

HAUSDORFF DISTANCE-MIXNET: A HYBRID CONVOLUTIONAL NEURAL NETWORK-TRANSFORMER NETWORK WITH BOUNDARY-AWARE MECHANISMS FOR POLYP SEGMENTATION

A THESIS REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE

OF

MASTER OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

KARTIK PRUTHI

(24/CSE/10)

UNDER THE SUPERVISION OF

PROF. ANIL SINGH PARIHAR

PROFESSOR AND HEAD OF DEPARTMENT

Department of Computer Science & Engineering

Delhi Technological University



COMPUTER SCIENCE & ENGINEERING DEPARTMENT

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi 110042

MAY, 2026



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to **Prof. Anil Singh Parihar** for his continuous guidance and mentorship that he provided us during the project. He showed us the path to achieve our targets by explaining all the tasks to be done and explained to us the importance of this project as well as its industrial relevance. He was always ready to help and clear doubts regarding any hurdles in this research. Without his constant support and motivation, this project would not have been successful.

Place: Delhi

Date: 30.05.2026

A handwritten signature in black ink, appearing to read 'K. Pruthi'.

Kartik Pruthi

(24/CSE/10)



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, KARTIK PRUTHI (Roll No – 24/CSE/10) student of M.Tech in Dept. of Computer Science & Engineering, hereby declare that the thesis titled “**Hausdorff Distance-Mixnet: A Hybrid Convolutional Neural Network-Transformer Network With Boundary-Aware Mechanisms For Polyp Segmentation**” which is submitted by me to the Computer Science & Engineering Department, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of degree of Master of Technology in the period from August 2024 to May 2026 under the supervision of Dr. Anil Singh Parihar, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

A handwritten signature in black ink, appearing to read "K Pruthi", is written over a light blue horizontal line.

Candidate's Signature

This is to certify that the student has incorporated all the correctness suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor

Signature of External Examiner



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE BY THE SUPERVISOR

I hereby certify that the thesis titled “**Hausdorff Distance-Mixnet: A Hybrid Convolutional Neural Network-Transformer Network With Boundary-Aware Mechanisms For Polyp Segmentation**” which is submitted by Kartik Pruthi (24/CSE/10) Dept. of Computer Science & Engineering Department, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology in Computer Science and Engineering, is a record of the research work carried out by the students under my supervision during the academic session 2024-26.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Signature

Prof. Anil Singh Parihar

Professor

Department of Computer Science and Engineering

Delhi Technological university

Date : 30.05.2026

Hausdorff Distance-Mixnet: A Hybrid Convolutional Neural Network-Transformer Network with Boundary-Aware Mechanisms for Polyp Segmentation

Kartik Pruthi

ABSTRACT

Colorectal cancer is one of the most preventable cancers. Most colorectal cancers develop from benign polyps that can be removed if found early on colonoscopy. Automated polyp segmentation helps in this regard, but current approaches suffer from a long-standing dilemma that convolutional networks can model fine texture information but fail to model long-range dependency, while transformers can model global structure but ignore fine boundaries. Most hybrid methods simply add these representations in a naive manner and are trained with region-overlap losses like Dice that are insensitive to the worst-case boundary errors of interest in the clinical setting. In this thesis, we propose a hybrid dual-branch architecture, namely HD-MixNet, for boundary-aware polyp segmentation. The model consists of a convolutional branch with Res2Net and a Swin Transformer branch, both streams running in parallel. They are combined with a Boundary-Aware Mix-Fusion (BAMF) module, which uses learned edge-gated blending to prevent the boundary suppression problem in attention-based fusion. We explicitly supervise boundaries to YOHO-inspired Edge-Detect and Boundary-Enhance blocks, and train with a composite objective including a differentiable distance-transform approximation of the Hausdorff distance, aligning optimization to worst-case boundary risk, rather than simply overlap. HD-MixNet achieved a mean dice score of 0.896 on Kvasir-SEG benchmark and competitive region-level segmentation with a compact dual-branch design. We also critically analyse the model boundary behaviour and investigate the relation between the overlap and the Hausdorff performance of the model and the factors that influence the worst-case boundary error.

TABLE OF CONTENTS

ACKNOWLEDGEMENT	i
CANDIDATE’S DECLARATION	ii
CERTIFICATE BY THE SUPERVISOR	iii
ABSTRACT.....	iv
TABLE OF CONTENTS.....	v
LIST OF TABLES.....	viii
LIST OF FIGURES.....	ix
LIST OF ABBREVIATIONS.....	x
Chapter 1 INTRODUCTION	1
1.1 Background and Motivation.....	1
1.2 The Architectural Dichotomy and Its Limitations	2
1.3 The Metric–Objective Misalignment	3
1.4 Problem Statement	4
1.5 Contributions.....	5
1.6 Research Objectives	6
Chapter 2 LITERATURE REVIEW	7
2.1 The Architectural Evolution of Segmentation Networks.....	7
2.1.1 Convolutional Architectures	7
2.1.2 The Limits of Locality and the Motivation for Global Modeling.....	9
2.2 The Self-Attention Mechanism and Vision Transformers.....	9
2.2.1 Scaled Dot-Product Attention	9
2.2.2 The Vision Transformer and Its Cost.....	11
2.2.3 The Swin Transformer	11
2.2.4 Hybrid CNN–Transformer Architectures	12
2.3 Explicit Boundary Modeling.....	13
2.4 Loss Functions: From Region Overlap to Boundary Distance	14

2.4.1	Region-Based Losses	14
2.4.2	Boundary-Aware Losses	15
2.4.3	Distance-Based Losses and the Hausdorff Distance	16
2.5	Training Strategies and Evaluation Metrics	17
2.5.1	Data Augmentation and Optimization	17
2.5.2	Evaluation Metrics	18
2.6	Benchmark Datasets for Polyp Segmentation	18
Chapter 3	METHODOLOGY	20
3.1	Overview of the HD-MixNet Architecture	20
3.2	Res2Net-Based Convolutional Branch	21
3.2.1	The Res2Net Block	22
3.3	Swin Transformer Branch	23
3.3.1	Patch Embedding and Hierarchical Stages	23
3.4	Explicit Boundary Supervision: Edge-Detect Blocks	25
3.5	Boundary-Aware Mix-Fusion (BAMF) Module	25
3.5.1	Working Mechanism	25
3.6	Decoder and Output Heads	27
3.7	Training Objective	28
3.7.1	Component Losses	28
3.7.2	Hausdorff-DT Loss (as implemented)	29
3.7.3	Total Objective	29
Chapter 4	EXPERIMENTS AND RESULTS	30
4.1	Experimental Setup	30
4.1.1	Hardware and Software	30
4.1.2	Dataset and Split	30
4.1.3	Preprocessing and Augmentation	30
4.2	Training Configuration	31
4.3	Evaluation Metrics	31
4.4	Quantitative Results	31
4.5	Qualitative Results	33
4.6	Analysis of the Semantic–Boundary Trade-off	33
4.7	Analysis with the Help of Graphs Depicting Change per Epoch	34

Chapter 5 CONCLUSION AND FUTURE WORK	37
5.1 Summary of Contributions.....	37
5.2 Discussion	38
5.3 Limitations	38
5.4 Future Work	39
5.5 Concluding Remarks	40
REFERENCES.....	41

LIST OF TABLES

Table 2.1	Comparative summary of representative segmentation methods on Kvasir-SEG	19
Table 3.1	Res2Net branch feature dimensions	22
Table 4.1	Comparison with representative methods on Kvasir-SEG	32

LIST OF FIGURES

Figure 2.1	Canonical U-Net encoder–decoder.....	9
Figure 2.2	Three fusion paradigms side by side: (a) serial CNN→Transformer (TransUNet-style); (b) late concatenation of two encoder outputs; (c) the parallel, bidirectional, boundary-gated fusion used by HD-MixNet	13
Figure 3.1	Overall architecture of HD-MixNet.....	21
Figure 3.2	Res2Net based CNN branch for HD-MixNet architecture	23
Figure 3.3	The Edge-Detect block	25
Figure 3.4	The BAMF module in three phases	27
Figure 4.1	Qualitative results panel: Original image Ground-truth mask Predicted probability map Overlay (GT in green, prediction in red)	33
Figure 4.2	Trend for Dice score per epoch	35
Figure 4.3	Trend for IoU score per epoch.....	35
Figure 4.4	Trend for training loss over epochs	35
Figure 4.5	Trend for various other metrics: (a) training loss (b) validation Dice score (c) Hausdorff-95 distance (d) loss vs. Dice correlation (e) learning rate (f) optimal threshold evolution (g) per-epoch improvement rate (h) all normalized metrics (i) convergence analysis	36

LIST OF ABBREVIATIONS

Abbreviation	Full Form
BAMF	Boundary-Aware Mix-Fusion
BCE	Binary Cross-Entropy
BE	Boundary-Enhance
CADx	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
CRC	Colorectal Cancer
DLF	Double-Level Fusion
ED	Edge-Detect
FCN	Fully Convolutional Network
HD	Hausdorff Distance
HD95	95th-Percentile Hausdorff Distance
HD-DT	Hausdorff Distance Transform (loss)
IoU	Intersection over Union (Jaccard Index)
mDice	Mean Dice Similarity Coefficient
MSA	Multi-head Self-Attention
SE	Squeeze-and-Excitation
SW-MSA	Shifted-Window Multi-head Self-Attention
TIF	Transformer Interactive Fusion
ViT	Vision Transformer
W-MSA	Window Multi-head Self-Attention
YOHO	You-Only-Have-One (framework)
ASPP	Atrous Spatial Pyramid Pooling
DT	Distance Transform
GPU	Graphics Processing Unit
VRAM	Video Random-Access Memory

Chapter 1

INTRODUCTION

1.1 Background and Motivation

Colorectal cancer (CRC) is one of the most common malignancies and the second leading cause of cancer death globally. But it is also one of the most preventable of the major cancers because it almost always progresses through a slow well-understood pathway. Benign growths on the wall of the bowel, called polyps, acquire changes over years before a minority become malignant. The long pre-malignant window is why screening works so well. Finding and removing a polyp while it is still benign means preventing cancer, not treating cancer.

The gold standard for this screening is colonoscopy. The clinician introduces a flexible endoscope through the colon. This enables him to directly visualize the surface of the mucosa and to excise suspicious lesions simultaneously. However, human factors restrict the effectiveness of the colonoscopy. The accuracy of detection depends on the operator and also on the session with the same operator, due to fatigue, time of withdrawal and quality of bowel preparation. Some lesions are particularly difficult to visualize flat and sessile serrated polyps are elevated very little above surrounding tissue; low-contrast lesions are difficult to differentiate from the mucosa; and polyps situated at the edge of the field of view or behind folds are easily missed. Missing a polyp is not a harmless mistake. It is a lesion which may progress as a residual lesion. It is an established factor in interval cancers that occur between scheduled screenings.

In this context, computer aided diagnosis (CADx) has emerged to assist the endoscopist and not to replace him. In CADx two related tasks need to be distinguished. Detection tells us if there is a polyp. Segmentation deals with the more difficult spatial problem of determining exactly which pixels are part of the polyp. Detection and resection are clinical

procedures but resection is not simply about knowing that a lesion exists but knowing the extent and margin and this is where this thesis is about segmentation. A segmentation model is like a second pair of eyes: it can pick up a region that the eyes missed and, more importantly, delineate the boundary of a lesion, delineating the tissue to be excised.

