Fortifying Medical Image Segmentation: Adversarially Robust and Trustworthy Deep Learning Solutions

Ph.D. Thesis

By
Sneha Shukla

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE

May 2025

Fortifying Medical Image Segmentation: Adversarially Robust and Trustworthy Deep Learning Solutions

A Thesis

Submitted in partial fulfillment of the
requirements for the award of the degrees
of
Doctor of Philosophy

by

Sneha Shukla

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE

May 2025

INDIAN INSTITUTE OF TECHNOLOGY INDORE

I hereby certify that the work which is being presented in the thesis entitled **Fortifying Medical Image Segmentation: Adversarially Robust and Trustworthy Deep Learning Solutions** in the partial fulfilment of the requirements for the award of the degree of **DOCTOR OF PHILOSOPHY** and submitted in the **Department of Computer Science & Engineering**, Indian Institute of Technology Indore, is an authentic record of my own work carried out during the time period from **May 2021** to **May 2025** under the supervision of **Dr. Puneet Gupta**.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

06/02/2026

Signature of the student with date

(Sneha Shukla)

This is to certify that the above statement made by the candidate is correct to the best of my/our knowledge.

6/2/26

Signature of Thesis Supervisor with date

(Dr. Puneet Gupta)

Sneha Shukla has successfully given her Ph.D. Oral Examination held on *06-02-2026*.

6/2/26

Signature of Thesis Supervisor with date

(Dr. Puneet Gupta)

ACKNOWLEDGEMENTS

First of all, I would like to thank my thesis supervisor, **Dr. Puneet Gupta**, for his consistent support and invaluable guidance throughout my PhD journey. His mentorship has been the cornerstone of my academic growth, under which I have gained a profound insight into in-depth research, emerging technologies, paper writing, and presentation skills. His deep knowledge and perceptive guidance have been instrumental in shaping me into an independent researcher. The way he tackles complex research problems is genuinely remarkable. His dynamic supervision and inspiring leadership boosted my research knowledge and taught me the importance of diligence, patience, and compassion. I am incredibly fortunate to have such an inspiring and supportive mentor. Thanks to him, I can confidently present my research anywhere with pride. Lastly, I would like to say, *Thank you, sir, for being a stepping stone in my PhD journey and for your unwavering belief in my potential.*

Further, I express my sincere gratitude to my cumulative evaluation of research progress committee members (PSPC) **Prof. Surya Prakash** and **Dr. Hareskrishna Yadav** for their insightful suggestions, constructive feedback, and periodic validation of the research work. I would also like to thank **Prof. Somnath Dey** and **Dr. Ranveer Singh**, Head of the Department, Computer Science and Engineering, for their immense help and support.

My heartfelt gratitude to the director of the Indian Institute of Technology Indore, **Prof. Suhas Joshi** for fostering a competitive and enriching research environment at the university. My profound thanks go out to the staff of the Indian Institute of Technology Indore, specifically **Mr. Shailendra Verma** and **Mrs. Ujavala Gorakh Langhi** for all the academic assistance and always being supportive with official tasks. I am also grateful to the Indian Institute of Technology Indore, for giving me the opportunity to pursue a PhD in Computer Science and Engineering.

A heartfelt thank you to my best friend, **Rupendra Pratap Singh Hada**, for being a constant source of support, encouragement, and joy throughout my PhD journey. Your fascinating conversations, friendship, and shared celebration of every achievement made this journey truly special.

I am profoundly grateful to be a part of the **Deep Intelligence Lab** and to have the sup-

port of its incredible team. I sincerely appreciate my senior, **Dr. Lokendra Birla**, and my colleagues-turned-friends **Anup Kumar Gupta, Rupesh Kumar, Aditya Dixit, Trishna Saikia, Ashutosh Dhamaniya, Rajesh Kumar, Aravind Ramagiri, Anirban Nath,** and **Adit Srivastava** for their support, encouragement, collaboration, and, above all, friendship. A special thanks to my friends **Anup Kumar Gupta** and **Rupendra Pratap Singh Hada** for their help and motivation during my initial days at IIT. I am also grateful to my juniors-cum-friends **Drishti Sharma** and **Prasanna Bairagi** for making my PhD journey truly memorable.

Finally, yet most importantly, I am deeply grateful to my parents, **Mr. Mahendra Kumar Shukla** and **Mrs. Pushpa Shukla**, whose belief in me has been my greatest source of strength. Their resilience, hard work, and independent nature have been a constant inspiration in fueling my ambition. Their unconditional love, endless prayers, and countless sacrifices have been the backbone of my PhD journey, providing me with the support and courage to persevere through every challenge. I am greatly thankful to my beloved sister, **Dr. Neha Mishra** for encouraging me at every step and constantly supporting me throughout my PhD. Her shared knowledge and valuable insights about the medical domain helped me a lot to better understand my work. A heartfelt thank you to my brother **Ankit Shukla** and my adorable nephew **Aviral Mishra** for their love, which has been an integral part of my journey. Once again, thank you all for being my pillars of strength.

Sneha Shukla

Dedicated

to

My Family, Friends
and Respected Teachers

ABSTRACT

In the era of digital innovation, Deep Learning (DL) has achieved remarkable success across various fields, including healthcare, where its decisions directly impact human lives. The DL-based medical models are versatile and capable of handling a wide range of tasks. Medical Image Segmentation (MIS) is one such critical task wherein a disease-afflicted Region of Interest (ROI) is separated from the unintended regions. Such ROIs primarily refer to any organs, cancerous cells, tissues, or lesions. Unfortunately, the DL-based MIS models are highly vulnerable to intelligently curated adversarial attacks, where small and imperceptible perturbations are imposed on the input that drastically mislead the predictions. This concern is more pervasive with the medical images, as their rich textural details can easily divert the focus of the model towards irrelevant regions, ultimately diminishing their performance and robustness. Moreover, the predictions of such models struggle with anomalous samples like adversarial and Out-Of-Distribution (OOD). While adversarial samples are generated by adding small and imperceptible perturbations to the input, OOD samples signify input data with a shifted distribution. Both of these samples significantly degrade the performance of the model. Furthermore, the opaque nature of DL models offers non-trustworthy predictions, causing conflicts among users over accepting the model prediction. All these problems could result in catastrophic consequences in the healthcare domain.

To bridge all these research gaps, this thesis aims to fortify the DL-based MIS model by making it adversarially robust, ensuring its trustworthiness, and simultaneously advancing its performance. In this direction, we initially propose a novel adversarial attack, ***DECEIT***, to perceive the effectiveness of such an attack on the DL-based MIS model. The attack is performed by backpropagating the loss function in relation to the input. Unfortunately, MIS models exhibit several non-differentiable layers and non-differentiable loss functions that hinder backpropagation, disrupting attacks. ***DECEIT*** excludes these non-differentiable layers and employs differentiable approximations instead. It also uses a surrogate loss function to smoothly attack the model. However, different surrogate losses behave differently for the same input-target pair. Thus, ***DECEIT*** performs parallel fusion by conducting attack operations on several surrogate loss functions and choosing the optimal one, which

imposes minimum perturbation while ensuring a successful attack. After ***DECEIT***, we subsequently propose a unified detection method, ***DISCERN***, which identifies adversarial and OOD samples from the clean ones in the DL-based MIS model while avoiding network retraining and Ground Truth (GT). Empirical detection methods are predominantly centred on Medical Image Classification (MIC) tasks, which struggle to translate well for MIS tasks. Moreover, it adapts either a probability distribution or a threshold-based distance approach to perform detection. In contrast, our proposed method, ***DISCERN***, benefits from the observations that clean samples are highly consistent with their rotated variants, as the MIS model ensures rotation-invariant predictions. Conversely, adversarial samples show reduced consistency with their respective variants since rotation weakens the efficacy of perturbations. Likewise, the consistency drops for OOD samples due to their inherent randomness in model predictions. Without relying on existing approaches, ***DISCERN*** performs consistency-based detection by analysing the similarity between samples' predictions and their respective variants to significantly identify adversarial and OOD samples.

To alleviate the impression of the adversarial attack on the DL-based MIS model, we eventually propose an adversarial defence, ***RELIVE*** (Cont***R***astiv***E*** Mu***LtI***tasking Ad***V***ersarial D***E***fence), elevating the adversarial robustness of the model with noteworthy performance gain. ***RELIVE*** exhibits contrastive learning, multitask learning, and their fusion-based defence. Initially, the contrastive learning-based defence builds on the insight that keeping the features of clean, adversarial, and augmented samples closer to each other during training substantially advances the adversarial robustness of the model. Subsequently, the multitask learning-based defence represents generalised features and significantly mitigates the impact of perturbation by selecting auxiliary tasks based on the weak correlation with the main task. Eventually, we analyse the individual advantages of contrastive and multitask learning and propose their fusion-based defence, wherein contrastive learning is exclusively applied to the main task in the proposed multitask architecture, which results in further enhancement in the model's resilience and performance. Since the non-transparency of DL-based MIS models fails to provide trustworthy predictions, which limits their applicability. Thus, we introduce a novel method, ***TrustMedIS*** (***Trust***worthy ***Med***ical ***I***mage ***S***egmentation), aiming to investigate the trustworthiness of the DL-based

MIS models while simultaneously advancing their performance. It works in three folds: *ET* (*E*xamining *T*rustworthiness), *ENT* (*E*levating *N*on-*T*rustworthy predictions), and *CSM* (*C*lassifier *S*election *M*ethod). *ET* observes the characteristics of input and output, followed by computing the consistency between these outputs and their respective rotated variants. It measures the confidence score and employs a prespecified threshold to identify the MIS prediction's trustworthiness. *ENT* employs *ET* and addresses non-trustworthy predictions by taking advantage of the insight that rotation can diminish erroneous impact in such predictions. *CSM* utilises *ENT* method and scrutinises multiple MIS models to select the optimal one that offers the most trustworthy prediction, ultimately enhancing the effectiveness of the DL-based MIS model.

In a summarised way, the overall thesis contributions are as follows: (1) Our proposed adversarial attack, ***DECEIT***, dynamically selects the optimal surrogate loss function while adding minimum perturbation, enabling a deeper understanding of the attack on the DL-based MIS models. (2) Our proposed method, ***DISCERN***, effectively detects the adversarial and OOD samples by analysing consistent behaviour of the input samples' predictions and their relative variants, advancing model robustness. (3) Our proposed adversarial defence, ***RELIVE***, leverages contrastive and multitask learning to develop a fusion-based defence, significantly mitigating adversarial perturbations while slightly improving model performance. (4) Our proposed method, ***TrustMedIS***, comprising *ET*, *ENT* and *CSM*, substantially evaluates the trustworthiness of DL-based MIS models and advances the performance of the non-trustworthy predictions. Experimental results on publicly available datasets across several state-of-the-art DL-based MIS models reveal that our proposed works in the thesis surpass the existing studies and successfully address all research gaps by providing comprehensive solutions to enhance the robustness, trustworthiness, and performance of DL-based MIS models, ultimately improving their effectiveness.

LIST OF PUBLICATIONS

The publications listed below have emerged from this doctoral dissertation (as of February 2026):

(A) From PhD thesis work:

A1. Journal Articles:

Published:

J1. **Sneha Shukla**, Anup Kumar Gupta, and Puneet Gupta, "Exploring the feasibility of adversarial attacks on medical image segmentation", *Multimedia Tools and Applications*, 83(4), pp. 11745-11768, 2024. (Impact Factor : 3.0)

J2. **Sneha Shukla**, Lokendra Birla, Anup Kumar Gupta, and Puneet Gupta, "Trustworthy medical image segmentation with improved performance for in-distribution samples", *Neural Networks*, 166, pp. 127-136, 2023. (Impact Factor : 6.0)

J3. **Sneha Shukla**, and Puneet Gupta, "EVADE: A Novel Method to Detect Adversarial and OOD Samples in Medical Image Segmentation", *Expert Systems with Applications*, 127319, 2025. (Impact Factor : 7.5)

J4. **Sneha Shukla**, and Puneet Gupta, "Elevating Adversarial Robustness by Contrastive Multitasking Defence in Medical Image Segmentation", *Neural Networks*, 108182, 2025. (Impact Factor : 6.0)

(B) Other publications during PhD:

B1. Journal Articles:

Published:

J1. Anirban Nath, **Sneha Shukla**, and Puneet Gupta, "MTMedFormer : Multi-Task Vision Transformer for Medical Imaging with Federated Learning", *Medical & Biological Engineering & Computing*, pp. 1-14, 2025. (Impact Factor: 2.6)

Under Review:

J2. **Sneha Shukla**, Puneet Gupta, and Esa Rahtu, "A Comprehensive Survey of Recent Transformer-based Attentions for Computer Vision Applications", *ACM Computing Surveys*, 2025. (Impact Factor : 23.8)

B2. Conference Articles:

Published:

C1. Lokendra Birla, **Sneha Shukla**, Anup Kumar Gupta, and Puneet Gupta, "ALPINE: Improving remote heart rate estimation using contrastive learning." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023.

C2. Lokendra Birla, **Sneha Shukla**, Trishna Saikia, and Puneet Gupta, "HR-TRACK: An rPPG Method for Heartrate Monitoring Using Temporal Convolution Networks." *International Conference on Pattern Recognition*, Cham: Springer Nature Switzerland, 2024.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AMAT	Adaptive Margin Adversarial Training
ARL	Adversarial Robust Learning
ASMA	Adaptive Segmentation Mask Attack
AT	Adversarial Training
Avg	Averaging
BCE	Binary Cross Entropy
CaraNet	Context Axial Reverse Attention Network
CNN	Convolutional Neural Network
CSM	Classifier Selection Method
CT	Computed Tomography
DA	Data Augmentation
DAG	Dense Adversarial Generation
DECEIT	Dynamic Loss SEleCtion based AdvErsarIal ATtack
DL	Deep Learning
ENT	Elevating Non Trustworthy prediction
ERNN	Evidence Reconciled Neural Network
ET	Examining Trustwothiness
FGSM	Fast Gradient Sign Method
GMM	Gaussian Mixture Model
GT	Ground Truth
ID	In Distribution
IGSM	Iterative Gradient Sign Method
IOU	Intersection Over Union
IPP	Image Part Permutation
ISBI	International Symposium on Biomedical Imaging
KL	Kullback-Leibler
LCDAE	Lung Cancer Data Augmentation Ensemble
MahD	Mahalanobis Distance

MFP	Maximum Foreground Pixels
MIC	Medical Image Classification
mIOU	mean Intersection Over Union
MIS	Medical Image Segmentation
MRI	Magnetic Resonance Imaging
MSP	Maximum Softmax Probability
MV	Majority Voting
NLCE	Non-Local Context Encoder
OOD	Out Of Distribution
PGD	Projected Gradient Decent
PPD	Parallel Partial Decoder
PRANet	Parallel Reverse Attention Network
RA	Reverse Attention
RELIVE	contRastivE muLtItasking adVersarial dEfence
ROI	Region Of Interest
SEViT	Self Ensembling Vision Transformer
SS	Spectral Signature
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TGANet	Text Guided Attention Network
TrustMedIS	Trustworthy Medical Image Segmentation
TTA	Test Time Augmentation
UACANet	Uncertainty Augmented Context Attention Network
UAD	Unsupervised Adversarial Detection
UQ	Uncertainity Quantification
VVC	Volume Variation Coefficient
wBCE	weighted Binary Cross Entropy
wIOU	weighted Intersection Over Union
XAI	eXplainable Artificial Intelligence

Chapter 1

Introduction

The revolutionary era of Artificial Intelligence (AI) has emphasised automation by enabling machines to imitate human comprehension, facilitate data-driven decision-making, and devise intelligent systems. Among its most effective subsets, Deep Learning (DL) [1] has emerged as a promising tool, aiding influential achievements. Inspired by the human brain, DL works in multi-layered neural networks to uncover intricate patterns from the extensive dataset. The availability of high-performing computing resources, usability of vast amounts of data, and development of cutting-edge learning algorithms are several key aspects for the great success of DL evolution [2]. This tremendous growth has led to breakthroughs in solving a plethora of life-critical healthcare problems, such as cancer diagnosis [3], tumour detection [4], heart rate monitoring and estimation [5, 6], pathology segmentation from coarse-level organs to fine-level cells [7], and clinical research. The DL in healthcare incorporates the knowledge from physician-level diagnostics to perform multiple crucial tasks such as classification, detection, segmentation, registration, drug discovery, clinical trial optimisation, and medical report analysis.

Medical Image Segmentation (MIS) [3] is one such important and challenging task wherein the disease-afflicted Region Of Interest (ROI) is localised and differentiated from

the healthy region. It involves segmenting anatomical structures [8], including vital organs such as the liver, brain, lungs, and heart, as well as identifying lesions and tumors like lung nodules [9] and skin lesions [10]. Moreover, it encompasses the segmentation of cells and tissues [11], such as cancerous cells in biopsy images, and integrates multi-modality data [12] by combining information from various imaging modalities like Computed Tomography (CT) scans and Magnetic Resonance Imaging (MRI). The DL evolution has developed several plausible MIS models, including Convolution Neural Network (CNN) based U-Net [13], the Transformer-based SSFormer [14], and the hybrid CNN-Transformer-based U-NetR [15]. These models significantly assist clinicians in improving diagnostic accuracy and treatment planning. Specifically, these can be utilised in segmenting polyps to detect colon cancer [3], heart localisation for cardiac disease treatment [16], identifying lungs for early diagnosis of COVID-19, pneumonia, and lung cancer [17], detecting tumour present in the brain [18].

Despite their critical importance, the DL-based MIS models are severely disrupted by several challenges, leading to their questionable acceptability in real-world scenarios. In essence, medical images are inherently difficult to process [19] due to excessive texture bias, complex anatomical structure, scarcity of expert-annotated data, class imbalance issues, and variations across imaging modalities. Moreover, these models are prone to adversarial attacks [20] and behave erroneously in the presence of anomalous samples generated by adversarial perturbations and distribution shifts in input, ultimately compromising their performance and robustness. Furthermore, the opaque nature of such models offers non-trustworthy results [2] that cause conflicts in end-users over accepting the model. All these challenges are briefly discussed in Section 1.1.

1.1 Research Gaps and Motivations

This thesis aims to overcome several research gaps encountered in the DL-based MIS model to elevate its performance, robustness, and trustworthiness. These gaps pose significant challenges that hinder the applicability of such models in real-world instances. All these research problems and motivations behind our thesis are broadly explained in the following subsections:

1.1.1 Difficulties in Medical Image Processing

In any DL model, the processing of medical images is way more difficult than that of natural images. Medical images often struggle with *texture bias* [21], wherein the DL-based MIS model is highly dependent on fine-grained textural features rather than meaningful shapes or structural information. Unlike natural images, where the images have well-defined edges and shapes, medical images often lack distinct edges or sharp boundaries [3], leading to redundant learning of textural noises. This results in a lack of generalisation and degradation in model performance due to overfitting on image-specific artefacts. Another major challenge with medical images is *data scarcity* [21], which affects the model performance due to restricted data availability. This problem arises from several factors, including limited expert-annotated datasets [22], strict restrictions on data sharing for patient privacy and ethical concerns (HIPAA[1], GDPR[2]) [23], class imbalance due to small lesions/tumors [24], the occurrence of rare diseases [25], and cost-effective high-resolution medical images [26].

Medical images also suffer from extensive *variation across imaging modalities* [27], such as differences in MRI, CT, and ultrasound data. Moreover, their anatomical structure changes across different patients and distinct diseases may present with distinct symptoms.

[1] https://www.cdc.gov/phlp/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html

[2] https://gdpr-info.eu/

Unlike natural images, medical images are independent of orientation and consider absolute pixel value as they hold vital diagnostic information rather than taking their relative measure. The scaling or batch normalisation of such pixels may lose critical features, which ultimately degrades the model performance [21]. Thus, it is essential to develop a novel DL-based MIS model that addresses all these aforementioned challenges to significantly enhance performance and drive clinically robust DL solutions.

1.1.2 Vulnerability to Adversarial Attacks

As the DL-based MIS model is extensively employed in several life-critical healthcare applications, their security is a major concern. Unfortunately, these models are sensitive to intelligently curated adversarial attacks [20], which are deliberately crafted by imposing small and imperceptible perturbations into the input, misleading the model predictions [28]. This challenge is particularly pronounced in healthcare, where medical images heavily depend on textural features. These features can inadvertently shift the model's attention to unintended ROIs, resulting in inaccurate predictions [29]. For instance, disease detection DL models can be fooled to perform misclassification (refer Figure 1.1), which thereby results in severe distress and serious implications [29]. Consequently, active research is carried out to understand and mitigate the generation of adversarial attacks so that secure and robust DL models can be designed [28] with minimal performance degradation. In this direction, adversarial attacks are popularly studied for Medical Image Classification (MIC) to understand and address security vulnerabilities [30], but their impact on DL-based MIS models still requires thorough investigations. Thus, attacking MIS models is more challenging than MIC models, as MIC needs only one target to be misclassified for a successful attack, whereas MIS considers perturbing every pixel in the input image. Further, the adversarial attack involves backpropagation of the loss function. However, their effectiveness can

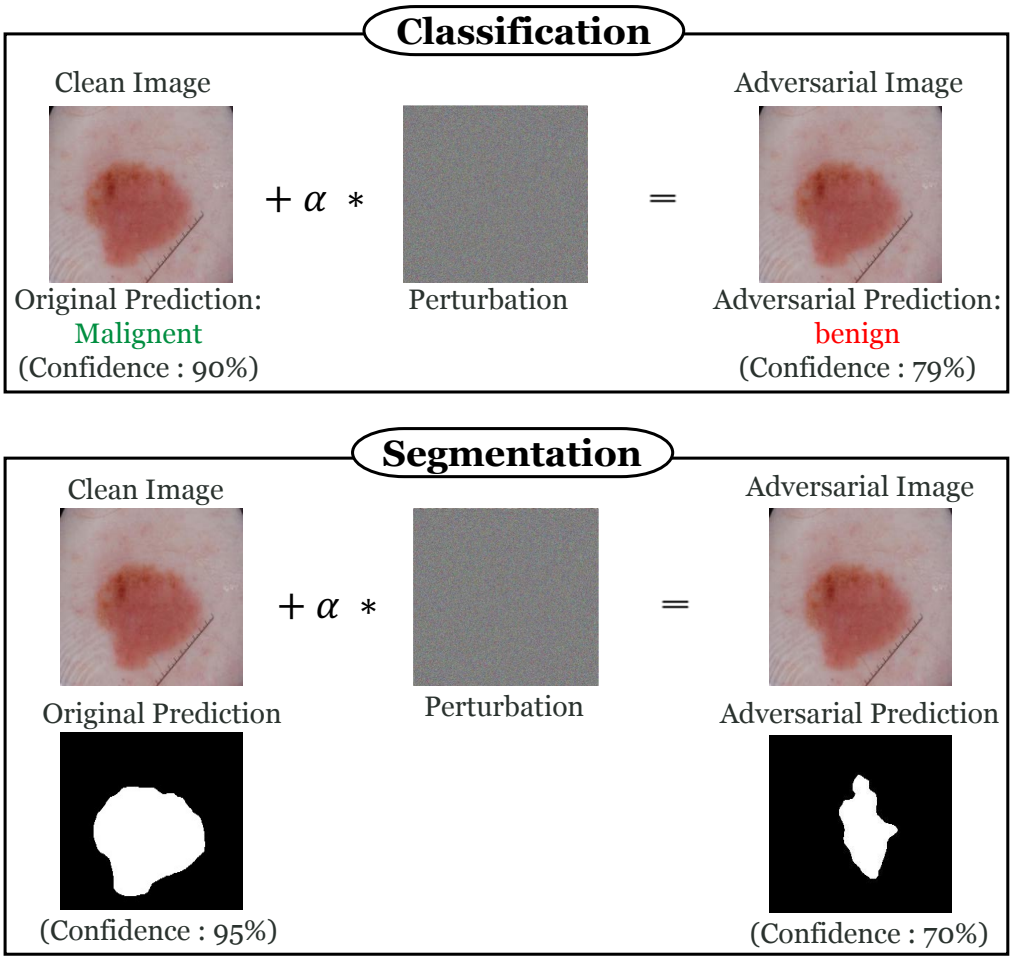

Figure 1.1: Illustration of an adversarial attack, deceiving a DL-based skin-lesion model. **Classification:** A *Malignant* scar is misclassified as *Benign* with high confidence. **Segmentation:** The model fails to preserve the shape and boundary of the diseased ROI. α is a perturbation multiplier, ensuring small perturbation.

be diminished by inhibiting backpropagation using either a non-differentiable neural network layer [31] or a non-differentiable loss function [32]. Specifically, the DL-based MIS models are extensively affected by such non-differentiability issues, which remain underexplored in existing literature. Conversely, the DL-based MIC models exhibit differentiable layers and incorporate differentiable loss functions for successful attacks [33].

The issue of non-differentiable loss functions in MIS is resolved by surrogate loss functions [34], which are differentiable approximations of the true loss functions. However, these approximations do not always guarantee successful attacks, as different losses be-

have uniquely for the same input and target. Thus, the selection of an optimal surrogate loss function is a significant challenge for a successful attack operation. This drives the development of a novel adversarial attack for the DL-based MIS model that tackles the non-differentiability constraints and chooses an optimal surrogate loss function for effective attack operations. Thus, understanding the development of adversarial attacks is important to examine their efficacy on DL-based MIS models so that one can defend them in order to offer a robust DL solution.

Existing adversarial defences primarily focus on increasing training samples to alleviate the impact of adversarial perturbations. In particular, Data Augmentation (DA) [35] incorporates training samples with their geometrical and pixel-level variations, while Adversarial Training (AT) [36] includes the adversarial samples into training. Unfortunately, AT fails to retain model performance due to extensive dependence on adversarial features, struggling to generalise well on clean samples [37]. Similarly, DA struggles to capture adversarial features as it heavily relies on augmented patterns, restricting the model's robustness. Nevertheless, both of these defences can mitigate the adversarial perturbation impact up to some extent, particularly in non-medical applications.

Since contrastive learning [38] enforces the model to behave consistently across clean and synthetically altered samples, showing the potential to substantially reduce the impact of adversarial attacks. However, this investigation for DL-based MIS models remains unexplored. Likewise, multitask learning [39, 40] strengthens adversarial robustness by training a model on a primary task alongside carefully chosen auxiliary tasks. However, improper auxiliary task selection may create a false sense of robustness, and its impact in the medical domain still requires deep exploration. All these aforementioned challenges become motivations to propose an effective defence that can leverage the benefits of multitasking and contrastive learning to significantly diminish the efficacy of adversarial attacks on the

DL-based MIS model without compromising model performance.

1.1.3 Confronting Anomalous Samples

DL-based MIS models yield incorrect results when faced with anomalies like adversarial or OOD samples. While adversarial samples are produced by imposing well-crafted imperceptible perturbations into the input [31], OOD samples are test samples whose distributions are misaligned with the training samples [41, 42]. Consequently, both types of samples significantly impair the performance and robustness of DL-based MIS models, leading to catastrophic consequences in healthcare. The rich textural features of medical images sometimes distract the model to emphasise non-ROIs [29], on which the impact of perturbations is too high, which causes the model to mislead. Likewise, the OOD effect disrupts model performance by exhibiting unfamiliar distributions with the training sample. Thus, it is crucial to precisely detect such anomalous samples that can mitigate their adverse impact on DL-based MIS models.

Several existing detection methods have greatly centred on DL-based MIC models, often requiring network retraining [43]. However, detection methods developed for MIC tasks often exhibit limited generalisation when applied to MIS tasks, offering inadequate outcomes [44]. Moreover, all the existing methods primarily emphasise either probability distribution [45, 46, 47] or a distance-based approach [43, 48]. The probability distribution-based method initially extracts the feature maps from clean samples and examines their probability density function. It observes the reliance of test sample distribution on this function, rejects those samples that show deviation in distribution, and classifies them as anomalous (adversarial/OOD) [45]. However, such methods are highly dependent on network logits, exhibiting similar characteristics for clean as well as adversarial and OOD samples, thus restricting detection efficacy [49, 48].

Similarly, the distance-based detection methods calculate diverse statistical measures such as the covariance matrix and mean from the extracted features of training samples. These measures are then utilised to compute the Mahalanobis Distance (MahD) between test sample features, following the comparison of this distance with a prespecified threshold to effectively determine whether a sample is clean or anomalous [43, 48]. Unfortunately, this method requires the additional burden of hyperparameter tuning the prespecified threshold, which is challenging and restricts detection performance if not optimised correctly [2]. To the end, there is a crucial requirement for a unified detection method that does not incorporate probability or distance-based approaches and precisely distinguishes clean and anomalous (adversarial/OOD) samples, specifically in a DL-based MIS model.

1.1.4 Trustworthiness Issue

The DL models exhibit a non-transparent nature that leads to distrusted predictions and limits their applicability. This raises a concern about the trustworthiness of DL-based MIS model prediction [50] (refer Figure 1.2). A non-trustworthy prediction could be erroneous and affect the model's performance, whereas a trustworthy prediction fosters confidence among end users, improving the adaptability of the model. However, assessing trustworthiness or determining model performance requires Ground Truth (GT), which is unavailable during testing, making the issue difficult. To resolve this, several empirical studies [51, 52, 53, 54, 55] quantifies uncertainty and detect OOD samples. OOD samples arise when test distribution is not properly associated with training distribution. Conversely, In-Distribution (ID) samples are well aligned with training distribution and expected to yield correct results. Unfortunately, a decline in performance has been observed in ID samples due to insufficient learning of the model, called *epistemic uncertainty* or presence of redundant noise in training data, termed *aleatoric uncertainty* [56]. Both of these uncertainties

could provide unjustified outcomes, alarming the requirement for a deeper exploration of trustworthiness problems in ID samples.

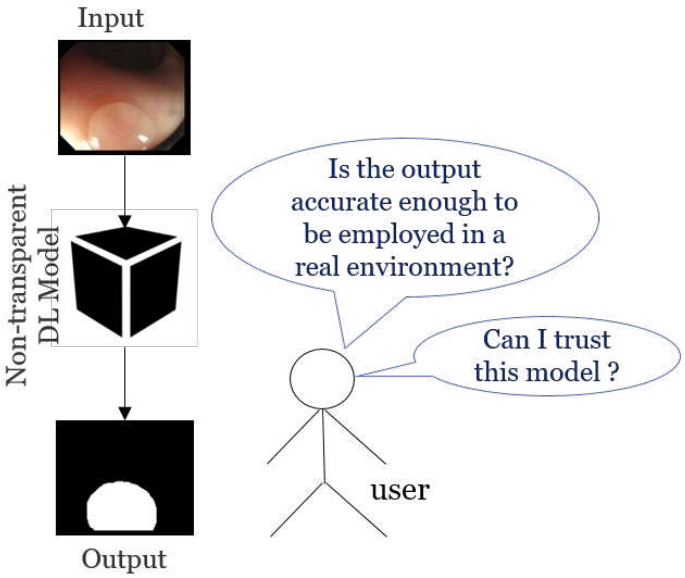

Figure 1.2: User's conflict over trusting the output of the non-transparent DL model.

A trustworthy MIS model detects pitfalls with non-trustworthy predictions, helping to prevent potentially adverse faults in the disease diagnosis process. Further, most of the existing UQ methods incorporate pixel-wise measures wherein trustworthiness is estimated across each pixel in the image. In contrast, research on measuring structure-wise trustworthiness (corresponding to an image's ROI or object) is still limited [57]. Thus, these research gaps motivate us to propose a novel method that can effectively investigate the structure-wise trustworthiness in DL-based MIS models for ID samples, along with enhancing the model's performance and robustness. This research will surely open a novel research direction in the areas of explainable artificial intelligence (XAI) [58, 59], contributing to the improved reliability and deployment of MIS in critical healthcare applications through responsible decision-making.

1.2 Contributions

To address all the aforementioned research gaps (kindly see Section 1.1), the objective of this thesis is to strengthen the DL-based MIS models by developing adversarially robust and trustworthy solutions while simultaneously advancing the model performance. In this direction, our first contribution introduces a novel adversarial attack on the DL-based MIS models with the aim of comprehending the efficacy of such attacks on these models. This work overcomes challenges related to non-differentiability and the selection of surrogate loss functions. Its detailed description is provided in Section 1.2.1. In our second contribution, we propose a unified anomaly detection method that differentiates between adversarial/OOD and clean samples in DL-based MIS models. It performs consistency analysis between input and their variants, eliminating the utilisation of GT and network retraining. Kindly refer to Section 1.2.2 for further information about our proposed detection method.

After getting the efficacy of adversarial attacks and identifying anomalous samples, our next objective is to defend against these attacks for DL-based MIS models. To the end, our third contribution proposes a novel adversarial defence, advancing the DL-based MIS models' robustness with negligible reduction in their performance. This work integrates contrastive and multitask learning, leveraging their combined strengths for a more effective defence. A thorough explanation of the proposed defence is demonstrated in Section 1.2.3. Likewise, our final contribution focuses on investigating the trustworthiness and enhancing the MIS model performance. For this purpose, we introduce a method that examines the characteristics of input and output, followed by computing consistency to determine the trustworthiness of the MIS prediction. It also refines the non-trustworthy predictions and selects the optimal MIS model, which offers the most trustworthy prediction. Kindly refer to Section 1.2.4 for an in-depth depiction of this work.

1.2.1 Understanding the efficacy of adversarial attacks in DL-based MIS models

To assess the impact of the adversarial attack on DL-based MIS models, we propose a novel attack called ***DECEIT*** (***D***ynamic Loss S***E***le***C***tion based Adv***E***rsar***I***al A***T***tack). The presence of non-differentiable layers and non-differentiable loss functions in MIS networks obstruct adversarial attacks by hindering the backpropagation process. Moreover, diverse surrogate loss functions possess varying behaviour for the same input and target, limiting attack capability. Thus, ***DECEIT*** overcomes these challenges to effectively attack the existing DL-based MIS models. Below are the primary contributions of ***DECEIT***:

1. Our proposed attack, ***DECEIT***, is specifically designed to attack DL-based MIS models. Attacking the MIS model is particularly challenging, as it requires altering each pixel in the image to effectively mislead the MIS outcomes.

2. ***DECEIT*** considers multiple surrogate loss functions and dynamically chooses the optimal one, which ensures a successful attack with minimum addition of perturbation. For this purpose, it performs parallel fusion to successfully attack the DL-based MIS model.

3. Usually, MIS models consist of non-differentiable preprocessing layers that disrupt gradient backpropagation, ultimately restricting adversarial attacks. Thus, ***DECEIT*** outlines several methods for approximating non-differentiable layers using differentiable functions, enabling a successful attack.

1.2.2 Detecting adversarial and OOD samples in DL-based MIS models

The adversarial and OOD samples degrade the performance of MIS models. To identify such samples, we propose a novel detection method called ***DISCERN***, which accurately distinguishes clean and adversarial/OOD samples from clean samples across DL-based MIS models. Without requiring network retraining and GT, it simply observes the characteristics of input and output, followed by assessing the consistency (or similarity) between the MIS predictions of input and their respective variants. Notably, variants indicate the rotated versions of the input sample. Hence, strong and high consistency is observed when the input is a clean sample, signifying mutually coherent characteristics. However, this consistency drops for the input as adversarial and OOD samples, showing non-coherent behaviour. Leveraging these observations, our proposed method, ***DISCERN***, performs consistency-based analysis for identification of adversarial and OOD samples. The core contributions made by ***DISCERN*** include:

1. ***DISCERN*** identifies adversarial samples by analysing the similarity (or consistency) between MIS predictions of input and their corresponding rotated variants. Since the MIS model is learned to offer rotation-agnostic predictions, it is assumed that clean samples' predictions and their respective variants depict similar characteristics with strong consistency [2]. Conversely, this assumption is breached when input is an adversarial sample, where a noticeable drop in consistency is observed between the adversarial samples' predictions and their related variants, signifying dissimilar behaviour. This occurs because rotation can diminish the impact of adversarial perturbation [60].

