: a simulation study," *Computational Statistics*, vol. 26, no. 1, pp. 31–54, 2011.

[155] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[156] S. Wu, "Construction of asymmetric copulas and its application in two-dimensional reliability modelling," *European Journal of Operational Research*, vol. 238, no. 2, pp. 476–485, 2014.

[157] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.

[158] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE transactions on systems, man, and cybernetics*, vol. 22, no. 3, pp. 418–435, 1992.

[159] Y. Xu, R. Quan, W. Xu, Y. Huang, X. Chen, and F. Liu, "Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches," *Bioengineering*, vol. 11, no. 10, p. 1034, 2024.

[160] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie, "Covid-ct-dataset: a ct scan dataset about covid-19," *arXiv preprint arXiv:2003.13865*, 2020.

[161] Y. Yu, C. Wang, Q. Fu, R. Kou, F. Huang, B. Yang, T. Yang, and M. Gao, "Techniques and challenges of image segmentation: A review," *Electronics*, vol. 12, no. 5, p. 1199, 2023.

[162] K.-H. Yuan, K. Hayashi, and P. M. Bentler, "Normal theory likelihood ratio statistic for mean and covariance structure analysis under alternative hypotheses," *Journal of Multivariate Analysis*, vol. 98, no. 6, pp. 1262–1282, 2007.

[163] C.-B. Zhang, P.-T. Jiang, Q. Hou, Y. Wei, Q. Han, Z. Li, and M.-M. Cheng, "Delving deep into label smoothing," *IEEE Transactions on Image Processing*, vol. 30, pp. 5984–5996, 2021.

[164] L. Zhang, R. Tanno, K. Bronik, C. Jin, P. Nachev, F. Barkhof, O. Ciccarelli, and D. C. Alexander, "Learning to segment when experts disagree," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. Springer, 2020, pp. 179–190.

[165] S. Zhang, B. Geng, P. K. Varshney, and M. Rangaswamy, "Fusion of deep neural networks for activity recognition: A regular vine copula based approach," in *2019 22th International Conference on Information Fusion (FUSION)*. IEEE, 2019, pp. 1–7.

[166] H. Zhao, H. Li, and L. Cheng, "Improving retinal vessel segmentation with joint local loss by matting," *Pattern Recognition*, vol. 98, p. 107068, 2020.

[167] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," Tech. Rep. [Online]. Available: https://github.com/hszhao/PSPNet

[168] ——, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2017, pp. 2881–2890.

[169] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[170] Y. Zhou, F.-J. Chang, H. Chen, and H. Li, "Exploring copula-based bayesian model averaging with multiple anns for pm2. 5 ensemble forecasts," *Journal of Cleaner Production*, vol. 263, p. 121528, 2020.

[171] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.