So it is not a theoretical question, what quality is expected of that boundary. If the resection line is too close, there is the risk of leaving malignant cells behind, if the resection line is too far away healthy mucosa is unnecessarily removed. The clinically dangerous failure mode is not a small uniform shrinkage of the predicted region, but a localized large deviation at some point on the contour. The difference between average agreement and worst-case deviation is a recurring theme throughout this thesis and eventually provides the motivation for its central methodological choice.

1.2 The Architectural Dichotomy and Its Limitations

Deep learning has dominated recent advances in medical image segmentation, and these advances have been shaped by a decades-old divide between two families of feature extractors, each with a distinct -- and complementary -- mode of failure.

The family of U-Net is the prototype of the Convolutional Neural Networks (CNNs) based on two inductive biases: locality, which assumes that each output depends only on a small spatial neighborhood, and translation equivariance, which assumes that the same filter is applied everywhere. These biases make CNNs efficient and good at learning local structure, e.g. texture and edges, from limited data. But they are constrained by locality too. The effective receptive field, the region of the input that actually influences a given output, grows slowly with depth and is in practice much smaller than the theoretical maximum. In practice, this means that for colonoscopy, CNNs simply can't 'see' far enough when the polyp occupies too much of the frame, or when you need to relate one side of the image to the other to disambiguate it from a confusing fold, resulting in fragmented masks and false positives.

The Vision Transformers (ViTs) introduced overcame this limitation. Self-attention replaces convolution and allows each spatial location to (in principle) attend to all other locations, providing a global receptive field from the very beginning. Transformer-based segmenters therefore display better global semantic consistency. But this power of the transformer has a cost for the CNN. Self-attention is not biased towards local high-frequency detail. Transformers operate on coarse patches, and aggregate globally; resulting in semantically correct predictions in the large, but blurred at the edge, exactly where flat and low-contrast polyps are located, and exactly the edge detail needed for clinical use.

So the two paradigms fail in opposite directions. The CNN sees sharply but narrowly, the transformer broadly but softly. The conceptual basis for hybrid architectures is this symmetry. We find that models that combine CNN and transformer components, such as TransUNet, HiFormer, DS-TransUNet and MixFormer, consistently outperform single-paradigm baselines, confirming the merit of the combination. But most have two structural faults. Conveniently, serial pipelines (feeding CNN features to a transformer) or late concatenation are lossy, introducing semantic misalignment between two different representational spaces and diluting the high-frequency edge information that the CNN branch is best placed to contribute. Second, in what they learn: the boundary is not often a clean target. The transition zone of the predicted probability map is implicit only, and is supposed to be a byproduct of getting the region right. It is a remarkable omission to not supervise the boundary, especially for a task where the clinical value is concentrated at the boundary.

1.3 The Metric–Objective Misalignment

The third problem is the training goal. This is not the same as architecture, and may be more primitive. Most segmentation models, including hybrid models, are trained with region-overlap losses, mainly Dice and Cross Entropy. These losses are stable and effective at maximizing the area of agreement between prediction and ground truth, but

they have a property that is benign in many applications and dangerous in this one: they are largely insensitive to boundary shape. Even if one predicted mask has a clean contour and the other has a spur of error protruding far outside the true boundary, two predicted masks overlapping the ground truth equally will get almost the same Dice scores. One concrete example gives a sense of the stakes involved. Take two predictions about the same polyp. The first one matches the lesion almost perfectly, except for a uniform one-pixel under segmentation around the whole border. The second agrees well with the lesion over most of its extent, but generates a single tongue of false-positive tissue extending many pixels into healthy mucosa at one point. These could be almost the same as Dice. They are worlds apart by the clinically relevant criterion, the first safe, the second with exactly the large, localized excursion that would mislead a resection. The model trained and selected only on Dice has no reason to prefer the first. The thesis is fundamentally about the misalignment of metrics and objectives. Clinicians do not judge segmentation safety by average overlap, but by worst case boundary error, formalized as the Hausdorff Distance – the largest distance between any point on one contour and the nearest point on the other. The Hausdorff Distance is the correct metric for the clinical question, but not a good training objective. The definition includes non-differentiable max and min functions and an expensive exact computation of it . Therefore, even in works that report Hausdorff Distance as an evaluation metric, it is seldom optimized for, as they are trained on Dice and then measured on Hausdorff, leaving the quantity they care about outside the loop the model learns from.

1.4 Problem Statement

Modern deep learning has greatly progressed automated polyp segmentation, yet none of the current methods satisfy three key clinical requirements concurrently:

- i. robust extraction of global semantic context for distinguishing polyps from visually similar gastrointestinal structures,
- ii. accurate delineation of lesion boundaries for safe and complete resection, and

- iii. direct optimization to minimize worst-case boundary error, the clinically relevant risk factor.

The CNN architectures are constrained by their small receptive fields and have difficulty with large or obscured polyps. Transformer based designs reason globally but coarse boundaries. Current hybrids only partially address these: their fusion rarely explicitly deals with boundaries and they are still trained with overlap-based losses that do not penalize small but clinically dangerous deviations. This thesis addresses the problem of unavailability of a segmentation framework that combines a boundary-aware objective, explicit boundary-supervision and complementary CNN-Transformer feature learning into a single model.

1.5 Contributions

Contributions of this thesis are:

- i. **HD-MixNet**, a hybrid dual-branch architecture The parallel encoder consists of a **Res2Net-based CNN** branch for learning fine-grained texture and edge information, and a **Swin Transformer branch** for long-range semantic context. Both branches are kept in parallel during the whole encoding process unlike serial hybrids, so that no spatial detail or global context is lost.
- ii. **A Boundary-Aware Mix-Fusion (BAMF) module.** a learned, gated blend to fuse the CNN and Transformer streams, and modulate the fused features with explicit edge maps, thereby reducing semantic misalignment between branches and alleviating the boundary suppression common in attention-based fusion.
- iii. Explicit boundary supervision through **Edge-Detect (ED)** and **Boundary-Enhance (BE)** blocks. In the YOHO framework, Sobel initialized edge-detection blocks are used in the encoder and a boundary-enhancement block is used in the decoder to use the boundary structure as an explicit signal, rather than an implicit by-product of region prediction.
- iv. A clinically relevant **Hausdorff-DT** loss term. We incorporate a differentiable approximation of the Hausdorff distance, based on distance transforms, into a

composite training objective such that the optimization penalizes errors that are spatially distant more than errors that are local, bringing the training signal closer to the clinical safety criteria.

1.6 Research Objectives

The main objective is the development of an automatic accurate and clinically-reliable segmentation of polyps from endoscopic images. The goals are

- i. To design a hybrid CNN-Transformer architecture which learns global context and fine texture simultaneously.
- ii. To develop a fusion mechanism which aligns complementary representations while preserving high-frequency boundary information.
- iii. To introduce explicit boundary learning via dedicated edge-focused modules which maintain structural integrity from encoder to decoder.
- iv. To provide a clinically aligned objective via a differentiable Hausdorff-DT loss which penalizes worst-case boundary deviation.
- v. To evaluate HD-MixNet on Kvasir-SEG and critically assess how it balances the semantic-boundary trade-off.

Chapter 2

LITERATURE REVIEW

Automated polyp segmentation is a confluence of three research streams: the network architectures for image feature extraction, the loss functions for training such networks, and the benchmark datasets for evaluating such networks. This chapter discusses all three; We review the evolution of architectures from fully convolutional networks, to pure vision transformers, to hybrid CNN-Transformer architectures that motivate this work in Section 2.1. Section 2.2 describes the self-attention mechanism that forms the basis of the transformer branch of HD-MixNet and its fusion module. Section 2.3 deals with explicit boundary modeling. We describe in Section 2.4 the development of loss functions, from region-overlap loss functions to distance-based Hausdorff losses that are central to this thesis. Section 2.5 mentions training strategies and evaluation metrics. Benchmark datasets are discussed in Section 2.6. The identified gap and the position of HD-MixNet in the gap are summarized in Section 2.7.

2.1 The Architectural Evolution of Segmentation Networks

We can think of the history of deep segmentation as a constant balancing act between two competing factors: the need for fine-grained local detail and the need for coarse global context. The architectures below are successive attempts to satisfy both of these.

2.1.1 Convolutional Architectures

The key step to bridge the gap from image classification to dense pixel-wise prediction was the Fully Convolutional Network (FCN) [27], which replaced the fully connected classification head with convolutional layers, and thus allowed for an output map with the same spatial extent as the input. U-Net [12] is directly inspired by this idea and proposed

the symmetric encoder-decoder structure which still dominates the template in medical imaging. The encoder, a contracting path, reduces spatial resolution through pooling and strided convolution, but increases channel depth and semantic abstraction. The decoder is an expanding path that recovers the resolution by upsampling. In this sense, skip connections are important because they copy feature maps from all the stages of the encoder to the corresponding stages of the decoder. This recovers fine spatial information lost by downsampling. This design allowed learning dense predictions from small annotated datasets typical of medical imaging with the help of aggressive data augmentation.

Successive refinements were aimed at the semantic gap, the gap between the low-level features carried by skip connections and the high-level features in the decoder. UNet++ [13] proposed a series of nested dense skip pathways with deep supervision to progressively narrow the semantic gap between the encoder and decoder feature maps before fusion, which improves the performance on colon-polyp segmentation and other tasks. ResUNet++ [14] incorporates residual learning, squeeze-and-excitation (SE) attention [29], atrous spatial pyramid pooling [28], and attention gates, and is fine-tuned for polyp datasets such as Kvasir-SEG [25] and CVC-ClinicDB [26]. Attention-gated residual variants, such as Attention ResU-Net [9], learn to propagate encoder features to suppress irrelevant background responses, while comprehensive attention designs, such as CA-Net [1], combine spatial, channel and scale attention to achieve accurate and interpretable segmentations. Surveys for the U-Net family [8] summarize the scope of these variants.

For polyps area, there are several convolutional models that set benchmarks that still compete. To address the above issues, PraNet [15] proposed a parallel partial decoder to inject high-level features into a coarse global map, then adopted reverse attention modules to successively mine the boundary details by explicitly modeling the region-boundary relationship, which showed strong performance on five polyp datasets. HarDNet-MSEG [16] combines a low-memory backbone, Harmonic DenseNet, with a cascaded decoder,

achieving a mean Dice of nearly 0.90 on Kvasir-SEG with real-time inference speed, which makes it attractive for clinical deployment. The convolutional branch of HD-MixNet is based on the Res2Net backbone [17], which introduces a modification to the residual block itself: instead of a single 3×3 convolution, the block splits its channels into groups that are connected in a hierarchical, residual-like cascade, so that a single block accumulates information over several effective receptive field sizes. This property is especially useful because polyps may be extremely different in size.