2. ***DISCERN*** can also distinguish the OOD samples from clean ones by assessing the

input and output characteristics to measure consistency in MIS predictions. It draws upon the observation that OOD samples exhibit dissimilarity with their respective variants as they tend to provide random predictions. This results in less similarity (or consistency) between the OOD samples' prediction and their respective variants.

3. ***DISCERN*** is a unified method designed specifically for MIS tasks to detect both types of anomalous samples, which are adversarial and OOD. In contrast to empirical studies, it does not rely on any probability distribution or threshold-based distance approach. Additionally, it does not incorporate GT and network retraining throughout the detection process.

1.2.3 Defending against adversarial attacks in DL-based MIS models

To alleviate the efficacy of adversarial attacks on DL-based MIS models, we devise an adversarial defence called ***RELIVE***, which stands for Cont***R***astiv***E*** Mu***LtI***tasking Ad***V***ersarial D***E***fence, with the aim of enhancing adversarial robustness of DL-based MIS models without compromising their performance. It employs clean, adversarial, and augmented samples and performs contrastive learning, which enforces the model to learn similar features of such samples, ultimately advancing model robustness. Moreover, it introduces a multitask model wherein the selection of auxiliary tasks depends upon their lower correlation to the main task, which benefits in enhancing adversarial robustness. Furthermore, it proposes a contrastive multitasking model that is built by fusing the proposed multitask model with contrastive learning, effectively diminishing the effect of adversarial perturbation while mildly improving the model performance. The major contributions of ***RELIVE*** are as follows:

1. Our novel defence, ***RELIVE***, strengthens the adversarial robustness of DL-based MIS

models while achieving a marginal performance gain. It first determines the individual importance of contrastive and multitask learning, followed by consolidating these approaches to mitigate the implications of adversarial attacks on model performance. Notably, it is the first MIS-specific approach that combines contrastive learning, multitask learning, and their fusion to improve adversarial resilience.

2. While some non-medical studies utilise multitask learning to counter adversarial attacks, effective mitigation occurs only when the auxiliary and main tasks have low correlation. To address this, our proposed defense, ***RELIVE***, incorporates a multitask model specifically for MIS, ensuring the careful selection of auxiliary tasks. This strategic selection enhances adversarial robustness while simultaneously improving overall model performance by providing generic feature representation.

1.2.4 Investigating trustworthiness and improving performance of DL-based MIS models

Due to the opaque nature of DL models, their predictions often lack trust, which limits their adaptability. To address this, we introduce a method called ***TrustMedIS*** (***Trust***worthy ***Med***ical ***I***mage ***S***egmentation), which employs ID samples and performs input-output behavioural analysis to compute the confidence measure between the input and their corresponding variants. ***TrustMedIS*** operates in three folds: *ET* (*E*xamining *T*rustworthiness), *ENT* (*E*levating *N*on-*T*rustworthy predictions), and *CSM* (*C*lassifier *S*election *M*ethod). The *ET* investigates the DL-based MIS models' trustworthiness. The *ENT* enhances the performance of non-trustworthy predictions. Given that the MIS performance of an individual model is often constrained, it can potentially produce erroneous results and fall short of clinical demands. Thus, this limitation can be addressed by incorporating several MIS models [61]. Building on this insight, our proposed *CSM* method considers multiple MIS

models and selects the optimal one, offering better predictions. The key contributions of ***TrustMedIS*** are summerised as follows:

1. The proposed *ET* method evaluates structure-wise trustworthiness of the DL-based MIS models employing ID samples. It measures the consistency (or similarity) between MIS predictions of input and their variants by leveraging the insight that an input image and its invert-rotated variant should produce similar predictions. Consequently, it identifies whether the MIS prediction is trustworthy or not.

2. Our novel *ENT* method refines the non-trustworthy prediction, benefitting from the insight that an input image that initially produces erroneous MIS prediction can be corrected by the same model simply by the rotation operation.

3. Our proposed *CSM* method improves model effectiveness by evaluating the trustworthiness of predictions from multiple MIS models and selecting the optimal one that provides the most reliable outcome.

1.3 Organisation

The thesis is divided into seven chapters, as depicted in Figure 1.3. Their organisation are as follows,

Chapter 1 (the current chapter) provides a detailed introduction to DL models and their significance in MIS. It then highlights key research gaps and motivations concerning the performance, robustness and trustworthiness of DL-based MIS models. Finally, the chapter outlines the thesis contributions and its overall organisation.

Chapter 2 presents a review of related work on DL-based MIS models. It offers a concise literature survey covering adversarial attacks, adversarial and OOD sample detection, adversarial defences, and trustworthiness in these models.

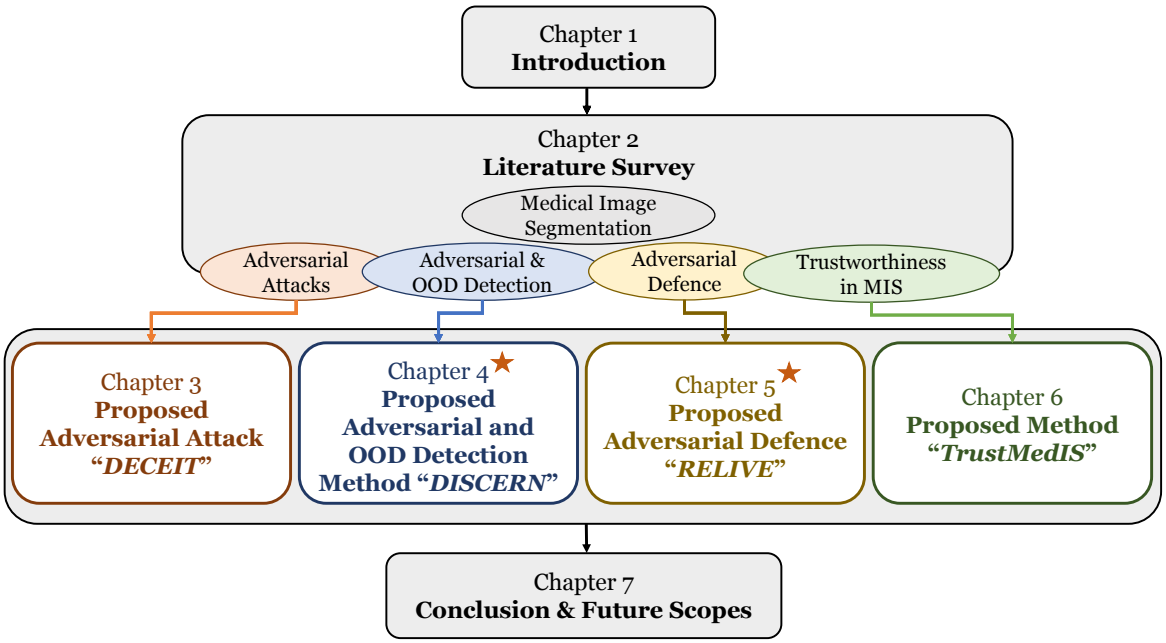

Figure 1.3: Thesis Organisation.

Chapter 3 describes the comprehensive working of our proposed adversarial attack, ***DECEIT***, on DL-based MIS models. It broadly explains the resolution of non-differentiability issues, the selection of surrogate loss function, and the parallel fusion technique. The chapter also presents a detailed experimental evaluation of the attack, conducted on open-sourced datasets across multiple existing MIS models.

Chapter 4 delivers a thorough analysis of our proposed unified detection method, ***DISCERN***, in DL-based MIS models. The chapter thoroughly describes consistency-based adversarial and OOD detection, highlighting their experimental results, including comparative analysis and ablation studies on publicly available datasets across multiple cutting-edge MIS models.

Chapter 5 presents the proposed adversarial defence, ***RELIVE***, for DL-based MIS models. It provides deep insights into contrastive learning, multitask learning and their consolidation-based defence. Additionally, the chapter includes a comprehensive exper-

imental evaluation of the defence, tested on publicly available datasets for several MIS models.

Chapter 6 offers the detailed working of the proposed method, ***TrsutMedIS***, for ID samples in DL-based MIS models. The chapter explains the meticulous functioning of the proposed *ET*, *ENT* and *CSM* methods, emphasising their significance in developing trustworthy MIS models. Additionally, the chapter presents comprehensive experimental evaluations to examine the impact of the proposed method.

Chapter 7 presents the concluding remarks on the overall research work described in this thesis and highlights potential avenues for future research.

Chapter 2

Literature Review

This chapter provides a holistic survey of empirical studies relevant to this thesis. It covers existing works on MIS, emphasising their evolution from early Convolution Neural Network (CNN)-based architectures to recent transformer-based models, as depicted in Section 2.1. Moreover, it explores fundamental adversarial attacks and related studies that highlight their significance in the medical domain, as given in Section 2.2. The chapter also demonstrates the previous works on anomalous sample detection (refer Section 2.3) and adversarial defences (see Section 2.4) in DL-based MIS models. Furthermore, it offers an in-depth exploration of trustworthiness in MIS models, as described in Section 2.5.

2.1 Medical Image Segmentation (MIS)

Over the past few years, CNN models have been greatly employed for diverse visual recognition tasks in several domains. Unfortunately, they demand huge training data and complex networks [62], hindering their adaptability in various tasks, specifically in MIS, where data scarcity is a major issue (see Section 1.1.1). To resolve this problem, U-Net [13] is proposed, which is characterised as an extension of the CNN model [62]. U-Net exhibits a U-shaped encoder-decoder architecture and employs less data, still providing accurate MIS

prediction. The encoder captures global contextual knowledge, while the decoder upsamples the extracted feature representation. The spatial information at each stage is preserved by skip connections, which connect the encoder and decoder. However, there is a semantic dissimilarity observed in the feature map, which affects MIS performance. This issue is resolved by U-Net++ [63], which is a variant of U-Net [13]. U-Net++ exhibits multiple nested skip connections to learn local information.

Since the ROI structure in MIS predictions is complicated and exhibits unclear edges. U-Net++ [63] considers the complete polyp region for segmentation but ignores the constraints of areas and boundaries, ultimately degrading segmentation performance. To mitigate this issue, PraNet [3] is proposed, which stands for Parallel Reverse Attention Network. PraNet is specifically designed for segmenting polyps in colon parts to detect colorectal cancer, taking into account both the areas and the boundaries. It employs a Parallel Partial Decoder (PPD) to capture high-level extracted features for predicting the coarse area of the input image. Additionally, it incorporates Reverse Attention (RA) to achieve refined boundaries, offering accurate MIS results. The boundaries of segmentation masks are also emphasised by an active contour-based model [64] for MIS. It learns the blurred and concave boundaries by using scalable regional statistics and the Gaussian function. Removing such boundary blurriness will provide a precise segmentation mask.

Likewise, the UACANet (Uncertainty Augmented Context Attention Network) [65] emphasises uncertain regions of the saliency map which exhibit boundary information. UACANet follows U-Net architecture [13], employing several encoders and decoders to extract uncertain ROIs and feature aggregation modules to predict accurate MIS outcomes. Sometimes, the DL-based MIS model avoids small ROIs, which leads to class imbalance problems and is important for the detection of severe diseases. To the end, CaraNet [66] is proposed, which stands for Context Axial Reverse Attention Network. It employs an axial

reverse attention network, a channel-wise feature pyramid module, and a partial decoder to handle the segmentation of small medical objects. Since the medical images exhibit variation in structure, leading to model overfitting. Thus, SSformer [14] is proposed, which incorporates a pyramidal transformer module for the encoder to learn generic features and enhance generalisation. Additionally, it employs a partial locality decoder to focus on the local details, leading to advanced MIS outcomes. Notably, UACANet [65] and SSFormer are available in two versions: Standard (S) and Large (L), based on a variety of encoder scales.

However, the complexity of the model increases when working for size variation, background area, dense skip connections, and boundary curves. This will diminish the generalisation ability of the model by employing additional layers and connections in the network. Thus, TGANet (Text Guided Attention Network) [67] is proposed, which considers the number and size of polyps to learn features using text-guided attention. It extracts features in the encoder and performs multi-scale feature aggregation to achieve an accurate MIS outcome in the decoder. This will ultimately advance the model's generalisation and performance.

2.2 Adversarial Attack on Medical Imaging

The adversarial attacks are executed by imposing intelligently engineered and imperceptible perturbations on the input in such a way that the model provides incorrect predictions. The first and basic attack is the Fast Gradient Sign Method (FGSM) [68], which is a one-step attack. In FGSM, the process of producing adversarial samples involves adding a perturbation (ϵ) to the gradient of the loss function (J). Mathematically,

$$X_{adv} = X + \epsilon * \text{sign}(\nabla_X J\left(f(X), y\right)) \tag{2.1}$$

where $f(.)$ depicts the DL model (also called a threat model) that accepts X input and predicts y output. X_{adv} is the generated adversarial input. Unfortunately, the FGSM attack adds larger perturbations as it is a single-step attack. As a solution, it is employed in an iterative manner by imposing a small amount of perturbation at each iteration. Such an attack is known as the Iterative Gradient Sign Method (IGSM) [69], formulated as,

$$X^0_{adv} = X \tag{2.2}$$

$$X^{i+1}_{adv} = clip\left(X^i_{adv} + \epsilon * \text{sign}(\nabla_{X^i_{adv}} J\left(f(X^i_{adv}), y)\right)\right) \tag{2.3}$$

where the superscript of X indicates iteration count and $clip$ limits the reliance of X_{adv} in specified range [20]. Another popular adversarial attack is Projected Gradient Decent (PGD) [70], which is a variant of the IGSM attack. In PGD, projection operation restricts the perturbation to not surpass some prespecified limits.

There can be two settings of adversarial attack, which are targeted and untargeted. The untargeted attack emphasises just misleading the prediction, while the targeted attack focuses on fooling the model by enabling it to predict some prespecified new target. Nevertheless, the objective of both of these attack settings is to enforce the model to offer inaccurate predictions. Kindly note that Equation (2.3) represents the mathematical expression of an untargeted attack. For targeted attacks, we modify Equation (2.3) with prespecified target y_t as,

$$X^{i+1}_{adv} = clip\left(X^i_{adv} - \epsilon * \text{sign}(\nabla_{X^i_{adv}} J\left(f(X^i_{adv}), y_t)\right)\right) \tag{2.4}$$

Existing work [29] reveals that DL-based models are more prone to adversarial attacks in the medical domain than in the non-medical domain, as the medical images exhibit more texture knowledge (see Section 1.1.1). This work [29] applies an adversarial attack on the MIC model with the aim of fooling the diagnoses of diabetic retinopathy from fundoscopy

images, thorax diseases from chest x-ray images, and melanoma from dermoscopic images. Likewise, [71] applies the FGSM attack to fool lung classification. In contrast to the MIC-based adversarial attacks, only a few studies have been conducted for MIS-based adversarial attacks. The reason is that the MIS architecture usually demands high computational preprocessing that leads to non-differentiable neural network layers. Moreover, some MIS error metrics (or loss functions), such as Intersection Over Union (IOU), are non-differentiable. This non-differentiability problem can hinder the backpropagation and ultimately block the adversarial attack (see Section 1.1.2). To address this problem, surrogate loss functions are utilised for MIS tasks, which are characterised by the differential approximation of the true loss function [32]. However, this approach restricts the applicability of the attack as multiple surrogate loss functions offer variation in output for the same input-target pair.

Nevertheless, some MIS-based adversarial attacks have incorporated differentiable layer-based models. However, such attacks have minimal effectiveness as they utilise a single surrogate loss function to address the problem of non-differentiability. In this direction, Adaptive Segmentation Mask Attack (ASMA) [72] is devised for fooling the glaucoma optic disc and skin lesions segmentation. The MIC-based attacks focus on misleading the single output label, whereas MIS-based attacks alter the multiple output pixels. Since IGSM [69] is a MIC-based attack and considers only a single target, it is not applicable directly to segmentation. Therefore, ASMA [72] modifies IGSM by incorporating all the pixels with distinct prediction and target labels. Likewise, Dense Adversarial Generation (DAG) [73, 74] is another MIS-based attack that is built on modifying IGSM. In [75], a modified version of FGSM is proposed, called inverse FGSM, aiming to attack brain tumour segmentation.

2.3 Adversarial and OOD Sample Detection in Medical Imaging

Existing studies have detected the adversarial and OOD samples based on two types of approaches: (1) assessing the probability distribution of features [45, 46, 47], (2) adapting the distance-based approach [43, 48]. In this direction, the Unsupervised Adversarial Detection (UAD) method [45] initially predicts the probability distribution map of clean sample features using the Gaussian Mixture Model (GMM) [76]. It then observes the reliance of test sample distribution on this training distribution map during inference and classifies those test samples as adversarial, whose distribution does not exhibit on the training distribution map. Likewise, the Self Ensembling Vision Transformer (SEViT) [77] is another adversarial detection method that leverages the observation that the final ViT [78] layer predictions are severely infected by adversarial attacks rather than the intermediate layer predictions in the initial blocks of ViT. To detect adversarial samples, SEViT considers that adversarial sample predictions vary in nature, while clean samples exhibit consistency in predictions.

To simultaneously identify adversarial and OOD samples, [43] computes the Mahalanobis Distance (MahD) score of features and compares it with the prespecified threshold for efficient detection. The OOD samples are identified by observing the spectral characteristics of the extracted feature maps [46]. These spectral characteristics are called Spectral Signature (SS), which detects OOD by calculating the distance between the SS of samples and classifying it using a prespecified threshold. The Maximum Softmax Probability (MSP) [47] identifies OOD by leveraging the observation that OOD samples tend to have smaller values of maximum softmax probability than ID samples. Similarly, the Evidence Reconciled Neural Network (ERNN) [79] seeks to determine OOD samples that are near the ID samples. Such OOD samples depict similar characteristics to the training samples but

contain diverse distributions.

The probability distribution-based methods restrict the detection ability, as they are often dependent on network logits, showcasing similar properties for clean samples even when the input is adversarial or OOD samples [49, 48]. Moreover, a higher structural similarity is observed in clean and adversarial samples because the adversarial sample possesses subtle and imperceptible perturbations. Consequently, their probability distributions exhibit no substantial differences, which leads to degradation in the performance of the detection method. Furthermore, the detection methods based on distance computation require a prespecified threshold, which is difficult to calibrate and ultimately restricts detection capabilities [2]. Additionally, some existing detection [43] emphasises only MIC tasks, requiring network retraining. However, such methods fail to convert effectively on MIS, affecting detection performance [44].

2.4 Adversarial Defences

To boost the robustness of DL-based MIS models against adversarial attacks, several defences have been proposed. It includes the Data Augmentation (DA) [80] method, wherein geometric and pixel-level transformed images are initially generated from clean samples, which are further utilised along with clean samples to train the model. This defence is employed in the medical domain [35], which performs JPEG compression for generating augmented samples. Unfortunately, such DA-based defences excessively rely on geometric and pixel-level transformed (or augmented) images, which are distinct from perturbation-added (or adversarial) images. As a result, they struggle to effectively capture adversarial features, leading to offering inadequate robustness results. Another popular defence is Adversarial Training (AT) [81] wherein the model is trained by incorporating adversarial samples along with clean samples. [36] and [71] utilise AT with the medical images. Since the

AT is primarily centered on learning adversarial features by the model, although it enhances adversarial robustness, it fails to maintain model performance [37].

Thus, a variant of AT is proposed, called Adaptive Margin Adversarial Training (AMAT) [82], which simultaneously processes an equal number of clean and adversarial samples and averages their respective losses to train the model. However, it is unable to substantially reduce the effect of adversarial perturbation on the model but retains model performance. Likewise, the Non-Local Context Encoder (NLCE) [83] defence captures short- and long-range spatial dependencies to learn global contexts, empowering feature activation using channel-wise attention to enhance the adversarial robustness of MIS models. Notably, the defences [83] and [71] are restricted to only CNN-based architectures. Since medical images have texture richness, blurry edges, and complex structure, several non-medical defences cannot be employed in the medical domain [19].

Multitask learning [39] is one such non-medical defence that mitigates the efficacy of adversarial perturbation by performing model training with the main task along with multiple auxiliary tasks. Unfortunately, it randomly selects the auxiliary tasks for model training that could adversely impact the model's performance and robustness. An auxiliary task selection should be objective-specific and correspond to the main task for advanced robustness. Likewise, contrastive learning [38] has also been used as a defence to reduce the impact of the adversarial attack on the model. It compels the model to learn similar features for clean and adversarial samples using contrastive loss. The contrastive learning defence is further explored by Adversarial Robust Learning (ARL) [84], which applies contrastive loss between pairs of adversarial and augmented samples to train the model. Since ARL avoids clean samples for training, the model struggles to learn clean sample features, which is mandatory to preserve model performance while enhancing adversarial robustness.

2.5 Trustworthiness in MIS

The trustworthiness of MIS models is examined by numerous existing Uncertainty Quantification (UQ)-based pixel-level measures [52], which is defined as the pixel-wise predictive probability distribution [85, 86]. Such distributions are optimised by network retraining using GT, which results in average optimisation if the model provides inaccurate outcomes. Hence, [87] ignores network retraining and iteratively applies the input to a perfectly trained CNN model, followed by some dropout layers. Similarly, [57] utilises Test Time Augmentation (TTA) in which uncertainty estimation is performed using entropy between the model's output and its respective augmented variants.

The Lung Cancer Data Augmentation Ensemble (LCDAE) [88] performs DA in several ways, incorporating six diverse fine-tuned transfer learning models to elevate the performance of MIC models with significantly large confidence values. Further, the pixel-based UQ metrics are also incorporated to detect OOD samples in MIS [54]. Several existing works exhibit structure-level measures to examine the trustworthiness of the model employing ID samples. The Volume Variation Coefficient (VVC) [57] executes it by computing the standard deviation between the volume of models' output and its respective augmented variants. When a comparison is performed between these two predictions, they exhibit similar volumes even though they are characterised as diverse shapes and their ROIs occupy different spatial locations, leading to high uncertainty. However, VVC ignores these crucial insights, thereby lowering the uncertainty. Additionally, VVC neglects to estimate uncertainty constrained to a finite range, ultimately failing to enhance MIS performance.

Chapter 3

Adversarial Attack on MIS models

Despite the critical importance of DL models in healthcare applications, they are vulnerable to well-crafted adversarial attacks. The effectiveness of such attacks depends on loss function backpropagation. Accordingly, these attacks are prevented by prohibiting the backpropagation with a non-differentiable layer in the network [31] or a non-differentiable loss function [32]. Unfortunately, these problems are commonly encountered in MIS. As a result, there have been limited studies suggested in this regard. Conversely, the majority of DL-based MIC models allow several differentiable loss functions and are made up of differentiable layers. Therefore, extensive research has been conducted on adversarial attacks targeting classification models [33]. The problem of non-differentiability in the loss function for MIS is addressed by employing an approximated actual loss function, which is differentiable, called the surrogate loss function [34]. This approximation is limited to only a few scenarios of adversarial attack, as it exhibits distinct behaviour for similar input-target pairs. Consequently, selecting the right surrogate loss function is vital for conducting effective adversarial attacks.

This chapter introduces an adversarial attack for MIS, depicted as ***DECEIT***, which stands for ***D***ynamic Loss S***E***le***C***tion-based Adv***E***rsar***I***al A***T***tack. It addresses the challenges

posed by the following two factors: (1) surrogate loss functions, (2) non-differentiable neural network layers to successfully attack the models. The chapter is organised as follows: Section 3.1 presents our proposed attack, ***DECEIT***. Section 3.2 demonstrates experimental details related to the proposed attack. Section 3.3 outlines the discussion corresponding to the proposed attack. Section 3.4 summarises the overall chapter.

3.1 Proposed Adversarial Attack: *DECEIT*

This section presents the proposed attack, ***DECEIT***, aiming to mislead the output of DL-based MIS models for a given input. Figure 3.1 illustrates the flow diagram of ***DECEIT*** attack for a visual representation. Initially, it modifies the state-of-the-art MIS models (known as "threat models") by substituting the non-differentiable neural network layers with their differentiable approximation (refer Section 3.1.1). Subsequently, the modified threat model is adversarially attacked using a surrogate loss function. These losses mitigate the issue of non-differentiability in commonly used MIS losses like IOU. Unfortunately, their reliance on approximating true loss functions introduces variability in behaviour for the same input-target pair, limiting their universal applicability across adversarial attacks. To address this, the proposed attack, ***DECEIT***, employs several existing surrogate loss functions and devises a new loss, called *Compound Loss* (details are given in Section 3.1.2). Each of these surrogate loss functions is utilised to attack DL-based MIS models (refer Section 3.1.3). Thus, for a given input, multiple adversarial samples are generated across different considered surrogate loss functions. Eventually, a parallel fusion operation is performed to identify the optimal adversarial sample that achieves a successful attack with minimal perturbation (kindly see Section 3.1.4). These detailed steps are outlined in Algorithm 3.1.

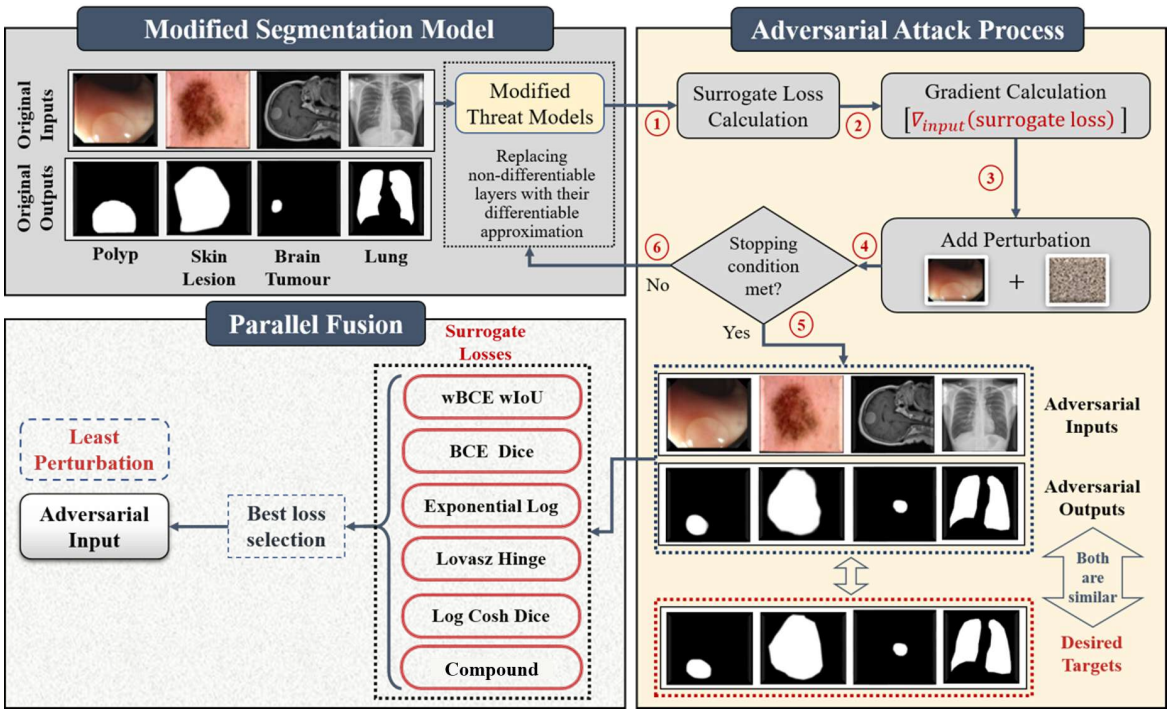

Figure 3.1: The complete work-flow of the proposed attack, ***DECEIT***. Initially, the threat model, exhibiting non-differentiable neural network layers, is modified by substituting these layers with their approximated differentiable layers. Subsequently, the modified threat model is then subjected to adversarial attacks using multiple surrogate loss functions. Eventually, parallel fusion is performed to choose the optimal surrogate loss function, enabling an effective attack with minimal perturbation.

3.1.1 Employing Differentiable Approximation Functions

Gradient calculation is required for the adversarial attack, which is possible only when the model exhibits differentiable layers. However, existing MIS models comprise multiple non-differentiable layers such as argmax, RGB to grayscale colour conversion, resizing, and normalisation. These layers obstruct the adversarial attack by obfuscating gradient backpropagation [31]. To address this problem, this subsection demonstrates the utilisation of differentiable approximations instead of non-differentiable layers, offering successful adversarial attacks.

Generally, MIS performs pixel-wise classification. Thus, a pixel's probability of falling into every class is calculated, and the class with the highest probability is allocated to that

Algorithm 3.1 Step-by-step description of the ***DECEIT*** attack.

Require: A set of surrogate loss functions; threat model; input image; desired target mask; a small step size for updates.

Ensure: A generated adversarial image that fools the threat model.

1: Replace non-differentiable layers in the threat model with differentiable approximations to obtain a modified model
2: **for** each surrogate loss function in the list **do**
3: Initialise iteration count to zero.
4: Initialise similarity score between prediction and target to zero.
5: **while** iteration count ≤ 10 **and** mIOU score ≤ 0.5 **do**
6: Update the input image by moving it in the direction that reduces the surrogate loss.
7: Compute the mIOU between the updated image's prediction and the target mask.
8: Increment the iteration count.
9: **end while**
10: Save the adversarial image generated using this loss function.
11: **end for**
12: From all generated adversarial images across different surrogate losses, choose the one that adds least perturbation.
13: Return this image as the final adversarial sample.

pixel. It is repeated for every input image's pixels using the argmax operation to obtain the final segmentation mask. Let n_p^c be the log probability of p^{th} pixel corresponds to c class, then the respective segmentation m_p for p^{th} pixel is demonstrated by:

$$m_p = \arg\max_c \left(n_p^c\right) \tag{3.1}$$

The outcome of argmax is a non-differentiable discrete categorical value, which offers zero gradients, thereby inhibiting the overall adversarial attack operation. As a solution, this non-differentiable argmax layer is substituted with a differentiable approximate layer for a successful attack. Inspired by [89] in which an approximation is made for classification, this is the first work that leverages a differentiable approximation layer for the adversarial attacks on segmentation models. Thus, we include Gumbel noise into the log class probabilities n_p^c

by following [89]. Mathematically, this can be expressed as,

$$q_p^c = n_p^c + g_p^c \tag{3.2}$$

where g_p^c is Gumbel noise at p^{th} pixel of c class, given by,

$$g_p^c \sim -\log(-\log(U(0,1)) \tag{3.3}$$

here $U(0,1)$ indicates the uniform distribution, exhibiting mean and variance as 0 and 1, respectively. The softmax layer can be employed as a differentiable approximation of the argmax layer with valid gradients, enabling smooth backpropagation during adversarial attack operations. Hence, we replace the argmax layer n_p^c with the softmax layer on the altered prediction of log probability q_p^c. The softmax layer applied to the p^{th} pixel of c^{th} class, thus, their mathematical output, s_p^c, is formulated as,

$$s_p^c = \frac{\exp\left(q_p^c/\tau\right)}{\sum_{c \in C} \exp\left(q_p^c/\tau\right)} \tag{3.4}$$

where τ depicts the non-negative scalar value, called *softmax temperature constant* (refer Section 3.2.3 for hyperparameter setting). The superscript C indicates the classes. We consider binary segmentation, which belongs to either class 0 or class 1. Consequently, the outputs acquired upon imposing the softmax layer are s_p^0 and s_p^1 corresponding to classes 0 and 1, respectively. It is noteworthy that the proposed substitution of argmax layers with approximated softmax layers in MIS models is executed for each pixel of the input, achieving the final segmentation mask.

Besides argmax, several other non-differentiable layers, such as colour conversion, resizing, and normalisation, are encountered in DL-based MIS models. To replace these layers, we used the open-sourced PyTorch library called *Kornia* [90], which contains numerous

differentiable modules to address the problem of non-differentiable layers in adversarial attack operations. An important criterion for models using approximated layers is that the performance of the modified and original MIS or threat models should be similar. We ensure this by comparing their performances before and after modifications (refer Section 3.3). Further, the resultant adversarial images produced by such modified threat models achieve successful attacks against the original threat models.

3.1.2 Exploring Surrogate Loss Functions

One of the crucial steps in an adversarial attack is loss computation. Unfortunately, many segmentation models exhibit non-differentiable loss functions, which result in zero gradient and fail to attack the model. As a remedy, MIS models employ surrogate loss functions [91], which are differentiable approximations of true loss, facilitating smoother attack operation. However, surrogate loss functions exhibit inconsistent behaviour with similar input-target pairs, limiting the effectiveness of attacks on MIS models. Our proposed attack, ***DECEIT***, overcomes this limitation by incorporating numerous widely recognised surrogate loss functions and introducing a novel loss function named *Compound Loss*. The specifics of these losses are detailed below.

3.1.2.1 Existing Surrogate Losses

The details of several existing surrogate loss functions are given below,

- **Log Cosh Dice Loss:** The non-convex nature of dice loss results in uncertain curvature with imprecise local minima that leads to inaccurate MIS predictions. The *Log Cosh Dice Loss* [91] is a dice loss variant that reduces the uncertainty in the loss curve and tackles the problem of class imbalance in MIS. It uses *log* to limit the range of the loss function and the hyperbolic function *cosh* to attain differentiability.

- **BCE Dice Loss:** It is a weighted addition of dice and Binary Cross Entropy (BCE) loss [91], leveraging their individual benefits. Dice loss emphasises foreground pixels, while its integration with BCE loss ensures that model training remains unaffected by background pixels. This enables the model to effectively capture the behaviour of small medical objects [92].

- **Exponential Log loss:** This loss is a blend of the exponential and logarithmic transformations of cross-entropy and dice loss [91]. It focuses primarily on weakly predicted pixels, strengthening them by leveraging the advantages of finer decision boundaries [93].

- **Weighted BCE and Weighted IOU loss:** This loss depicts the weighted combination of BCE and IOU loss [3]. The local restrictions of the prediction are managed by weighted BCE loss, while the weighted IoU loss focuses on global restrictions. Despite offering equal importance to every pixel, this loss prioritises weakly predicted pixels, strengthening the capability of the model to produce balanced outputs.

- **Lovasz Hinge loss:** It is a variant of the IOU loss function, exhibiting differentiable and tractable properties. To achieve a soft predicted mask, this loss encourages positive scores for the foreground pixels [94].

3.1.2.2 Proposed Loss : *Compound*

In order to achieve the combined effects of all the previously mentioned surrogate loss functions, we propose a novel surrogate loss function named, *Compound* loss. This loss is calculated as a weighted addition of the aforementioned surrogate losses. Let $l \in l_1, \cdots, l_k$ be all the considered existing surrogate loss functions, then the *Compound* loss $\mathcal{L}$ is mathe-

matically defined as,

$$\mathcal{L} = \sum_{j=1}^{k} w_j * l_j \tag{3.5}$$

where k signifies the total count of existing surrogate losses and $w \in w_1, \cdots, w_k$ are the assigned weights for each loss, reflecting its relative contribution to $\mathcal{L}$. Kindly refer to Section 3.2.3 for hyperparameter settings of w.

3.1.3 Developing Adversarial Attack

The DL-based MIS models are attacked by integrating imperceptible perturbations into the input image. In this direction, the adversarial samples are produced with respect to each considered surrogate loss function. The proposed attack, ***DECEIT***, is one of the variants of the existing adversarial attack, *IGSM* [69] (refer Section 2.2), in which untargeted and targeted attacks are achieved using equations (2.3) and (2.4), respectively. It is noteworthy that these equations need only one pixel as a target to attack, which is applicable to classification tasks. In contrast, the proposed attack, ***DECEIT***, fools segmentation models, which demand all pixels within an image to be attacked. Thus, for segmentation purposes, the equations (2.3) and (2.4) are altered as,

$$X_{adv}^{i+1} = clip\left(X_{adv}^{i} + \epsilon * \text{sign}(\nabla_{X_{adv}^{i}} J\left(f(X_{adv}^{i}), Y\right))\right) \tag{3.6}$$

$$X_{adv}^{i+1} = clip\left(X_{adv}^{i} - \epsilon * \text{sign}(\nabla_{X_{adv}^{i}} J\left(f(X_{adv}^{i}), T\right))\right) \tag{3.7}$$

where Y and T signify the MIS prediction and desired target, respectively. The process is repeated iteratively until a stopping condition is met, which occurs either when the mean IOU (mIOU) surpasses a predefined threshold or the highest number of iterations is reached. Notably, the adversarial images that satisfy the stopping condition will be categorised as

successful cases, while those that do not are deemed failure cases. All the hyperparameter settings related to adversarial attacks are provided in Section 3.2.3.

3.1.4 Performing Parallel Fusion

We develop multiple adversarial samples for a single input by applying different surrogate loss functions. We characterise the optimal adversarial sample as the one that induces the least perturbation while still achieving misguided predictions. Thus, we select the optimal adversarial sample by employing parallel fusion. It analyses the adversarial perturbations corresponds to each sample to choose one surrogate loss function that produces the least perturbed adversarial samples. Such adversarial perturbation is obtained by calculating the L_∞ distance using the adversarial and clean samples. Notably, for certain inputs, a surrogate loss function might not be able to attack the MIS model. In such situations, the resulting samples can not be regarded as adversarial samples. Thus, they are excluded in finding the optimal adversarial sample in parallel fusion. Similarly, there is a chance of getting identical adversarial samples from two different surrogate loss functions. Considering their equal importance, we select any of those samples. Suppose $\{l_1, \cdots, l_i\}$ are some of the surrogate loss functions that successfully attack the DL-based MIS model for input $\bar{X}$, offering respective adversarial samples as $\left\{\bar{X}_1, \cdots, \bar{X}_i\right\}$. Mathematically, the optimal adversarial sample $\bar{X}_s$ and the respective surrogate loss function l_s can be expressed as,

$$l_s = l_p \quad \text{and} \quad \bar{X}_s = \bar{X}_p \tag{3.8}$$

where the index p can be represented as,

$$p = \underset{j \in (1,\cdots,i)}{argmin}(\|\bar{X} - \bar{X}_j\|_\infty) \tag{3.9}$$

3.2 Experimental Results

3.2.1 Datasets and Metrics

We carried out experiments on open sourced datasets across four distinct healthcare applications: polyp segmentation, skin lesion segmentation, brain tumour segmentation, and lung segmentation. In this direction, we employed 602 images of the following four polyp datasets: CVC-300 [95], CVC-ClinicDB [96], CVC-ColonDB [97], and Kvasir [98]. Moreover, we used 600 dermoscopic images from the International Symposium on Biomedical Imaging (ISBI) Challenge 2017 (Part 1) dataset [10], 300 brain tumour images[1] and 100 chest X-ray images[2] for skin lesion, brain tumour, and lung segmentation, respectively.

We employed the following metrics to get the ***DECEIT***'s efficacy on DL-based MIS models.

1. **Attack Success Rate (ASR):** It refers to the fraction of input samples that were successfully attacked relative to the total sample count.

2. **Average Distortion (δ_∞):** It is measured by taking the average of L_∞ distance between the clean and adversarial sample across all successfully generated adversarial samples.

3.2.2 Threat Models

We measured the efficacy of ***DECEIT*** on several well-known pre-trained MIS models, referred to as threat models. In this direction, we considered PraNet [3] [3] and TGANet [4][67], both of which are specifically designed for polyp segmentation. PraNet focuses on

[1] https://github.com/sdsubhajitdas/Brain-Tumor-Segmentation
[2] https://github.com/IlliaOvcharenko/lung-segmentation
[3] https://github.com/DengPingFan/PraNet
[4] https://github.com/nikhilroxtomar/tganet

coarse area prediction, followed by refining the boundaries of the segmentation mask, while TGANet accounts for the number and size of polyps in the input image. In addition, we performed transfer learning [99] in PraNet by fine-tuning it for skin lesion segmentation, considering it a threat model. For brain tumour and lung segmentation, we utilised U-Net [13] as another threat model. U-Net processes input features by gradually downsampling them using an encoder and subsequently upsampling them to the original resolution with a decoder. It also employs skip connections to retain crucial feature information. It is worth noting that in our experiments, TGANet is attacked using CVC-ClinicDB [96] and Kvasir [98] polyp datasets.

3.2.3 Experimental Settings

The experiments are performed on a server equipped with an Nvidia V100 GPU and an Intel Xeon Gold 6132 CPU, along with 192 GB of RAM. Our proposed attack, ***DECEIT***, is examined on untargeted (U) and four distinct targets (T_1, T_2, T_3, T_4) for each image. We consider the T_1 to be a randomly selected segmentation mask from the test dataset. T_2 is acquired from the test images by considering the highest mean IOU across the predictions. We select complete white and black images as the T_3 and T_4, respectively.

For a thorough investigation, ***DECEIT*** employs step size (ϵ) of 1/255. If this value is set below 1/255, quantisation errors in the image pixels prevent the modifications from being correctly preserved in the original image. Conversely, if the value exceeds 1/255, large perturbations are introduced, which ultimately reduce the attack's effectiveness [100, 101]. We limit the maximum iteration count to 10, beyond which no adversarial perturbation can be applied to input image[69]. Following [69], we set mean IOU threshold as 0.5 for the stopping condition and second target (T_2) selection. Notably, the mean IOU values for the absent classes in the target are invalid, leading to incorrect threshold mean IOU. This issue

occurs with third (T_3) and fourth targets (T_4), where the images consist entirely of white (1s) and black (0s) pixels, respectively. In such a scenario, we compute the IOU for T_3 and T_4 with a threshold of 0.5 [102].

For rigours evaluation, we set the weighting parameters w for the proposed *Compound* loss (refer to equation (3.5)) as 0.21, 0.18, 0.20, 0.15, and 0.26 with respect to $w_1, w_2, w_3, w_4,$ and w_5, respectively. These values are chosen by applying a grid search operation on training data. In correspondence with [89], the *softmax temperature constant*, τ, is set to 0.1, which offers optimal results for differentiable approximation.

3.2.4 Comparative Evaluation

Table 3.1 represents the comparative analysis of our proposed attack, ***DECEIT***. It depicts that ***DECEIT*** outperforms the existing attacks, ASMA [72] and FGSM [75] by obtaining a high ASR and low δ_∞. This stems from the optimal selection of the surrogate loss function and parallel fusion operation to select optimal adversarial samples. However, Table 3.1 also reveals that in some instances, ASMA achieves a lower δ_∞ than ***DECEIT***. In these scenarios, ASMA considers a smaller count of samples to compute δ_∞, as it achieves a lower ASR than ***DECEIT***, signifying that only a small subset of samples successfully execute the attack. On the other hand, FGSM provides higher δ_∞ than ***DECEIT*** as it adds excess perturbations in one step. Moreover, FGSM fails to perform an attack successfully with the target T_3 (exhibiting all white pixels) across all the datasets, excluding polyp segmentation on the TGANet threat model [67] and skin lesion segmentation on the fine-tuned PraNet threat model [3].

In essence, our proposed attack, ***DECEIT***, demonstrates superior performance over the existing MIS attacks, achieving higher ASR , as shown in Figure 3.2, which highlights some successful attack examples. Additionally, Table 3.1 indicates ***DECEIT*** achieves rela-

Table 3.1: Comparative analysis of our proposed attack, ***DECEIT***.

Dataset	Threat Model	Attack	U		T_1		T_2		T_3		T_4	
			ASR **(in %)**	δ_∞	ASR **(in %)**	δ_∞	ASR **(in%)**	δ_∞	ASR **(in %)**	δ_∞	ASR **(in %)**	δ_∞
$Polyp_1$ CVC-300	PraNet	ASMA	96.67	3.793	38.98	3.913	100.00	1.000	5.00	4.333	100.00	1.000
		FGSM	100.00	5.000	10.17	5.000	100.00	5.000	0.00	n/a	100.00	5.000
		***DECEIT*(Ours)**	**100.00**	**3.100**	**100.00**	**3.780**	**100.00**	**1.000**	**100.00**	**4.183**	**100.00**	**1.000**
$Polyp_2$ CVC-ClinicDB	PraNet	ASMA	64.52	4.000	29.51	**3.389**	96.77	1.150	3.23	5.500	100.00	1.000
		FGSM	100.00	5.000	16.39	5.000	93.55	5.000	0.00	n/a	100.00	5.000
		***DECEIT*(Ours)**	**100.00**	**3.645**	**93.44**	3.667	**100.00**	**1.065**	**100.00**	**4.419**	**100.00**	**1.000**
$Polyp_3$ Kvasir	PraNet	ASMA	82.00	5.976	37.37	3.595	99.00	1.051	9.00	4.556	100.00	1.030
		FGSM	100.00	5.000	25.25	5.000	96.00	5.000	3.00	5.000	98.00	5.000
		***DECEIT*(Ours)**	**100.00**	**3.810**	**100.00**	**3.343**	**100.00**	**1.040**	**100.00**	**4.330**	**100.00**	**1.020**
$Polyp_4$ CVC-ColonDB	PraNet	ASMA	93.68	3.671	19.53	3.473	82.14	**1.043**	0.79	4.667	100.00	1.000
		FGSM	100.00	5.000	1.32	5.000	79.76	5.000	0.00	n/a	100.00	5.000
		***DECEIT*(Ours)**	**100.00**	**2.671**	**100.00**	**3.251**	**100.00**	1.277	**100.00**	**4.471**	**100.00**	**1.000**
Skin Lesion	PraNet	ASMA	59.67	4.466	71.95	1.734	97.50	1.070	37.17	3.556	97.83	1.261
		FGSM	100.00	5.000	66.94	5.000	94.33	5.000	9.50	5.000	90.83	5.000
		***DECEIT*(Ours)**	**100.00**	**3.483**	**100.00**	**1.651**	**100.00**	**1.067**	**100.00**	**3.048**	**100.00**	**1.113**
Brain Tumour	UNet	ASMA	99.33	3.862	29.10	**1.000**	99.67	1.017	0.00	n/a	100.00	1.000
		FGSM	100.00	5.000	29.10	5.000	99.67	5.000	0.00	n/a	100.00	5.000
		***DECEIT*(Ours)**	**100.00**	**3.462**	**44.48**	1.895	**100.00**	**1.003**	0.00	n/a	**100.00**	**1.000**
Lung	UNet	ASMA	81.00	**1.000**	20.00	**1.000**	19.00	**1.000**	0.00	n/a	100.00	1.000
		FGSM	100.00	5.000	20.00	5.000	19.00	5.000	0.00	n/a	100.00	5.000
		***DECEIT*(Ours)**	**100.00**	1.230	**76.00**	2.297	**76.00**	2.162	0.00	n/a	**100.00**	**1.000**
$Polyp_2$ CVC-ClinicDB	TGANet	ASMA	100.00	**3.629**	24.59	1.533	100.00	1.016	0.00	n/a	100.00	1.000
		FGSM	100.00	5.000	18.03	5.000	98.39	5.000	0.00	n/a	100.00	5.000
		***DECEIT*(Ours)**	**100.00**	4.081	**100.00**	**1.262**	**100.00**	**1.000**	**96.77**	**6.500**	**100.00**	**1.000**
$Polyp_3$ Kvasir	TGANet	ASMA	100.00	**3.840**	36.36	**1.972**	100.00	1.000	23.00	3.565	100.00	1.120
		FGSM	100.00	5.000	25.25	5.000	100.00	5.000	4.00	5.000	96.00	5.000
		***DECEIT*(Ours)**	**100.00**	3.960	**100.00**	2.232	**100.00**	**1.000**	**100.00**	**3.250**	**100.00**	**1.030**

Note: n/a: The attack failed in some cases where the target was pure white. Bold values represent the best result.

tively lower ASR values for brain tumour segmentation across U-Net threat model [13] on target T_1. This is because the segmentation mask for brain tumour segmentation exhibits fewer foreground pixels than background pixels, making it more difficult for the attack to successfully infer the randomly selected target (T_1). Likewise, any attack on brain tumour and lung segmentation across a complete white target (T_3) results in 0% ASR values. Such failure cases can be visualised in Figure 3.3.

3.2.5 Ablation Study

To thoroughly examine the significance of surrogate loss functions and parallel fusion, we conducted an ablation study of our proposed attack, ***DECEIT***. In this direction, the untargeted and four different targeted attacks are applied to considered threat models, employing surrogate loss functions across all datasets and their performance with ***DECEIT***. The results, given in Table 3.2, depict that the parallel fusion-based ***DECEIT*** attack ob-

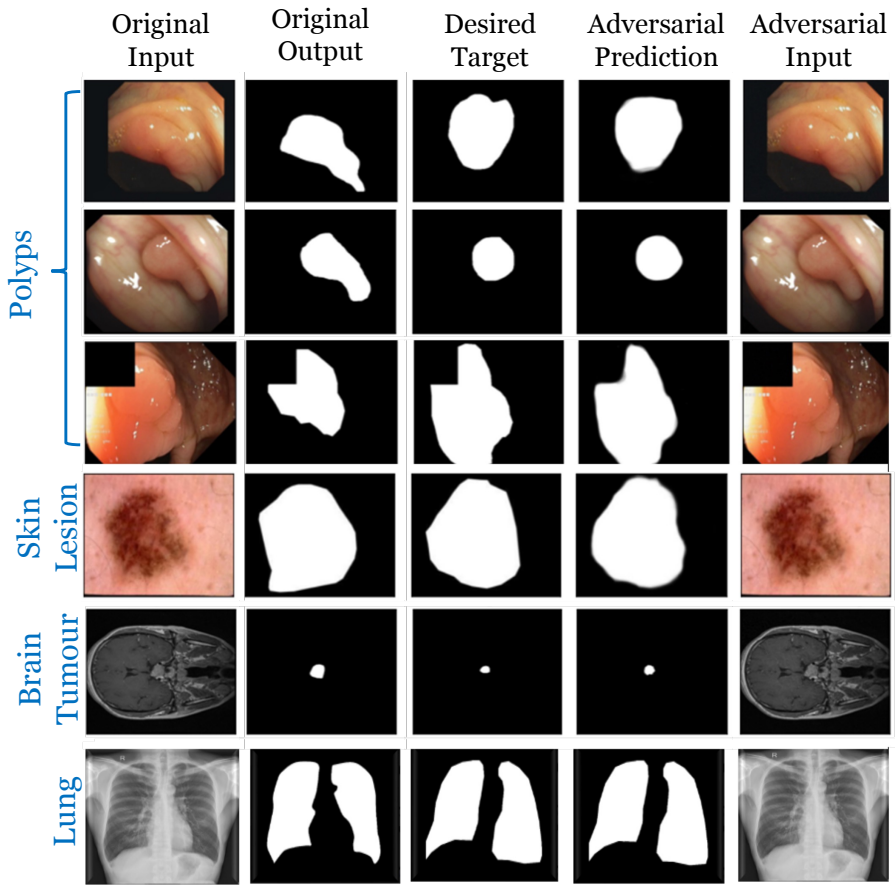

Figure 3.2: Examples showcasing successful cases of ***DECEIT*** attack, where adversarial predictions closely match the desired target while diverging from the original clean image prediction and imperceptible perturbations added in adversarial images.

tained the highest ASR in all scenarios. Additionally, it demonstrated lower δ_∞ in most cases, such as:

- All U attack cases except brain tumour segmentation on U-Net and Polyp_2 segmentation on TGANet threat models.
- All T_1 attacks except skin lesion, Polyp_1, and Polyp_2 segmentation on PraNet and brain tumour segmentation on UNet threat models.
- All T_2 attacks except Polyp_3 and Polyp_4 segmentation on the PraNet threat model.
- All T_3 attacks except lung and brain tumour segmentation on the U-Net threat model.

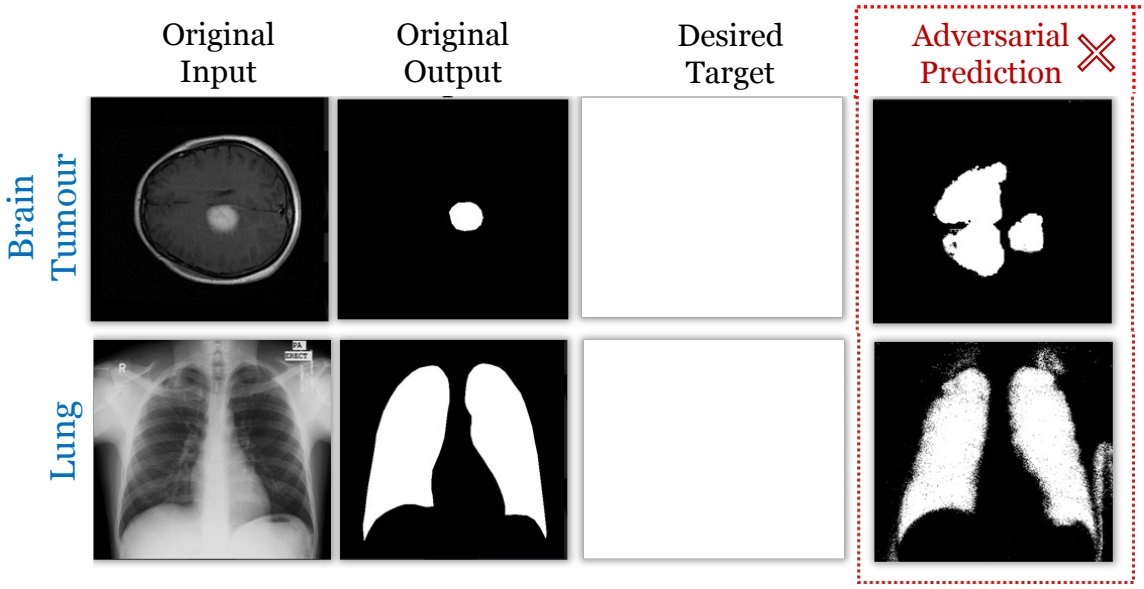

Figure 3.3: Examples showcasing the failure cases of ***DECEIT*** attack, across T_3 (complete white image) for lung and brain tumour segmentation.

These results indicate that, in such scenarios, the ASR across different surrogate loss functions is lower than that of ***DECEIT***. Thus, only a smaller number of successfully attacked samples contribute to the calculation of δ_∞.

Table 3.2 also reveals that there are consistent results reflected from T_2 and T_3 attacks. Unfortunately, ***DECEIT*** fails to attack U-Net for a brain tumour and lung segmentation under T_3 attack (Kindly refer Section 3.2.4 and Figure 3.3). Notably, our proposed surrogate loss function, *Compound* loss, offers better or comparable results than other surrogate loss functions, ensuring its effectiveness as a viable surrogate loss function. Moreover, we analyse that all the threat models can be easily attacked in a single step under T_4. Furthermore, we have also examined the loss function used in the existing ASMA attack [72] by integrating it in our proposed attack, ***DECEIT***. However, this did not lead to any advancement in attack results, and therefore, it was excluded from our experiments.

We carried out further evaluations to understand the relationship between ASR and varying L_∞ values, as shown in Figure 3.4. It illustrates the responses of the considered

Table 3.2: Ablation study of our proposed attack, ***DECEIT***.

Dataset	Threat Model	Surrogate Loss	U		T_1		T_2		T_3		T_4	
			ASR (in %)	δ_∞	ASR (in %)	δ_∞	ASR (in %)	δ_∞	ASR (in %)	δ_∞	ASR (in %)	δ_∞
$Polyp_1$ CVC-300	PraNet	Log Cosh Dice	100.00	3.533	100.00	4.373	100.00	1.000	93.33	6.875	100.00	1.000
		BCE Dice	95.00	6.386	79.66	5.021	100.00	1.000	100.00	4.383	100.00	1.000
		Exponential Log	36.67	4.091	33.90	3.750	100.00	1.000	95.00	5.000	100.00	1.000
		wBCE wIOU	100.00	3.750	72.88	5.186	100.00	1.000	90.00	6.407	100.00	1.000
		Lovasz Hinge	100.00	3.917	37.29	**3.682**	100.00	1.000	86.67	6.615	100.00	1.000
		Compound	50.00	4.933	44.07	4.423	100.00	1.000	100.00	4.450	100.00	1.000
		Parallel Fusion (Ours)	**100.00**	**3.100**	**100.00**	3.780	**100.00**	**1.000**	**100.00**	**4.183**	**100.00**	**1.000**
$Polyp_2$ CVC-ClinicDB	PraNet	Log Cosh Dice	100.00	4.097	83.61	3.941	100.00	1.065	91.94	6.526	100.00	1.000
		BCE Dice	69.35	5.744	55.74	4.588	100.00	1.081	100.00	4.597	100.00	1.000
		Exponential Log	22.58	4.429	37.70	**3.522**	96.77	1.100	100.00	5.065	100.00	1.000
		wBCE wIOU	100.00	4.032	62.30	3.921	100.00	1.065	95.16	6.475	100.00	1.000
		Lovasz Hinge	100.00	4.935	47.54	4.103	100.00	1.097	91.94	6.491	100.00	1.000
		Compound	33.87	4.857	36.07	3.545	100.00	1.226	100.00	4.645	100.00	1.000
		Parallel Fusion (Ours)	**100.00**	**3.645**	**93.44**	3.667	**100.00**	**1.065**	**100.00**	**4.419**	**100.00**	**1.000**
$Polyp_3$ Kvasir	PraNet	Log Cosh Dice	100.00	4.360	96.97	3.948	100.00	1.050	94.00	5.819	98.00	1.000
		BCE Dice	63.00	5.683	87.88	4.138	99.00	1.091	100.00	4.460	100.00	1.020
		Exponential Log	23.00	5.043	83.84	3.807	98.00	**1.031**	89.00	5.112	100.00	1.020
		wBCE wIOU	100.00	4.470	94.95	4.117	100.00	1.050	92.00	5.728	100.00	1.020
		Lovasz Hinge	100.00	4.550	85.86	4.435	100.00	1.040	94.00	5.777	100.00	1.020
		Compound	35.00	4.914	92.93	4.011	98.00	**1.031**	100.00	4.540	100.00	1.020
		Parallel Fusion (Ours)	**100.00**	**3.810**	**100.00**	**3.343**	**100.00**	1.040	**100.00**	**4.330**	**100.00**	**1.000**
$Polyp_4$ CVC-ColonDB	PraNet	Log Cosh Dice	100.00	3.026	96.31	4.890	100.00	1.315	84.21	6.856	100.00	1.000
		BCE Dice	83.16	4.120	98.42	3.976	97.92	1.356	100.00	4.768	100.00	1.000
		Exponential Log	49.21	3.112	87.60	3.976	89.58	1.399	94.47	5.365	100.00	1.000
		wBCE wIOU	100.00	3.071	96.31	4.704	96.13	1.356	83.42	6.688	100.00	1.000
		Lovasz Hinge	98.42	3.380	94.20	4.966	93.45	1.261	83.95	6.781	100.00	1.000
		Compound	60.00	3.395	97.89	3.873	91.07	1.307	100.00	4.829	100.00	1.000
		Parallel Fusion (Ours)	**100.00**	**2.671**	**100.00**	**3.251**	**100.00**	**1.277**	**100.00**	**4.471**	**100.00**	**1.000**
Skin Lesion	PraNet	Log Cosh Dice	100.00	3.672	100.00	1.776	100.00	1.073	100.00	3.833	90.83	**1.000**
		BCE Dice	99.50	4.859	100.00	1.810	100.00	1.075	100.00	3.085	100.00	1.118
		Exponential Log	25.67	4.182	88.98	**1.469**	98.83	1.081	99.33	3.468	100.00	1.118
		wBCE wIOU	100.00	3.687	100.00	1.768	100.00	1.077	100.00	3.802	100.00	1.128
		Lovasz Hinge	100.00	3.817	98.66	1.926	100.00	1.090	100.00	3.832	100.00	1.168
		Compound	37.67	4.173	98.33	1.803	100.00	1.097	100.00	3.098	100.00	1.118
		Parallel Fusion (Ours)	**100.00**	**3.483**	**100.00**	1.651	**100.00**	**1.067**	**100.00**	**3.048**	**100.00**	1.113
Brain Tumour	UNet	Log Cosh Dice	92.33	4.552	36.45	1.376	100.00	1.007	0.00	n/a	100.00	1.000
		BCE Dice	98.00	4.452	34.11	1.402	100.00	1.003	0.00	n/a	100.00	1.000
		Exponential Log	4.67	**3.857**	34.78	1.519	99.67	1.000	0.00	n/a	100.00	1.000
		wBCE wIOU	93.67	4.598	32.78	1.276	100.00	1.007	0.00	n/a	100.00	1.000
		Lovasz Hinge	67.67	5.005	31.44	**1.160**	100.00	1.003	0.00	n/a	100.00	1.000
		Compound	41.67	4.912	35.45	1.557	100.00	1.003	0.00	n/a	100.00	1.000
		Parallel Fusion (Ours)	**99.33**	3.862	**44.48**	1.895	**100.00**	**1.003**	0.00	n/a	**100.00**	**1.000**
Lung	UNet	Log Cosh Dice	100.00	1.230	73.00	2.311	74.00	2.176	0.00	n/a	100.00	1.000
		BCE Dice	100.00	1.250	70.00	2.301	72.00	2.171	0.00	n/a	100.00	1.000
		Exponential Log	100.00	1.250	71.00	2.317	72.00	2.177	0.00	n/a	100.00	1.000
		wBCE wIOU	100.00	1.230	72.00	2.309	73.00	2.175	0.00	n/a	100.00	1.000
		Lovasz Hinge	100.00	1.250	70.00	2.322	71.00	2.189	0.00	n/a	100.00	1.000
		Compound	100.00	1.250	72.00	2.315	73.00	2.176	0.00	n/a	100.00	1.000
		Parallel Fusion (Ours)	**100.00**	**1.230**	**76.00**	**2.297**	**76.00**	**2.162**	0.00	n/a	**100.00**	**1.000**
$Polyp_2$ CVC-Clinic DB	TGANet	Log Cosh Dice	100.00	4.484	100.00	1.475	100.00	1.000	33.87	8.143	100.00	1.000
		BCE Dice	85.48	4.887	100.00	1.279	100.00	1.000	93.55	6.724	100.00	1.000
		Exponential Log	23.23	**1.500**	100.00	1.328	100.00	1.000	61.29	7.132	100.00	1.000
		wBCE wIOU	100.00	4.290	100.00	1.393	100.00	1.000	35.48	8.227	100.00	1.000
		Lovasz Hinge	83.87	6.712	100.00	1.393	100.00	1.000	35.48	8.227	100.00	1.000
		Compound	33.27	**1.500**	100.00	1.262	100.00	1.000	95.16	6.780	100.00	1.000
		Parallel Fusion (Ours)	**100.00**	4.081	**100.00**	**1.262**	**100.00**	**1.000**	**96.77**	**6.500**	**100.00**	**1.000**
$Polyp_3$ Kvasir	TGANet	Log Cosh Dice	100.00	4.484	100.00	3.303	100.00	1.000	99.00	5.697	97.00	1.000
		BCE Dice	90.00	5.244	100.00	2.323	100.00	1.000	100.00	3.310	100.00	1.030
		Exponential Log	24.05	3.975	98.99	2.469	100.00	1.000	100.00	3.530	100.00	1.030
		wBCE wIOU	100.00	4.480	97.98	3.196	100.00	1.000	99.00	5.576	100.00	1.040
		Lovasz Hinge	100.00	4.900	91.92	3.330	100.00	1.000	99.00	5.687	100.00	1.030
		Compound	34.87	6.500	100.00	2.293	100.00	1.000	100.00	3.330	100.00	1.030
		Parallel Fusion (Ours)	**100.00**	**3.960**	**100.00**	**2.232**	**100.00**	**1.000**	**100.00**	**3.250**	**100.00**	**1.030**

Note: n/a: The attack failed in some cases where the target was pure white. Bold values represent the best result.

threat models and surrogate loss functions. The results imply that, when compared to alternative loss functions, our proposed attack, ***DECEIT***, leveraging parallel fusion, consistently obtains the highest ASR .

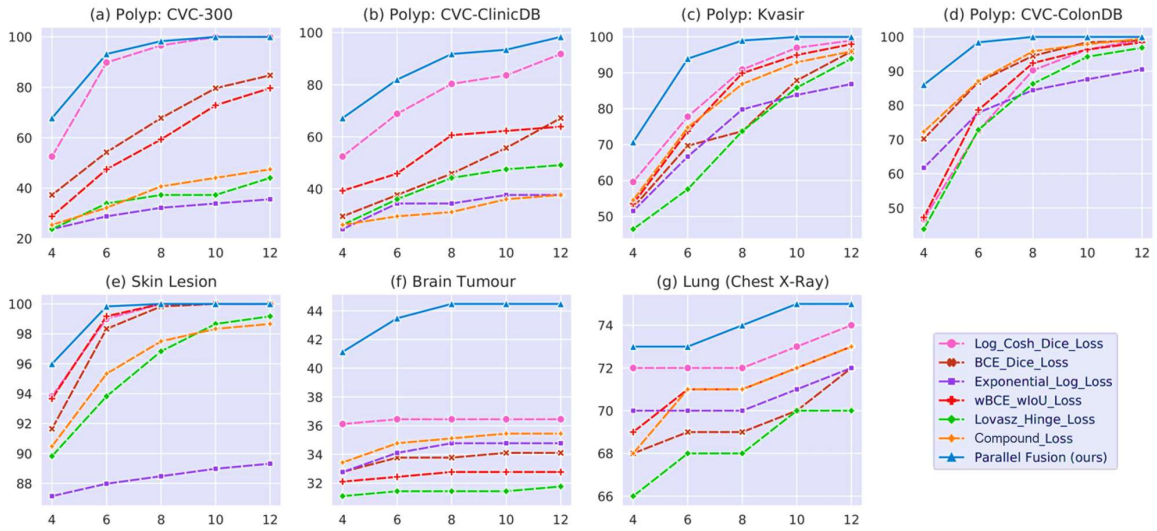

Figure 3.4: Performance of ***DECEIT*** across various maximum permissible distortions for all datasets, with the x-axis and y-axis representing L_∞ distortion and the ASR, respectively. It demonstrates that our parallel fusion-based approach surpasses all the existing surrogate losses.

3.3 Discussion

The modified threat models (or DL-based MIS models) in the proposed attack, ***DECEIT*** employ differentiable approximates in place of non-differentiable layers under the condition that the outcomes of the modified and original models should exhibit similarity. To validate this, we conducted experiments to differentiate the results of the threat model prior to and later than applying differentiable approximation. Specifically, the evaluation metrics mean IOU and dice are computed on all the datasets across each considered threat models, as presented in Table 3.3. The results confirm that the modified threat models exhibit similar performance to the original ones, validating their suitability for adversarial attack opera-

tions.

Table 3.3: Comparison between the threat models' performance before and after substituting non-differentiable layers.

Dataset	Threat Model	Original		Modified	
		mean IOU	Dice	mean IOU	Dice
$Polyp_1$ CVC-300	PraNet	0.797	0.871	0.797	0.870
$Polyp_2$ CVC-ClinicDB	PraNet	0.849	0.899	0.851	0.899
$Polyp_3$ Kvasir	PraNet	0.841	0.898	0.840	0.896
$Polyp_4$ CVC-ColonDB	PraNet	0.640	0.709	0.640	0.708
Skin Lesion	PraNet	0.831	0.837	0.831	0.836
Brain Tumour	UNet	0.772	0.745	0.772	0.743
Lung	UNet	0.927	0.961	0.926	0.959
$Polyp_2$ CVC-ClinicDB	TGANet	0.899	0.946	0.899	0.946
$Polyp_3$ Kvasir	TGANet	0.833	0.898	0.833	0.898

Table 3.4: Efficiency analysis of the proposed attack, ***DECEIT***. GPU memory usage and model parameters are computed in MiB and millions, respectively.

Dataset	Threat Model	GPU Memory Usage	Model Parameters	Time (in seconds)				
				U	T_1	T_2	T_3	T_4
$Polyp_1$ CVC-300	PraNet	1979	32.55	24.5445	29.8462	23.2724	26.9557	26.7528
$Polyp_2$ CVC-ClinicDB	PraNet	1979	32.55	24.1215	25.3992	24.8152	25.2190	25.8978
$Polyp_3$ Kvasir	PraNet	1979	32.55	29.6956	34.0440	30.9220	31.0632	31.7451
$Polyp_4$ CVC-ColonDB	PraNet	1987	32.55	20.6782	23.0043	20.9246	21.8113	22.0598
Skin Lesion	PraNet	1981	32.55	16.6190	26.8238	22.6270	19.9621	25.1174
Brain Tumour	UNet	1547	1.94	9.2064	11.6381	12.1330	11.9106	9.7289
Lung	UNet	3133	22.98	37.3741	39.6809	40.5019	39.0110	38.0752
$Polyp_2$ CVC-ClinicDB	TGANet	2011	19.84	17.3710	24.6613	17.2903	24.1613	14.2742
$Polyp_3$ Kvasir	TGANet	2011	19.84	34.1452	48.9839	33.5968	48.3871	37.5645

Besides this, we have conducted an efficiency analysis of the ***DECEIT*** attack to evaluate GPU memory usage and overall time complexity, as given in Table 3.4. The result shows that attacking the U-Net threat model utilised the least memory for brain tumour segmentation and the most memory for lung segmentation compared to other threat models. Further, the attacks on the PraNet model exhibited moderate execution times across all attack settings. Conversely, attacking TGANet for $Polyp_3$ segmentation and U-Net for lung segmentation required relatively additional time for most of the settings. Notably, attacking the PraNet threat model employs the highest number of parameters among all the considered threat models.

3.4 Summary

This chapter has proposed a novel adversarial attack, ***DECEIT***, for DL-based MIS models. Several existing MIS models contain non-differentiable layers or non-differentiable loss functions, which result in zero gradients during backpropagation and impede the attack process. Our proposed attack, ***DECEIT***, has addressed this problem by substituting these non-differentiable neural network layers with its approximated differentiable layers. Additionally, it utilises several surrogate loss functions, which are approximated true loss functions and is differentiable for MIS. The optimal surrogate loss is selected among all considered ones using parallel fusion that offers the least adversarial perturbation-added adversarial samples, ensuring a successful attack. Thus, the dynamic selection of loss functions through parallel fusion, along with the substitution of non-differentiable layers, enabled the effective production of adversarial samples with minimal perturbations. Our experimental outcomes revealed that our proposed attack, ***DECEIT***, surpasses the existing attacks, achieving high ASR in low δ_∞.

Chapter 4

Detection of Adversarial and OOD samples in MIS

The DL models have shown enormous achievement in healthcare applications. Unfortunately, their performance deteriorates when they encounter anomalies or corrupted data, including adversarial and OOD samples. While adversarial samples add a small amount of imperceptible perturbations [31], OOD samples signify a substantial distribution shift in input [41, 42]. Such samples can have devastating impacts on life-critical healthcare applications, requiring robust detection methods to mitigate their detrimental impact. Existing detection methods focused on MIC tasks [43], which perform inadequately for MIS applications. Such methods require network retraining and typically utilise either distance-based approaches [43, 48] or probability-distribution-based functions [45, 46, 47]. Distance-based methods often rely on predetermined thresholds, which are challenging to fine-tune [2] and obstruct detection effectiveness. Similarly, the probability-distribution-based methods heavily depend on network logits, which behave similarly for adversarial, OOD, and clean samples, thereby constraining detection performance [49, 48].