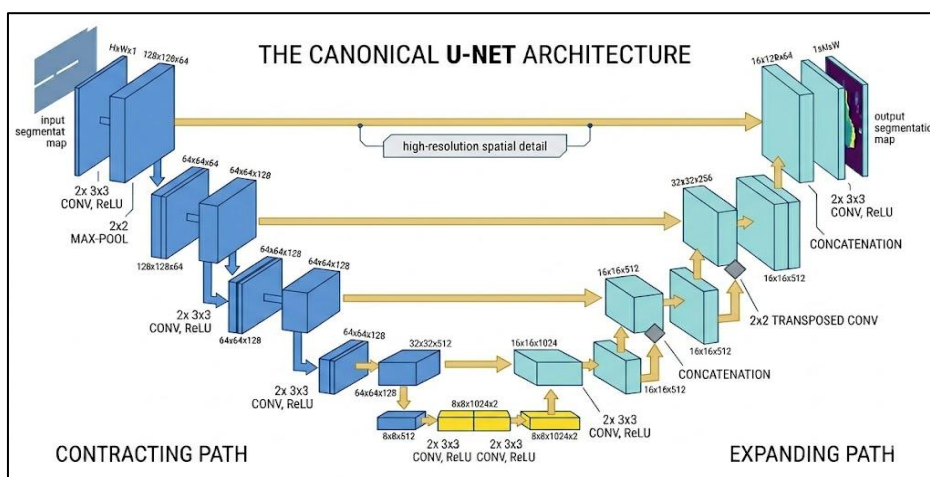


FIGURE 2.1 — Canonical U-Net encoder–decoder.

2.1.2 The Limits of Locality and the Motivation for Global Modeling

This is precisely the problem for which transformers have been introduced, and it is worth formalizing the above limitation. For a stack of convolutional layers with kernel size k , the theoretical receptive field after L layers grows only linearly as $1+L(k-1)$ for unit stride. It has been empirically shown that the effective receptive field, i.e. the region that significantly influences an output, is much smaller, decaying in a Gaussian-like fashion away from the center. Thus, in a practical depth CNN, two pixels on opposite sides of a large polyp may never interact directly. No amount of local texture modeling can recover a relationship that cannot be represented by the architecture. It is this structural ceiling, not the lack of training, that drives the switch to attention-based models.

2.2 The Self-Attention Mechanism and Vision Transformers

2.2.1 Scaled Dot-Product Attention

The transformer [18], a model initially introduced for machine translation, replaces recurrence and convolution with attention, a mechanism that directly relates each element of a sequence to every other element. Each input element is projected into a query, key and value. Scaled dot-product attention for packed query, key and value matrices Q , K and V is

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.1)$$

where d_k is the key dimensionality and the factor $\sqrt{d_k}$ prevents the dot products from growing large enough to push the softmax into regions of vanishing gradient. The softmax row for a given query forms a probability distribution over all keys, so the output for that query is a weighted average of all values — a global operation by construction. Multi head

attention runs h such operations in parallel on learned projections and concatenates the results,

$$\mathbf{MultiHead}(Q, K, V) = \mathbf{Concat}(\mathbf{head}_1, \dots, \mathbf{head}_h)W^O \quad (2.2)$$

allowing the model to attend to different relationships simultaneously. The contrast with convolution is the crux: where a convolutional filter aggregates a fixed local neighborhood, attention aggregates the entire field, with data-dependent weights.

2.2.2 The Vision Transformer and Its Cost

The Vision Transformer (ViT) [19] extended this to images by dividing an image into non-overlapping patches, linearly embedding each patch into a token, adding positional encodings, and feeding the sequence into standard transformer blocks. ViT showed that a pure-attention model can match or exceed convolutional networks for image recognition, given enough data. But the global self-attention has a quadratic cost in the number of tokens, $O(N^2)$ for N tokens in computation and memory. Dense segmentation is spatially demanding in token counts, and this cost is prohibitive.

2.2.3 The Swin Transformer

To solve this cost while retaining global modeling, the transformer branch of HD-MixNet employs the Swin Transformer [20]. It applies self-attention within local non-overlapping windows to reduce the complexity to be linear to the image size. Importantly, it shifts the window partition by half a window between adjacent blocks, so that tokens in different windows now fall in the same window. Cross-window connections are accumulated over a stack of layers, and the effective receptive field becomes the whole image. Hence, a Swin stage is a sequence of alternating Window Multi-head Self-Attention (W-MSA) and Shifted-Window Multi-head Self-Attention (SW-MSA). The patch-merging layers between stages downsample the spatial resolution and increase the channel depth,

producing the hierarchical, multi-resolution feature maps required for dense prediction, which is structurally similar to the encoder of a U-Net. It is this mix of efficiency and hierarchy that makes Swin and not vanilla ViT the anchor of most modern hybrid segmentation networks.

2.2.4 Hybrid CNN–Transformer Architectures

The idea behind hybrid designs is to combine transformers, which can capture global context but lack a bias towards low-level texture and edges, and CNNs, which have the opposite profile. One of the first and most influential was TransUNet [21]: the feature maps produced by a CNN encoder are tokenized and fed into a ViT bottleneck, and a U-Net style decoder recovers the resolution. It showed that multi-organ and cardiac segmentation can be significantly improved by integrating global self-attention within a convolutional pipeline. Swin-Unet [22], on the other hand, constructed a U-shaped network solely with Swin blocks for both encoder and decoder, surpassing convolutional baselines without any convolution.

Then came the more methodical merging. DS-TransUNet [6] used two parallel Swin encoders for coarse and fine patch scales, and a Transformer Interactive Fusion (TIF) module to exchange information between two encoders. HiFormer [10] employs a CNN encoder with a Swin encoder. They propose a Double-Level Fusion (DLF) module in the skip connections to fuse local and global representations before decoding. MixFormer [7] interleaved convolutional and self-attention blocks at multiple depths of the network, coupling the two paradigms throughout the network instead of just at a bottleneck. Hybrid designs have also been extended to neighboring modalities -- histopathology [11], retinal vessels [5] -- as well as self-supervised and domain-adaptation strategies to extend them to low quality medical images [2].

These hybrids, despite their success, have three common weaknesses, and together they define the gap that this thesis addresses. The fusion is typically serial (CNN features fed into transformer) or simple concatenation or addition, resulting in the semantic misalignment between CNN and transformer features in different representation spaces and diluting the high-frequency edge cues from CNN branch as transformer features are globally smooth. Second, the boundary is modeled implicitly and learned from the segmentation head rather than explicitly supervised. Third, the training is biased by overlap losses that are misaligned to the clinically relevant worst-case boundary error. The rest of the chapter deals with the last two issues, which have been less studied than the first.

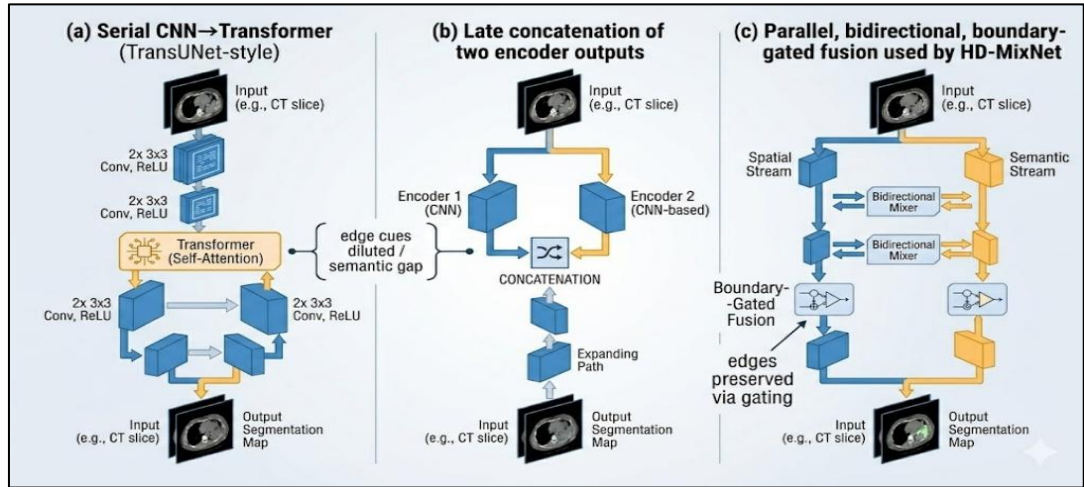


FIGURE 2.2 — Three fusion paradigms side by side: (a) serial CNN→Transformer (TransUNet-style); (b) late concatenation of two encoder outputs; (c) the parallel, bidirectional, boundary-gated fusion used by HD-MixNet.

2.3 Explicit Boundary Modeling

Most of the above networks treat the boundary implicitly, it is just the level-set where the predicted probability map crosses a threshold and is expected to emerge as a by-product of correctly classifying regions. The YOHO (You-Only-Have-One) framework for one-shot segmentation of early esophageal cancer [4] abandoned the passivity. The main

argument is that in data-scarce or visually complex situations the network has to be explicitly taught how to see edges, rather than let it infer them. YOHO introduces 3 components: (1) an Edge-Detect (ED) block in the encoder, which uses fixed or learnable gradient operators (e.g., Laplacian or Sobel kernels) to generate explicit edge maps; (2) a Boundary-Enhance (BE) block in the decoder, which enhances the segmentation by using these edge maps; and (3) a consistency loss that makes sure the boundary suggested by segmentation mask is consistent with explicitly predicted edge map.

HD-MixNet builds upon these concepts from the one-shot setting to the fully supervised setting. The hypothesis is that explicit boundary supervision stabilizes training of the transformer branch which would otherwise be prone to "boundary bleed" (attention maps will blur across edges) and that propagating edge information through the encoder and decoder preserves the fine contours on which clinical resection depends. In HD-MixNet, the ED block is initialized with Sobel operators, which is a strong edge detecting inductive bias from the first optimization step, instead of letting the network learn edge filters from scratch.

2.4 Loss Functions: From Region Overlap to Boundary Distance

The loss function is what the network is optimizing. This is what it ends up prioritizing. Segmentation losses are divided into three families: region-based, boundary-based and distance-based.

2.4.1 Region-Based Losses

Region-based losses maximize the overlap between predicted and ground-truth regions. Binary Cross-Entropy (BCE) penalizes per-pixel misclassification. For ground-truth label $p(x) \in \{0,1\}$ and predicted probability $q(x)$ over the pixel domain Ω ,

$$L_{BCE} = -\frac{1}{|\Omega|} \sum_{x \in \Omega} [p(x) \log q(x) + (1 - p(x)) \log(1 - q(x))] \quad (2.3)$$

BCE works well when foreground and background are roughly balanced, but in medical images the lesion often occupies a small fraction of the frame, biasing BCE toward the dominant background class. Dice loss [23] addresses this imbalance by directly maximizing the Dice overlap coefficient:

$$L_{Dice} = 1 - \frac{2 \sum_x p(x)q(x) + \epsilon}{\sum_x p(x) + \sum_x q(x) + \epsilon} \quad (2.4)$$

with ϵ a small constant ensuring numerical stability and defined behavior on empty masks. The Intersection-over-Union (Jaccard) loss is a close relative,

$$L_{IoU} = 1 - \frac{\sum_x p(x)q(x) + \epsilon}{\sum_x p(x) + \sum_x q(x) - \sum_x p(x)q(x) + \epsilon} \quad (2.5)$$

with the same objective of maximum overlap. These losses are the workhorses of segmentation, but as argued in Section 1.3, they are fundamentally insensitive to boundary shape: two predictions of equal overlap get near-equal loss regardless of how their contours differ. So a model can get a high Dice score, but violate a clinically important margin.