This chapter proposes a novel unified detection method for MIS, called ***DISCERN***, designed to detect both adversarial and OOD samples and accurately distinguish them from

clean samples. Unlike existing detection methods, ***DISCERN*** eliminates the need for network retraining and reliance on distance and probability distribution-based approaches. Instead, it measures the similarity between the MIS predictions of an input and its rotational versions to achieve precise detection. The chapter is structured as follows: Section 4.1 presents our proposed detection method, ***DISCERN***. Section 4.2 demonstrates experimental details associated with ***DISCERN***. Section 4.3 provides a discussion about the proposed method. Section 4.4 summarises the overall chapter.

4.1 Proposed Detection Method: *DISCERN*

This section demonstrates the proposed detection method, ***DISCERN***. It analyses the consistent relation between the predictions of the input image and its respective rotated variants to effectively differentiate the adversarial and OOD samples from clean samples in DL-based MIS models. For visual comprehension, the complete workflow of ***DISCERN*** is depicted in Figure 4.1. Initially, the adversarial and OOD samples are constructed by adversarially attacking the DL-based MIS models and employing the Image Part Permutation (IPP) method [103], respectively (Refer to Section 4.1.1). ***DISCERN*** considers the following three key observations for detection.

1. There is a high consistency (similar behaviour) between the clean input MIS prediction and their respective variant predictions (rotated versions of input) [2], as the MIS model is often learned to offer orientation-independent outcomes.

2. The adversarial variant predictions (the output of rotated versions of adversarial input) are minimally influenced by adversarial perturbation [60]. Consequently, it shows low consistency or dissimilar behaviour with the adversarial input prediction.

3. Likewise, the OOD samples behave randomly, which leads to inaccurate MIS predictions. Hence, it is obvious that their respective rotated variants also provide some other random MIS predictions. As a result, there are inconsistent MIS predictions for OOD samples and their rotated versions.

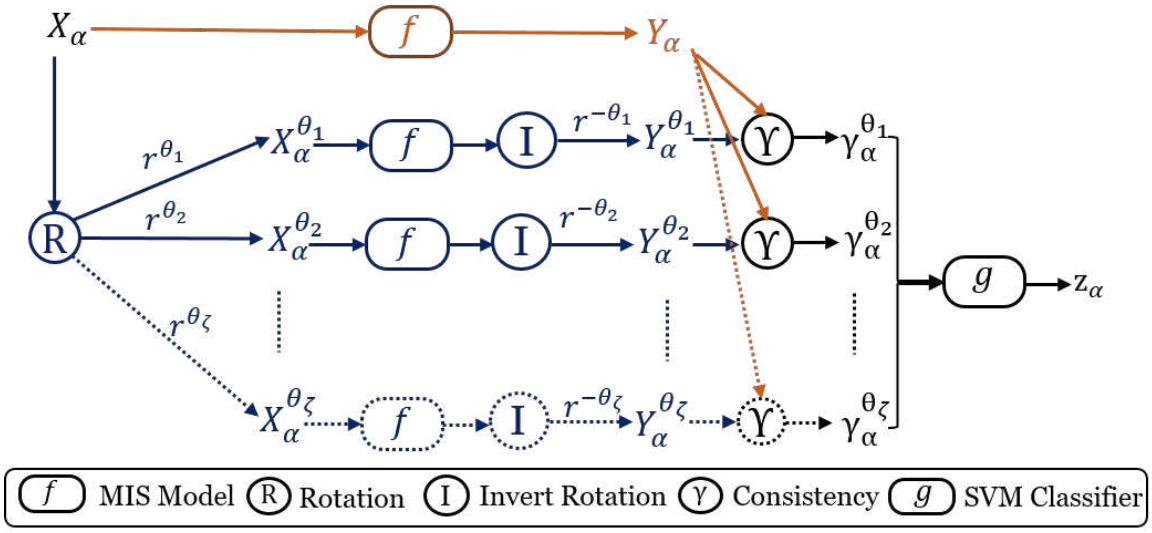

Figure 4.1: The complete work-flow of our proposed detection method, ***DISCERN***. The MIS model f accepts input X_α and their respective variants $X_\alpha^{\theta_1}$, $X_\alpha^{\theta_2}$ $\cdots$ $X_\alpha^{\theta_\zeta}$ to produce their corresponding predictions (Y_α, $Y_\alpha^{\theta_1}$, $Y_\alpha^{\theta_2}$ $\cdots$ $Y_\alpha^{\theta_\zeta}$). The measured consistencies $\Upsilon_\alpha^{\theta_1}$,$\Upsilon_\alpha^{\theta_2}$, $\cdots$, $\Upsilon_\alpha^{\theta_\zeta}$ is fed to the SVM classifier g to achieve sample detection output z_α.

Algorithm 4.1 Proposed adversarial and OOD detection method, ***DISCERN***.

Require: Input sample X_α; MIS model $f(\cdot)$; SVM classifier $g(\cdot)$.
Ensure: Detection output z_α.

$Y_\alpha = f(X_\alpha)$ ▷ prediction of X_α, refer Eq. (4.5)
$X_\alpha^\theta = r^\theta(X_\alpha)$ ▷ creating input variants, refer Eq.(4.6)
$Y_\alpha^\theta = r^{-\theta}(f(X_\alpha^\theta))$ ▷ prediction of X_α^θ, refer Eq. (4.7)
$\Upsilon_\alpha^\theta = \frac{2*\left|Y_\alpha \cap Y_\alpha^\theta\right|}{|Y_\alpha|+\left|Y_\alpha^\theta\right|}$ ▷ calculate consistency, refer Eq. (4.9).
Compute Υ_α^θ for different values of θ.
$z_\alpha = g(\Upsilon_\alpha^{\theta_1}, \Upsilon_\alpha^{\theta_2} \cdots, \Upsilon_\alpha^{\theta_\zeta})$ ▷ detection output, refer Eq. (4.10).
return z_α

DISCERN leverages the aforementioned observations by obtaining the MIS prediction of input variants (refer Section 4.1.2) and examining their consistency with actual MIS

prediction (kindly see Section 4.1.3). A detailed step-by-step explanation of ***DISCERN*** is provided in Algorithm 4.1.

4.1.1 Generating Adversarial and OOD Samples

4.1.1.1 Adversarial Samples

Adversarial attacks present substantial risks to the security and robustness of DL-based MIS models [31]. Hence, it becomes essential to comprehend the generation of adversarial samples for their precise detection, which ultimately advances the MIS models' robustness. For this reason, we perform iterative adversarial attacks [104] operations on several existing MIS models ("threat models") to craft adversarial samples. A visual representation of adversarial sample generation is given in Figure 4.2. Unlike attacking the classification model, wherein the single class (or target) is required to be incorrectly categorised, the MIS model is complicated and challenging to attack. This complexity arises because segmentation performs multi-class classification, demanding adversarial perturbation to affect every pixel of the segmentation mask [105]. Further, several existing MIS models exhibit non-differentiable preprocessing layers that constrain attack operations [106]. We overcome this limitation by adapting threat models with the substitution of such layers with their differential approximation using Kornia [90].

Let the adversarial sample X_{adv} be constructed by imposing the perturbation ϵ in the gradient of loss J. Mathematically,

$$X^0_{adv} = X \tag{4.1}$$

$$X^{i+1}_{adv} = clip\left(X^i_{adv} - \epsilon * \text{sign}(\nabla_{X^i_{adv}} J\left(f(X^i_{adv}), T\right))\right) \tag{4.2}$$

$$X^{i+1}_{adv} = clip\left(X^i_{adv} + \epsilon * \text{sign}(\nabla_{X^i_{adv}} J\left(f(X^i_{adv}), Y\right))\right) \tag{4.3}$$

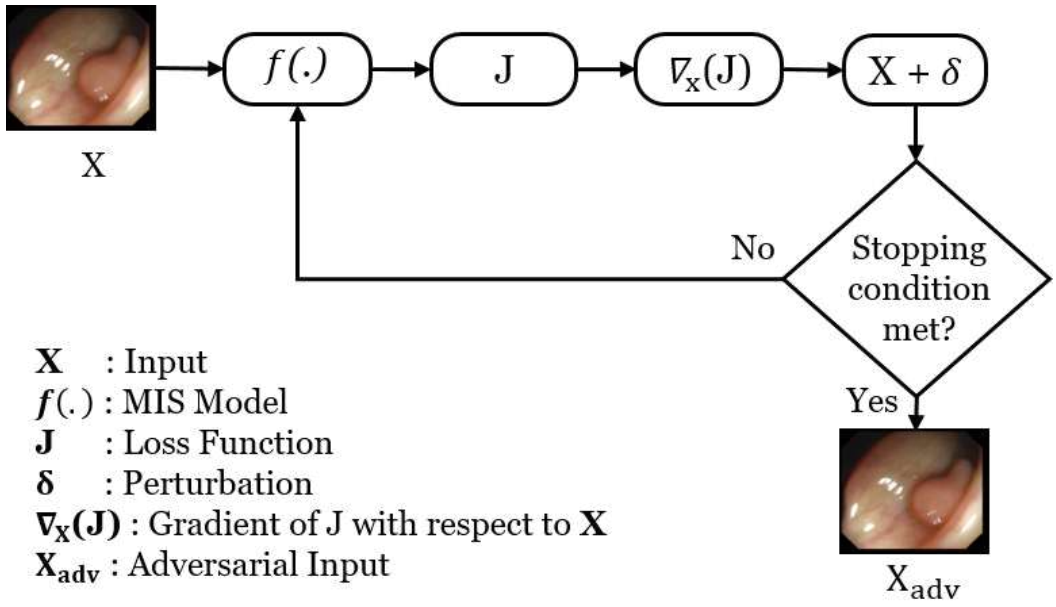

Figure 4.2: Generation of adversarial samples, X_{adv}.

where $f(\cdot)$, X, Y, and T represent the modified threat model, input, corresponding MIS output, and target mask, respectively. The clipping operation, $clip$, ensures that X_{adv} remains within the specified range [106]. The superscript i indicates iteration count. Incorporating the previous work [106][1], we iteratively perform Equations (5.25) and (5.26) operations until it reaches the stopping condition. This condition is determined by the maximum iteration count i^{max} and prespecified target T. Samples will only be chosen as X_{adv} if they fulfil the stopping condition, which is defined as a target-based measure m_T that must surpass the prespecified threshold v or i^{max} be reached. The hyperparameter selection of i^{max}, m_T and v is provided in Subsection 4.2.3.

4.1.1.2 OOD Samples

The DL-based MIS models often perform incorrectly due to the presence of OOD samples in test datasets [107]. The OOD samples are the test samples whose distribution deviates from that of the training samples, which results in the prediction of random outputs.

[1] https://github.com/SnehaShukla937/MEDIS_ATTACK

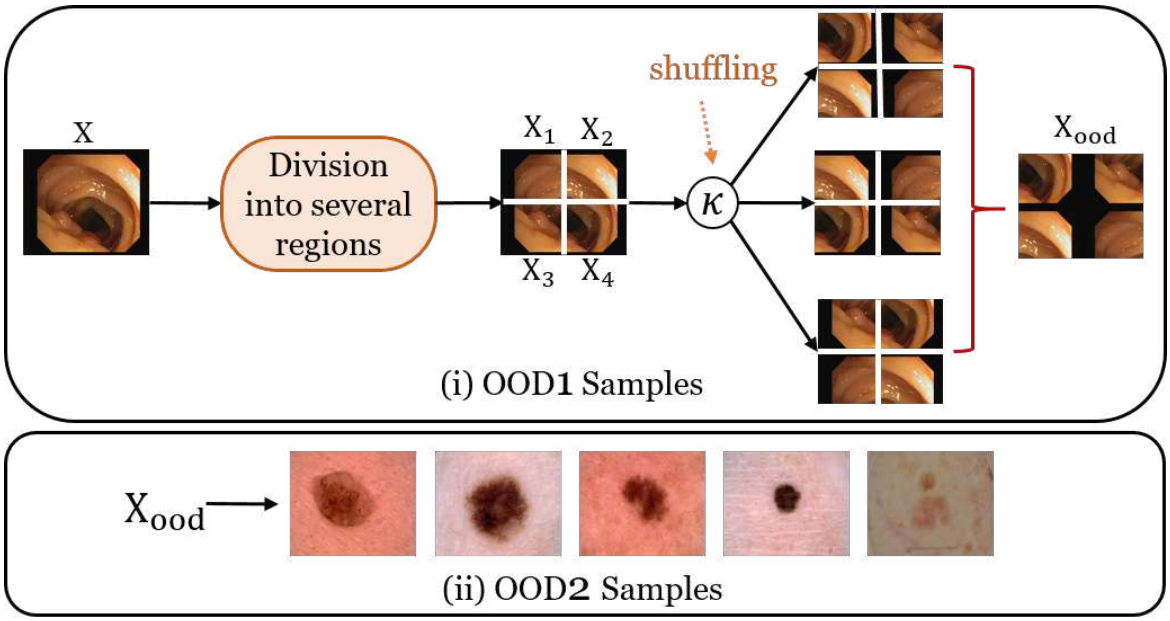

Figure 4.3: OOD samples (X_{ood}) utilised by ***DISCERN***. (i) Generated X_{ood} from X using IPP method. (ii) X_{ood} taken from the skin-lesion datasets.

Thus, it is crucial to analyse and explore the development of OOD samples for their accurate detection. To this end, the IPP [103] method is used, which employs ID samples to produce OOD samples. Notably, ID samples are those samples whose distribution aligns with training sample distribution, usually providing correct predictions. Thus, the IPP method divides the ID sample into multiple regions and applies a random shuffling operation, thereby altering the spatial structure. The output image exhibits the shifted distribution of input, effectively representing the OOD sample. Mathematically, the constructed OOD sample X_{ood} can be expressed as,

$$X_{ood} = \kappa(X_1, X_2, ...X_q) \tag{4.4}$$

where X represents the ID sample, which is partitioned into q identical regions such that $(X_1, X_2, ...X_q) \in X$. κ depicts the random shuffling operation applied to these regions. It is noteworthy that our proposed method, ***DISCERN***, categorises such OOD samples as OOD1. Additionally, it considers samples from different modalities, such as skin-lesion

segmentation, which do not participate in the training process. ***DISCERN*** classifies these samples as OOD2 samples. A visualisation of all these considered OOD samples can be found in Figure 4.3

4.1.2 Obtaining MIS Predictions of Input Variants

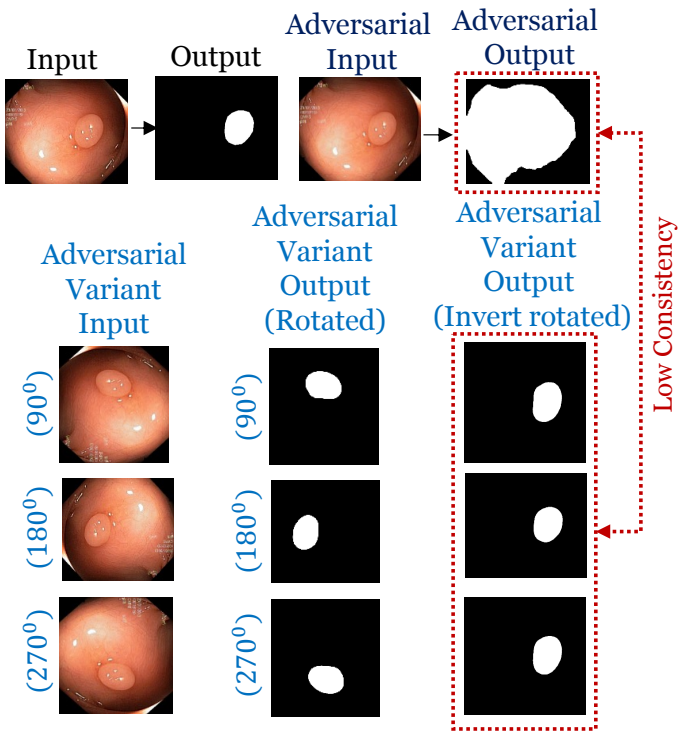

Figure 4.4: In this example, the adversarial output and its invert-rotated variant outputs exhibit low consistency, as the adversarial perturbation has little effect on the variant output.

The invert-rotated MIS predictions of rotated clean samples exhibit similarity or strong consistency with that of non-rotated input sample [2]. This happens because the MIS model is expected to offer predictions irrespective of orientation. However, this observation is irrelevant for adversarial or OOD input samples. In essence, when these types of input samples are rotated by a specific angle and then invert-rotated back, their MIS predictions show noticeable differences or reduced consistency compared to the predictions for the original, non-rotated inputs. This observation can be seen in Figure 4.4 and Figure 4.5. Leveraging

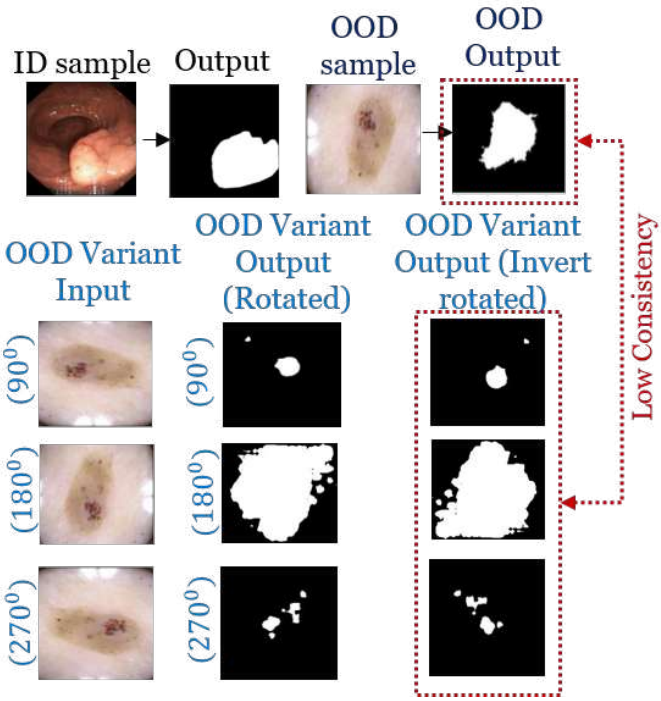

Figure 4.5: In this example, the OOD output and its respective invert-rotated variant outputs depict low consistency since both the samples consistently produce random predictions across all scenarios.

these insights, we generate input variants by performing a rotational transformation on input samples at diverse angles.

The input samples and their respective variants are fed into the MIS model, producing predictions of each sample. The input variants' MIS predictions are invert-rotated by the same angle to ensure consistency and comparability in the later detection steps. Mathematically,

$$Y_\alpha = f(X_\alpha) \tag{4.5}$$

where $f(\cdot)$ indicates MIS model that generates prediction Y_α. X_α depicts the input sample, which can be clean (X) / adversarial (X_{adv}) / OOD (X_{ood}). To create input variants X_α^θ, the X_α is rotated at an angle θ, represented as,

$$X_\alpha^\theta = r^\theta(X_\alpha) \tag{4.6}$$

where $r^{\theta}(\cdot)$ denotes the rotation operation at θ. The MIS prediction of X_{α}^{θ} is achieved by applying it to $f(\cdot)$ and rotating it at an angle of $-\theta$. Mathematically, this variant prediction Y_{α}^{θ} is given by,

$$Y_{\alpha}^{\theta} = r^{-\theta}(f(X_{\alpha}^{\theta})) \tag{4.7}$$

where $r^{-\theta}(\cdot)$ indicates invert-rotation (rotation at $-\theta$). Kindly refer to Subsection 4.2.3 for hyperparameter tuning of θ.

4.1.3 Detecting Samples through Consistency Analysis

The MIS prediction of clean inputs are anticipated to behave similarly with their respective variants, while the these prediction of the adversarial or OOD samples shares dissimilarity with their respective variants. Mathematically, the observation can be expressed as,

$$\Upsilon_{\alpha}^{\theta} = \begin{cases} High, & X_{\alpha} \textit{ is clean sample} \\ Low, & X_{\alpha} \textit{ is adversarial/OOD sample} \end{cases} \tag{4.8}$$

where $\Upsilon_{\alpha}^{\theta}$ depicts consistency, which is computed between Y_{α} and Y_{α}^{θ}. When Y_{α} is highly consistent with Y_{α}^{θ}, then X_{α} is identified as a clean sample; otherwise, X_{α} belongs to an adversarial/OOD sample. The proposed detection method, ***DISCERN***, consider $\Upsilon_{\alpha}^{\theta}$ as a dice similarity score [108], formulated as,

$$\Upsilon_{\alpha}^{\theta} = \frac{2 * \left| Y_{\alpha} \cap Y_{\alpha}^{\theta} \right|}{|Y_{\alpha}| + |Y_{\alpha}^{\theta}|} \tag{4.9}$$

Nevertheless, we observe that $\Upsilon_{\alpha}^{\theta}$ can sometimes yield high values even when the sample is adversarial or OOD for a certain angle θ. To mitigate such issues, our proposed method, ***DISCERN***, incorporates several input variants with diverse values of θ. Specifically, we compute $\Upsilon_{\alpha}^{\theta}$ for different rotation angles θ (refer Subsection 4.2.3). For a better understand-

ing, let us assume that the input is rotated in ζ diverse angles such that $(\theta_1, \theta_2, \cdots, \theta_\zeta)$. Thus, their respective consistency values are represented as $(\Upsilon_\alpha^{\theta_1}, \Upsilon_\alpha^{\theta_2}, \cdots, \Upsilon_\alpha^{\theta_\zeta})$. To achieve detection output z_α, these consistency values are fed into the Support Vector Machine (SVM) classifier [109] $g(\cdot)$, which predicts whether the input sample X_α is clean or adversarial/OOD. Mathematically,

$$z_\alpha = g(\Upsilon_\alpha^{\theta_1}, \Upsilon_\alpha^{\theta_2}, \cdots, \Upsilon_\alpha^{\theta_\zeta}) \tag{4.10}$$

The input sample X_α categories the detection output z_α using:

$$z_\alpha = \begin{cases} 0, & X_\alpha \textit{ is clean sample} \\ 1, & X_\alpha \textit{ is adversarial / OOD sample} \end{cases} \tag{4.11}$$

The detection-related experimental settings are provided in Subsection 4.2.3.

4.2 Experimental Results

4.2.1 Datasets and Metric

Our proposed method, ***DISCERN***, employed open-source datasets, used for polyp segmentation tasks for all experiments. The total numnber of these samples are given in Table 4.1 and Table 4.2 for adversarial and OOD detection, respectively. As observed in Table 4.1, the number of adversarial samples is different from the number of clean samples. It happens due to the fact that adversarial samples are created by imposing an adversarial attack on well-known DL-based MIS models; specifically the samples that satisfy the specified stopping conditions (refer to sub-sections 4.1.1.1), are considered adversarial.

The complete training samples are acquired from CVC-ClinicDB [96] and Kvasir [98] datasets, excluding the OOD2 training samples. Likewise, all the test samples are taken from the CVC-300 [95], CVC-ClinicDB [96], CVC-ColonDB [97], ETIS-LaribPolypDB

[110], and Kvasir [98] datasets, excluding the OOD2 test samples. The OOD2 samples are sourced from the skin-lesion dataset [10] for both training and testing purposes.

Table 4.1: Sample count for adversarial sample detection.

Model	Operation	Clean	Adversarial	*Total*
PraNet [3]	Train	1450	1450	*2900*
	Test	798	658	*1456*
SSFormer-S [14]	Train	1450	1449	*2899*
	Test	798	723	*1521*
SSFormer-L [14]	Train	1450	1450	*2900*
	Test	798	753	*1551*
UACANet-S [65]	Train	1450	1450	*2900*
	Test	798	724	*1522*
UACANet-L [65]	Train	1450	1450	*2900*
	Test	798	724	*1522*
CaraNet [66]	Train	1450	1443	*2893*
	Test	798	658	*1456*

Note: The number of training and testing samples are consistent across the all considered attack settings.

Table 4.2: Sample count for OOD sample detection.

OOD Type	Operation	Clean	OOD	*Total*
OOD1 samples	Train	1450	1450	*2900*
	Test	798	798	*1596*
OOD2 samples	Train	1450	2000	*3450*
	Test	723	600	*1323*

Note: The number of samples are consistent across all evaluated MIS models. OOD1 samples represent OOD data produced from the IPP method, while OOD2 samples refer to OOD data sourced from different segmentation tasks.

The performance of ***DISCERN*** is assessed by the Detection Success Rate (DSR) metric, which defines the percentage of correctly detected samples to the total samples. DSR

can be formulated as,

$$DSR = \frac{|clean_c| + |anomaly_c|}{|clean| + |anomaly|} \times 100 \tag{4.12}$$

where $clean$ and $anomaly$ demonstrate the total count of clean and adversarial/OOD samples, respectively. The subscript c indicates the samples that were precisely detected.

4.2.2 Considered MIS Models

Our proposed method, ***DISCERN***, incorporates diverse existing DL-based MIS models such as PraNet[2] [3], SSFormer[3] [14], UACANet[4] [65] and CARANet[5] [66]. All these models are specifically devised for polyp segmentation. The Pranet [3] model exhibits PPD and RA that help to effectively distinguish the polyp area and boundaries from images, ensuring precise segmentation. The SSFormer [14] model aims to provide generalised polyp segmentation by leveraging a pyramidal transformer encoder and a progressive locality decoder. It effectively learns general ROI and obtains a precise segmentation mask. The UACANet [65] model, a variant of the U-Net [13] model, exhibits a feature aggregation method to create a perfect segmentation mask by step-by-step calculating feature maps and augmenting inconsistent ROI. Meanwhile, the CaraNet [66] model is designed for segmenting small medical objects. It utilises a pyramidal channel feature module to refine high-level extracted features and axial reverse attention to achieve the final segmentation mask. Notably, SSFormer and UACANet models are available in Standard (S) and Large (L) variants based on diverse scales of encoder. We utilise all versions to comprehensively evaluate the effectiveness of ***DISCERN***.

[2]Model details are available at: `https://github.com/DengPingFan/PraNet`
[3]Model details are available at: `https://github.com/Qiming-Huang/ssformer`
[4]Model details are available at: `https://github.com/plemeri/UACANet`
[5]Model details are available at: `https://github.com/AngeLouCN/CaraNet`

4.2.3 Experimental Settings

All the experiments are executed on a server, featuring an Nvidia V100 GPU, an Intel Xeon Gold 6132 CPU, and 192 GB of RAM. ***DISCERN*** utilises publicly available pre-trained weights of all the existing MIS models [3, 14, 65] except CaraNet [66]. Following [3], we trained the CaraNet MIS model using 1450 polyp images from CVC-ClinicDB [96] and Kvasir [98] datasets. For this training, epochs and batch size are tuned as 125 and 6, respectively.

For a comprehensive assessment, the adversarial samples (X_{adv}) are produced using untargeted (U) (refer Equation (5.26)) and two distinct targets (T_w and T_b) (refer Equation (5.25)) attack settings. These two targets (T_w and T_b) are images having all pixels as 0 and 1, respectively. In essence, T_w is considered a pure white image, while T_b is a completely black image. The step size, ϵ, is set as $1/255$ for attacking the models. While the smaller values of ϵ lead to erroneous saving of modified images due to the quantisation effect, the larger ϵ enhance the amount of perturbation added to X that limits the efficacy of the attack [111]. Moreover, the maximum iteration count, i^{max}, is restricted to 8, controlling the extent of perturbation applied to X.

A successfully attacked adversarial sample (X_{adv}) must meet the stopping condition, which depends upon a proper choice of target-based measure (m_T). Our proposed method, ***DISCERN*** selects dice [108] as m_T for U and T_w settings, while mIOU [112] for specific T_b settings. For T_w and U, dice measures the white pixel overlap between the predicted output and the target. However, for any prediction with target black (T_b), the dice would always be zero due to the absence of white pixels in the black image. Thus, we employ mIOU as m_T in the stopping condition for black targets (T_b) [106]. Moreover, the threshold, v, is tuned as 0.5 and 0.7 for mIOU and dice, respectively, as these values yield optimal results. Furthermore, the loss function J is the fusion of weighted Binary Cross Entropy (wBCE) and

weighted Intersection Over Union (wIOU) [3] to effectively generate adversarial samples.

To generate OOD1 samples, ***DISCERN***, employs the existing IPP method [103], requiring a maximum count of dividing areas (q). We set q as 4 [103], since smaller q reduces the efficacy of shifting distribution, while a larger q leads to enhanced computation complexity. To effectively learn the ID sample's shifting distribution in the IPP method, region shuffling (κ) is performed in horizontal, vertical, and diagonal directions. To deeply investigate the importance of rotation angles for sample detection, the samples are rotated at multiple degrees. Specifically, we consider rotation angles of 90^0, 180^0 and 270^0 [113]. Our outcomes reveal that omitting any considered angle ignores valuable insights about the input, ultimately lowering detection performance. Moreover, we found that adding additional angles to the already considered angles marginally enhances the detection performance but also increases the complexity.

Table 4.3: Kernel assignment.

Model Name	**Kernel Type**
UACANet-S [65]	RBF
UACANet-L [65]	Linear
PraNet [3]	RBF
CaraNet [66]	Polynomial
SSFormer-S [14]	RBF
SSFormer-L [14]	Polynomial

DISCERN determines consistency by running each MIS model four times for a single input sample. This includes one run for the actual input and three runs for its variants, generating four MIS predictions in total. The computed consistency values are labelled as 0 for clean samples and 1 for adversarial or OOD samples and fed into the SVM classifier as a training dataset to perform binary classification. Thus, the database containing consistency values is utilised to train the SVM classifier across multiple settings of kernel, including linear, Radial Basis Function (RBF), and polynomial. Among them, one kernel is selected

for each considered MIS model as shown in Table 4.3, which is decided on the basis of the highest DSR during training. Additionally, the SVM cost function is fine-tuned to 0.1 [114] using grid-search method.

4.2.4 Comparative Evaluation

Table 4.4 and 4.5 present the comparative analysis of ***DISCERN***, to identify adversarial and OOD samples, respectively. The experimental outcomes are obtained for all considered MIS models, incorporating the allocated SVM kernel as depicted in Table 4.3.

It is evident from Table 4.4 that ***DISCERN*** surpasses the existing adversarial detection methods, UAD [45], MahD [43] and SEViT [77] across all attack settings (U, T_w, T_b) in all datasets. Notably, these existing methods were designed primarily for classification purposes. To ensure a equitable comparison with ***DISCERN***, we altered these methods for MIS tasks. UAD is an unsupervised adversarial detection method that calculates the probability distribution of training samples from input features, employing the GMM. It then investigates the reliance of test samples on this training distribution. If the test sample falls inside the training distribution, the sample is identified as clean; otherwise, it is adversarial. Unfortunately, the distribution of adversarial samples often resembles the clean samples' distribution and shows strong structure similarity due to the inclusion of minimal perturbation. As a result, UAD struggles to distinguish adversarial samples accurately, leading to misclassification and lower DSR values.

Similarly, MahD calculates the Mahalanobis distance-based score and differentiates it with the prespecified threshold for the detection of adversarial samples. Further, SEViT works on the fact that predictions from the initial ViT blocks are more resistant to adversarial samples than the final ViT prediction. Leveraging this observation, SEViT focuses on the non-diagonal elements of the Kullback-Leibler (KL) divergence matrix, anticipating

Table 4.4: Comparative results of ***DISCERN*** for adversarial sample's detection.

Attack Settings	Methods	D1	D2	D3	D4	D5
U	UAD [45]	70.56	69.65	68.89	75.90	69.92
	MahD [43]	40.00	29.18	43.44	38.33	44.71
	SEViT [77]	55.16	51.47	50.52	50.57	57.88
	DISCERN	**74.02**	**86.78**	**86.25**	**83.86**	**72.81**
T_w	UAD [45]	65.00	75.65	83.67	80.87	62.76
	MahD [43]	50.00	32.16	35.17	40.41	52.50
	SEViT [77]	58.15	54.47	51.59	53.99	58.62
	DISCERN	**78.63**	**86.53**	**86.97**	**89.01**	**72.37**
T_b	UAD [45]	58.76	70.76	65.90	68.87	54.00
	MahD [43]	50.00	28.16	53.45	48.33	34.71
	SEViT [77]	51.16	50.47	48.61	49.98	55.63
	DISCERN	**66.83**	**77.99**	**79.52**	**78.73**	**60.59**

Note: The values signify DSR (in %). Bold entries represent the highest DSR. **D1**:CVC-ColonDB, **D2**:Kvasir, **D3**:CVC-ClinicDB, **D4**:CVC-300, **D5**:ETIS.

that they will be lower for clean samples and much higher for adversarial/OOD samples. However, this method does not perform well when handling small perturbations due to the insignificant input changes. In addition to that, the existing methods, MahD and SEViT, demand an additional burden of tuning the prespecified threshold, hindering detection capability.

Likewise, Table 4.5 illustrates that ***DISCERN*** performs better than existing OOD detection methods, SS [46], MSP [47], MahD [43] and ERNN [79]. The SS method involves computing the L_2-norm of Singular Value Decomposition (SVD) of extracted features at some specific layers in the model. The distance between a test sample's feature vector and its adjecent neighbor in the training set is computed and contrasts with a prespecified threshold to effectively identify out-of-distribution (OOD) samples. The MSP method relies on the features' softmax distribution and considers their maximum value to differentiate between clean and OOD samples. ERNN employs evidence reconcile block to mitigate contradicting

Table 4.5: Comparative results of ***DISCERN*** for OOD detection.

Methods	OOD1 Samples					OOD2 Samples
	D1	**D2**	**D3**	**D4**	**D5**	**D6**
MahD [43]	10.78	14.00	19.67	23.33	29.18	22.83
SS [46]	32.11	16.00	14.52	26.67	21.94	68.98
ERNN [79]	43.95	58.00	52.91	56.67	56.66	n/a
MSP [47]	55.89	54.00	53.83	50.00	27.78	65.55
DISCERN	**60.44**	**61.58**	**63.21**	**62.92**	**64.29**	**75.38**

Note: The values signify DSR (in %). Bold entries represent the highest DSR. The label 'n/a' denotes the inapplicability of the ERNN method to OOD2 samples. Dataset details: **D1**:CVC-ColonDB, **D2**:Kvasir, **D3**:CVC-ClinicDB, **D4**:CVC-300, **D5**:ETIS, **D6**:Skin-Lesion.

testimony from the results and minimises errors in uncertainty quantification to significantly identify *near* OOD samples. Some of the existing OOD detection methods [49, 48] utilise network logits, which restrict detection capability as they behave identically for all samples. Moreover, SS and MahD methods require distance-related prespecified thresholds, which makes it challenging to tune and learn the different spatial properties of OOD test images, ultimately reducing detection accuracy. Furthermore, ERNN method performs classification tasks and is constrained to identify only *near* OOD. Hence, it is applicable for comparison with only OOD1 samples in ***DISCERN***.

In contrast to the previous methods, our proposed method, ***DISCERN***, achieves consistently high DSR values across all attack settings for each dataset. This improvement is attributed to the consideration of input variants and analysis of their behaviour in relation to the original input through consistency computation. The result of ***DISCERN*** is visually illustrated in Figure 4.6. It depicts how ***DISCERN*** determines the unknown sample type of a new sample by examining consistent relation between the input samples' prediction and their respective variants. Strong consistency signify clean samples, whereas weak consistency indicate adversarial or OOD samples.

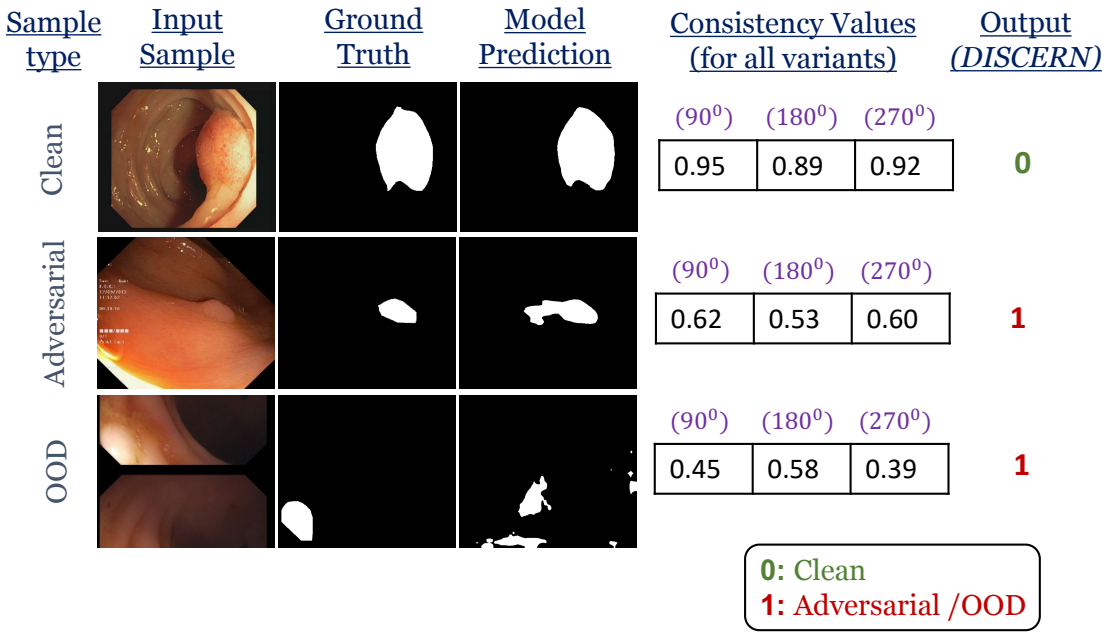

Figure 4.6: Figure depicts the result visualization of our proposed detection method, ***DISCERN***. Notably, 'Sample type' and 'Ground Truth' are indicated here only for better clarity and understanding. ***DISCERN*** itself does not rely on these elements.

4.2.5 Ablation Study

For rigorous evaluation, several ablation studies have been performed related to our proposed method, ***DISCERN***, as outlined below,

- To comprehensively evaluate the importance of diverse settings of attacks in ***DISCERN***, the DSR is calculated for U, T_w, and T_b attacks as demonstrated in Table 4.6. The outcomes demonstrate that ***DISCERN*** achieves higher DSR values across T_w compared to both U and T_b. In particular, during the T_w setting, the optimal performance is obtained in 70% of cases, while it is reduced to 25% and 5% for U and T_b, respectively. This is due to the fact that in the case of T_w (white target), the prediction is substantially distorted by adversarial perturbation, whereas during U and T_b, the prediction is slightly affected by perturbations. Consequently, we observe that T_w is more resistant to attacks but easier to detect, whereas U and T_b are more vulnerable

Table 4.6: ***DISCERN*** performance across different attack settings for adversarial detection.

Models	**Data**	U	T_w	T_b
		$DSR \uparrow$	$DSR \uparrow$	$DSR \uparrow$
PraNet [3]	CVC-300 [95]	93.333	94.167	80.000
	CVC-ClinicDB [96]	89.167	88.333	67.500
	CVC-ColonDB [97]	81.050	82.216	69.825
	ETIS [110]	65.663	65.361	57.229
	Kvasir [98]	92.929	90.909	74.747
SSFormer-S [14]	CVC-300 [95]	78.151	86.555	85.714
	CVC-ClinicDB [96]	87.705	91.803	81.967
	CVC-ColonDB [97]	70.588	74.370	67.367
	ETIS [110]	76.294	77.384	61.853
	Kvasir [98]	80.402	81.910	75.377
SSFormer-L [14]	CVC-300 [95]	75.000	81.667	76.667
	CVC-ClinicDB [96]	84.426	88.525	85.246
	CVC-ColonDB [97]	63.401	74.830	67.075
	ETIS [110]	77.005	79.144	67.112
	Kvasir [98]	85.000	87.000	77.500
UACANet-S [65]	CVC-300 [95]	90.833	93.333	82.500
	CVC-ClinicDB [96]	85.366	93.496	89.431
	CVC-ColonDB [97]	79.586	81.655	70.207
	ETIS [110]	73.803	72.676	59.155
	Kvasir [98]	87.940	85.930	84.422
UACANet-L [65]	CVC-300 [95]	82.500	91.667	86.667
	CVC-ClinicDB [96]	87.097	86.290	87.903
	CVC-ColonDB [97]	76.028	81.560	73.901
	ETIS [110]	79.625	76.944	68.365
	Kvasir [98]	89.500	89.500	81.000
CaraNet [66]	CVC-300 [95]	83.333	86.667	60.833
	CVC-ClinicDB [96]	83.740	85.366	65.041
	CVC-ColonDB [97]	73.490	77.172	52.577
	ETIS [110]	64.478	62.687	49.851
	Kvasir [98]	84.925	83.920	74.874

Note: The DSR values are in $\%$.

to attacks than T_w yet harder to detect.

- To thoroughly analyse the efficacy of diverse OOD samples in ***DISCERN***, the DSR

Table 4.7: ***DISCERN*** performance across various OOD samples for OOD detection.

Models	OOD1 Samples					OOD2 Samples
	D1	**D2**	**D3**	**D4**	**D5**	**D6**
UACANet-S [65]	63.289	60.000	62.903	55.833	64.082	72.865
UACANet-L [65]	54.211	60.590	60.806	60.833	65.867	73.772
PraNet [3]	58.553	57.500	60.484	64.167	65.612	85.231
CaraNet [66]	55.789	56.500	61.677	65.000	55.102	75.351
SSFormer-S [14]	63.684	66.500	68.871	62.500	66.122	72.487
SSFormer-L [14]	67.105	68.659	64.516	69.167	68.929	72.562

Note: The values depict DSR (in %). **D1**:CVC-ColonDB, **D2**:Kvasir, **D3**:CVC-ClinicDB, **D4**:CVC-300, **D5**:ETIS, **D6**:Skin-Lesion.

Table 4.8: Performance of ***DISCERN*** across diverse types of SVM kernel.

Models	RBF	Polynomial	Linear
UACANet-S [65]	**75.12**	72.26	72.83
UACANet-L [65]	72.75	70.36	**74.83**
PraNet [3]	**72.90**	71.20	69.83
CaraNet [66]	66.45	**69.10**	64.06
SSFormer-S [14]	**73.13**	71.30	69.11
SSFormer-L [14]	72.94	**75.25**	69.17

Note: The digits represent average DSR (in %).

Table 4.9: Overall computational time taken by ***DISCERN***.

Models	MIS Prediction Time	Additional Steps Time	Total Time
UACANet-S [65]	125.21	17.86	143.07
UACANet-L [65]	148.89	26.73	175.62
PraNet [3]	106.78	62.38	169.16
CaraNet [66]	164.95	13.28	178.23
SSFormer-S [14]	120.49	17.89	138.38
SSFormer-L [14]	137.23	20.51	157.74

Note: Additional Steps: creating variants, invert rotation of MIS prediction, measuring consistency values, and classifier-based final detection. This experiment is conducted on CVC-300 [95] datasets. The measured time is in seconds.

is computed specifically for OOD1 and OOD2 samples, as depicted in Table 4.7. The outcomes indicate that ***DISCERN*** obtains higher DSR values for OOD2 samples

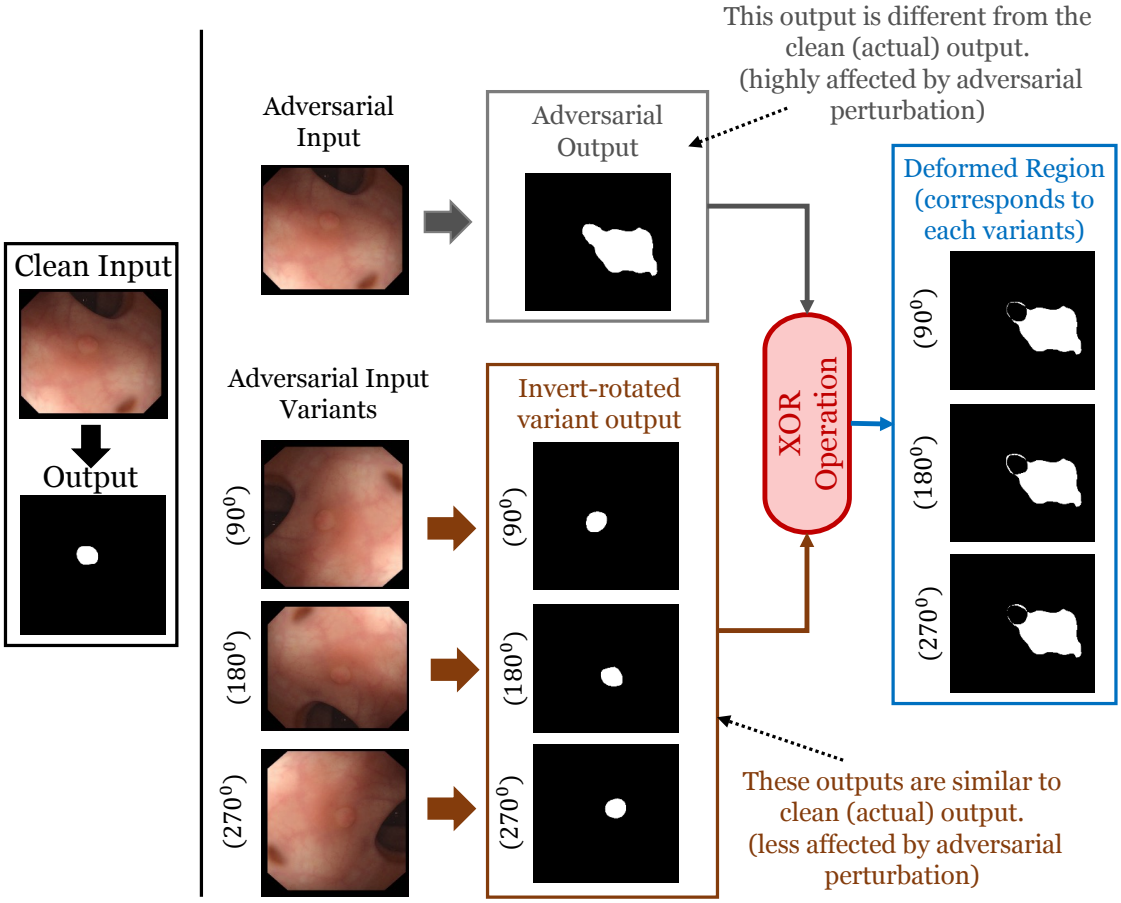

Figure 4.7: The deformed region is identified by employing the XOR function on MIS outputs of adversarial samples and their variants. Since the adversarial variants exhibit similar behaviour with the clean samples' output, their XOR with the corresponding adversarial outputs, which substantially deviate from clean ones, shows areas impacted by adversarial perturbations.

compared to OOD1 samples. It is noteworthy that OOD2 samples pertain to different MIS tasks (this work employs skin-lesion segmentation), while OOD1 samples are created using the IPP method [103]. Since the OOD1 samples are curated from ID samples, they are categorised as *near* OOD, encompassing a purely semantic variation from ID samples. In contrast, the OOD2 samples are classified as *far* OOD, exhibiting both semantic and domain shifts from the ID samples [115]. Consequently, there is strong consistent relation between the OOD1 samples' prediction and their respective variants, resulting in small DSR values.

Table 4.10: Comparison of MIS model performance after removing distorted regions.

Models	Original[a]	Adverasrial[b]	Removed Deformation[c]
UACANet-S [65]	0.90	0.39	0.59
UACANet-L [65]	0.91	0.27	0.73
PraNet [3]	0.87	0.43	0.61
CaraNet [66]	0.87	0.18	0.78
SSFormer-S [14]	0.88	0.21	0.64
SSFormer-L [14]	0.89	0.28	0.70

Note: The digits signify dice score measured between: a: the MIS prediction of clean sample and GT. b: the MIS prediction of adversarial sample and GT. c: the MIS prediction of adversarial sample (after eliminating the deformed region) and GT.

- ***DISCERN*** selects the optimal SVM kernel across each considered MIS model by leveraging the highest DSR values during training (kindly see Table 4.3). For better understanding, the average DSR values for each model are shown in Table 4.8, which is computed on test datasets considering all three kernels. The results demonstrate that selecting the right kernel based on training data enables achieving the highest DSR on the test dataset. In essence, Table 4.8 depicts that the best performance, which is over 70% DSR, is obtained by the ***RBF*** kernel across PraNet, SSFormer-S and UACANet-S models. Moreover, the *Polynomial* kernel yields the highest DSR values of 75.25% and 69.10% across SSFormer-L and CaraNet models, respectively. Furthermore, *Linear* kernel is chosen for the UACANet-L model, achieving a DSR of 74.83%.

- Table 4.9 represents the overall computational time, required for ***DISCERN***. This is analysed by computing each step time of ***DISCERN*** for each considered MIS model employing the CVC-300 polyp dataset [95]. The table indicates that the step of obtaining prediction from the MIS model takes substantially longer than the remaining steps, such as variant creation, invert rotation, consistency computation, and final

detection. As the ***DISCERN*** needs four runs of the MIS models to provide four predictions, one for input and three for its corresponding variants, the MIS predictions demand large computation. On the other hand, the remaining steps required little time, indicating the effectiveness and flexibility of ***DISCERN***.

- An important insight observed by ***DISCERN*** is that MIS prediction of adversarial variants reduces perturbation impact. On top of this observation, we examine the potential of extracting the deformed areas of MIS prediction caused by adversarial perturbation. This is accomplished by applying an XOR logic between the model's prediction on the adversarial input and its corresponding variants, resulting in deformed regions, as visualised in Figure 4.7. Further, the deformed regions are then eliminated from the adversarial prediction, and the refined output is evaluated against the GT. Their results are represented in Table 4.10 across all the considered MIS models, depicting enhanced model performance after the removal of deformed regions from adversarial prediction. It is noteworthy that this performance differs slightly from the actual one since the rotational variant prediction lessens the influence of perturbations while retaining certain residual effects.

4.3 Discussion

An important observation employed by ***DISCERN*** is that there is a strong consistency in clean samples with their corresponding variants. This insight has been previously employed by [2] to advance the MIS model performance. However, ***DISCERN*** incorporates this observation to effectively identify the adversarial or OOD samples while eliminating the GT and network retraining. Nevertheless, a notable limitation of ***DISCERN*** is the lower DSR observed in certain cases during the detection of OOD1 samples. As OOD1 samples

are generated from ID samples, they demonstrate similar properties to ID samples. Consequently, they are difficult to distinguish from clean samples, offering lower DSR. Another constraint of *DISCERN* is that its computational time depends upon the number of input variants. Since it requires input's MIS prediction and all relative variants to compute consistency, the model must be executed separately for each prediction, leading to increased computational time.

4.4 Summary

This chapter has introduced a method, ***DISCERN***, with the goal of detecting adversarial and OOD samples in MIS. A key observation underlying this method is that clean samples' MIS predictions show strong consistency with their relative rotated variants, while adversarial and OOD samples are less consistent. Based on this insight, we have concluded that such findings can be useful for the effective identification of adversarial and OOD samples. Thus, we have initially produced the input variants by a rotational transformation of input and obtained their predictions from well-known MIS models. Subsequently, consistency has been measured between the input samples' MIS prediction and their respective variants. Eventually, the computed consistencies have been employed to classify samples as clean, adversarial and OOD. The experiments conducted on open-sourced datasets have demonstrated that ***DISCERN*** outperform all the existing methods, offering high DSR values.

Chapter 5

Contrastive Multitasking Adversarial Defence on MIS

The effectiveness of DL-based MIS models is significantly compromised by well-engineered adversarial attacks. Such attacks introduce minute and imperceptible perturbations into the input that mislead prediction. This problem is particularly severe in medical images, as their complicated texture may cause the model to emphasise irrelevant ROIs. To improve model robustness, empirical defence methods [35, 80, 36, 81] primarily increase training samples. In this direction, the DA [35, 80] employs clean samples with their geometrical and pixel-level variants, while AT [36, 81] mixes the clean samples with adversarial ones. Since DA excessively relies on augmented features, it becomes challenging for the DL model to precisely learn adversarial features. Meanwhile, AT struggles to generalise well on clean samples; thus, it fails to maintain the model's performance [37]. Although these existing defences have shown promise in non-medical fields, their influence in the medical domain still requires thorough investigation. Additionally, contrastive [38] and multitask [39, 40] learning have been utilised in improving adversarial robustness. Unlike contrastive learning [38], there is a lack of enforced learning between the synthetically added samples (adversarial and augmented) that contrast with clean ones. Moreover, multitask learning

predominantly depends on auxiliary tasks, which should be selected objectively [116] to prevent false robustness [117]. These research gaps require a comprehensive exploration of whether contrastive learning can substantially diminish the efficacy of adversarial attacks and how its integration with multitask learning influences adversarial robustness in DL-based MIS models.

This chapter proposes a novel adversarial defence, ***RELIVE***, which stands for Cont***R***astiv***E*** Mu***LtI***tasking Ad***V***ersarial D***E***fence. It aims to enhance the adversarial robustness of the DL-based MIS model while substantially improving their performance. ***RELIVE*** comprises contrastive learning, multitask learning, and their fusion-based defence. The contrastive learning-based defence enforces the model to capture the similarity between clean, adversarial, and augmented samples. The multitask learning-based defence identifies multiple auxiliary tasks based on their weak correlation with the main task, offering generalised feature representation. The contrastive multitask fusion-based defence integrates the proposed generic architecture of a multitask model with contrastive learning, further empowering the model's adversarial robustness. The chapter is arranged as follows: Section 5.1 demonstrates the workflow of our proposed defence, ***RELIVE***. Section 5.2 presents their experimental outcomes. Section 5.3 summarises the complete chapter.

5.1 Proposed Adversarial Defence: *RELIVE*

This section describes our proposed defence, ***RELIVE***, with the aim of strengthening the adversarial robustness of DL-based MIS models. Initially, it explores the importance of contrastive learning in reducing the efficacy of adversarial perturbation in DL-based MIS models (kindly refer Section 5.1.1). It leverages the observation that when training involves clean, adversarial, and augmented samples to capture similar feature representations, the model implicitly learns to diminish adversarial perturbations. Since [39] suggests that care-

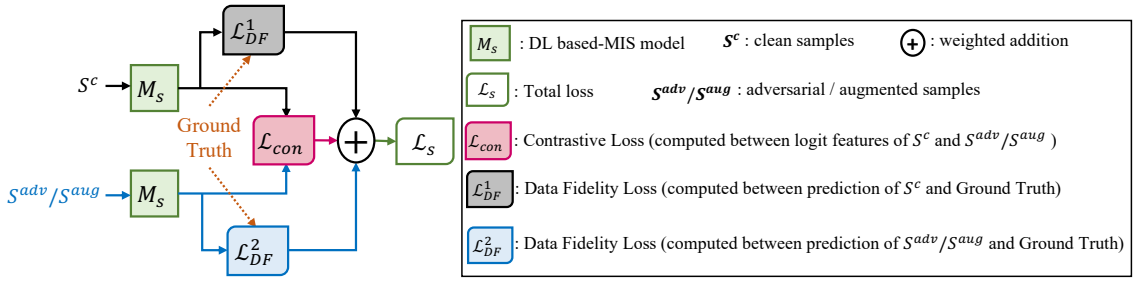

Figure 5.1: Our proposed contrastive learning-based defence. Initially, S^c and S^{adv}/S^{aug} samples are fed into the M_s. Subsequently, the contrastive loss ($\mathcal{L}_{con}$) captures their similar features, and data fidelity losses ($\mathcal{L}_{DF}^1$ and $\mathcal{L}_{DF}^2$) assure accurate model learning. Eventually, the final MIS loss ($\mathcal{L}_s$) is a weighted addition of all these losses.

ful selection of auxiliary tasks in any multitasking model will help to eliminate the impact of adversarial attack as it learns generic feature representation. Thus, we subsequently examine the significance of multitask learning for improving adversarial robustness, as given in Section 5.1.2. Eventually, we leverage the individual benefits of contrastive and multitask learning by integrating them to enhance the robustness of DL-based MIS models. In particular, the fusion-based defence specifically applies contrastive learning to the MIS tasks within the proposed multitask model, as presented in Section 5.1.3. The generation of adversarial samples is provided in Section 5.1.4. The complete workflows for contrastive learning-based, multitask learning-based, and contrastive multitasking fusion-based defences are visualised in Figures 5.1, 5.2, and 5.3, respectively and their steps are outlined in Algorithm 5.1, 5.2 and 5.3, respectively.

5.1.1 Adversarial Defence by Contrastive Learning

Contrastive learning [118] boosts the efficacy of the DL-based MIS model by keeping the positive samples closer to the input feature (or clean samples). Some existing studies [119] also incorporate negative samples, pushing them farther from clean ones to improve MIS model performance. It is noteworthy that our proposed defence, ***RELIVE***, only utilises positive samples to examine the importance of contrastive learning in strengthening the

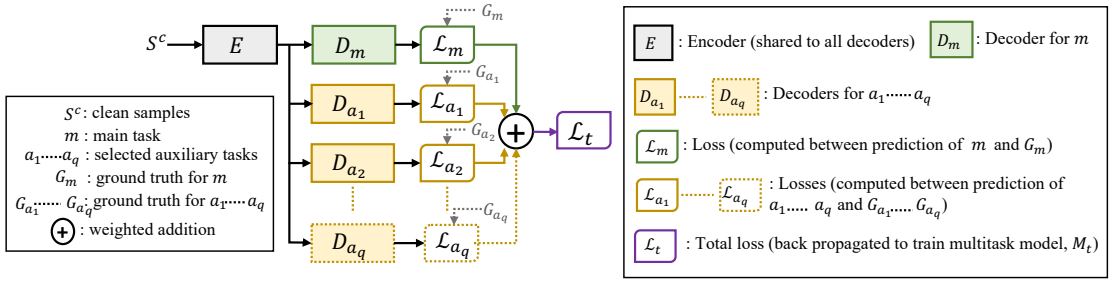

Figure 5.2: Our proposed multitask-learning-based defence. The encoder (E) accepts S^c, shared with several decoders ($D_m, D_{a_1}, \cdots, D_{a_q}$). The loss for each task ($\mathcal{L}_m, \mathcal{L}_{a_1}, \cdots, \mathcal{L}_{a_q}$) is computed against its ground truth, added, and backpropagated, enabling the trained model to be minimally vulnerable to adversarial perturbations.

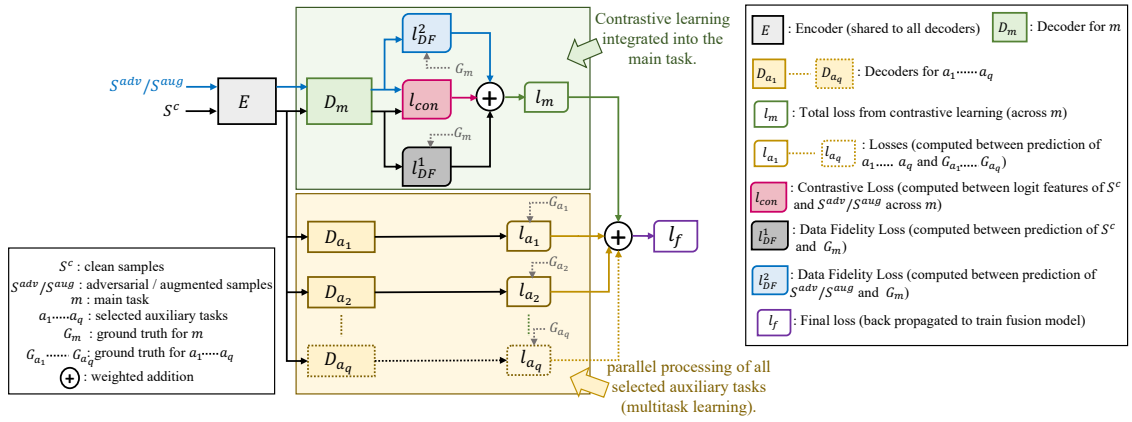

Figure 5.3: Our proposed contrastive multitask fusion-based defence. S^c and S^{adv}/S^{aug} pass through E and linked to $D_m, D_{a_1}, \cdots, D_{a_q}$. All the losses related to contrastive learning ($\mathcal{L}_{con}$, $\mathcal{L}^1_{DF}$ and $\mathcal{L}^2_{DF}$) are computed at D_m for the main task to get l_m, which is further added to $l_{a_1}, \cdots, l_{a_q}$ for auxiliary tasks. The final loss (l_f) sums all decoder losses and backpropagated, making the model highly resistant to adversarial attacks.

Algorithm 5.1 Proposed contrastive learning-based defence.

Require: Clean Sample S^c; Synthetically Generated Sample $S^{'}$; MIS model M_s.
Ensure: Trained M_s, resistant to adversarial attack.
for each epoch **do** ▷ loop for each epoch
  for each batch-wise sample **do** ▷ loop for each batch
    $f^c = M_{s_p}(S^c)$ and $f^{'} = M_{s_p}(S^{'})$ ▷ output at logit layer p
    $Y^c = M_s(S^c)$ and $Y^{'} = M_s(S^{'})$ ▷ output at final layer
    $\mathcal{L}_{con}(f^c, f^{'}) = \frac{1}{N}\sum_{j=1}^{N}\left(f^c(j) - f^{'}(j)\right)^2$ ▷ contrastive loss computation
    $\mathcal{L}_{DF} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{BCE}}$ ▷ refer Eq. (5.6),(5.7),(5.8)
    $\mathcal{L}^1_{DF} = \mathcal{L}_{DF}(Y^c)$ and $\mathcal{L}^2_{DF} = \mathcal{L}_{DF}(Y^{'})$ ▷ data fidelity loss computation
    $\mathcal{L}_s = \sigma * \mathcal{L}_{con} + \mu_1 * \mathcal{L}^1_{DF} + \mu_2 * \mathcal{L}^2_{DF}$ ▷ final training loss
    $\mathcal{L}_s$ is backpropagated to train the model.
  end for
end for
return Trained M_s, resistant to adversarial attack.

Algorithm 5.2 Proposed multitask learning-based defence.

Require: Clean Sample S^c; Multitask model M_t; Main Task m; Auxiliary Task a.
Ensure: Trained M_t, resistant to adversarial attack.
Following Section 5.1.2.1 to select a.
for each epoch **do** ▷ loop for each epoch
  for each batch-wise sample **do** ▷ loop for each batch
    $EU = M^e_t(S^c)$ ▷ encoder output
    $Y^m = M^{d^m}_t(EU)$ ▷ decoder output for m
    for $z \in 1\ to\ q$ **do** ▷ loop for each a
      $Y^{a_z} = M^{d^{a_z}}_t(EU)$ ▷ decoder output for a
    end for
    $\mathcal{L}_t = \mathcal{L}_m(Y^m, G_m) + \sum_{z=1}^{q}\mathcal{L}_{a_z}(Y^{a_z}, G_{a_z})$ ▷ final training loss (refer Eq. (5.16))
    $\mathcal{L}_t$ is backpropagated to train the model.
  end for
end for
return Trained M_t, resistant to adversarial attack.

Algorithm 5.3 Proposed contrastive multitask fusion-based defence.

Require: Clean Sample S^c; Synthetically Generated Sample $S^{'}$; Multitask model M_t; Main Task m; Auxiliary Task a.
Ensure: Adversarially robust trained fusion model.
Following Section 5.1.2.1 to select a.
for each epoch **do** ▷ loop for each epoch
for each batch-wise sample **do** ▷ loop for each batch
$EU^c = M_t^e(S^c)$ and $EU^{'} = M_t^e(S^{'})$ ▷ encoder output (refer Algorithm 5.2)
$F^c = M_{t_p}^{d^m}(EU^c)$ and $F^{'} = M_{t_p}^{d^m}(EU^{'})$ ▷ decoder output at logit layer p for m
$Y_{s^c}^m = M_t^{d^m}(EU^c)$ and $Y_{s^{'}}^m = M_t^{d^m}(EU^{'})$ ▷ decoder output at final layer for m
$l_{con} = \mathcal{L}_{con}(F^c, F^{'})$ ▷ contrastive loss computation (refer Algorithm 5.1)
$l_{DF}^1 = \mathcal{L}_{DF}(Y_{s^c}^m)$ and $l_{DF}^2 = \mathcal{L}_{DF}(Y_{s^{'}}^m)$ ▷ data fidelity loss computation (refer Algorithm 5.1)
$l_m = \sigma * l_{con} + \mu_1 * l_{DF}^1 + \mu_2 * l_{DF}^2$ ▷ Total loss for m (refer Algorithm 5.1 and Eq. (5.22))
Compute $\mathcal{L}_a$ from Algorithm 5.2 for all selected auxiliary tasks a.
$l_f = l_m + \mathcal{L}_a$ ▷ final training loss (refer Eq. (5.23))
l_f is backpropagated to train the model.
end for
end for
return Adversarily robust trained fusion model.

robustness of DL-based MIS models. Specifically, we enforce the model to capture the similarity between the feature representations of clean and synthetically generated positive samples, which is achieved by optimising the contrastive loss between these samples (refer Figure 5.1). Notably, ***RELIVE*** treats adversarial and augmented samples as synthetically generated positive samples. Further, it computes the data fidelity loss individually for clean and positive samples with GT to ensure that the model is producing accurate predictions, guiding the model to produce outputs closer to the GT. While the contrastive loss encourages the model to learn identical features for clean, adversarial, and augmented samples, the data fidelity loss ensures correct predictions for each sample. Thus, we examine these two losses inherently in contrastive learning to effectively diminish the efficacy of adversarial attacks in the DL-based MIS model while maintaining its performance.

5.1.1.1 Contrastive Loss

Let M_s be the DL-based MIS model, which performs contrastive learning by using synthetically generated samples $S^{'}$ and clean samples S^c. We refer logit features of each sample for contrastive loss computation because these features helps the model to adapt more detailed representations, which are essential to capture generalised and well-featured embeddings. These logit features f^c and $f^{'}$ (corresponds to S^c and $S^{'}$, respectively) can be expressed as,

$$f^c = M_{s_p}(S^c) \tag{5.1}$$

$$f^{'} = M_{s_p}(S^{'}); \quad \text{where, } S' \in \{S^{adv}, S^{aug}\} \tag{5.2}$$

where p indicates the logit layer. S^{adv} and S^{aug} depict the adversarial and augmented samples, respectively (refer to Sections 5.1.4 and 5.2.3 for generation of these samples). The logit features, f^c and $f^{'}$, should be similar for defending against adversarial attacks. Thus, we calculate contrastive loss, $\mathcal{L}_{con}$, between these features. ***RELIVE*** employs L_2 loss [120] as contrastive loss, formulated as,

$$\mathcal{L}_{con}(f^c, f^{'}) = \frac{1}{N}\sum_{j=1}^{N}\Big(f^c(j) - f^{'}(j)\Big)^2 \tag{5.3}$$

where N indicates the total number of samples. $\mathcal{L}_{con}$ pushes the model to better align f^c with $f^{'}$, advancing adversarial robustness.

5.1.1.2 Data Fidelity Loss

The data fidelity loss, $\mathcal{L}_{DF}$, verifies that the model is providing true prediction. It is calculated separately for each sample between their prediction and GT. Mathematically,

$$Y^c = M_s(S^c) \tag{5.4}$$

$$Y^{'} = M_s(S^{'}); \quad \text{where } S' \in \{S^{adv}, S^{aug}\} \tag{5.5}$$

where Y^c and $Y^{'}$ are the model predictions across S^c and $S^{'}$, respectively. These predictions should be similar to GT (G) for precise model learning. Thus, ***RELIVE*** defines the $\mathcal{L}_{DF}$ as a joint dice BCE loss [3], expressed for any prediction, O, as,

$$\mathcal{L}_{DF}(O) = \mathcal{L}_{\text{Dice}}(O) + \mathcal{L}_{\text{BCE}}(O) \tag{5.6}$$

where

$$\mathcal{L}_{\text{Dice}}(O) = 1 - \frac{2\sum_{i=1}^{v} O(i)G(i)}{\sum_{i=1}^{v} O(i) + \sum_{i=1}^{v} G(i)} \tag{5.7}$$

and

$$\mathcal{L}_{\text{BCE}}(O) = -\frac{1}{N}\sum_{i=1}^{N} (G(i)\log(O(i)) + (1 - G(i))\log(1 - O(i))) \tag{5.8}$$

hence, the $\mathcal{L}_{DF}$ for S^c and S' is denoted by $\mathcal{L}^1_{DF}$ and $\mathcal{L}^2_{DF}$, respectively. Mathematically (refer Equation (5.6)),

$$\mathcal{L}^1_{DF} = \mathcal{L}_{DF}(O = Y^c) \tag{5.9}$$

$$\mathcal{L}^2_{DF} = \mathcal{L}_{DF}(O = Y^{'}) \tag{5.10}$$

These losses help to enhance the model performance by keeping Y^c and $Y^{'}$ closer to G.

5.1.1.3 Total Loss

The proposed adversarial defence by contrastive learning trains M_s, employing S^c, S^{adv}, and S^{aug} and learns their associated features, followed by $\mathcal{L}_{con}$ and $\mathcal{L}_{DF}$ computation. Mathematically,

$$\mathcal{L}_s = \sigma * \mathcal{L}_{con} + \mu_1 * \mathcal{L}^1_{DF} + \mu_2 * \mathcal{L}^2_{DF} \tag{5.11}$$

where $\mathcal{L}_s$ depicts the total loss, which is the weighted addition of all the losses. These weights are denoted as σ, μ_1 and μ_2 (refer Section 5.2.3). The trained model receives the

new sample during inference, producing a prediction that is resistant to adversarial perturbation.

5.1.2 Adversarial Defence by Multitask Learning

Multitask learning allows the model to simultaneously execute several tasks, enabling it to offer generalised feature representation by capturing a variety of features associated with these tasks. Moreover, it reduces the efficacy of adversarial perturbations on the model trained for separate tasks [39]. Our proposed defence, ***RELIVE***, leverages this observation by modifying the single-task DL-based MIS model (M_s) to a multitasking model. In essence, M_s is designed in such a way that it can perform various selected auxiliary tasks. Thus, our proposed multitask model selects the MIS to be the main task, followed by several selected auxiliary tasks (refer Figure 5.2). The explanations of auxiliary task selection and generic model representation are provided as follows:

5.1.2.1 Selecting Auxiliary Tasks

The arbitrary inclusion of auxiliary tasks in any multitask model could be fatal, leading to a deceptive impact of robustness [117]. Thus, the selection of auxiliary tasks should be based on their relevance to the main task. In this direction, our proposed multitask model examines the dependencies between the individual performance of the main task and all the auxiliary tasks. Let us assume that m be a main task and a indicate several auxiliary tasks such that $a \in \{a_1, a_2, \cdots a_q\}$. Mathematically,

$$\rho = \frac{\sum_{i=1}^{n}(Y^m(i) - \bar{Y^m})(Y^a(i) - \bar{Y^a})}{\sqrt{\sum_{i=1}^{n}(Y^m(i) - \bar{Y^m})^2 \sum_{i=1}^{n}(Y^a(i) - \bar{Y^a})^2}} \tag{5.12}$$

where Y^m and Y^a represent the predictions of m and a, respectively. ρ depicts the dependencies between Y^m and Y^a. The proposed multitask model considers the Pearson

correlation coefficient [121, 2] as ρ. n indicates total sample counts and $\bar{h}$ represents the mean value of h. If m and a exhibit low correlation, it enhances the model's robustness. In such cases, the a can be considered as one of the auxiliary tasks for the model. Conversely, highly correlated tasks share similar characteristics and may acquire redundant information, which can hinder performance and increase the model's vulnerability to attacks, ultimately weakening its adversarial robustness. Thus, our proposed multitask model highly depends upon the value of ρ, which is compared by the prespecified threshold λ for final auxiliary task selection. Mathematically,

$$\varphi = \begin{cases} a \textit{ is selected} \ ; & \textit{if } \rho <= \lambda \\ a \textit{ is rejected} \ ; & \textit{if } \rho > \lambda \end{cases} \tag{5.13}$$

where φ depicts the condition to choose a. The hyperparameter selection of λ is explained in Section 5.2.3.