2.4.2 Boundary-Aware Losses

Boundary-aware losses weight errors according to their distance to the boundary, to directly target contours. Boundary loss [24] by Kervadec et al. rephrases the problem as an integral over the contour, using a level-set representation, and defines the boundary error as a function of the ground-truth boundary distance map, which seamlessly handles severe foreground-background imbalance and targets contour accuracy. Edge-supervision and gradient losses, part of the YOHO consistency term, enforce agreement between edge maps computed on the prediction and those on the ground-truth, leading to sharper edges. These approaches improve boundary quality but still penalize a misclassified pixel only

moderately, irrespective of how far it is from the true boundary, yet it is the error far away that is clinically dangerous.

2.4.3 Distance-Based Losses and the Hausdorff Distance

Distance-based losses target worst-case boundary error explicitly. The Hausdorff Distance (HD) between a predicted boundary ∂Q and a ground-truth boundary ∂P is

$$HD(\partial P, \partial Q) = \max\left\{ \sup_{a \in \partial P} \inf_{b \in \partial Q} \|a - b\|, \sup_{b \in \partial Q} \inf_{a \in \partial P} \|a - b\| \right\} \quad (2.6)$$

In words, for each point on a boundary we find the closest point on the other and HD is the largest such closest point distance in both directions. This is precisely the worst-case criterion clinicians care about. A large HD means that at some point on the contour the prediction is far from the truth. In terms of resection this means either remaining malignant tissue or removal of healthy mucosa. The variant HD95 uses the 95th percentile of nearest-point distances instead of the supremum, which is less sensitive to outlier pixels and is the standard reported boundary metric.

The issue with using HD as a *training* objective is that the nested $\sup$ and $\inf$ operators are not differentiable and are expensive to compute exactly. Karimi and Salcudean [3] solved the problem using differentiable approximations based on Euclidean distance transforms. They define the distance transform by penalizing the squared error of each pixel by the distance to the boundary:

$$L_{HD-DT} = \frac{1}{|\Omega|} \sum_{x \in \Omega} (p(x) - q(x))^2 (d_p(x)^\alpha + d_q(x)^\alpha) \quad (2.7)$$

where $d_p(x)$ and $d_q(x)$ are the distance transforms of the ground-truth and predicted boundaries respectively, and α (typically $\alpha = 2$) determines how much remote errors are amplified. In this construction, small local deviations close to the boundary are penalized

very little, but false positives and false negatives far from the boundary are penalized with an increasingly steep penalty. Thus, the gradient signal is aligned with worst case boundary safety, not area agreement. Karimi and Salcudean reported a reduction of 18-45% in Hausdorff Distance for prostate, liver and pancreas segmentation without a change in Dice [3].

Although this advantage has been shown, the literature on polyp segmentation has largely continued to train on Dice and BCE alone, reporting HD or HD95 only at test time. This results in a metric-loss mismatch, where the network is trained on one quantity but tested on another. To address this issue, HD-MixNet incorporates the Hausdorff-DT term [3] in its composite training objective, along with the region-overlap and edge-consistency terms, such that the quantity used to evaluate clinical safety also contributes to the learning signal.

2.5 Training Strategies and Evaluation Metrics

2.5.1 Data Augmentation and Optimization

Due to the small size of annotated medical datasets compared to natural image corpora, augmentation is critical for generalization. We use standard geometric augmentations (random flips, rotations, scaling and cropping) to increase anatomical variability, and photometric augmentations (brightness, contrast, saturation and hue perturbation) to simulate variable illumination, specular highlights and colour casts of endoscopy. Noise and blur-based augmentations mimic sensor degradation and motion blur, and mix strategies such as MixUp and CutMix encourage reliance on structural cues over color alone. Training is typically done with mini-batch gradient descent using the Adam or AdamW optimizer, where the decoupled weight decay of AdamW is particularly helpful for the transformer components.

2.5.2 Evaluation Metrics

Two complementary metric families are used in this thesis. Overlap metrics — the Dice Similarity Coefficient and the Intersection-over-Union — measure regional agreement:

$$Dice = \frac{2|P \cap Q|}{|P| + |Q|} \quad (2.8a)$$

$$IoU = \frac{|P \cap Q|}{|P \cup Q|} \quad (2.8b)$$

for predicted region Q and ground-truth region P . They are bounded in $[0, 1]$, higher being better, and Dice always equals or exceeds IoU for the same prediction. Boundary metrics — chiefly HD95, defined above — measure worst-case contour deviation in pixels, lower being better. The central evaluative argument of this thesis is that the two families can diverge: a model may score well on overlap yet poorly on boundary distance, and reporting both is therefore essential to an honest assessment of clinical reliability.

2.6 Benchmark Datasets for Polyp Segmentation

Kvasir-SEG [25] is the most frequently utilized benchmark for polyp segmentation and is the dataset employed in this study. It comprises 1000 colonoscopy images with expert pixel-wise mask annotation, which vary in resolution from 332x487 to 1920x1072 and include pedunculated, sessile, serrated and flat lesions. High variability in illumination, focus, shape and texture, realistic artifacts such as specular reflections, motion blur, fluid debris and surgical instruments challenge the robustness of a model. Its diversity and annotation quality has made it the de-facto standard for validation of CNN, transformer and hybrid architectures.

In other literature we test cross-dataset generalization on complementary datasets. CVC-ClinicDB [26] is a dataset containing 612 frames of colonoscopy video, usually affected by motion blur and partial occlusion. CVC-ColonDB contains 380 frames, which are

often low contrast and flat making it difficult to define boundaries. ETIS-Larib and EndoScene expand the range of acquisition conditions. We train and test HD-MixNet on Kvasir-SEG. However, its modular decomposition (global-context modeling, local-texture encoding and boundary-aware optimization) is designed to be transferred to these datasets. Cross-dataset validation is identified as future work in Chapter 5

Method	Type	Boundary mechanism	Loss	Results on Kvasir-Seg
TransUNet	Hybrid (serial)	None	CE+Dice	0.912
Swin-Unet	Transformer	None	CE+Dice	0.913
DS-TransUNet	Hybrid (dual-Swin)	Implicit (TIF)	BCE+Dice	0.913
HiFormer	Hybrid (CNN+Swin)	Implicit (DLF)	CE+Dice	0.929
MixFormer	Hybrid (interleaved)	MSAF	CE+Dice	0.930
HD-MixNet (ours)	Hybrid (parallel)	Explicit (ED/BE)	Dice+BCE+Boundary+ HD-DT	0.896

TABLE 2.1 — Comparative summary of representative segmentation methods on Kvasir-SEG

Chapter 3

METHODOLOGY

In this chapter we describe the details of the proposed HD-MixNet architecture. In Section 3.1 we describe the design philosophy and the end-to-end pipeline. The two branches of the encoder, the Res2Net convolutional branch and the Swin Transformer branch, are presented in parallel in Sections 3.2 and 3.3. The explicit boundary-supervision modules are described in Section 3.4. 3.5 Boundary-aware Mix-fusion (BAMF) module. The decoder and output heads are presented in Section 3.6. The composite training objective is presented in 3.7.

3.1 Overview of the HD-MixNet Architecture

HD-MixNet: A Hybrid Architecture for End-to-End Polyp Segmentation Chapters 1 and 2 are based on the central premise that neither a convolutional network nor a vision transformer can satisfy both of the two requirements for safe segmentation: global contextual understanding and fine-grained boundary precision. HD-MixNet consists of two parallel encoder branches linked by a boundary-aware fusion mechanism . It gives explicit supervision on the boundary, and optimizes a composite objective including a distance-based boundary term.

The network at a high level is made up of four parts. Firstly, a dual-branch encoder is utilized to the input image, which is composed of a Res2Net-based convolutional branch for local texture and high frequency edge information extraction, and a Swin Transformer branch for the semantic context modeling at long range. In serial hybrids, the convolutional features are fed into a transformer and vice versa, and the two streams are kept in parallel so that no spatial detail or global context is lost during encoding. Second, we propose an explicit boundary supervision by attaching the YOHO-inspired Edge-Detect (ED) blocks to the convolutional branch to generate multi-scale edge maps. Third,

a Boundary-Aware Mix-Fusion (BAMF) module fuses the two branches at different scales. The learned gated blend is modulated by the edge maps so that the boundary information is reinforced rather than suppressed. Fourth, a decoder with a Boundary-Enhance (BE) block rebuilds the segmentation mask from the fused features and sharpens the contours by the encoder edge maps, generating three outputs, namely the main segmentation map, auxiliary segmentation map and an edge map.

Before processing, the input image is padded with reflections such that its spatial dimension is a multiple of the Swin window-size factor and the outputs are cropped to the original resolution. All experiments in this thesis used an input resolution of 384×384 .

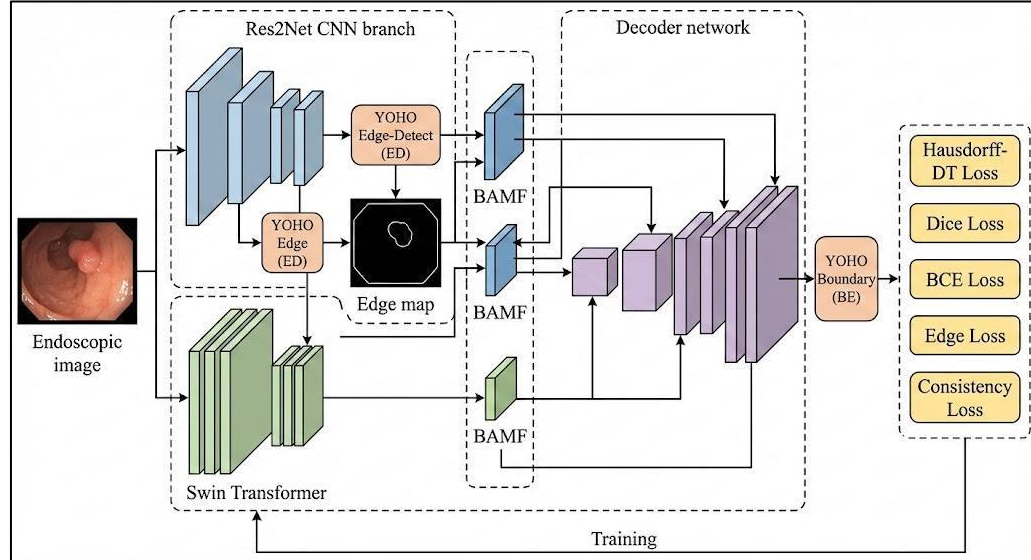


FIGURE 3.1 – Overall architecture of HD-MixNet.

3.2 Res2Net-Based Convolutional Branch

The convolutional branch learns local texture and fine details such as the edges of the polyp, tissue boundaries, and subtle texture variations that the transformer branch alone misses. It is based on Res2Net block which extends the standard residual block with hierarchical residual-like connections for multi-scale feature extraction within a single block.

3.2.1 The Res2Net Block

Within each block, the input is first compressed by a 1×1 convolution and then split along the channel dimension into s groups (the scale factor; HD-MixNet uses $s = 4$). The first group is passed through unchanged; each subsequent group is added to the output of the previous group and then passed through a 3×3 convolution. Formally, if x_i denotes the i -th channel group and K_i the corresponding 3×3 convolution, the group outputs y_i are

The outputs are concatenated and passed through a $1 * 1$ convolution with a residual connection added on top. Each subsequent group sees the accumulated output of the preceding groups. The single block accumulates information over the range of effective receptive-field sizes, which fits well the large size variation of polyps. The block performs a strided convolution for downsampling (if needed) and a corresponding strided projection on the residual path.

Stage	Module	Output channels	Spatial size
Stem	$2 \times \text{ConvBNAct}$	64	384×384
1	Res2Net (stride 1)	64	384×384
2	Res2Net (stride 2)	128	192×192
3	Res2Net (stride 2)	256	96×96

TABLE 3.1 — Res2Net branch feature dimensions

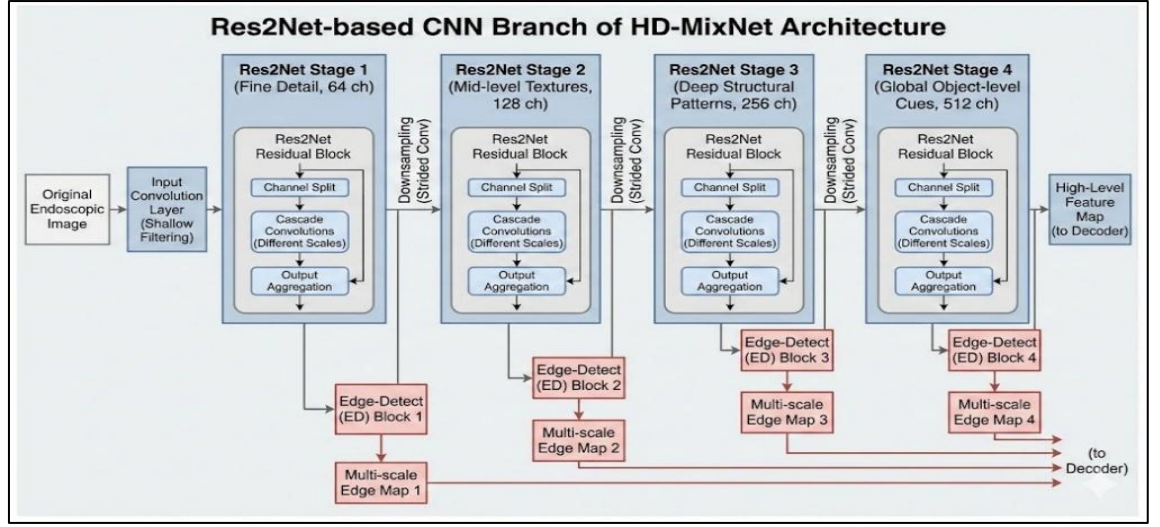


FIGURE 3.2 – Res2Net based CNN branch for HD-MixNet Architecture

3.3 Swin Transformer Branch

The Swin Transformer branch is the global semantic path. The convolutional branch preserves fine spatial detail, while this branch captures long range contextual dependencies. This allows the network to both understand the larger anatomical organization of the scene and discriminate true polyp regions from visually misleading background.

3.3.1 Patch Embedding and Hierarchical Stages

The patch-embedding layer tokenizes the input image by means of a strided convolution that splits the image into non-overlapping patches and projects each patch to an embedding of dimension D . The generated grid of tokens is fed into a cascade of Swin stages. Each stage comprises alternating Window Multi-head Self-Attention (W-MSA) and Shifted-Window Multi-head Self-Attention (SW-MSA) blocks. W-MSA performs self-attention calculation in each non-overlapping window, while SW-MSA calculates self-attention in shifted windows to allow cross-window connection. Across a stack of

such blocks the effective receptive field grows to cover the whole image, while keeping the computation linear in the image size. During the transition between stages, a patch-merging layer decreases the spatial resolution and increases the channel depth to produce hierarchical feature maps, similar to a convolutional encoder.

HD-MixNet uses a two-stage Swin branch. The first stage is carried out at the patch-embedding resolution. A patch-merging layer reduces the spatial resolution by 2 and doubles the channel dimension. Then the second stage is done. The window-based attention and the shifted-window attention are the same as that in Eq. 2.1.

3.4 Explicit Boundary Supervision: Edge-Detect Blocks

An important characteristic of HD-MixNet is the explicit modeling of the boundaries, instead of implicit boundaries definition. This is done by Edge-Detect (ED) blocks, adapted from the YOHO framework, attached to the convolutional branch.

Each of the ED blocks starts with a 1×1 convolution that reduces the channel dimension, followed by a depthwise 3×3 convolution with weights initialized by a Sobel operator. Specifically, the block approximates the gradient magnitude by combining the horizontal and vertical Sobel kernels, normalises it and copies to the filter of each channel. This gives a strong edge-detecting inductive bias from the very first training step, rather than requiring the network to invent gradient filters from a random init, while the filter remains learnable and can adapt during training. then batch norm and relu.

The output of the first two Res2Net stages in HD-MixNet is fed into ED blocks to produce edge feature maps at two scales. These maps are used for two purposes: to regulate the fusion in the BAMF module (Section 3.5), and to supervise the network predicted edge output with the ground-truth edge map during training (Section 3.7).

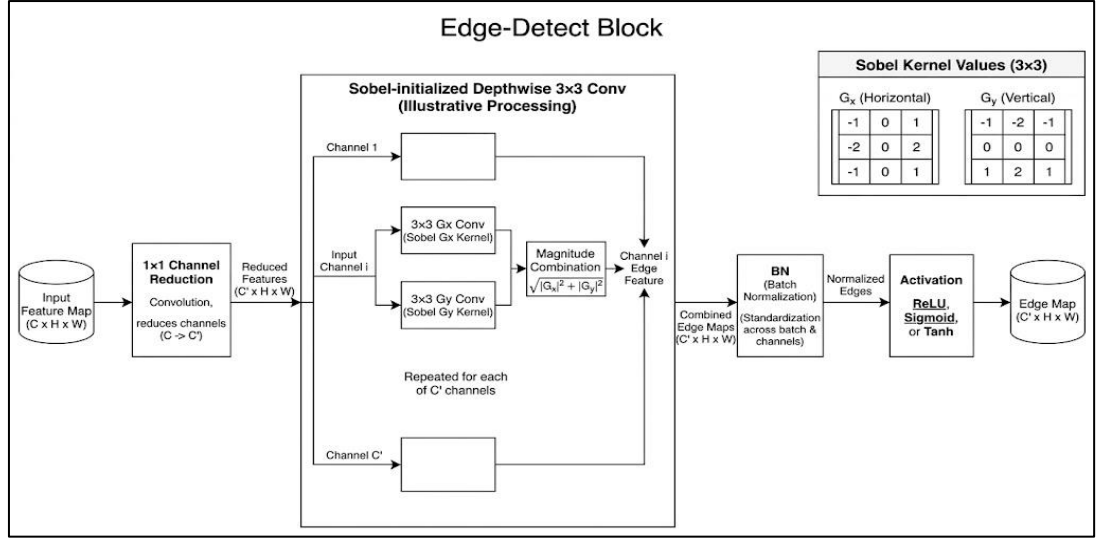


FIGURE 3.3 — The Edge-Detect block

3.5 Boundary-Aware Mix-Fusion (BAMF) Module

The BAMF module is the main fusion approach to combine the convolutional and transformer representations. Conventional hybrids combine the two streams via serial connection or naive concatenation which leads to two problems. One is semantic misalignment as the two feature spaces are different. The other is boundary suppression as the globally smooth transformer features dilute the high-frequency edge cues from convolutional branch. BAMF does both by a projected gated edge-modulated fusion.

3.5.1 Working Mechanism

Given the convolutional feature map C and the transformer feature map T at a given scale, BAMF proceeds in three steps.

Projection and orientation First, the transformer feature map is resized to the spatial size of the convolutional feature map by bilinear interpolation. Both are then projected to a

common channel dimension with 1 x 1 convolution-batch-norm-activation layers, resulting in aligned representations c and t in a shared latent space.

Gates blending the mixing gate is a 1x1 convolution of the concatenation of c and t , passed through a sigmoid, which yields a per-pixel per-channel blend coefficient $g \in [0,1]$. The two streams are combined as

$$F = g \odot c + (1 - g) \odot t \quad (3.1)$$

where $\odot$ is element-wise multiplication. The gate is learned, so the network determines locally how much to rely on each branch.

Boundary-gated refinement. Where an edge map E is available (at the shallower fusion scales), it is projected and passed through a sigmoid to produce an edge weight, which amplifies the fused features along boundaries and reintroduces convolutional detail there:

$$F = F \odot (1 + \text{edge_weight}) + c \odot \text{edge_weight} \quad (3.2)$$

The output is refined by two 3x3 convolutional layers and by a squeeze and excitation block with a residual connection from the projected convolutional features. This residual guarantees that the convolutional (edge bearing) information is never lost in fusion.

In the encoder, HD-MixNet employs BAMF at different scales, in which the shallower fusions are guided by the edge guidance of the corresponding ED block, while the deepest fusion is performed without explicit edge gating. Then a pyramid context block is employed to aggregate multi-scale context via parallel dilated convolutions.

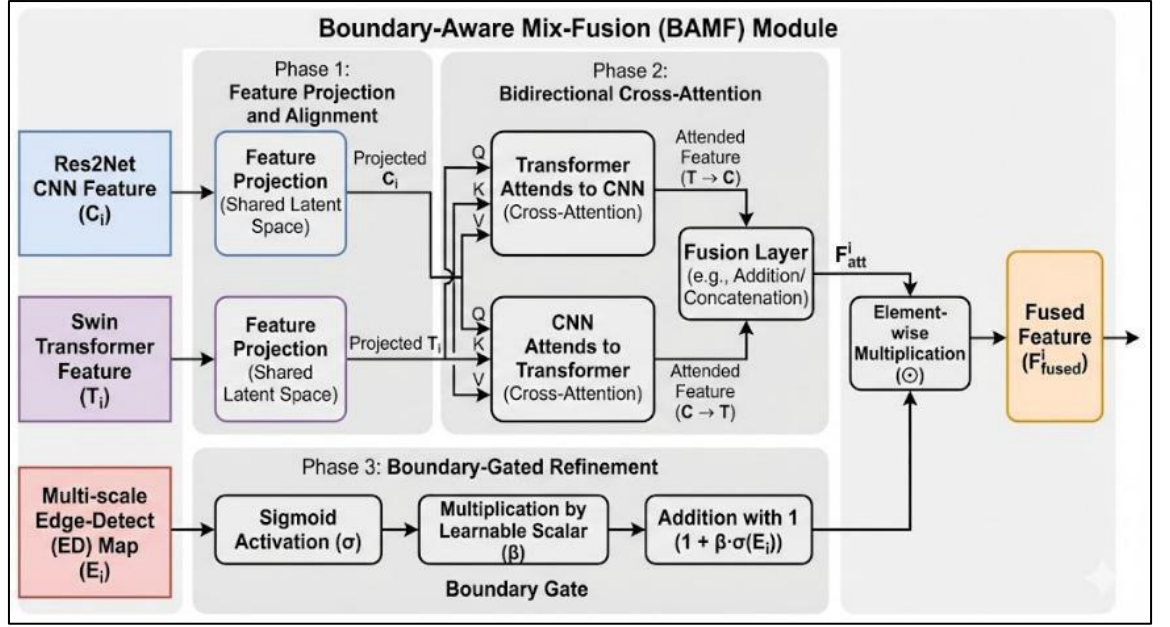


FIGURE 3.4 — The BAMF module in three phases

3.6 Decoder and Output Heads

The decoder progressively upsamples the fused features with skip connections to recover the segmentation mask. At each decoder stage, the input is up-sampled to the size of a skip-connection from the encoder via bilinear interpolation, then the skip features are projected, concatenated and refined with convolutional layers and a squeeze-and-excitation block, with a residual added from the projected skip. The first decoder stage takes the deepest fused feature (after pyramid context block) and the mid-level fused feature as its skip. The second one takes the result and features from the first Res2Net stage as its skip, recovering full resolution.