5.1.2.2 Generalised Multitask Model

Our proposed defence, ***RELIVE***, introduces a generalised multitask model to diminish the effect of adversarial attacks on the DL-based MIS model. It consists of a shared encoder with diverse task-oriented decoders. The encoder permits the model to adapt general features suitable for all the tasks, while the decoder enables the model to adaptively function for various objectives simultaneously. Such encoder-decoder-based models handle multiple tasks built into a single model. Thus, our proposed multitask model is designed by integrating diverse decoders into existing single-task DL-based MIS models. These decoders correspond to several selected auxiliary tasks. Suppose the proposed multitask model is denoted by M_t, exhibiting a shared encoder M_t^e and task-oriented decoders, M_t^d. The output

EU , at M_t^e is given by,

$$EU = M_t^e(S^c) \tag{5.14}$$

where EU is applied to all the task-specific decoders. Let $M_t^{d^m}$ and $M_t^{d^a}$ are the decoders for m and a, respectively. Their corresponding predictions, Y^m and Y^a, are given by,

$$Y^m = M_t^{d^m}(EU) \quad \text{and} \quad Y^a = M_t^{d^a}(EU) \tag{5.15}$$

where Y^a is calculated for all selected auxiliary task decoders such that $Y^a \in \{Y^{a_1}, Y^{a_2}, \cdots Y^{a_q}\}$. The model is trained by optimising predictions for each task by computing individual losses, compared with their respective GTs, and taking a summation of them to achieve total loss, $\mathcal{L}_t$. Mathematically,

$$\mathcal{L}_t = \mathcal{L}_m(Y^m, G_m) + \sum_{z=1}^{q} \mathcal{L}_{a_z}(Y^{a_z}, G_{a_z}) \tag{5.16}$$

where $\mathcal{L}_m$ and G_m denote the ground truth and loss for the main tasks, respectively. Likewise, $\mathcal{L}_a \in \{\mathcal{L}_{a_1}, \mathcal{L}_{a_2}, \cdots \mathcal{L}_{a_q}\}$ and $G_a \in \{G_{a_1}, G_{a_2}, \cdots G_{a_q}\}$ depicts losses and ground truth for each auxiliary task, respectively. q indicates the total number of considered auxiliary tasks. This $\mathcal{L}_t$ is backpropagated to train M_t. Since the resulting trained multitask model, M_t, is capable of capturing different features across several tasks, it is highly resilient to adversarial attacks. It offers generic feature representation, which ultimately advances the adversarial robustness of the DL-based MIS model. All the experimental settings related to M_t are presented in Section 5.2.3.

5.1.3 Fusing Contrastive and Multitask Learning

As contrastive and multitask learning advances the robustness of DL-based MIS models, Our proposed defence, ***RELIVE***, leverages their individual benefits to make the model ad-

versarially resistant. To the end, we consolidate these two learnings to devise a contrastive multitask fusion-based defence (refer Figure 5.3). While contrastive learning accumulates identical characteristics of clean, adversarial, and augmented samples to enhance adversarial robustness without diminishing model performance, multitask learning allows the model to learn generic features, helping to increase their performance and robustness. Our proposed fusion-based defence simultaneously fed clean, adversarial, and augmented samples into the multitask model, which performs the main task along with multiple selected auxiliary tasks (refer Section 5.1.2.1).

As our primary objective is to elevate the adversarial robustness of the MIS task, we exclusively apply contrastive learning to the MIS task (main task). Thus, for the main task, we share the encoder with two decoders to simultaneously process clean and synthetically generated adversarial and augmented samples. Let the encoder, M_t^e, accepts S^c and $S^{'}$ as input, then their respective outputs, EU^c and $EU^{'}$, are given by (refer Equation (5.14)),

$$EU^c = M_t^e(S^c) \quad \text{and} \quad EU^{'} = M_t^e(S^{'}) \tag{5.17}$$

where U^c and $U^{'}$ are applied to the decoder $M_t^{d^m}$ across each samples. Their respective features, F^c and $F^{'}$, at the logit layer p (refer Section 5.1.1) can be represented as (refer Equations (5.1) and (5.2)),

$$F^c = M_{t_p}^{d^m}(EU^c) \quad \text{and} \quad F^{'} = M_{t_p}^{d^m}(EU^{'}) \tag{5.18}$$

here $M_{t_p}^{d^m}$ defines the decoder across m at p for M_t. The F^c and $F^{'}$ should be consistent with each other for the advancement of adversarial robustness. For this purpose, we compute the contrastive loss between these features, represented as (refer Equation (5.3)),

$$l_{con} = \mathcal{L}_{con}(F^c, F^{'}) \tag{5.19}$$

Likewise, the data fidelity loss is computed to maintain the model performance. It is calculated between individual sample predictions $Y^m_{s^c}$ and $Y^m_{s'}$ for S^c and $S^{'}$, respectively, and corresponding GT. From Eq. (5.15), these predictions can be expressed as,

$$Y^m_{s^c} = M^{d^m}_t(EU^c) \quad \text{and} \quad Y^m_{s'} = M^{d^m}_t(EU^{'}) \tag{5.20}$$

Following Equations (5.9) and (5.10), data fidelity losses, l^1_{DF} and l^2_{DF} are calculated as,

$$l^1_{DF} = \mathcal{L}_{DF}(O = Y^m_{s^c}) \quad \text{and} \quad l^2_{DF} = \mathcal{L}_{DF}(O = Y^m_{s'}) \tag{5.21}$$

In our proposed fusion-based defence, the objective function l_m for the main task (m) can be represented as (refer to Equation (5.11)),

$$l_m = \sigma * l_{con} + \mu_1 * l^1_{DF} + \mu_2 * l^2_{DF} \tag{5.22}$$

where σ, μ_1, μ_2 are the weights for each loss (see Section 5.2.3). The final loss, l_f, of the proposed fusion-based defence is the summation of the loss corresponds to the main task (l_m) and the considered auxiliary task losses ($\mathcal{L}_a$) (refer Equation (5.16)). Thus, we replace $\mathcal{L}_m$ with l_m in Equation (5.16) to achieve l_f, expressed as,

$$l_f = l_m + \sum_{z=1}^{q} \mathcal{L}_{a_z} \tag{5.23}$$

where l_f is the combination of l_m and $\mathcal{L}_a \in \{\mathcal{L}_{a_1}, \mathcal{L}_{a_2}, \cdots \mathcal{L}_{a_q}\}$. This fusion of contrastive and multitask features greatly enhances adversarial robustness, strengthening the model's resistance to adversarial attacks.

5.1.4 Generating Adversarial Samples

The DL-based MIS models are extensively impacted by adversarial attacks [31]. Thus, it is essential to comprehend the production of adversarial samples to defend against these attacks. Thus, we attack the DL-based MIS models (threat model) to devise adversarial samples. Attacking the segmentation model is challenging as it involves pixel-wise classification in which each pixel is considered a target to be misclassified [20]. Conversely, the classification model has only one target to attack [105]. Thus, the adversarial sample, S^{adv}, is crafted by imposing a small amount of imperceptible perturbation ϵ in the gradient direction of adversarial loss l^{adv}. Mathematically,

$$S_0^{adv} = S^c \tag{5.24}$$

$$S_{i+1}^{adv} = clip\left(S_i^{adv} - \epsilon * \text{sign}(\nabla_{S_i^{adv}} l^{adv}\left(M_s(S_i^{adv}), T\right))\right) \tag{5.25}$$

$$S_{i+1}^{adv} = clip\left(S_i^{adv} + \epsilon * \text{sign}(\nabla_{S_i^{adv}} l^{adv}\left(M_s(S_i^{adv}), Y^c\right))\right) \tag{5.26}$$

where M_s, T and Y^c are the threat model, desired target, and MIS prediction, respectively. The subscript i denotes step count, and $clip$ depicts the clipping operation, which assures S_{adv} will not go beyond the defined range [20]. The targeted and untargeted attacks are represented by Equations (5.25) and (5.26), respectively. Following [106][1], these attacks are iteratively executed until the predefined stopping condition are satisfied, based on T and the maximum number of iterations, i^{max}. Samples fulfilling the stopping condition are selected as S_{adv}. The stopping condition is defined as either a target-dependent measure exceeding a threshold or the maximum iteration count achieved. Kindly see Section 5.2.3 for a comprehensive overview of the attack settings.

[1] https://github.com/SnehaShukla937/MEDIS_ATTACK

5.2 Experimental Results

5.2.1 Dataset and Metrics

The proposed defence, ***RELIVE***, is experimented on the v7-Darwin Chest X-Ray dataset [122] for lung segmentation. We use 2000 training and 1000 testing samples for the model. The robustness of the model is investigated by employing 1000 samples for attack performance. ***RELIVE*** utilises the following evaluation metrics:

- **Attack Success Rate (ASR) :** It is defined as the proportion of the count of successfully attacked samples to the total count of samples [106].
- **Average distortion (δ_∞) :** Over the total count of successful S^{adv}, it is the average of L_∞ distance, which is computed between S^c and S^{adv} [106].
- **Dice (dsc) :** It examines the overlap between MIS prediction and ground truth [108, 2].
- **Pearson Correlation (ρ) :** It quantifies the linear relation between variables lies in the ranges from 0 to 1 [121, 2] (Kindly refer Equation (5.12)).

Our proposed defence, ***RELIVE***, examines the adversarial robustness using ASR and δ_∞ while assessing the MIS model performance by dsc. The auxiliary task is selected by using ρ in the proposed multitask model.

5.2.2 Utilised MIS Models

The efficacy of ***RELIVE*** is tested on multiple existing MIS models. The first model is SSFormer [2] [14], which is designed to segment polyps from colonoscopic images and

[2] https://github.com/Qiming-Huang/ssformer

characterised by an encoder-decoder architecture. It employs a pyramidal encoder and progressive locality decoder to learn generic ROI and achieve an accurate segmentation mask. The SSFormer comes in two publicly available versions: Standard (S) and Large (L), based on varying scale sizes of encoders. Our proposed defence, ***RELIVE*** employs both versions of SSFormer. The second model is PraNet [3] [3], which is based on the popular UNet model [13]. PraNet provides the output of the coarse ROI of the colonoscpic image by employing a PPD and improves its boundary with a RA module. Likewise, the third model is CaraNet [4] [66], which utilises channel-wise future pyramid modules to extract high-level features and apply them to the axial reverse attention. This will generate segmentation masks of small medical objects.

5.2.3 Experimental Settings

All the experiments of ***RELIVE*** are executed on a server having a Nvidia V100 GPU, an Intel Xeon Gold 6132 CPU, and 192 GB of RAM. The adversarial robustness is evaluated on well-known PraNet [3], SSFormer [14] and CaraNet [66] MIS models. To attack these models, we use untargeted (U) (refer Equation 5.26) and two different targeted (T_w and T_b) (refer Equation 5.25) settings. The T_w exhibits all the pixel values as 1 (pure white image), while T_b contains the entire pixel values as 0 (pure black image). The step size (ϵ) for the attack is tuned to $1/255$. Smaller ϵ leads to quantisation error whereas larger ϵ includes large perturbation into input, limiting the efficacy of attack [111]. The maximum iteration (i^{max}) is set as 10, restricting the amount of added adversarial perturbation into the input. The target-dependent measure is chosen as dice with a threshold of 0.7 for T_w and U attacks, while it is mIOU with a threshold of 0.5 for T_b attack [106]. The loss (l^{adv}) for the attack is selected as an addition of wBCE and wIOU [3]. Notably, these aforementioned settings are

[3] https://github.com/DengPingFan/PraNet
[4] https://github.com/AngeLouCN/CaraNet

utilised for examining the adversarial robustness of DL-based MIS models and the creation of adversarial samples (S^{adv}).

Our proposed multitask model's efficacy depends upon the appropriate choice of auxiliary tasks. This is achieved by computing the correlation between the performances of the main and different auxiliary tasks. As the object detection and image classification tasks are less correlated with the MIS task, we select them as auxiliary tasks for the multitask model. The correlation threshold (λ) is set as 0.5 based on performing a grid search on the training dataset, achieving optimal robustness results (refer to Table 5.3). The architecture details of multitask model are as follows: the encoder incorporates existing PVT architecture [123], the MIS decoder follows SSFormer architecture [14], and the object detection decoder considers DETR architecture [124]. For the classification decoder, the encoder output is applied to several Conv2D layers with ReLU activation function and linear layers, followed by the softmax layer. We employ 2000 images of the lung dataset [122] to train the multitask model, which executes MIS, object detection, and classification tasks simultaneously. The training is performed by setting epoch, batch size, and learning rate as 200, 6, and 0.0001, respectively. We impose openCV operations on the MIS masks to generate GT bounding boxes for object detection. The training loss for the multitask model is considered as follows: BCE Dice loss [3] for segmentation, a combination of GIOU, BCE, and L_2 loss [124] for object detection, and BCE loss [125] for classification.

For rigorous analysis, we generate the augmented sample (S^{aug}) using diverse methods such as brightness enhancement, inpainting, horizontal and vertical Sobel filtering, colour blending, histogram equalisation, noise addition, shifting in all directions, and JPEG compression. We tune the contrastive loss weights σ, μ_1, and μ_2 as 0.5, 0.25, and 0.25, respectively, using the grid search operation.

5.2.4 Comparative Evaluation

Table 5.1 and 5.2 represent the comparative analysis of our proposed defence, ***RELIVE***. It shows that ***RELIVE*** surpasses the existing defences across various existing MIS models, successfully mitigating the efficacy of adversarial attacks with little improvement in model performance. In particular, it decreases the attack success rate (ASR) up to 0% within the maximum average perturbation (δ_∞), ultimately advancing the model's robustness. Moreover, it obtains a notable dice score (dsc), signifying improvement in model performance. It is observed from Table 5.1 that if the trained MIS model on S^c is adversarially attacked, then adversarial perturbation significantly compromises the model's robustness with ASR up to 100% at minimum δ_∞. It happens particularly across all the models for targeted (T_w and T_b) attacks except CaraNet in T_w and the PraNet model in untargeted (U) attacks. Consequently, we observe from Table 5.2 that a substantial decline occurs in dsc, with the average value over 90% to below 60% across all models.

Likewise, when the model training involves augmented samples (S^{aug}) (refer Table 5.1), the ASR is lowered to below 70% with moderate increment in δ_∞ for most of the models across all attack settings, except PraNet in the U attack. It is noteworthy that ASR decreases to below 40% across SSFormer in T_w and T_b settings with large δ_∞. Further, Table 5.2 signifies that the augmentation also handles the model performance, as indicated by a slight advancement in dsc for all models except SSFormer-L. Since the augmentation [35] imposes geometric and pixel-level transformation in the input, which is substantially different from adversarial perturbations, resulting in the avoidance of adversarial features. Hence, augmentation successfully performs model generalisation but fails to enhance adversarial robustness. Additionally, the excessive dependence on augmented data may lead to model overfitting to these transformations. As a result, this will provide the deceptive perception of enhanced robustness, in which the MIS model might be susceptible to adversarial samples

Table 5.1: Comparative adversarial performance of ***RELIVE***.

	Methods	Models	Samples			U		T_w		T_b	
			S^c	S^{adv}	S^{aug}	ASR	δ_∞	ASR	δ_∞	ASR	δ_∞
	No Defence	M1	✓	×	×	90	6.10	100	4.18	100	4.59
		M2				87	6.50	100	4.02	100	4.70
		M3				100	5.23	100	4.98	100	4.21
		M4				95	5.98	97	5.34	100	4.95
Existing Defence	Augmentation	M1	×	×	✓	54	8.82	26	9.31	37	8.95
		M2				59	8.43	21	9.45	39	8.49
		M3				72	7.20	59	8.40	62	8.05
		M4				65	7.93	45	8.86	42	8.90
	AT	M1	✓	✓	×	32	9.32	24	9.35	19	9.73
		M2				38	8.94	23	9.47	28	8.98
		M3				57	7.24	43	8.54	39	8.79
		M4				48	7.98	34	9.01	27	9.22
	AMAT	M1	✓	✓	×	79	8.01	60	8.38	36	9.41
		M2				82	7.94	54	6.24	42	8.32
		M3				89	7.23	76	8.87	68	7.93
		M4				80	8.12	72	8.45	53	8.85
	Contrastive	M1	✓	✓	×	19	9.66	14	9.80	3	9.95
		M2				23	9.21	11	9.91	7	8.94
		M3				52	8.73	36	8.48	21	8.78
		M4				38	8.21	25	9.17	14	9.36
		M1	✓	×	✓	17	9.70	12	9.89	5	9.78
		M2				20	8.90	9	9.95	13	8.92
		M3				37	8.18	22	8.97	19	8.83
		M4				26	9.49	17	9.26	9	9.73
	ARL	M1	×	✓	✓	35	8.24	23	8.98	20	9.04
		M2				42	7.83	19	9.18	25	8.85
		M3				71	7.96	48	7.99	39	8.69
		M4				59	6.01	32	8.78	23	8.92
***RELIVE* (Ours)**	Contrastive (Ours)	M1	✓	✓	✓	16	9.71	2	9.92	0	9.91
		M2				18	9.28	1	9.95	0	9.97
		M3				28	8.26	16	9.79	9	9.52
		M4				20	9.06	8	9.83	5	9.79
	Multitask (Ours)	M1	✓	×	×	65	8.34	43	8.94	62	8.64
		M2				72	7.84	35	9.12	67	8.19
		M3				81	6.98	58	8.40	52	8.96
		M4				69	8.30	40	9.02	30	9.20
	Contrastive Multitask Fusion (Ours)	M1	✓	✓	✓	12	9.73	**0**	**10.00**	**0**	**10.00**
		M2				10	9.92	**0**	**10.00**	**0**	**10.00**
		M3				19	9.98	1	10.00	2	10.00
		M4				**8**	**10.00**	**0**	**10.00**	**0**	**10.00**

Note: ASR and δ_∞ represent attack success rate (%) and average distortion, respectively. S^c, S^{adv}, and S^{aug} are clean, adversarial, and augmented samples, respectively. Bold values indicate the best defence result. M1:SSFormer-S, M2:SSFormer-L, M3:PraNet, M4:CaraNet.

Table 5.2: Comparative model performance of ***RELIVE***.

	Methods	Models	Samples			$dsc \uparrow$
			S^c	S^{adv}	S^{aug}	
	No Defence	M1	✓	×	×	0.5679* (0.9408)
		M2				0.5229* (0.9579)
		M3				0.4572* (0.9105)
		M4				0.5091* (0.9356)
Existing Defence	Augmentation	M1	×	×	✓	**0.9450**
		M2				0.9504
		M3				**0.9110**
		M4				**0.9498**
	AT	M1	✓	✓	×	0.9225
		M2				0.9371
		M3				0.8934
		M4				0.9226
	AMAT	M1	✓	✓	×	0.9389
		M2				0.9500
		M3				**0.9128**
		M4				**0.9380**
	Contrastive	M1	✓	✓	×	0.9220
		M2				0.9471
		M3				0.9025
		M4				0.9336
		M1	✓	×	✓	**0.9429**
		M2				**0.9598**
		M3				0.8949
		M4				0.9127
	ARL	M1	×	✓	✓	0.8942
		M2				0.9064
		M3				0.8524
		M4				0.8994
***RELIVE* (Ours)**	Contrastive (Ours)	M1	✓	✓	✓	**0.9514**
		M2				0.9562
		M3				**0.9210**
		M4				**0.9380**
	Multitask (Ours)	M1	✓	×	×	**0.9525**
		M2				**0.9589**
		M3				0.9002
		M4				0.9329
	Contrastive Multitask Fusion (Ours)	M1	✓	✓	✓	**0.9510**
		M2				**0.9590**
		M3				**0.9197**
		M4				**0.9374**

Note: dsc represents dice between prediction and GT. Bold values indicate the improved dsc of the model. Under No Defence, the value in bracket () depicts actual model performance while * indicates model performance after attack. M1:SSFormer-S, M2:SSFormer-L, M3:PraNet, M4:CaraNet.

even though it performs exceptionally well on augmented data.

AT [36, 81] mixes minimal S^{adv} with S^c to strengthen model robustness. Table 5.1 demonstrates that, as compared to augmentation defence, AT substantially decreases ASR with increased δ_∞ for all the models across U and T_b attack settings. However, it struggles to preserve model performance by providing lower dsc than actual model performance across all the models, as depicted in Table 5.2. This occurs due to the fact that AT overfits in the direction of S^{adv} and fails to provide generalised output on S^c, resulting in restricted applicability. A variant of AT, called AMAT [82], involves an equal number of S^c and S^{adv} samples simultaneously apply to the model, wherein training averages the individual losses. It can be attributed from Table 5.1 that even though AMAT decreases ASR, it does not surpass AT and augmentation in robustness improvement across all attacks and models. Meanwhile, it offers better dsc than AT across CaraNet and PraNet models, as indicated in Table 5.2.

To advances the adversarial robustness of DL-based MIS models, previous studies on contrastive learning-based defences [38] are only limited to non-medical applications, training on S^c with the combination in any of S^{adv} and S^{aug}. The results of these combinations are depicted in Tables 5.1 and 5.2, revealing that existing contrastive defence substantially diminishes the ASR with large δ_∞. In particular, with S^c-S^{adv}, ASR decreases to below 25% for all the models across T_b attacks and both SSFormer versions across U and T_w attacks. Likewise, with S^c-S^{aug}, ASR drops to 25% across each model in T_w and T_b attack settings, both SSFormer versions across U attacks. It is noteworthy that only the combination with S^{aug} manages to retain model performance with the small enhancement in dsc for SSFormer-L and SSFormer-S, as shown in Table 5.2. Table 5.1 and 5.2 signifies that the ARL [84] calculates contrastive loss between S^{aug} and S^{adv}, significantly reduces ASR and dsc across all attack settings and models. In essence, it enhances the model's robustness

while degrading its performance because ARL neglects S^c during model training; thus, the model avoids learning clean features and reflects a decreased model performance.

As opposed to existing defences, our proposed defence, ***RELIVE***, incorporates contrastive learning, multitask learning, and integration of contrastive multitask approaches to effectively diminish the efficacy of adversarial attacks on several MIS models and marginally boosts their performance. Table 5.1 and Table 5.2 demonstrate the following observations of ***RELIVE***:

- The proposed contrastive learning-based defence, in which the model training is performed using S^c, S^{adv} and S^{aug} samples, outperforms all the existing defences with minimum ASR upto 0% and maximum δ_∞ in all the attack settings across each model (refer Table 5.1). Ironically, it achieves exceptional performance across the SSFormer model in T_b attack settings. Moreover, it effectively advances the model performance by providing a higher dsc value than all the model performances under no attack, except with SSFormer-L (refer Table 5.2). It happens because the proposed contrasive loss enables the model to effectively learn and advance the consistency between clean, adversarial, and augmented features. Further, the data fidelity loss assures that the model is providing accurate predictions across each sample.

- Our proposed multitask learning-based defence represents the mild decline in ASR with advanced δ_∞ for all the attack settings and models (refer Table 5.1). Moreover, it enhances the model performance for the SSFormer model (refer Table 5.2). Since task selection in multitask learning is important as it can weaken the emphasis on resilient feature learning, we select the auxiliary tasks depending upon the lower correlation with the main task in the proposed multitask learning-based defence. This will empower the model's robustness while preserving its performance.

- Our proposed fusion-based defence, leveraging the benefits of multitask and contrastive learning, surpasses all the existing defences and the proposed individual learning (contrastive or multitask) based defences (refer Table 5.1). The proposed fusion-based defence, which trains the model by incorporating all the samples (S^c, S^{adv} and S^{aug}), successfully drops the ASR up to 0% with a high δ_∞ as 10 for all the models, except PraNet for T_w and T_b attacks. We also observe a reduction in ASR with maximum δ_∞ under U attack. Further, our proposed fusion-based defence maintains all the model performances by obtaining high dsc values (refer Table 5.2). Since the model emphasises capturing both augmented and adversarial features and combining the advantages of contrastive and multitask learning, its fusion elevates the overall adversarial robustness.

5.2.5 Ablation Study

This section presents different ablation studies of ***RELIVE***. To better comprehend the efficacy of auxiliary task selection, we compare the adversarial performance of our proposed multitask-based defence by setting multiple combinations of different auxiliary tasks, having the main task as segmentation. Thus, Table 5.3 demonstrates that when the main task (S) is paired with any of the auxiliary tasks, such as SD or SC, it fails to substantially drop ASR across T_b attack settings for all the models except SSFormer-L under SC. Likewise, these combinations struggle to reduce the ASR across T_w attack settings across SSFormer-S and PraNet models under SD and across the SSFormer-S model under SC, as well as across U attack settings across SSFormer-L in SC. This occurs because the auxiliary tasks, object detection (D) and classification (C), signify moderate similarity with the main task, segmentation (S). It is evidenced by the ρ values in Table 5.3, which range from 0.3 to 0.6 for all the models.

Moreover, the combination of the main task (S) and boundary detection task (as SB) results in the minimum ASR across T_w attack settings for all models except PraNet, whereas for SSFormer-S and L models, SB achieves the maximum ASR up to 100% across U attack settings with reduced δ_∞. In this combination, the ASR under T_b attack settings remain unaffected but lower δ_∞ across all models, indicating a strong correlation between S and B tasks, achieving ρ values up to 0.9 for all models. Furthermore, pairing all the tasks with the main task (denoted as SDCB) provides a similar performance as SB. We found a small decline in ASR under T_w attack settings across all the models in SDCB. This combination depicts the strong positive correlation between the main task (S) and all other auxiliary tasks (DCB), having ρ values ranging from 0.5 to 0.7 for all models. Since the auxiliary task B and its combination with other tasks (denoted as DCB) exhibit a significant correlation with the main task (S), it degrades the adversarial performance in the context of robustness. Thus, our proposed multitask model does not consider the boundary detection task (B) as an auxiliary task. In essence, it combines the main task (S) with the two auxiliary tasks, object detection (D) and classification (C). Table 5.3 indicates that SDC combination substantially drops ASR up to 30% under T_b, 35% under T_w and 65% under U attack settings for CaraNet, SSFormer-L and SSFormer-S models, respectively. It happens due to the lower correlation between S, D and C, evidenced by ρ values of 0.187 in the SSFormer-S model and ranges from 0.2 to 0.3 across other models.

To rigorously analyse the impact of contrastive loss in the proposed contrastive learning-based and fusion-based defences, we compare their adversarial performances by considering multiple contrastive losses. This involves contrastive loss, which is similar to and different from data fidelity loss, defined as dice bce (kindly refer Equations (5.6), (5.7) and (5.8)) and L_2 (kindly refer Equation (5.3)) loss, respectively. Table 5.4 indicates that L_2 loss surpasses dice bce loss across all the cases in the proposed contrastive learning-based defence and in

Table 5.3: Adversarial performance of proposed multitask model considering MIS as the main task with several auxiliary tasks. S→Segmentation, D→Object Detection, C→Classification, B→Boundary Detection.

Tasks	S	D	C	B	Models	U		T_w		T_b		$\rho \downarrow$
						$ASR \downarrow$	$\delta_\infty \uparrow$	$ASR \downarrow$	$\delta_\infty \uparrow$	$ASR \downarrow$	$\delta_\infty \uparrow$	
S	✓	×	×	×	M1	90	6.10	100	4.18	100	4.59	n/a
					M2	87	6.50	100	4.02	100	4.70	
					M3	100	5.23	100	4.98	100	4.21	
					M4	95	5.98	97	5.34	100	4.95	
SD	✓	✓	×	×	M1	86	7.54	100	4.35	100	6.38	0.492
					M2	89	6.98	97	4.45	100	6.97	0.332
					M3	95	7.87	100	5.15	100	5.99	0.407
					M4	90	6.59	95	4.89	100	6.63	0.419
SC	✓	×	✓	×	M1	82	7.25	100	4.23	100	5.64	0.595
					M2	85	7.03	94	5.89	98	6.82	0.504
					M3	90	6.89	98	5.31	100	5.95	0.515
					M4	88	6.95	95	5.68	100	5.21	0.484
SB	✓	×	×	✓	M1	100	2.24	95	3.83	100	1.30	0.887
					M2	100	3.15	98	2.18	100	1.78	0.879
					M3	96	4.59	100	4.23	100	2.87	0.989
					M4	94	5.01	98	3.35	100	2.56	0.915
SDCB	✓	✓	✓	✓	M1	100	4.44	92	7.39	100	6.95	0.746
					M2	100	4.89	86	7.56	97	7.94	0.597
					M3	100	4.21	89	7.89	100	6.54	0.689
					M4	100	4.76	94	7.25	100	6.21	0.656
SDC (Ours)	✓	✓	✓	×	M1	**65**	**8.34**	43	8.94	62	8.64	**0.187**
					M2	72	7.84	**35**	**9.12**	67	8.19	0.379
					M3	81	6.98	58	8.40	52	8.96	0.389
					M4	69	8.30	40	9.02	**30**	**9.20**	0.215

Note: ASR, δ_∞, and ρ depict attack success rate (in %), average distortion, and correlation coefficient between respective tasks' prediction and MIS prediction, respectively. Bold values indicate the best result. n/a signifies that correlation is not applicable. M1:SSFormer-S, M2:SSFormer-L, M3:PraNet, M4:CaraNet.

most of the cases in the proposed fusion-based defence. In essence, it effectively decreases ASR up to 0% within high δ_∞. The outcomes reveal that contrastive loss and data fidelity loss should be different from each other, as they concentrate on distinct purposes in training. While contrastive loss captures the relative similarity between samples, the data fidelity loss

Table 5.4: Adversarial performance of the proposed defences across different contrastive losses for MIS.

Defence	Models	Contrastive Loss	U		T_w		T_b	
			$ASR\downarrow$	$\delta_\infty\uparrow$	$ASR\downarrow$	$\delta_\infty\uparrow$	$ASR\downarrow$	$\delta_\infty\uparrow$
Contrastive	M1	Dice BCE	30	8.54	20	8.97	18	8.23
		L_2	**16**	**9.71**	**2**	**9.92**	**0**	**9.91**
	M2	Dice BCE	38	7.49	17	8.41	12	7.21
		L_2	**18**	**9.28**	**1**	**9.95**	**0**	**9.97**
	M3	Dice BCE	56	6.98	48	7.45	17	9.10
		L_2	**28**	**8.26**	**16**	**9.79**	**9**	**9.52**
	M4	Dice BCE	48	6.95	19	8.76	15	8.94
		L_2	**20**	**9.06**	**8**	**9.83**	**5**	**9.79**
Contrastive Multitask Fusion	M1	Dice BCE	10	9.86	2	9.88	0	10.00
		L_2	12	9.73	**0**	**10.00**	**0**	**10.00**
	M2	Dice BCE	16	9.01	6	9.73	5	9.86
		L_2	**10**	**9.92**	**0**	**10.00**	**0**	**10.00**
	M3	Dice BCE	12	10.00	4	10.00	0	10.00
		L_2	19	9.98	**1**	**10.00**	2	10.00
	M4	Dice BCE	16	9.43	2	9.99	2	9.87
		L_2	**8**	**10.00**	**0**	**10.00**	**0**	**10.00**

Note: ASR and δ_∞ depict attack success rate (in %) and average distortion, respectively. Bold digits indicate the instances where L_2 outperforms Dice BCE loss. M1:SSFormer-S, M2:SSFormer-L, M3:PraNet, M4:CaraNet.

focuses on model performance in correspondence with GT.

We further analyse the importance of diverse weight assignments to the contrastive loss relative to the data fidelity loss. In particular, we evaluate the adversarial performance of several DL-based MIS models by giving equal weights to both the losses and providing maximum weight to contrastive loss. It can be observed from Table 5.5 that adversarial results on maximum weighted contrastive loss surpass equal weights on both losses across all the attack settings for each model in both the proposed defences. In essence, it significantly lowers the ASR up to 0% in targeted attacks and an average decline in ASR for U attacks for all the models. The primary objective of our proposed defence, ***RELIVE***, is to elevate

Table 5.5: Adversarial performance of the proposed defence across different weights of contrastive loss for MIS.

Defence	Models	Weights of Contrastive Loss	U		T_w		T_b	
			$ASR\downarrow$	$\delta_\infty\uparrow$	$ASR\downarrow$	$\delta_\infty\uparrow$	$ASR\downarrow$	$\delta_\infty\uparrow$
Contrastive	M1	Equal	67	6.52	35	5.23	20	8.43
		Maximum	**16**	**9.71**	**2**	**9.92**	**0**	**9.91**
	M2	Equal	69	6.01	38	4.98	26	6.75
		Maximum	**18**	**9.28**	**1**	**9.95**	**0**	**9.97**
	M3	Equal	72	5.87	21	6.67	34	5.83
		Maximum	**28**	**8.26**	**16**	**8.79**	**9**	**9.52**
	M4	Equal	70	5.95	17	7.90	29	6.21
		Maximum	**20**	**9.06**	**8**	**9.83**	**5**	**9.79**
Contrastive Multitask Fusion	M1	Equal	24	8.49	14	8.21	12	10.00
		Maximum	**12**	**9.73**	**0**	**10.00**	**0**	**10.00**
	M2	Equal	36	7.94	17	8.38	9	9.21
		Maximum	**10**	**9.92**	**0**	**10.00**	**0**	**10.00**
	M3	Equal	58	6.03	20	7.20	15	8.29
		Maximum	**19**	**9.98**	**1**	**10.00**	**2**	**10.00**
	M4	Equal	43	6.41	12	8.49	5	9.94
		Maximum	**8**	**10.00**	**0**	**10.00**	**0**	**10.00**

Note: ASR and δ_∞ depict attack success rate (in %) and average distortion, respectively. Bold values indicate the best result. M1:SSFormer-S, M2:SSFormer-L, M3:PraNet, M4:CaraNet.

the adversarial robustness of the DL-based MIS model, wherein contrastive loss is crucial to capture robust adversarial features during training. Therefore, we prioritise allocating the maximum weight to contrastive loss in our proposed defence.

To better understand the impact of S^c, S^{adv} and S^{aug} samples on our proposed fusion-based defence, we compare the adversarial robustness and model performance on different combinations of these samples. To the end, we incorporate three different combinations as $S^c - S^{adv}$, $S^c - S^{aug}$, and $S^c - S^{adv} - S^{aug}$. Table 5.6 reveals that adversarial results on $S^c - S^{adv}$ surpass $S^c - S^{aug}$ against U attack settings by significantly diminishing ASR up to 0% in maximum δ_∞. Moreover, model performance on $S^c - S^{aug}$ depicts little improvement

Table 5.6: Contrastive Multitask Fusion across different samples.