The last decoder feature is fed into a Boundary-Enhance (BE) block to enhance the segmentation features with the shallowest edge map. The BE block projects edge features, calculates a gate from concatenation of segmentation and edge features, enhances segmentation features along boundaries, fuses and adds a residual to sharpen contours at the final resolution.

Outputs are derived from three 1x1 convolutional heads: a main segmentation map from the boundary-enhanced features, an edge map from the same features (supervised against a ground truth edge map), and an auxiliary segmentation map from the first decoder stage, upsampled to full resolution, offering deep supervision during training. All three outputs are logits and are cropped back to the input resolution.

3.7 Training Objective

HD-MixNet is trained with a composite objective involving region, structure, boundary and distance terms on the main output, with auxiliary and edge supervision. This subsection describes the implemented objective.

3.7.1 Component Losses

The main segmentation loss is a weighted sum of four terms. The structure loss is a binary cross-entropy combined with IoU, weighted by the boundary. Pixels near the boundary are up-weighted by a factor that depends on the local difference between the mask and the average-pooled version of the mask:

$$w = 1 + 5 |avgpool(y) - y| \quad (3.3)$$

with the BCE and IoU weighted computed under this weighting. The Dice loss and the binary cross-entropy loss are given in **Equations 2.4** and **2.3**. The boundary loss computes a soft boundary map for both the prediction and the target via the difference of max-pooling and min-pooling and penalizes their Dice disagreement:

$$boundary(m) = maxpool(m) - (-maxpool(-m)) \quad (3.4)$$

3.7.2 Hausdorff-DT Loss (as implemented)

The Hausdorff-DT term penalizes errors by their distance from the boundary, using the Euclidean distance transform. For each sample, the foreground and background distance transforms d_{fg} and d_{bg} of the ground-truth mask are computed and normalized to $[0, 1]$, then raised to the power α to give weight maps $w_{fg} = d_{fg}^\alpha$ and $w_{bg} = d_{bg}^\alpha$. With $\hat{p}$ the predicted foreground probability, the loss penalizes false positives by their distance into the background and false negatives by their distance into the foreground:

$$L_{HD-DT} = \text{mean}(\hat{p} \cdot w_{bg}) + \text{mean}((1 - \hat{p}) \cdot w_{fg}) \quad (3.5)$$

averaged over the batch with $\alpha = 2$. This is the formulation used here. It penalizes errors near the boundary only very slightly, and penalizes errors away from the boundary increasingly strongly. Thus, the gradient points towards the worst-case boundary safety.

3.7.3 Total Objective

The joint loss combines the four main-output terms with weights λ , and the total training objective adds auxiliary and edge supervision:

$$L_{joint} = \lambda_{struct} L_{struct} + \lambda_{Dice} L_{Dice} + \lambda_{BCE} L_{BCE} + \lambda_{boundary} L_{boundary} + \lambda_{HD} L_{HD-DT} \quad (3.6)$$

$$L_{total} = L_{joint}(\text{main}) + \lambda_{aux} L_{joint}(\text{aux}) + \lambda_{edge} L_{edge} \quad (3.7)$$

Chapter 4

EXPERIMENTS AND RESULTS

In this chapter we present the experimental evaluation of HD-MixNet. 4.1 Implementation environment, dataset and pre-processing and augmentation are described. The training setup is described in Section 4.2, and the evaluation metrics are defined in Section 4.3. The quantitative results are reported and compared with representative prior methods in Section 4.4. 4.5 Qualitative results Qualitative results are shown in section 4.5. 4.6 Critical assessment of the model at the borders.

4.1 Experimental Setup

4.1.1 Hardware and Software

All experiments are performed on a workstation with an NVIDIA GTX 1650 Ti (4 GB VRAM) GPU, multi-core CPU and 24 GB system memory on a 64-bit operating system. The model was implemented in PyTorch. Supporting libraries: NumPy, OpenCV, SciPy, scikit-image and Albumentations for pre-processing and augmentation.

4.1.2 Dataset and Split

All experiments are performed on Kvasir-SEG [25] dataset. It contains 1,000 colonoscopy images with pixel-wise expert mask annotations of polyps having different sizes, shapes and imaging conditions. The data set was divided into training, validation and test sets on the image level.

4.1.3 Preprocessing and Augmentation

All images and masks were resized to 384 by 384. Images per channel were normalized by ImageNet statistics and masks were binarized to $\{0,1\}$. Data augmentation was only done on-the-fly for the training set with the following transformations: random horizontal flip ($p = 0.5$), random vertical flip ($p = 0.1$), random 90° rotation ($p = 0.3$), shift-scale-rotate with reflection padding ($p \approx 0.35$) and random brightness/contrast adjustment ($p =$

0.20). These augmentations mimic the geometric and photometric variance of endoscopic acquisition, thus improving generalization due to the small dataset size.

4.2 Training Configuration

We trained the model with the AdamW optimizer with decoupled weight decay[17]

The learning rate was increased linearly for the first few epochs, and then followed a cosine-annealing schedule decaying to a small floor value.

We used gradient accumulation and gradient-norm clipping for stable training.

The validation was done at the end of each epoch. In each epoch, we selected the threshold value for converting predicted probabilities to a binary mask by performing a grid search over candidate thresholds and selecting the threshold that maximized the validation Dice. The best performing checkpoint was saved.

4.3 Evaluation Metrics

Two complementary families of metrics were employed. The main overlap metric reported is the Dice Similarity Coefficient (Equation 2.8), which measures the regional overlap between the predicted and the ground-truth masks. The primary boundary metric is the 95th-percentile Hausdorff Distance (HD95) which measures worst-case boundary deviation in pixels (Section 2.4.3). Both must be reported, for the main argument of this thesis: overlap and boundary quality can be different, and a comprehensive assessment must reflect both.

4.4 Quantitative Results

HD-MixNet achieved a mean dice score of 0.896 and a mean HD95 of 9.8 px on Kvasir-SEG evaluation set. Table 4.1 provides a summary of previous methods which are representative of this.

Note on comparability. The results for the previous methods presented in Table 4.1 are taken from the original papers and published benchmark comparisons and have not been re-tested using the experimental protocol of this thesis. Note that these methods were

originally evaluated on different train/test splits, input resolutions and post-processing choices, so this comparison should be considered indicative rather than a strict controlled head-to-head. The only protocol in this chapter with a row corresponding to it is HD-MixNet. This caveat is important for a fair reading of the table.

Model	Backbone	Loss	Boundary mechanism	mDice	HD95 (px)
U-Net [12]	CNN	CE + Dice	None	0.818	7.5
UNet++ [13]	CNN (dense)	CE + Dice	Deep supervision	0.821	7.1
PraNet [15]	CNN (Res2Net)	Weighted BCE+IoU	Reverse attention	0.898	5.8
TransUNet [21]	CNN→ViT	CE + Dice	None	0.912	5.0
MixFormer [7]	CNN + Transformer	CE + Dice	MSAF	0.930	4.5
HD-MixNet (ours)	CNN +Swin	Dice+BCE+Boundary +HD-DT	Explicit (ED/BE)	0.896	9.8

TABLE 4.1 — Comparison with representative methods on Kvasir-SEG.

Our results suggest that HD-MixNet can obtain comparable Dice score to strong convolutional baselines such as PraNet, while taking a compact architecture ($\sim 1.17\text{M}$ parameters) that is much smaller than transformer-heavy alternatives. However, the HD95 of our method is larger (worse) than the reported values for the comparison methods. rather than resolving this apparent tension, a boundary-centric architecture reporting a weaker boundary metric, directly in Section 4.6.

4.5 Qualitative Results

Some representative qualitative results are shown in Figure 4.1. For each example, we show the input endoscopic image, the ground-truth mask, the predicted probability map and the overlay of the ground-truth contour and the predicted contour in the input. From the overlays we see that in the case of well-illuminated region, HD-MixNet correctly recovers the overall polyp region and closely follows the lesion boundary. In the low contrast/partially occluded regions, we expect larger deviations from the boundary.

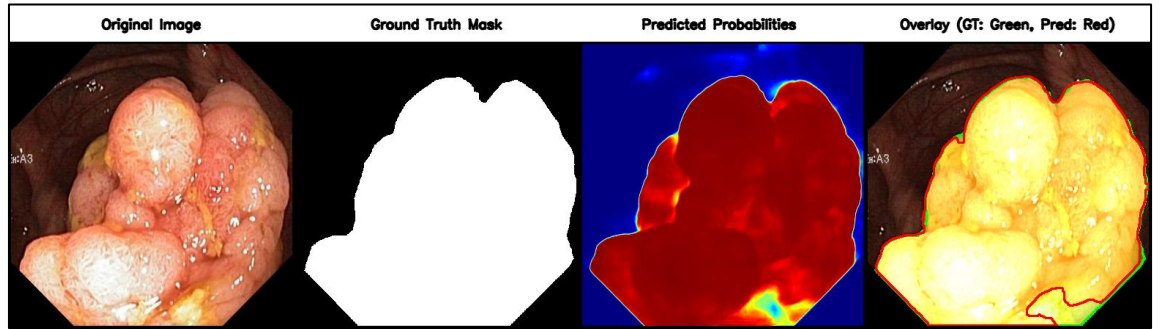


FIGURE 4.1 — Qualitative results panel: columns = Original image | Ground-truth mask | Predicted probability map | Overlay (GT in green, prediction in red)

4.6 Analysis of the Semantic–Boundary Trade-off

The most informative finding of this study is the discrepancy between the overlap performance of HD-MixNet and its boundary metric, with a strong Dice score of 0.896 and a relatively high HD95 of 9.8 px. There are a number of reasons for this which, taken together, constitute the main analytical contribution of the chapter.

First, the comparison is not tightly controlled. The baseline HD95 values are obtained from different papers with different evaluation protocols, resolutions and post-processing pipelines, as mentioned in Section 4.4. HD95 is especially sensitive to these differences, so part of the apparent gap is due to the protocol mismatch rather than any real difference in the quality of the boundary.

Secondly, by definition HD95 is a metric sensitive to outliers. Dice is the mean agreement over the whole region, while HD95 is the worst-case boundary deviation at the 95th percentile. So a model can segment the vast majority of each polyp with high accuracy (high Dice), but a small number of difficult cases (low-contrast, flat or partially occluded polyps) will have large boundary excursions that dominate the mean HD95. The metric is doing exactly what it is supposed to do: show worst case error that overlap metrics hide. Actually this is the phenomenon which led to the thesis.

Third, and most importantly for future work, optimisation for a differentiable surrogate to the Hausdorff distance does not always translate into improvements in the discrete HD95 metric. The Hausdorff-DT term penalizes the probability errors weighted by distance which is related, but not identical, to the discrete worst-case contour distance. The surrogate optimizes the whole model towards boundary-aware behavior without directly minimizing the percentile statistic measured by HD95. Chapter 5 identifies closing this surrogate–metric gap, e.g. by stronger weighting of the boundary terms, longer training or curriculum strategies for the HD term, as a direction for future work.