	Models	Training Samples			U		T_w		T_b		$dsc \uparrow$
		S^c	S^{adv}	S^{aug}	$ASR \downarrow$	$\delta_\infty \uparrow$	$ASR \downarrow$	$\delta_\infty \uparrow$	$ASR \downarrow$	$\delta_\infty \uparrow$	
No Defence	M1	✓	×	×	90	6.10	100	4.18	100	4.59	0.9408
	M2				87	6.50	100	4.02	100	4.70	0.9579
	M3				100	5.23	100	4.98	100	4.21	0.9105
	M4				95	5.98	97	5.34	100	4.95	0.9356
Contrastive Multitask Fusion	M1	✓	✓	×	16	9.45	0	9.97	**0**	**10.00**	0.9255
	M2				19	8.99	**0**	**10.00**	**0**	**10.00**	0.9332
	M3				25	7.96	2	9.93	**0**	**10.00**	0.8932
	M4				18	9.15	**0**	**10.00**	**0**	**10.00**	0.9302
	M1	✓	×	✓	38	9.62	0	9.99	2	9.95	0.9490†
	M2				23	8.95	0	9.95	0	9.90	0.9592†
	M3				37	7.98	9	10.00	7	10.00	0.9185†
	M4				29	9.76	2	10.00	4	10.00	0.9362†
	M1	✓	✓	✓	12	9.73	**0**	**10.00**	**0**	**10.00**	0.9510†
	M2				10	9.92	**0**	**10.00**	**0**	**10.00**	0.9590†
	M3				19	9.98	1	10.00	2	10.00	0.9197†
	M4				**8**	**10.00**	**0**	**10.00**	**0**	**10.00**	0.9374†

Note: ASR, δ_∞ and dsc depict attack success rate (in %), average distortion and dice between prediction and ground truth, respectively. S^c, S^{adv}, S^{aug} are the clean, adversarial, augmented samples, respectively. Bold values indicate the best defence result. † depicts improved dsc. M1:SSFormer-S, M2:SSFormer-L, M3:PraNet, M4:CaraNet.

compared to $S^c - S^{adv}$ across all the models. Conversely, the $S^c - S^{adv} - S^{aug}$ reduces ASR up to 0% with maximum δ_∞ against targeted attack settings for all the models except PraNet. This combination also maintains the model performance by achieving significant dsc values across each model. Hence, good adversarial performances are obtained from the $S^c - S^{adv}$ and $S^c - S^{adv} - S^{aug}$ combinations of samples against targeted attacks. However, $S^c - S^{adv}$ fails to retain model performance due to its too much dependency on adversarial features. Thus, we incorporate the combination of $S^c - S^{adv} - S^{aug}$ training samples in our proposed fusion-based defence.

5.3 Summary

This chapter has proposed a novel adversarial defence, ***RELIVE***, to effectively mitigate the effect of adversarial attacks on the DL-based MIS model while enhancing model performance. ***RELIVE*** exhibits contrastive learning, multitask learning and their fusion-based defence. Our proposed contrastive learning-based defence employs contrastive loss, which enables the model to capture the similarity between the clean, adversarial, and augmented features and data fidelity loss to ensure accurate model learning. Moreover, the proposed multitask learning-based defence encourages the model to capture a variety of task-specific features to elevate generalised feature representation. It selects the auxiliary tasks based on their correlation with the main task, demonstrating that less correlated tasks substantially advance the adversarial robustness of the model. Furthermore, our proposed fusion-based defence leverages the advantages of contrastive and multitask learning. In essence, contrastive learning is applied to the main task within the proposed multitask model. This approach is significantly effective in diminishing the efficacy of adversarial perturbation on the DL-based MIS model while preserving model performance. Experimental outcomes demonstrate that our proposed defence, ***RELIVE***, comprising contrastive, multitask, and their fusion-based defence, surpasses existing adversarial defences by reducing ASR up to 0% within large δ_∞. Thus, it has successfully elevated the adversarial robustness of the DL-based MIS models and achieved notable performance gain.

Chapter 6

Improving Trustworthiness and Performance of MIS models

DL-based MIS models have proliferated remarkable success in healthcare. However, their non-transparent nature often leads to erroneous and unreliable outcomes. These outcomes could have fatal consequences in life-critical medical applications, ultimately undermining the trustworthiness of MIS predictions [50]. Such predictions appear from inaccurate ID and OOD samples. The OOD samples diminish the model performance as their training and testing distributions exhibit dissimilarities. Conversely, the ID samples are associated with similar training and testing distributions; thus, they are anticipated to provide accurate results. Unfortunately, some ID samples still perform erroneously, compromising MIS performance. This results in a lack of reliability caused by aleatoric and epistemic uncertainties, which are influenced by insufficient model training and the inclusion of redundant noise in data [56]. Such uncertainties lead to distrusted outcomes, highlighting the need for a comprehensive exploration of the trustworthiness problem in ID samples. Moreover, GT is required to assess the trustworthiness and performance of the MIS model, which is unavailable during inference, making the problem inherently complex and challenging. Furthermore, prior research primarily focuses on quantifying pixel-wise trustworthiness, while

estimating the trustworthiness of image ROIs or structural components remains largely unexplored.

This chapter proposes a novel method, ***TrustMedIS***, which is ***Trust***worthy ***Med***ical ***I***mage ***S***egmentation, aiming to investigate the trustworthiness of MIS prediction and enhance the model performance for ID samples. It leverages the characteristics of input and output, followed by analysing the consistency between the MIS predictions of input and their variants. ***TrustMedIS*** comprises three key components: *ET* (*E*xamining *T*rustworthiness), *ENT* (*E*levating *N*on-*T*rustworthy predictions), and *CSM* (*C*lassifier *S*election *M*ethod). *ET* examines the trustworthiness of MIS prediction through consistency computation. *ENT* advances the performance of non-trustworthy predictions. *CSM* considers multiple MIS models and selects the optimal one, offering the most trustworthy prediction. The chapter is structured as follows: Section 6.1 provides a detailed explanation of our proposed method, ***TrustMedIS***. Section 6.2 presents the experimental results of the proposed method. Section 6.3 delivers the discussion related to the proposed method. Section 6.4 summarises the overall chapter.

6.1 Proposed Method: *TrustMedIS*

This section provides the description about the proposed method, ***TrustMedIS***. It examines the MIS predictions' structure-wise trustworthiness for ID samples without relying on GT or network retraining and enhances the MIS model performance, ultimately strengthening decision-making. ***TrustMedIS*** is composed of three novel methods: *ET* (refer Figure 6.1), *ENT* (refer Figure 6.2), and *CSM* (refer Figure 6.3).

The *ET* method leverages the observation that input samples' MIS prediction and their corresponding variants (rotated version of input) share a strong correlation. Building on this observation, the *ET* method initially determines the trustworthiness of the MIS model by

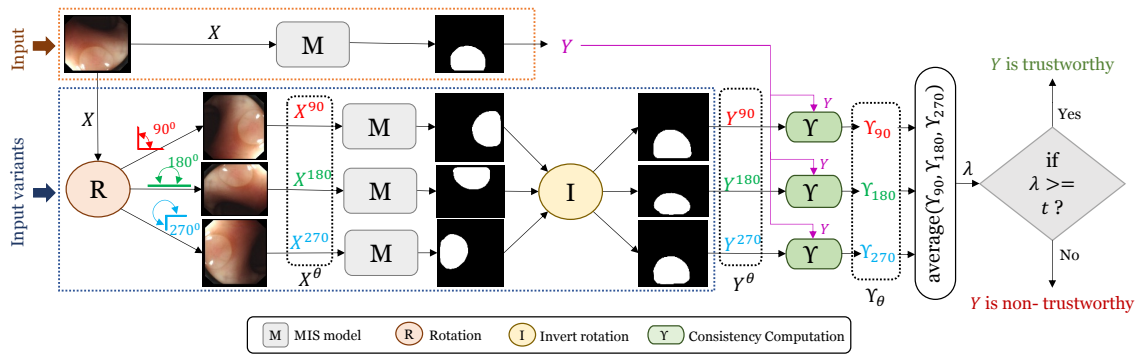

Figure 6.1: The complete work-flow of *ET* method. The input (X) and its variants (X^{θ}, where, θ is 90^0, 180^0, 270^0 rotation) are fed to the segmentation model to get predictions (Y and Y^{θ}). Using these predictions the consistencies Υ_{θ} are measured to evaluate confidence measure (λ), which is further compared by a pre-specified threshold (t) to decide whether Y is trustworthy or non-trustworthy.

getting predictions for input and their corresponding variants, analysing the consistent relation between these predictions, and finally calculating the confidence measure to classify predictions as trustworthy or non-trustworthy. Subsequently, the *ENT* method capitalises on the insights that rotational transformation on the input image can mitigate the erroneous impact in MIS prediction. Following this, the *ENT* method utilises the *ET* method to advance the performance of non-trustworthy MIS predictions. Eventually, the *CSM* method boosts the efficacy of model by observing the trustworthiness of MIS predictions from multiple existing MIS models using the *ENT* method and significantly selecting the optimal model, which offers the most trustworthy prediction. The detailed descriptions of *ET*, *ENT*, and *CSM* are demonstrated in Sections 6.1.1, 6.1.2, and 6.1.3, respectively and their steps are outlined in Algorithms 6.1, 6.2 and 6.3, respectively.

6.1.1 Examining MIS Model's Trustworthiness by *ET* method

The non-transparent behaviour of DL-based MIS models hinders the inherent understanding of the model architecture, which results in distrusted predictions from these models. This leads to catastrophic consequences in the medical domain. Thus, it is crucial to

Algorithm 6.1 The proposed *ET* method.

Require: Input X; MIS model $f(\cdot)$; Threshold t.
Ensure: Decision of trustworthiness (dot).
Create variants X^{θ} by rotating X at different angles θ.
$Y = f(X)$ ▷ prediction of X, refer Eq. (6.1).
$Y^{\theta} = r^{-\theta}(f(X^{\theta}))$ ▷ invert rotated prediction of X^{θ}, for several values of θ, refer Eq. (6.1).
Compute Υ_{θ} for all Y^{θ} using Eq. (6.2).
$\lambda = \underset{\theta}{\text{average}}(\Upsilon_{\theta})$ ▷ confidence measure using Eq. (6.3).
if $\lambda \geq t$ ▷ condition to investigate T or NT.
then $dot \rightarrow$ trustworthy ▷ dot signifies that Y is trustworthy.
else $dot \rightarrow$ non-trustworthy ▷ dot signifies that Y is non-trustworthy.
end if
return dot

Algorithm 6.2 The proposed *ENT* method.

Require: Input X and MIS model $f(\cdot)$.
Ensure: Prediction S.
$dot = ET(X, f)$ ▷ decision of trustworthiness using *ET* (refer Algo 6.1).
Y = $f(X)$ ▷ prediction of X (Algo 6.1).
if $dot \rightarrow$ trustworthy
then $S = Y$
else
Create variants,Y^{90}, Y^{180}, Y^{270} of prediction Y.
Apply *ET* (Algo 6.1) to Y , Y^{90},Y^{180},and Y^{270} and obtain the confidence measures z_1, z_2, z_3, and z_4, respectively.
Select the best two predictions Y_A and Y_B, using maximum two confidence measures.
if $A(Y_A) > A(Y_B)$ ▷ ref Eq. (6.5).
then $F = Y_A$
else $F = Y_B$
end if
if $A(F) > \beta \times T_p(F)$ ▷ ref Eq.(6.6).
then $S = F$
else $S = Y_A \cap Y_B$
end if
end if
return S ▷ improved performance of non-trustworthy prediction.

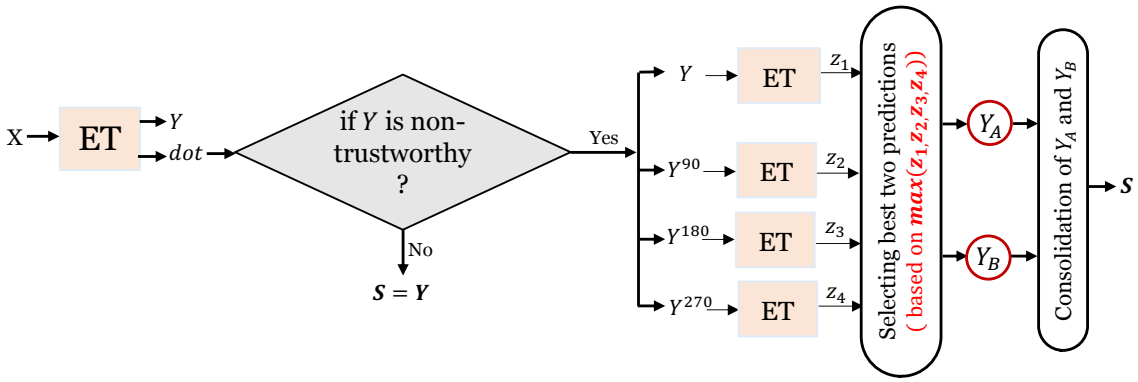

Figure 6.2: Work-flow of *ENT* method. Initially, the MIS prediction (Y) and decision of trustworthiness (dot) are obtained from *ET* method. If Y is trustworthy, S corresponds to Y; otherwise, Y and their respective variants (Y^{θ}) are applied to *ET* method for evaluating their respective confidence measures z_1, z_2, z_3,z_4. The maximum two values offer the optimal two predictions Y_A and Y_B, which are consolidated to get the improved prediction S.

Algorithm 6.3 The proposed *CSM* method.

Require: Input X and Multiple MIS models $M_1, \cdots M_n$.
Ensure: Prediction S.
  for $v \in 1$ *to* n **do** ▷ loop for each model M
    Apply *ENT* (Algo. 6.2) on M_v and X to get S, Y_A and Y_B.
    $\Upsilon_v = \Upsilon(Y_A, Y_B)$ ▷ compute consistency using Eq. (6.2)
  end for
  $j = \underset{k\in(1,\cdots,n)}{argmax}(\Upsilon_k)$ ▷ index of selected model ref Eq. (6.7)
  $S = ENT(X, M_j)$ ▷ applying *ENT* (Algo. 6.2) on model, M_j.
  return S

measure the trustworthiness of each individual prediction. To this end, a novel method, *ET*, is proposed, which utilises a confidence measure to examine the trustworthiness of DL-based MIS models. This can be visualised in Figure 6.1. For trustworthy predictions, when a rotated input is fed into the MIS model, its invert-rotated prediction shows high similarity with the actual (non-rotated) prediction. This occurs as the DL-based MIS models are designed to produce rotation-invariant predictions. However, this phenomenon is violated for non-trustworthy predictions in which MIS predictions of actual (non-rotated) input are erroneous, but the impact of errors is reduced when rotated input is applied to the model (refer Figure 6.4). Leveraging these observations, our proposed method, *ET*, rotates the input sam-

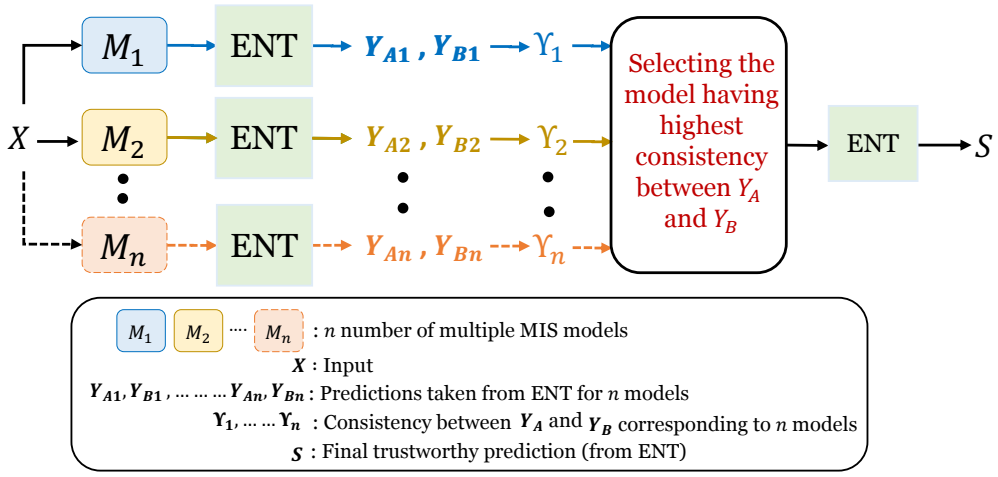

Figure 6.3: The work-flow of *CSM* method. Initially, the input (X) is fed to an MIS model, following *ENT* method to achieve the optimal predictions (Y_A, Y_B). Subsequently, the consistency (Υ) is calculated between them. These operations are performed across multiple MIS models ($M_1, M_2, \cdots M_n$). The model with maximum consistency is chosen, across which *ENT* is employed to achieve the most trustworthy prediction S.

ples at several angles and produces diverse variants of input, called input variants. While alternative augmentation methods such as colour modifications or noise addition could be used to create input variants, they would alter the texture and shift the distribution of input. This leads to the generation of OOD samples, providing inferior MIS predictions. Thus, *ET* method exclusively employs rotation operations for creating input variants. Such variants are applied to MIS models, and their predictions are further invert-rotated at the same rotation angle for comparison with actual (non-rotated) MIS predictions. Mathematically,

$$Y = f(X) \text{ and } Y^{\theta} = r^{-\theta}(f(X^{\theta})) \tag{6.1}$$

where $f(\cdot)$ represents the MIS model. X and Y indicate the input and their corresponding MIS prediction. X^{θ} depicts the input variants generated by rotating X with a rotation angle of θ and Y^{θ} is their corresponding invert-rotated variant predictions. $r^{-\theta}$ depicts rotation at $-\theta$. Further, the input and its respective variants exhibit high similarity; thus, their MIS predictions are anticipated to be strongly consistent with each other. This consistency, Υ_{θ},

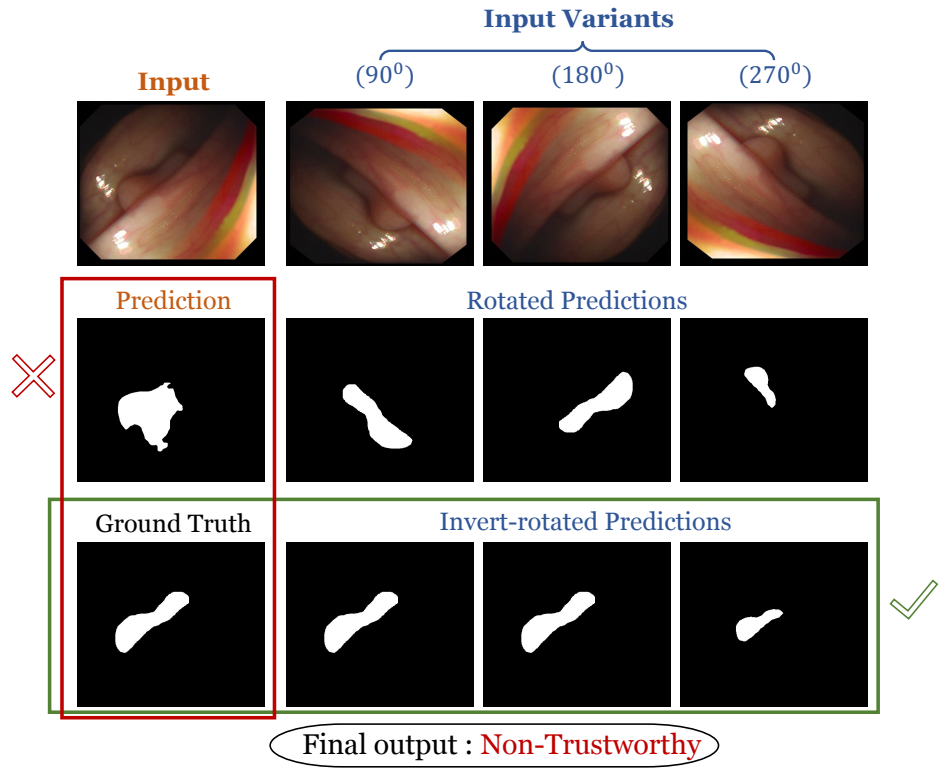

Figure 6.4: A case of non-trustworthy MIS prediction where the original prediction (enclosed in red rectangle) differs from the ground truth, while its invert-rotated variants (enclosed in green rectangle) align perfectly with the ground truth.

between the input samples' MIS prediction, Y, and its variants, Y^{θ}, is computed by dice metrics [108], formulated as,

$$\Upsilon_{\theta} = \frac{2 * \left| Y \cap Y^{\theta} \right|}{|Y| + |Y^{\theta}|} \tag{6.2}$$

It is noteworthy that a rotated input variant may lead to non-trustworthy predictions, highlighting low consistency with the actual (non-rotated) input. To address this, the input variants are created by considering diverse angles (Υ_{θ} is computed for various rotation angles θ) and averaging these consistency values. The averaged output is considered as a confidence measure λ, expressed as,

$$\lambda = \underset{\theta}{\text{average}}(\Upsilon_{\theta}) \tag{6.3}$$

Our proposed method, *ET*, employs three rotation angles, which are 90^0, 180^0 and 270^0 (refer Section 6.2.3). The λ is used to classify the prediction as trustworthy or non-trustworthy. Mathematically, the decision of trustworthiness, dot, is given by,

$$dot = \begin{cases} \text{trustworthy}, & \text{if } \lambda \geq t \\ \text{non-trustworthy}, & \text{otherwise} \end{cases} \tag{6.4}$$

ET considers the MIS prediction as trustworthy when λ is higher than or same as the pre-specified threshold value t; otherwise, the MIS prediction is classified as non-trustworthy. The value of t is selected by hyperparameter tuning, as provided in Section 6.2.3.

6.1.2 Elevating MIS Model's Performance by *ENT* method

The MIS prediction of actual (non-rotated) input is erroneous for non-trustworthy segmentation. Ironically, when the same input is rotated and processed by the same MIS model, it can yield a correct prediction [113] (refer Figure 6.4). Building on this insight, our proposed method, *ENT* (refer Figure 6.2), aims to enhance the MIS performance of non-trustworthy predictions. To achieve this, it first utilises the *ET* method (Section 6.1.1) to identify the trustworthiness of MIS prediction. If it is found to be trustworthy, the prediction is accepted as the final output; otherwise, the final MIS prediction is achieved by integrating the two most trustworthy predictions. This integration is performed by applying the MIS prediction Y and their corresponding invert-rotated predictions (variants) Y^{90}, Y^{180} and Y^{270} to *ET* method, in which a confidence measure is computed for each prediction. The best two predictions, Y_A and Y_B, with the highest confidence measures, are chosen. Among Y_A and Y_B, one prediction F, which exhibits a large number of foreground

pixels, will be selected (refer Section 6.2.3). Mathematically,

$$F = \begin{cases} Y_A, & \text{if } A\left(Y_A\right) > A\left(Y_B\right) \\ Y_B, & \text{otherwise} \end{cases} \tag{6.5}$$

where the operation $A\left(\cdot\right)$ computes the number of foreground pixels. However, there is a possibility that the chosen prediction may exhibit the fewest foreground pixels, resulting in erroneous segmentation. To mitigate this, we ensure that the chosen prediction, F, can have enough foreground pixels, that is, greater than or equal to β times the total number of pixels. If the condition is satisfied, then F is considered the final trustworthy prediction S; otherwise, S is determined by employing the intersection rule between Y_A and Y_B. Mathematically, the final prediction, S, is represented as,

$$S = \begin{cases} F, & \text{if } A\left(F\right) > \beta \times T_p\left(F\right) \\ Y_A \cap Y_B, & \text{otherwise} \end{cases} \tag{6.6}$$

where the operation $T_p\left(\cdot\right)$ gives the total count of pixels. The considered rule of intersection and β are determined by extensive experiments, and their descriptions are given in Section 6.2.3.

6.1.3 Selecting the Most Effective MIS Model by *CSM* method

A single MIS model could result in erroneous and non-trustworthy outcomes. Conversely, when several MIS models are incorporated in *ENT* method (Section 6.1.2), it enhances MIS performance and trustworthiness through the selection of the most optimal MIS model. This observation is leveraged by our proposed method, *CSM*, (kindly refer Figure 6.3), which considers various existing MIS models, observes their coherent characteristics, and chooses the model that offers the most effective trustworthy prediction. This approach

benefits from multiple models and strengthens the overall MIS performance compared to predictions achieved from a single MIS model.

Our proposed method, *CSM*, initially considers one MIS model and utilises *ENT* method (refer Section 6.1.2). The *ENT* method provides the predictions, Y_A and Y_B, and evaluates the consistency between them using Equation (6.2). This workflow is iteratively applied across each of the considered segmentation models. Among them, one model with the highest degree of consistency is selected, and its respective prediction from *ENT* is chosen as the final outcome of *CSM*.

Let *CSM* considers n number of existing MIS models, such that $(M_1, M_2, \cdots, M_n)$, with their respective consistency values between Y_A and Y_B is represented as $(\Upsilon_1, \Upsilon_2, \cdots, \Upsilon_n)$. Mathematically,

$$j = \underset{k \in (1,\cdots,n)}{argmax}(\Upsilon_k) \tag{6.7}$$

where j depicts the index of the selected MIS model. Thus, the final prediction, S, for input, X, can be expressed as,

$$S = S_j = ENT(X, M_j) \tag{6.8}$$

where M_j indicates the selected optimal model, providing the most effective and trustworthy predictions, S_j, from the *ENT* method.

6.2 Experimental Results

6.2.1 Datasets and Metrics

The experiments of ***TrsutMedIS*** are conducted on open-sourced MIS datasets, exhibiting colonoscopic images for polyp segmentation. We employed 798 images from CVC-300

[95], CVC-ClinicDB [96], CVC-ColonDB [97], ETIS-LaribPolypDB [110] and Kvasir [98] datasets. The performance of ***TrsutMedIS*** is evaluated by the following metrics:

- **Dice Score** (dsc): It measures the intersection between segmentation masks, assessing the similarity between them.
- **Pearson Correlation Coefficient** (ρ): It depicts how two variables are linearly related to each other. Its values range up to 1, wherein higher values indicate a strong correlation.

6.2.2 Considered MIS Models

TrustMedIS is evaluated on various existing MIS models. The PraNet [1] [3] model addresses area-boundary constraints and effectively segments these polyp regions. It detects the coarse areas using a PPD while refining the boundaries with RA mechanisms, providing accurate MIS prediction. The SSFormer [2] [14] model works on generalised MIS. It follows the architecture exhibiting a pyramidal encoder and progressive locality decoder for effectively capturing the genralised polyps in the image.

The UACANet [3] [65] model follows U-Net [13] architecture, incorporating additional encoders and decoders to learn uncertain ROIs at each stage. It performs feature aggregation to provide accurate MIS prediction. The CaraNet [4] [66] model incorporates feature pyramid modules and the reverse axial attention to furnish high-level extracted features and produce precise segmentation masks. It mainly concentrated on segmenting minute ROIs, which is crucial for several medical applications, including polyp and brain tumor segmentation. Notably, SSFormer and UACANet models are available in Standard (S) and Large

[1]Model details are available at `https://github.com/DengPingFan/PraNet`
[2]Model details are available at `https://github.com/Qiming-Huang/ssformer`
[3]Model details are available at `https://github.com/plemeri/UACANet`
[4]Model details are available at `https://github.com/AngeLouCN/CaraNet`

(L) versions, varying based on encoder scales. ***TrustMedIS*** utilises both versions for comprehensive evaluation.

6.2.3 Experimental Settings

We have conducted experiments on a server featuring an Nvidia V100 GPU, an Intel Xeon Gold 6132 CPU, and 192 GB of RAM. Following [3], we have trained the existing MIS model CaraNet [66] using 1450 polyp images acquired from CVC-ClinicDB [96] and Kvasir [98] datasets. Besides this, we employed open-sourced pre-trained weights of other considered MIS models, including PraNet [3], SSFormer [14], and UACANet [65].

To thoroughly investigate the significance of rotation angles in trustworthiness evaluation, we create the input variants at different rotation angles and utilise them in our proposed method, ***TrustMedIS***. We experience a reduction in MIS performance when any of the considered rotation angles (90^0, 180^0 and 270^0) is neglected, as most of the important information is lost. Moreover, if we incorporate additional angles along with considered rotation angles (90^0, 180^0 and 270^0), a negligible enhancement in MIS performance is observed. Therefore, ***TrustMedIS*** generates input variants using only 90^0, 180^0 and 270^0 rotation angles.

The effectiveness of ***TrustMedIS*** relies on cautious choice of hyperparameters. In this direction, the prespecified threshold (t) (refer Equation (6.4)) is determined by observing the *ENT* performance on the training dataset[1], employing a grid-search operation. Thus, t is tuned to 0.9 at which optimal MIS performance is achieved (kindly refer Table 6.5 for *ENT* performance on test dataset at different thresholds). Likewise, we set the weighing parameter, β, as 0.20 (refer Equation (6.6)).

Our proposed method, *ENT*, advances the MIS performance of non-trustworthy predictions by integrating the two most optimal MIS predictions (Y_A and Y_B) (refer to equation

(6.6)), incorporating the intersection rule for a few cases. To comprehensively evaluate the efficacy of the intersection rule in *ENT* method, we compare it against alternative operations such as union and maximum foreground pixels. The union operation involves all the pixels of Y_A and Y_B to get the final MIS prediction, while the maximum foreground pixel operation considers anyone from Y_A and Y_B, which exhibits a large count of foreground pixels as the MIS prediction. Consequently, the average performance of *ENT* method using intersection, union, and maximum foreground pixel operations is observed as 0.97, 0.95, and 0.92, respectively. As the intersection rule provides superior performance compared to other operations, we utilise it in our *ENT* method.

6.2.4 Comparative Evaluation

The comparative evaluation of our proposed method, ***TrustMedIS***, exhibiting *ET*, *ENT* and *CSM* methods, is presented in Tables 6.1, 6.2 and 6.4, respectively. The existing method [57] employs the VVC to determine structural uncertainty. In essence, this can be represented as the number of foreground pixels of MIS predictions of input and their respective variants. Any model with high uncertainty depicts low trustworthiness. To enable the fair comparison between the uncertainty measure (VVC) of the existing method [57] and the confidence measure (λ) of our proposed method, *ET*, we compute 1/VVC instead of VVC. In this direction, we compute the Pearson correlation coefficient between 1/VVC and dsc, depicted as ρ_e and between λ and dsc, indicated as ρ. The outcomes are given in Table 6.1, signifying that *ET* method provides a strong correlation (ρ). Notably, the trustworthiness of any prediction reflects the closeness of prediction with GT, even though GT itself is not included to measure this closeness. Therefore, it can be observed from Table 6.1 that a strong positive correlation (ρ) between λ and dsc ensures the reliability of a trustworthiness measure employed by *ET*. Further, ρ consistently surpasses ρ_e across all considered datasets

and models, demonstrating that *ET* offers a more effective trustworthiness evaluating measure as compared to VVC. This occurs as the VVC ignores the spatial position and shape of foreground areas, hindering its effectiveness in determining trustworthiness.

Table 6.1: Comparative results of *ET* method.

Dataset	Methods	M1	M2	M3	M4	M5	M6
CVC-300	ρ_e	0.27	0.32	0.34	0.46	0.32	0.36
	$\rho\ (Ours)$	**0.44**	**0.47**	**0.80**	**0.70**	**0.88**	**0.87**
CVC-ClinicDB	ρ_e	0.21	0.41	0.29	0.26	0.21	0.35
	$\rho\ (Ours)$	**0.89**	**0.81**	**0.55**	**0.65**	**0.75**	**0.95**
CVC-ColonDB	ρ_e	0.43	0.35	0.40	0.37	0.40	0.39
	$\rho\ (Ours)$	**0.86**	**0.91**	**0.90**	**0.89**	**0.72**	**0.66**
ETIS	ρ_e	0.36	0.44	0.41	0.34	0.40	0.33
	$\rho\ (Ours)$	**0.79**	**0.81**	**0.81**	**0.53**	**0.77**	**0.70**
Kvasir	ρ_e	0.38	0.30	0.31	0.23	0.31	0.30
	$\rho\ (Ours)$	**0.82**	**0.68**	**0.71**	**0.74**	**0.77**	**0.69**

Note: Existing Models - **M1**: UACANet-S, **M2**: UACANet-L, **M3**: PraNet, **M4**: CaraNet, **M5**: SSFormer-S, **M6**: SSFormer-L. All the values are the Pearson correlation coefficient, measured between: (i) TTA-based 1/VVC method and dice performance of original (non-rotated) MIS prediction with GT (dsc), denoted as ρ_e. (ii) our proposed confidence measure (λ) and dsc, depicted as ρ.

We perform a comparison between our proposed method, *ENT* and existing methods [57, 113], which incorporate input variants and apply majority voting to achieve the final prediction. The existing methods used in [57, 113] perform TTA and omit trustworthiness measures for advancing the MIS performance. The comparative results between the MIS performance from our proposed method, *ENT* (dsc^{ENT}) and the existing TTA-based method (dsc^{TTA}) are demonstrated in Table 6.2. It shows that dsc^{ENT} surpasses dsc^{TTA} in several cases, signifying that *ENT*, which relies solely on trustworthiness measures, offers a more effective way to enhance the MIS performance than TTA-based methods. A major limitation of TTA methods is that they assign equal importance to all predictions (input and its all variants) through majority voting. Consequently, even when the actual (non-rotated) pre-

Table 6.2: Comparative results of *ENT* method.

Dataset	Models	dsc	dsc^{TTA}	$dsc^{ENT}(Ours)$
CVC-300	PraNet [3]	0.8710	0.8998*	0.9064*†
	SSFormer-S [14]	0.8870	0.8872*	0.8939*†
	SSFormer-L [14]	0.8950	0.9051*	0.9095*†
	UACANet-S [65]	0.9020	0.8934	0.9037*†
	UACANet-L [65]	0.9100	0.9029	0.9136*†
	CaraNet [66]	0.8771	0.8593	0.8785*†
CVC-ClinicDB	PraNet [3]	0.8990	0.8363	0.8385†
	SSFormer-S [14]	0.9160	0.9155	0.9229*†
	SSFormer-L [14]	0.9060	0.9016	0.9092*†
	UACANet-S [65]	0.9160	0.9044	0.9183*†
	UACANet-L [65]	0.9260	0.9091	0.9140†
	CaraNet [66]	0.9016	0.8813	0.8942†
CVC-ColonDB	PraNet [3]	0.7090	0.7152*	0.7275*†
	SSFormer-S [14]	0.7720	0.7712	0.7744*†
	SSFormer-L [14]	0.8020	0.8066*	0.7984
	UACANet-S [65]	0.7830	0.7867*	0.7828
	UACANet-L [65]	0.7510	0.7520*	0.7673*†
	CaraNet [66]	0.7102	0.7273*	0.7326*†
ETIS	PraNet [3]	0.6280	0.6481*	0.6611*†
	SSFormer-S [14]	0.7670	0.7796*	0.7813*†
	SSFormer-L [14]	0.7960	0.8150*	0.8200*†
	UACANet-S [65]	0.6940	0.7149*	0.7264*†
	UACANet-L [65]	0.7660	0.7664*	0.7682*†
	CaraNet [66]	0.6355	0.6064	0.6129†
Kvasir	PraNet [3]	0.8980	0.8977	0.9043*†
	SSFormer-S [14]	0.9250	0.9161	0.9263*†
	SSFormer-L [14]	0.9170	0.9146	0.9158†
	UACANet-S [65]	0.9050	0.9084*	0.9093*†
	UACANet-L [65]	0.9120	0.8943	0.9135*†
	CaraNet [66]	0.9112	0.8821	0.9127*†

Note: The digits signify the dice score, measured between GT and : (i) actual MIS prediction (dsc), (ii) existing TTA prediction (dsc^{TTA}), (iii) the proposed *ENT* prediction (dsc^{ENT}). $*$ and † signifies values larger than dsc and dsc^{TTA}, respectively.

diction is accurate, the final MIS performance suffers a lot if any of the variant predictions are incorrect.

Further, Table 6.2 also illustrates that dsc^{ENT} outperforms dsc in several scenarios,

Table 6.3: Comparative results of *ENT* across existing PraNet model (employed rotation augmentation during training).

Dataset	dsc^{aug}	$dsc^{ENT}(Ours)$
CVC-ColonDB	0.7217	**0.7275**
Kvasir	0.9006	**0.9043**
CVC-ClinicDB	0.8264	**0.8385**
CVC-300	0.8989	**0.9064**
ETIS	0.6581	**0.6611**

Note: dsc^{aug} represents dice score, measured between the GT and the output of the existing PraNet model (trained by using rotation augmentation).

Table 6.4: Comparative results of *CSM* method.

Models	**Dataset**				
	CVC-300	**CVC-ClinicDB**	**CVC-ColonDB**	**ETIS**	**Kvasir**
M1	0.9020	0.9160	0.7830	0.6940	0.9050
M2	0.9100	**0.9260**	0.7510	0.7660	0.9120
M3	0.8710	0.8990	0.7090	0.6280	0.8980
M4	0.8771	0.9016	0.7102	0.6355	0.9112
M5	0.8870	0.9160	0.7720	0.7670	**0.9250**
M6	0.8950	0.9060	0.8020	0.7960	0.9170
M1$^{\text{aug}}$	0.8989	0.8264	0.7217	0.6581	0.9006
CSM (Ours)	**0.9166**	0.9214	**0.8062**	**0.8095**	0.9213

Note: Existing Models - **M1**: UACANet-S, **M2**: UACANet-L, **M3**: PraNet, **M4**: CaraNet, **M5**: SSFormer-S, **M6**: SSFormer-L, **M1$^{\text{aug}}$**: PraNet (using rotation augmentation during training).

highlighting that *ENT* method effectively advances the MIS performance of existing well-known models. It occurs because our proposed method, ***TrsusMedIS***, precisely examines the MIS models' trustworthiness by *ET* method and refines the non-trustworthy predictions using *ENT* method, leveraging input variants. Notably, dsc depicts the existing MIS model's performance trained with rotation augmentation, except for the PraNet model [3]. Thus, we perform training on the PraNet model by employing rotation augmentation and distinguish

their performance (dsc^{aug}) with dsc^{ENT}. The results are described in Table 6.3, which reveal that dsc^{ENT} performs better than dsc^{aug} across each dataset, ensuring the feasibility of *ENT* method.

Table 6.4 distinguishes the MIS performance of *CSM* against the existing MIS models across all the datasets. It reveals that *CSM* surpasses the performance of existing models, specifically on the CVC-300, CVC-ColonDB, and ETIS datasets. Moreover, *CSM* achieves competitive performance on the CVC-ClinicDB and Kvasir datasets, surpassing all models except UACANet-L on CVC-ClinicDB and SSFormer-S on Kvasir. Furthermore, we compare the *CSM* with the PraNet model, which we trained with rotation augmentation operation, denoted as M1^{aug}. It indicates that *CSM* offers better predictions than M1^{aug} for all the datasets. This improved performance by *CSM* is due to the fact that it incorporates similarity between the input's prediction and their corresponding variants and selects the optimal MIS model, generating the most trustworthy prediction. These performances also ensure that *ET* offers a robust trustworthiness indicator that can be effectively adapted for diverse MIS models.

6.2.5 Ablation Study

Several ablation studies have been performed by ***TrsutMedIS***. To comprehensively analyse the impact of a prespecified trustworthiness threshold t on *ENT*, we evaluate their performance by setting diverse values of t. The results are given in Table 6.5, which highlights that optimal performance is achieved on t as 0.9. It also depicts the possibility of tuning t to some different value at which the optimal results can be obtained.

Our proposed method, *CSM*, initially considers a model that offers the most effective and trustworthy prediction for each input. To thoroughly understand the *CSM* method, we employ several operations, such as averaging (*Avg*), Majority Voting (*MV*) and Maximum

Table 6.5: *ENT* performance on different thresholds.

Dataset	Models	Different Threshold Values				
		0.86	**0.88**	**0.9**	**0.92**	**0.94**
CVC-300	**M1**	0.905	0.904	0.904	0.904	0.905
	M2	0.913	0.913	0.914	0.913	0.913
	M3	0.905	0.905	0.906	0.906	0.907
	M4	0.874	0.874	0.878	0.880	0.883
	M5	0.894	0.893	0.894	0.895	0.894
	M6	0.908	0.909	0.909	0.910	0.911
CVC-ClinicDB	**M1**	0.916	0.916	0.918	0.918	0.918
	M2	0.914	0.914	0.914	0.913	0.913
	M3	0.839	0.839	0.838	0.838	0.834
	M4	0.895	0.895	0.894	0.894	0.893
	M5	0.923	0.923	0.923	0.923	0.923
	M6	0.909	0.909	0.909	0.908	0.908
CVC-ColonDB	**M1**	0.783	0.783	0.783	0.782	0.780
	M2	0.769	0.768	0.767	0.767	0.766
	M3	0.726	0.727	0.728	0.727	0.726
	M4	0.731	0.732	0.733	0.733	0.733
	M5	0.773	0.774	0.774	0.774	0.775
	M6	0.797	0.798	0.798	0.798	0.797
ETIS	**M1**	0.726	0.726	0.726	0.728	0.730
	M2	0.768	0.768	0.768	0.768	0.767
	M3	0.661	0.662	0.661	0.662	0.665
	M4	0.612	0.613	0.613	0.611	0.610
	M5	0.781	0.782	0.781	0.782	0.783
	M6	0.820	0.820	0.820	0.820	0.821
Kvasir	**M1**	0.909	0.910	0.909	0.908	0.908
	M2	0.915	0.915	0.913	0.913	0.913
	M3	0.905	0.904	0.904	0.905	0.906
	M4	0.912	0.912	0.913	0.913	0.914
	M5	0.920	0.920	0.922	0.922	0.922
	M6	0.910	0.914	0.916	0.915	0.915

Note: The digits indicate the dice values, measured between advanced trustworthy MIS prediction by *ENT* and GT. Existing Models - **M1**: UACANet-S, **M2**: UACANet-L, **M3**: PraNet, **M4**: CaraNet, **M5**: SSFormer-S, **M6**: SSFormer-L.

Foreground Pixels (*MFP*), to select the optimal MIS model. The performances of these operations are compared against *CSM*, as given in Table 6.6. *Avg* provides the final segmen-

Table 6.6: *CSM* performance against diverse operations.

Dataset	Diverse Operaions			CSM (Ours)
	MFP	***Avg***	***MV***	
CVC-ColonDB	0.755	0.753	0.774	**0.806**
Kvasir	0.908	0.912	0.916	**0.921**
CVC-ClinicDB	0.903	0.914	0.912	**0.921**
CVC-300	0.833	0.885	0.903	**0.917**
ETIS	0.625	0.684	0.779	**0.810**

Note: The digits signify the dice performance, measured between *CSM* prediction and GT.

Table 6.7: Sample distribution allocated by *CSM* method across various MIS models.

Dataset	Total Samples	M1	M2	M3	M4	M5	M6
CVC-ColonDB	380	38	64	8	77	100	93
Kvasir	100	10	11	0	34	28	17
CVC-ClinicDB	62	4	6	1	18	23	10
CVC-300	60	7	21	1	13	9	9
ETIS	196	16	35	6	37	60	42

Note: Existing Models - **M1**: UACANet-S, **M2**: UACANet-L, **M3**: PraNet, **M4**: CaraNet, **M5**: SSFormer-S, **M6**: SSFormer-L.

tation mask by averaging the individual performance of all the considered models, while *MV* applies majority voting on each input pixel for MIS. *MFP* selects a model based on the maximum count of foreground pixels. As depicted in Table 6.6, *CSM* surpasses all these operations, assuring its ability to choose the optimal model through trustworthy analysis. Further, our proposed method, *CSM*, offers flexibility by dynamically selecting the optimal model and does not depend upon one model. This adaptability is evidenced in Table 6.7, which highlights the distribution of sample counts chosen by *CSM* across all considered MIS models.

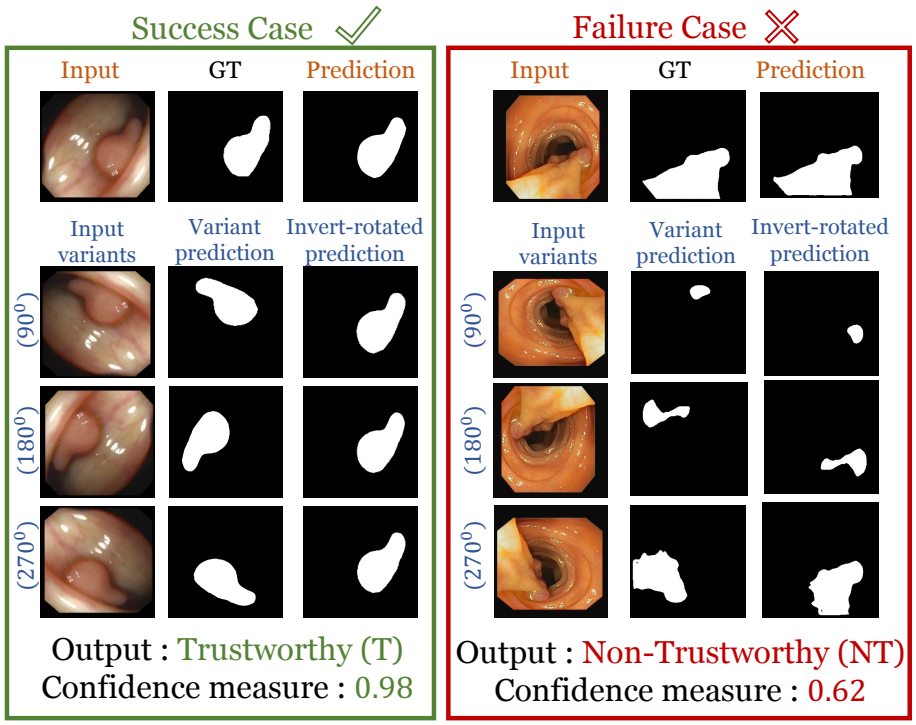

Figure 6.5: Qualitative results of *ENT* method, showcasing both success and failure cases.

6.3 Discussion

In Section 6.2.4, Table 6.1 justifies that *ET* method offers reliable trustworthiness metrics λ by showing significant high values of ρ. Likewise, Table 6.2 and 6.4 demonstrate that *ENT* method enhances the performance and efficacy of existing MIS models. Unfortunately, we observe the performance reduction in Table 6.2 and Table 6.4 in a few cases. This happens because the MIS predictions exhibit GT with additional erroneous blobs, leading to lower consistency between input and its respective variants, degrading model performance. This can be depicted in Figure 6.5, visualising the success and failure cases of our proposed *ENT* method. A notable limitation of ***TrustMedIS*** is that while *ENT* and *CSM* aim to advance MIS performance, they struggle to offer confidence metrics for the final output. Moreover, Table 6.5 demonstrates that the MIS performance can be increased by tuning diverse thresholds t.

6.4 Summary

This chapter has proposed a method, ***TrustMedIS***, that examines the trustworthiness of MIS models for ID samples through our proposed *ET* method. In particular, the *ET* method observes the consistency between the input samples' MIS predictions and their corresponding input variants, evaluates the confidence metric and compares it with a prespecified threshold to determine trustworthiness. Moreover, *TrustMedIS* investigates the relevance of a trustworthiness measure (achieved by *ET*) to boost the MIS model performance by our proposed *ENT* method and leveraging multiple models to achieve enhanced effectiveness by our proposed *CSM* method. Specifically, the *ENT* method has advanced the performance of non-trustworthy predictions by assessing the trustworthiness of input variant predictions. Likewise, *CSM* has considered multiple MIS models and adaptively selects the optimal one that offers the most trustworthy and optimal prediction, ultimately improving the efficacy of the model. Experimental outcomes revealed that ***TrustMedIS***, consisting of *ET*, *ENT* and *CSM*, has successfully accomplished its objectives, leading to enhanced MIS performance and robustness.

Chapter 7

Conclusion and Future Scopes

This chapter provides the conclusion and possible future directions of the overall thesis. The objective of this thesis is to strengthen the DL-based MIS models by advancing their performance, adversarial robustness, and trustworthiness. To achieve this, we have identified several research gaps in existing DL-based MIS models and proposed innovative solutions. Thus, our major thesis contributions are structured in four folds. Initially, we have proposed a novel adversarial attack, ***DECEIT***, to understand the efficacy of adversarial attacks in DL-based MIS models. Subsequently, we have developed a novel detection method, ***DISCERN***, to effectively identify the adversarial and OOD samples in DL-based MIS models through consistency analysis between input and their respective variants. Eventually, we have proposed a novel adversarial defence, ***RELIVE***, to defend against adversarial attacks by leveraging contrastive and multitask learning. Additionally, we have introduced a novel method, ***TrustMedIS***, to investigate the trustworthiness and improve the performance of DL-based MIS models. Collectively, all four of these contributions lead to an adversarially robust, trustworthy, and better-performing DL-based MIS model.

Our proposed attack, ***DECEIT***, dynamically selects the optimal surrogate loss function to effectively deceive the DL-based MIS models, misguiding their predictions. Several MIS

models exhibit non-differentiable layers and loss functions, which inhibit backpropagation and disrupt attack. As a solution, ***DECEIT***, has replaced these non-differentiable neural network layers with the approximated differential layers. Moreover, it has substituted non-differentiable loss functions with surrogate losses. However, the optimal selection of the surrogate loss function is crucial for an effective attack. Thus, ***DECEIT*** has performed parallel fusion by evaluating attacks on multiple surrogate loss functions and selecting the optimal one that induces minimal perturbation. Experimental outcomes on publicly available polyp datasets reveal that ***DECEIT*** has surpassed the well-known attacks across multiple well-known DL-based MIS models, achieving a high attack success rate with minimum average distortion.

The clean samples' MIS prediction are strongly consistent with their respective variants, while the adversarial and OOD samples exhibit significantly lower consistency. Our proposed detection method, ***DISCERN***, exploits this insights to precisely detect adversarial and OOD samples for DL-based MIS models. To achieve this, we have generated input variants by rotation and evaluated the consistency between the input samples' MIS prediction and their respective variants. These consistency values are further utilised to train the model that effectively differentiates adversarial/OOD and clean samples. Experimental results across several existing DL-based MIS models demonstrate that ***DISCERN*** has outperformed the existing detection method by obtaining the maximum detection success rate.

Our proposed adversarial defence, ***RELIVE***, advances the model's robustness by defending against adversarial attacks while providing minimal improvement in model performance. It comprises contrastive learning, multitask learning, and their fusion-based defence. Initially, our proposed contrastive learning-based defence has computed contrastive loss by enforcing the model to capture similar features for clean, adversarial, and augmented samples, mitigating the efficacy of adversarial perturbations. Additionally, it has also employed

data fidelity loss to enhance model performance across each sample type. Subsequently, our multitask learning-based defence has provided generic feature representation by learning diverse features across multiple tasks. It has additionally selected auxiliary tasks relative to the main task. We have observed that robustness improves when there is a lower correlation between the auxiliary and main tasks. Eventually, our proposed contrastive multitask fusion-based defence leverages the individual benefits of both contrastive and multitask learning. Following the proposed multitask model, contrastive learning is exclusively incorporated into the main task. This fusion-based defence further reinforces the model's resistance to adversarial attacks. As a result, ***RELIVE*** has surpassed the existing defences and achieved enhanced robustness of the DL-based MIS model by lowering the attack success rate up to 0% in high average distortion with slight elevation in the model performance.

Our proposed method, ***TrustMedIS***, assess the DL-based MIS models' trustworthiness for ID samples. It has operated in three key methods: *ET*, *ENT* and *CSM*. The *ET* method has evaluated the similarity between the MIS predictions of input and its rotated variants, followed by calculating a confidence measure to determine trustworthiness. This computed confidence measure from *IT* has further been utilised to advance the model performance through the *ENT* method and achieved improved efficacy by considering several MIS models using the *CSM* method. In particular, *ENT* has refined the non-trustworthy predictions, while *CSM* has flexibly picked the optimal model that offers the most trustworthy prediction. Experimental outcomes revealed that ***TrustMedIS***, consisting of *ET*, *ENT*, and *CSM*, has effectively achieved its objectives.

Future scopes include devising more resilient and explainable DL-based MIS models. To achieve this, several well-known model-based and post-hoc explainability methods can be explored to enrich trust in model decisions. Our proposed detection method, ***DISCERN***, can evaluate the effectiveness of any adversarial defence. In essence, ***DISCERN*** can be

iteratively applied to identify samples prior and after defence implementation to check the reduction in perturbation impact on the sample. Currently, ***DISCERN*** performs binary classification to detect samples as clean and adversarial/OOD. It can be extended to support multi-class classification, differentiating between clean, adversarial, *near* and *far* OOD samples, providing in-depth and interpretable results. Through these approaches, one can strive to foster greater robustness and transparency in DL-based MIS models.

Bibliography

[1] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, p. 221, 2017.

[2] S. Shukla, L. Birla, A. K. Gupta, and P. Gupta, "Trustworthy Medical Image Segmentation with improved performance for in-distribution samples," *Neural Networks*, vol. 166, pp. 127–136, 2023.

[3] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020, pp. 263–273.

[4] Q. Huang, D. Wang, Z. Lu, S. Zhou, J. Li, L. Liu, and C. Chang, "A novel image-to-knowledge inference approach for automatically diagnosing tumors," *Expert Systems with Applications*, vol. 229, p. 120450, 2023.

[5] L. Birla, S. Shukla, A. K. Gupta, and P. Gupta, "ALPINE: Improving remote heart rate estimation using contrastive learning," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5029–5038.

[6] L. Birla, S. Shukla, T. Saikia, and P. Gupta, "HR-TRACK: An rPPG Method for Heartrate Monitoring Using Temporal Convolution Networks," in *International Conference on Pattern Recognition*, 2024, pp. 370–385.

[7] A. Esteva *et al.*, "Deep learning-enabled medical computer vision," *Digital Medicine*, vol. 4, no. 1, pp. 1–9, 2021.

[8] T. Akinci D'Antonoli *et al.*, "TotalSegmentator MRI: Robust Sequence-independent Segmentation of Multiple Anatomic Structures in MRI," *Radiology*, vol. 314, no. 2, p. e241613, 2025.

[9] C. Gao, L. Wu, W. Wu, Y. Huang, X. Wang, Z. Sun, M. Xu, and C. Gao, "Deep learning in pulmonary nodule detection and segmentation: a systematic review," *European radiology*, vol. 35, no. 1, pp. 255–266, 2025.

[10] Y. Yuan, M. Chao, and Y. Lo, "Automatic skin lesion segmentation with fully convolutional-deconvolutional networks," *CoRR*, vol. abs/1703.05165, 2017.

[11] A. Srivastava, A. Ramagiri, P. Gupta, and V. Gupta, "SANGAM: Synergizing Local and Global Analysis for Simultaneous WBC Classification and Segmentation," in *International Conference on Pattern Recognition*, 2025, pp. 154–169.

[12] L. Huang, S. Ruan, P. Decazes, and T. Denœux, "Deep evidential fusion with uncertainty quantification and reliability learning for multimodal medical image segmentation," *Information Fusion*, vol. 113, p. 102648, 2025.

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

[14] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, and S. Song, "Stepwise feature fusion: Local guides global," in *Medical Image Computing and Computer Assisted Intervention*, 2022, pp. 110–120.

[15] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.

[16] H. Cui, Y. Wang, F. Zheng, Y. Li, Y. Zhang, and Y. Xia, "P2TC: A Lightweight Pyramid Pooling Transformer-CNN Network for Accurate 3D Whole Heart Segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[17] A. Das, "Adaptive UNet-based Lung Segmentation and Ensemble Learning with CNN-based Deep Features for Automated COVID-19 Diagnosis," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5407–5441, 2022.

[18] V. Sasank and S. Venkateswarlu, "An automatic tumour growth prediction based segmentation using full resolution convolutional network for brain tumour," *Biomedical Signal Processing and Control*, vol. 71, p. 103090, 2022.

[19] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, pp. 137–178, 2021.

[20] A. Arnab, O. Miksik, and P. H. Torr, "On the robustness of semantic segmentation models to adversarial attacks," in *Computer Vision and Pattern Recognition*, 2018, pp. 888–897.

[21] A. S. Panayides *et al.*, "AI in medical imaging informatics: current challenges and future directions," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 7, pp. 1837–1857, 2020.

[22] B. Hu and A. Qin, "Expert-Adaptive Medical Image Segmentation," in *IEEE Conference on Artificial Intelligence*, 2024, pp. 549–554.

[23] T. White, E. Blok, and V. D. Calhoun, "Data sharing and privacy issues in neuroimaging research: Opportunities, obstacles, challenges, and monsters under the bed," *Human Brain Mapping*, vol. 43, no. 1, pp. 278–291, 2022.

[24] N. Nasalwai, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Addressing the class imbalance problem in medical image segmentation via accelerated tversky loss function," in *Pacific-Asia conference on knowledge discovery and data mining*, 2021, pp. 390–402.

[25] J. Lee, C. Liu, J. Kim, Z. Chen, Y. Sun, J. R. Rogers, W. K. Chung, and C. Weng, "Deep learning for rare disease: A scoping review," *Journal of Biomedical Informatics*, vol. 135, p. 104227, 2022.

[26] N. Pradhan, V. S. Dhaka, G. Rani, and H. Chaudhary, "Transforming view of medical images using deep learning," *Neural Computing and Applications*, vol. 32, no. 18, pp. 15 043–15 054, 2020.

[27] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: an experimental study," *Medical Image Analysis*, vol. 89, p. 102918, 2023.

[28] P. Gupta and E. Rahtu, "MLAttack: Fooling semantic segmentation networks by multi-layer attacks," in *German Conference on Pattern Recognition*, 2019, pp. 401–413.

[29] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognition*, vol. 110, p. 107332, 2021.

[30] M. Xu, T. Zhang, Z. Li, M. Liu, and D. Zhang, "Towards evaluating the robustness of deep diagnostic models by adversarial attack," *Medical Image Analysis*, vol. 69, p. 101977, 2021.

[31] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International Conference on Machine Learning*, 2018, pp. 274–283.

[32] M. Cissé, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling Deep Structured Visual and Speech Recognition Models with Adversarial Examples," in *Advances in Neural Information Processing Systems*, 2017, pp. 6977–6987.

[33] F. Carrara, F. Falchi, R. Caldelli, G. Amato, and R. Becarelli, "Adversarial image detection in deep neural networks," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 2815–2835, 2019.

[34] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *Symposium on Security and Privacy*, 2017, pp. 39–57.

[35] S. B. U. Haque and A. Zafar, "Robust medical diagnosis: a novel two-phase deep learning framework for adversarial proof disease detection in radiology images," *Journal of Imaging Informatics in Medicine*, vol. 37, no. 1, pp. 308–338, 2024.

[36] Z. Liu, J. Zhang, V. Jog, P.-L. Loh, and A. B. McMillan, "Robustifying deep networks for medical image segmentation," *Journal of digital imaging*, vol. 34, pp. 1279–1293, 2021.

[37] L. Ma and L. Liang, "Increasing-margin adversarial (IMA) training to improve adversarial robustness of neural networks," *Computer Methods and Programs in Biomedicine*, vol. 240, p. 107687, 2023.

[38] C.-H. Ho and N. Nvasconcelos, "Contrastive learning with adversarial examples," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 081–17 093, 2020.

[39] C. Mao, A. Gupta, V. Nitin, B. Ray, S. Song, J. Yang, and C. Vondrick, "Multitask learning strengthens adversarial robustness," in *European Conference on Computer Vision*, 2020, pp. 158–174.

[40] A. T. Sandnes, B. Grimstad, and O. Kolbjørnsen, "Multi-task neural networks by learned contextual inputs," *Neural Networks*, vol. 179, p. 106528, 2024.

[41] J. Li, P. Chen, Z. He, S. Yu, S. Liu, and J. Jia, "Rethinking Out-of-distribution (OOD) detection: Masked Image Modeling is All You Need," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 578–11 589.

[42] Y. Wang, H. Xue, K. Zhou, T. Chen, and X. Wang, "A two-stage co-adversarial perturbation to mitigate out-of-distribution generalization of large-scale graph," *Expert Systems with Applications*, p. 124472, 2024.

[43] A. Uwimana and R. Senanayake, "Out of distribution detection and adversarial attacks on deep neural networks for robust medical image analysis," *arXiv preprint arXiv:2107.04882*, 2021.

[44] M. Angus, K. Czarnecki, and R. Salay, "Efficacy of pixel-level OOD detection for semantic segmentation," *arXiv preprint arXiv:1911.02897*, 2019.

[45] X. Li, D. Pan, and D. Zhu, "Defending against adversarial attacks on medical imaging ai system, classification or detection?" in *International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 1677–1681.

[46] D. Karimi and A. Gholipour, "Improving calibration and out-of-distribution detection in deep models for medical image segmentation," *IEEE Transactions on Artificial Intelligence*, 2022.

[47] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[48] C. González *et al.*, "Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation," *Medical Image Analysis*, vol. 82, p. 102596, 2022.

[49] M. Hein, M. Andriushchenko, and J. Bitterwolf, "Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 41–50.

[50] X. Zhang, B. Qian, S. Cao, Y. Li, H. Chen, Y. Zheng, and I. Davidson, "INPREM: An interpretable and trustworthy predictive model for healthcare," in *International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 450–460.

[51] J. Postels, M. Segu, T. Sun, L. D. Sieber, L. Van Gool, F. Yu, and F. Tombari, "On the practicality of deterministic epistemic uncertainty," in *International Conference on Machine Learning*, vol. 162, 2022, pp. 17 870–17 909.

[52] B. Lambert, F. Forbes, A. Tucholka, S. Doyle, H. Dehaene, and M. Dojat, "Trustworthy clinical AI solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis," *arXiv preprint arXiv:2210.03736*, 2022.

[53] M. S. Ayhan and P. Berens, "Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks," in *Medical Imaging with Deep Learning*, 2018.

[54] K. Zou, X. Yuan, X. Shen, Y. Chen, M. Wang, R. S. M. Goh, Y. Liu, and H. Fu, "EvidenceCap: Towards trustworthy medical image segmentation via evidential identity cap," *arXiv preprint arXiv:2301.00349*, 2023.

[55] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley *et al.*, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[56] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.

[57] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, 2019.

[58] K. Lekadir, R. Osuala, C. Gallin, N. Lazrak, K. Kushibar *et al.*, "FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging," *arXiv preprint*, 2021.

[59] P. Angelov and E. Soares, "Towards explainable deep neural networks (xDNN)," *Neural Networks*, vol. 130, pp. 185–194, 2020.

[60] A. Agarwal, R. Singh, M. Vatsa, and N. Ratha, "Image transformation-based defense against adversarial perturbation on deep learning models," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2106–2121, 2020.

[61] S. Aljahdali and E. A. Zanaty, "Combining multiple segmentation methods for improving the segmentation accuracy," in *IEEE Symposium on Computers and Communications*, 2008, pp. 649–653.

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

[63] 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*, 2018, pp. 3–11.

[64] A. Faisal and C. Pluempitiwiriyawej, "Active contour driven by scalable local regional information on expandable kernel," *Journal of Science and Applicative Technology*, vol. 4, no. 1, pp. 1–14, 2020.

[65] T. Kim, H. Lee, and D. Kim, "UACAnet: Uncertainty augmented context attention for polyp segmentation," in *ACM International Conference on Multimedia*, 2021, pp. 2167–2175.

[66] A. Lou, S. Guan, H. Ko, and M. H. Loew, "CaraNet: Context axial reverse attention network for segmentation of small medical objects," in *Medical Imaging 2022: Image Processing*, vol. 12032, 2022, pp. 81–92.

[67] N. K. Tomar, D. Jha, U. Bagci, and S. Ali, "TGANet: Text-guided attention for improved polyp segmentation," *arXiv preprint arXiv:2205.04280*, 2022.

[68] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *International Conference on Learning Representations*, 2015.

[69] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *International Conference on Learning Representations, Workshop Track Proceedings*, 2017.

[70] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," in *International Conference on Learning Representations, Conference Track Proceedings*, 2018.

[71] R. Paul, M. Schabath, R. Gillies, L. Hall, and D. Goldgof, "Mitigating adversarial attacks on medical image understanding systems," in *International Symposium on Biomedical Imaging*, 2020, pp. 1517–1521.

[72] U. Ozbulak, A. Van Messem, and W. De Neve, "Impact of adversarial examples on deep learning models for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 300–308.

[73] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in *International Conference on Computer Vision*, 2017, pp. 1369–1378.

[74] M. Paschali, S. Conjeti, F. Navarro, and N. Navab, "Generalizability vs. Robustness: Investigating medical imaging networks using adversarial examples," in *Medical Image Computing and Computer Assisted Intervention*, vol. 11070, 2018, pp. 493–501.

[75] M. Pervin, L. Tao, A. Huq, Z. He, L. Huo *et al.*, "Adversarial attack driven data augmentation for accurate and robust medical image segmentation," *arXiv preprint arXiv:2105.12106*, 2021.

[76] B. Zhao, X. Wen, and K. Han, "Learning semi-supervised gaussian mixture models for generalized category discovery," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 623–16 633.

[77] F. Almalik, M. Yaqub, and K. Nandakumar, "Self-ensembling vision transformer (sevit) for robust medical image classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022, pp. 376–386.

[78] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, 2021.

[79] W. Fu, Y. Chen, W. Liu, X. Yue, and C. Ma, "Evidence Reconciled Neural Network for Out-of-Distribution Detection in Medical Images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023, pp. 305–315.

[80] C. Chen, Z. Li, C. Ouyang, M. Sinclair, W. Bai, and D. Rueckert, "Maxstyle: Adversarial style composition for robust medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022, pp. 151–161.

[81] A. Shafahi, M. Najibi, M. A. Ghiasi *et al.*, "Adversarial training for free!" *Advances in neural information processing systems*, vol. 32, 2019.

[82] L. Ma and L. Liang, "Adaptive adversarial training to improve adversarial robustness of dnns for medical image segmentation and detection," *arXiv preprint arXiv:2206.01736*, 2022.

[83] X. He, S. Yang, G. Li, H. Li, H. Chang, and Y. Yu, "Non-local context encoder: Robust biomedical image segmentation against adversarial attacks," in *AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8417–8424.

[84] C. Chen, D. Ye, Y. He, L. Tang, and Y. Xu, "Improving Adversarial Robustness With Adversarial Augmentations," *IEEE Internet of Things Journal*, 2023.

[85] Y. Ding, J. Liu, X. Xu, M. Huang, J. Zhuang, J. Xiong, and Y. Shi, "Uncertainty-aware training of neural networks for selective medical image segmentation," in *Medical Imaging with Deep Learning*, 2020, pp. 156–173.

[86] G. Carannante, D. Dera, N. C. Bouaynaya, G. Rasool, and H. M. Fathallah-Shaykh, "Trustworthy Medical Segmentation with Uncertainty Estimation," *arXiv preprint*, 2021.

[87] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen, "Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps," *Medical Image Analysis*, vol. 60, p. 101619, 2020.

[88] Z. Ren, Y. Zhang, and S. Wang, "LCDAE: Data Augmented Ensemble Framework for Lung Cancer Classification," *Technology in Cancer Research & Treatment*, vol. 21, p. 15330338221124372, 2022.

[89] E. Jang, S. Gu, and B. Poole, "Categorical Reparameterization with Gumbel-Softmax," in *International Conference on Learning Representations*, 2017.

[90] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski, "Kornia: an open source differentiable computer vision library for pytorch," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 3674–3683.

[91] S. Jadon, "A survey of loss functions for semantic segmentation," in *Computational Intelligence in Bioinformatics and Computational Biology*, 2020, pp. 1–7.

[92] J. Su, Z. Liu, J. Zhang, V. S. Sheng, Y. Song, Y. Zhu, and Y. Liu, "DV-Net: Accurate liver vessel segmentation via dense connection model with D-BCE loss function," *Knowledge-Based Systems*, p. 107471, 2021.

[93] K. C. Wong, M. Moradi, H. Tang, and T. Syeda-Mahmood, "3D segmentation with exponential logarithmic loss for highly unbalanced object sizes," in *Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 612–619.

[94] 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 *Computer Vision and Pattern Recognition*, 2018, pp. 4413–4421.

[95] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach *et al.*, "A benchmark for endoluminal scene segmentation of colonoscopy images," *CoRR*, vol. abs/1612.00799, 2016.

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

[97] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2016.

[98] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-SEG: A segmented polyp dataset," in *International Conference on Multimedia Modeling*, 2020, pp. 451–462.

[99] H. Munusamy, J. Karthikeyan, G. Shriram, S. Thanga Revathi, and S. Aravindkumar, "FractalCovNet architecture for COVID-19 chest X-ray image classification and CT-Scan image segmentation," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 3, pp. 1025–1038, 2021.

[100] A. K. Gupta, P. Gupta, and E. Rahtu, "FATALRead-fooling visual speech recognition models," *Applied Intelligence*, pp. 1–16, 2021.

[101] S. Mishra, A. K. Gupta, and P. Gupta, "DARE: Deceiving audio–visual speech recognition model," *Knowledge-Based Systems*, vol. 232, p. 107503, 2021.

[102] J. Song, W. Ahn, S. Park, and M. Lim, "Failure detection for semantic segmentation on road scenes using deep learning," *Applied Sciences*, vol. 11, no. 4, p. 1870, 2021.

[103] M. Zaid, S. Ali, M. Ali, S. Hussein, A. Saadia, and W. Sultani, "Identifying out of distribution samples for skin cancer and malaria images," *Biomedical Signal Processing and Control*, vol. 78, p. 103882, 2022.

[104] C. Wan, F. Huang, and X. Zhao, "Average gradient-based adversarial attack," *IEEE Transactions on Multimedia*, vol. 25, pp. 9572–9585, 2023.

[105] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Annals of translational medicine*, vol. 8, no. 11, 2020.

[106] S. Shukla, A. K. Gupta, and P. Gupta, "Exploring the feasibility of adversarial attacks on medical image segmentation," *Multimedia Tools and Applications*, 2023.

[107] P. de Jorge, R. Volpi, P. H. Torr, and G. Rogez, "Reliability in Semantic Segmentation: Are We on the Right Track?" in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7173–7182.

[108] T. Eelbode, J. Bertels, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimization for medical image segmentation: Theory and practice when evaluating with Dice Score or Jaccard Index," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3679–3690, 2020.

[109] F. Nie, Z. Hao, and R. Wang, "Multi-class support vector machine with maximizing minimum margin," in *AAAI Conference on Artificial Intelligence*, vol. 38, no. 13, 2024, pp. 14 466–14 473.

[110] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[111] S. Mishra, A. K. Gupta, and P. Gupta, "DARE: Deceiving audio–visual speech recognition model," *Knowledge-Based Systems*, vol. 232, p. 107503, 2021.

[112] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *IEEE conference on computer vision and pattern recognition*, 2018, pp. 1857–1866.

[113] N. Moshkov, B. Mathe, A. Kertesz-Farkas, R. Hollandi, and P. Horvath, "Test-time augmentation for deep learning-based cell segmentation on microscopy images," *Scientific Reports*, vol. 10, no. 1, pp. 1–7, 2020.

[114] A. Roy and S. Chakraborty, "Support vector machine in structural reliability analysis: A review," *Reliability Engineering & System Safety*, vol. 233, p. 109126, 2023.

[115] J. Yang *et al.*, "Openood: Benchmarking generalized out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 598–32 611, 2022.

[116] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, and S. Savarese, "Which tasks should be learned together in multi-task learning?" in *International Conference on Machine Learning*, 2020, pp. 9120–9132.

[117] S. Ghamizi, M. Cordy, M. Papadakis, and Y. Le Traon, "Adversarial robustness in multi-task learning: Promises and illusions," in *AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 697–705.

[118] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, 2020, pp. 1597–1607.

[119] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[120] J. T. Barron, "A general and adaptive robust loss function," in *IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4331–4339.

[121] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.

[122] v7Labs, "Labeled COVID-19 Chest X-Ray Dataset," 2020, accessed on October, 2021. [Online]. Available: http://darwin.v7labs.com/

[123] W. Wang, E. Xie, X. Li, D.-P. Fan *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[124] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020, pp. 213–229.

[125] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE access*, vol. 8, pp. 4806–4813, 2019.