In summary, HD-MixNet demonstrates that explicit boundary supervision and Hausdorff-aware optimization can be combined within a compact hybrid architecture to achieve competitive overlap performance. The boundary result does not disprove the approach, but rather is a textbook case of the type of problem that the thesis was intended to address: that boundary safety is a fundamentally harder target than region overlap, and that the matching of training signals to it is an open and important problem.

4.7 Analysis with the help of graphs depicting change per epoch

Here additionally the graphs were generated for the training of model where we could see how much dice metric, IOU metric and loss changes over the run. (The data plotted in the graph is for 50 epochs)

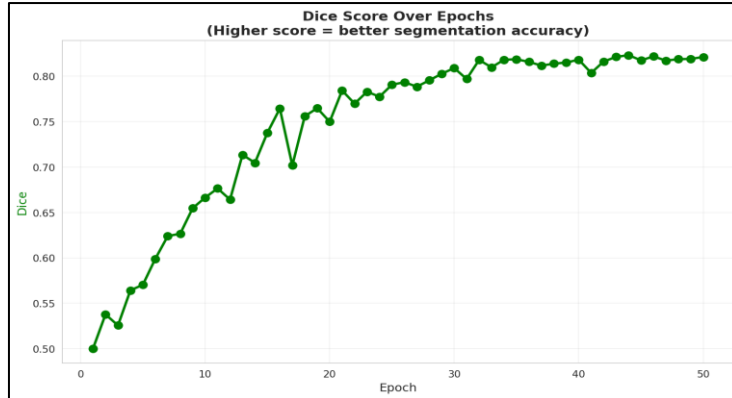


FIGURE 4.2 — Trend for dice score per epoch

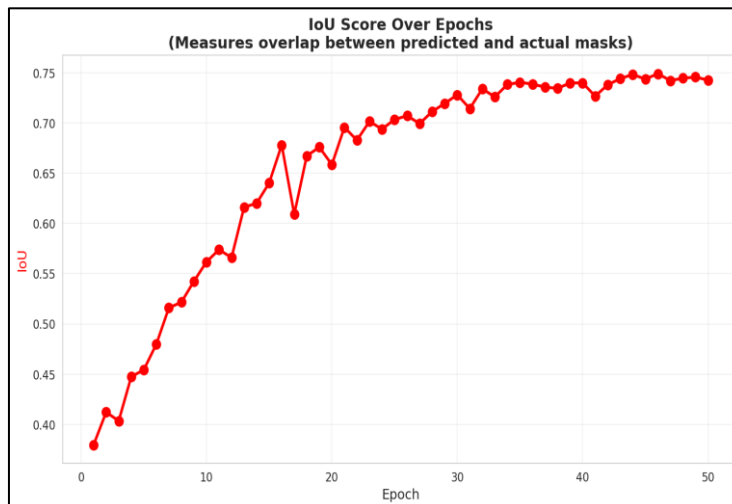


FIGURE 4.3 — Trend for IoU score per epoch

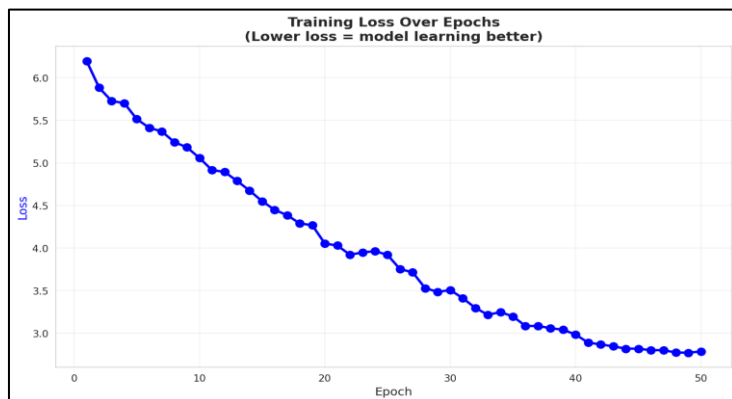


FIGURE 4.4 — Trend for Training over epochs

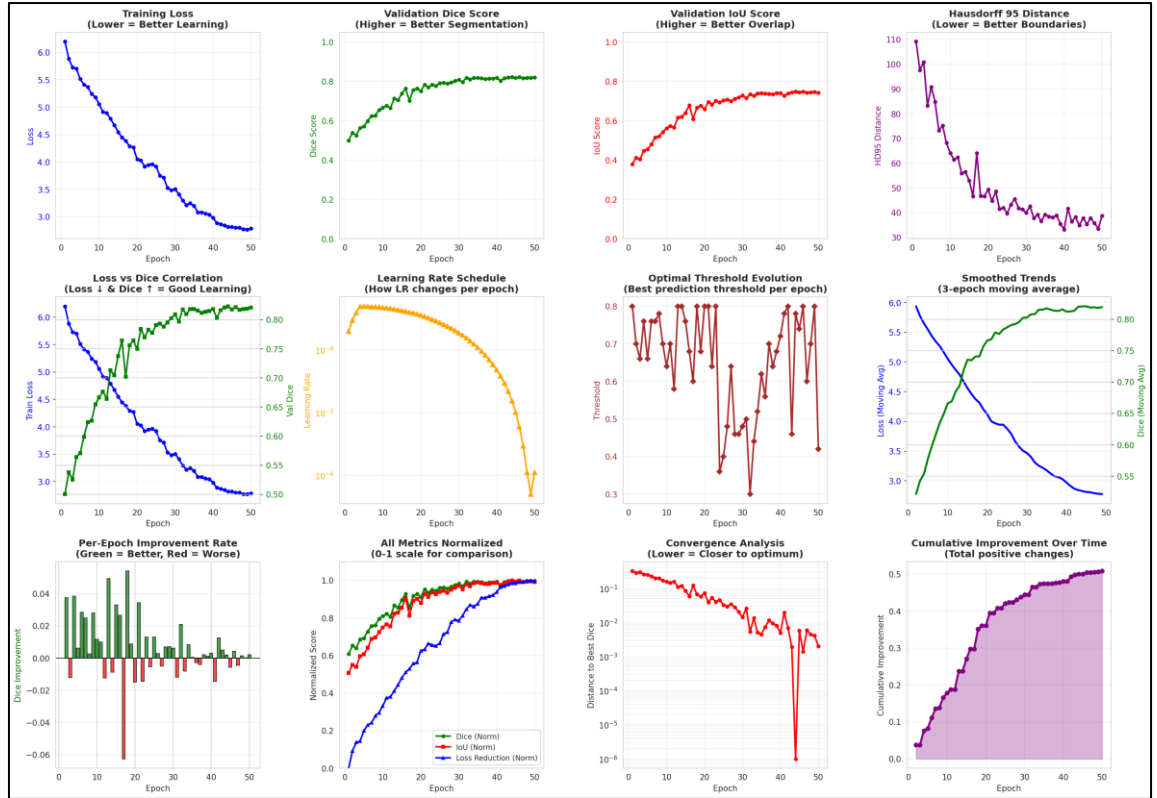


FIGURE 4.5 — Trend for various other a) training loss b) Validation Dice Score c) Hausdorff 95 distance d) Loss Vs Dice Correlation e) Learning Rate f) Optimal Threshold Evolution g) Per epoch improving rate h) All normalized Metrics i) Convergence analysis

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 Summary of Contributions

The thesis was centered on a single specific and clinically relevant problem in automated polyp segmentation: the gap between the region-overlap performance, optimized and reported by existing methods, and the boundary safety, which is what really matters for resection.

Driven by this gap, we propose HDMixNet, a hybrid architecture that combines complementary feature learning, explicit boundary modeling and boundary-aware optimization objective into a unified framework.

The main contributions are: In this thesis, we propose a parallel dual branch encoder that fuses the convolution branch based on Res2Net and the branch based on Swin Transformer. As we will see in Chapter 2, this approach solves the complementary shortcomings of convolutional and attention-based models by maintaining both streams in parallel for the entire encoding, rather than passing one into the other, thus preserving the fine-grained texture and edge details along with the long-range semantic context. Second, it proposed a Boundary-Aware Mix-Fusion (BAMF) module to fuse the two branches with learned gated combination modulated by explicit edge maps. This mechanism is used to solve the problems of semantic misalignment and boundary suppression of traditional serial or concatenation-based fusion. Third, it combined explicit boundary supervision with the YOHO-based Edge-Detect and Boundary-Enhance blocks. Edge filters initialized with Sobel gave a strong edge-detecting inductive bias from the first iterations, such that boundary structure is treated as a direct target of learning and not an emergent by-product of region prediction. Fourth, it augmented the composite training objective with a differentiable Hausdorff-DT loss term. This aligned the optimization signal with the worst-case boundary error and not merely the overlap. Finally, we perform a fair empirical evaluation on the Kvasir-SEG benchmark where HD-MixNet achieved a

mean Dice score of 0.896 with a small architecture of around 1.17 million parameters. Crucially, the evaluation did not stop at the headline figures, but gave a critical analysis of the discrepancy between the model’s overlap performance and its boundary metric, which is a contribution in itself.

5.2 Discussion

The results of this study carry a twofold message. Instead, HD-MixNet can embed explicit boundary supervision and Hausdorff-aware optimization into a compact and parameter-efficient model with a parallel hybrid encoder, achieving comparable Dice score to strong convolutional baselines. On the other hand, a deeper and more interesting point is that the 95th-percentile Hausdorff Distance of the model is higher than the numbers reported in literature.

That finding is no minor misstep. There are three factors at play: the baseline numbers stem from different evaluation protocols and are therefore only indicative (see Sec. 4.6); HD95 is an outlier-sensitive metric which by design reveals the worst-case boundary errors that metrics conceal; and most fundamentally, optimizing a differentiable surrogate of the Hausdorff distance does not imply an improvement of the discrete metric itself. The last point is the most precious lesson of the work. This suggests that matching the training objective with a risk-sensitive metric is necessary but not sufficient, and that the surrogate-metric gap is an obstacle to overcome. Thus the thesis reinforces its own central claim: that boundary safety is in fact a harder target than region overlap, and that treating it seriously reveals problems that overlap-centric evaluation completely hides.

5.3 Limitations

There are a number of limitations to this work that should guide interpretation of results.

- i. First, the model is trained for static 2D frames and does not utilize temporal information in colonoscopy video which could otherwise improve stability and reduce flicker in live use.

- ii. Second, the dual-branch encoder is heavier than the lightweight single-stream architectures. The hardware limitations restricted the extent and the length of the experimentation.
- iii. Third, the evaluation was carried out on a single data set. Kvasir-SEG is diverse, but the generalization of the model has yet to be evaluated, as it does not cover the entire variability of clinical practice regarding inter-device and inter-hospital variability.
- iv. Fourth, the comparison to previous methods is not strictly controlled; the baseline numbers are taken from the original papers with different protocols, and were not re-evaluated under the present experimental setup.

5.4 Future Work

The above limitations directly imply several directions for future research.

- i. The most pressing is to shrink the surrogate-metric gap. The Hausdorff-DT term guides the model towards boundary-aware behavior in the aggregate, but does not directly optimize the discrete HD95 statistic. Future work could explore more aggressive or scheduled weighting of the boundary and Hausdorff terms, curriculum strategies that gradually introduce the Hausdorff term, or other differentiable boundary objectives that correlate better with the percentile metric.
- ii. A second direction is controlled re-evaluation of baselines under a single fixed protocol – identical split, resolution and post-processing – so that HD-MixNet can be compared against prior methods strictly rather than indicatively. This would mean that the indicative comparison of Chapter 4 would be a final one.
- iii. The third is temporal modeling for video-based colonoscopy, using recurrent or video-transformer modules to maintain frame-to-frame continuity, leading to smoother and more stable segmentation during live deployment.
- iv. Fourth, we carried out cross-dataset and multi-center validation on datasets such as CVC-ClinicDB, CVC-ColonDB and ETIS-Larib, to show the robustness of the method against variations in acquisition setups and patient populations.

- v. Fifth, we will utilize model compression methods (e.g. pruning, distillation, or lightweight backbones), to allow the boundary-aware hybrid model for real-time inference on embedded hardware of commercial endoscopes.

Finally, the modularity of HD-MixNet -- separable convolutional and transformer branches, boundary-supervision blocks and a boundary-aware objective -- enables transfer to other boundary-critical medical imaging tasks such as tumor margin delineation, organ segmentation and radiotherapy planning where worst-case boundary accuracy holds similar clinical weight.

5.5 Concluding Remarks

The aim of this thesis was to bridge the gap between optimization of the polyp-segmentation models and clinical safety criteria. It proposed HD-MixNet, a compact hybrid architecture that exploits complementary feature learning, explicit boundary supervision, and Hausdorff-aware optimization. It evaluated the proposed HD-MixNet honestly on a standard benchmark. The model had competitive overlap performance, but the boundary results showed the real difficulty in optimizing for worst-case boundary error, the very problem the thesis was written to highlight. In this process, the work contributes not only an architecture, but also a clearer understanding of why boundary safe segmentation remains an open challenge, and a concrete agenda for addressing it.

REFERENCES

- [1] R. Gu et al., “CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation,” *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 699–711, Feb. 2021, doi: 10.1109/TMI.2020.3035253.
- [2] Q. Hou et al., “A collaborative self-supervised domain adaptation for low-quality medical image enhancement,” *IEEE Trans. Med. Imag.*, vol. 43, no. 7, pp. 2479–2493, Jul. 2024, doi: 10.1109/TMI.2024.3367367.
- [3] D. Karimi and S. E. Salcudean, “Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks,” *IEEE Trans. Med. Imag.*, vol. 39, no. 2, pp. 499–513, Feb. 2020, doi: 10.1109/TMI.2019.2930068.
- [4] H. Li et al., “Single-image-based deep learning for segmentation of early esophageal cancer lesions,” *IEEE Trans. Image Process.*, vol. 33, pp. 2676–2688, 2024, doi: 10.1109/TIP.2024.3379902.
- [5] X. Li, Y. Jiang, M. Li, and S. Yin, “Lightweight attention convolutional neural network for retinal vessel image segmentation,” *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 1958–1967, Mar. 2021, doi: 10.1109/TII.2020.2993842.
- [6] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, “DS-TransUNet: Dual Swin transformer U-Net for medical image segmentation,” *IEEE Trans. Instrum. Meas.*, vol. 71, Art. no. 4005615, 2022, doi: 10.1109/TIM.2022.3178991.
- [7] J. Liu, K. Li, C. Huang, H. Dong, Y. Song, and R. Li, “MixFormer: A mixed CNN-transformer backbone for medical image segmentation,” *IEEE Trans. Instrum. Meas.*, vol. 74, Art. no. 5001220, 2025, doi: 10.1109/TIM.2024.3497060.
- [8] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, “U-Net and its variants for medical image segmentation: A review of theory and applications,” *IEEE Access*, vol. 9, pp. 82031–82057, 2021, doi: 10.1109/ACCESS.2021.3086020.
- [9] J. Zhang, Z. Jiang, J. Dong, Y. Hou, and B. Liu, “Attention gate ResU-Net for automatic MRI brain tumor segmentation,” *IEEE Access*, vol. 8, pp. 58533–58545, 2020, doi: 10.1109/ACCESS.2020.2983075.
- [10] M. Heidari et al., “HiFormer: Hierarchical multi-scale representations using transformers for medical image segmentation,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, 2023, pp. 6191–6201, doi: 10.1109/WACV56688.2023.00614.

- [11] D. Kumar K, S. Pandi S, M. K, M. G, and P. R, “Integrating Unet segmentation and vision transformers for improved bone cancer diagnosis in histopathological images,” in Proc. Int. Conf. Comput. Commun. Technol. (ICCCT), Chennai, India, 2025, doi: 10.1109/ICCCT63501.2025.11020476.
- [12] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Med. Image Comput. Comput.-Assist. Interv. (MICCAI), 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
- [13] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A nested U-Net architecture for medical image segmentation,” in Proc. Deep Learn. Med. Image Anal. (DLMIA) Workshop, MICCAI, 2018, pp. 3–11, doi: 10.1007/978-3-030-00889-5_1.
- [14] D. Jha et al., “ResUNet++: An advanced architecture for medical image segmentation,” in Proc. IEEE Int. Symp. Multimedia (ISM), 2019, pp. 225–230, doi: 10.1109/ISM46123.2019.00049.
- [15] D.-P. Fan et al., “PraNet: Parallel reverse attention network for polyp segmentation,” in Proc. Med. Image Comput. Comput.-Assist. Interv. (MICCAI), vol. 12266, 2020, pp. 263–273, doi: 10.1007/978-3-030-59725-2_26.
- [16] C.-H. Huang, H.-Y. Wu, and Y.-L. Lin, “HarDNet-MSEG: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean Dice and 86 FPS,” arXiv preprint arXiv:2101.07172, 2021.
- [17] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, “Res2Net: A new multi-scale backbone architecture,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 2, pp. 652–662, Feb. 2021, doi: 10.1109/TPAMI.2019.2938758.
- [18] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 5998–6008.
- [19] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
- [20] Z. Liu et al., “Swin Transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.
- [21] J. Chen et al., “TransUNet: Transformers make strong encoders for medical image segmentation,” arXiv preprint arXiv:2102.04306, 2021.

- [22] H. Cao et al., “Swin-Unet: Unet-like pure transformer for medical image segmentation,” in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2022, pp. 205–218, doi: 10.1007/978-3-031-25066-8_9.
- [23] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in *Proc. Int. Conf. 3D Vis. (3DV)*, 2016, pp. 565–571, doi: 10.1109/3DV.2016.79.
- [24] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. Ben Ayed, “Boundary loss for highly unbalanced segmentation,” *Med. Image Anal.*, vol. 67, Art. no. 101851, 2021, doi: 10.1016/j.media.2020.101851.
- [25] D. Jha et al., “Kvasir-SEG: A segmented polyp dataset,” in *Proc. Int. Conf. Multimedia Modeling (MMM)*, vol. 11962, 2020, pp. 451–462, doi: 10.1007/978-3-030-37734-2_37.
- [26] 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,” *Comput. Med. Imag. Graph.*, vol. 43, pp. 99–111, 2015, doi: 10.1016/j.compmedimag.2015.02.007.
- [27] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.
- [28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.
- [29] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis HAUSDORFF DISTANCE MIXNET: A HYBRID CONVOLUTIONAL NEURAL NETWORK
- TRANSFORMER NETWORK WITH BOUNDARY - AWARE MECHANISMS FOR POLYP SEGMENTATION

Total Pages 54 Name of the Scholar KARTIK PRUTHI

Supervisor (s)

(1) Prof. Anil Singh Parihar

(2) _____

(3) _____

Department COMPUTER SCIENCE AND ENGINEERING

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin Similarity Index: 10%, Total Word Count:
10905

Date: _____

Candidate's Signature

Signature of Supervisor(s)

HD-MixNet

ORIGINALITY REPORT

10%

SIMILARITY INDEX

7%

INTERNET SOURCES

7%

PUBLICATIONS

3%

STUDENT PAPERS

PRIMARY SOURCES

1	arxiv.org Internet Source	1%
2	spectrum.library.concordia.ca Internet Source	<1%
3	www.mdpi.com Internet Source	<1%
4	Submitted to Delhi Technological University Student Paper	<1%
5	Submitted to University of Ulster Student Paper	<1%
6	assets-eu.researchsquare.com Internet Source	<1%
7	dongxiaozyu.github.io Internet Source	<1%
8	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2022", Springer Science and Business Media LLC, 2022 Publication	<1%
9	"Machine Learning in Medical Imaging", Springer Science and Business Media LLC, 2021 Publication	<1%
10	Submitted to Swinburne University of Technology Student Paper	<1%
11	Submitted to University of Warwick	

<1 %

12

5dok.net

Internet Source

<1 %

13

"Medical Image Computing and Computer Assisted Intervention – MICCAI 2019", Springer Science and Business Media LLC, 2019

Publication

<1 %

14

Submitted to Columbia University

Student Paper

<1 %

15

www.ysciei.com

Internet Source

<1 %

16

Haidar A. Almubarak. "Self-supervised learning for data augmentation in histopathology image segmentation", Scientific Reports, 2025

Publication

<1 %

17

ebin.pub

Internet Source

<1 %

18

repositorio.ufmg.br

Internet Source

<1 %

19

theses.hal.science

Internet Source

<1 %

20

Submitted to KJSCE

Student Paper

<1 %

21

Lei Yang, Hongyong Wang, Yuge Gu, Guibin Bian, Yanhong Liu, Hongnian Yu. "TMA-Net: A Transformer-Based Multi-Scale Attention Network for Surgical Instrument Segmentation", IEEE Transactions on Medical Robotics and Bionics, 2023

Publication

<1 %

22	Submitted to National University of Ireland, Galway Student Paper	<1 %
23	Submitted to University of KwaZulu-Natal Student Paper	<1 %
24	www.nazneenrajani.com Internet Source	<1 %
25	Submitted to University of Bradford Student Paper	<1 %
26	Submitted to University of Glasgow Student Paper	<1 %
27	ejde.math.txstate.edu Internet Source	<1 %
28	Idan Distelfeld, Shmuel Osovski. "Efficient multi-objective optimization of composite microstructures for thermal protection systems", Composite Structures, 2025 Publication	<1 %
29	Debesh Jha, Pia H. Smedsrud, Dag Johansen, Thomas de Lange, Havard D. Johansen, Pal Halvorsen, Michael A. Riegler. "A Comprehensive Study on Colorectal Polyp Segmentation With ResUNet++, Conditional Random Field and Test-Time Augmentation", IEEE Journal of Biomedical and Health Informatics, 2021 Publication	<1 %
30	Submitted to ICTS Student Paper	<1 %
31	www.um.edu.mt Internet Source	<1 %
32	Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, Anne L. Martel.	<1 %

"Loss odyssey in medical image segmentation", Medical Image Analysis, 2021

Publication

- 33 Xi Di, Yang Liu, Suoliang Chang, Wenbin Tian, Jiangtao Ma, Zilong Dong. "FaultEdgeFormer: an edge enhanced transformer model for 3D seismic fault detection", Journal of Geophysics and Engineering, 2026 <1 %

Publication

- 34 Xiaowei Liu, Yikun Hu, Jianguo Chen. "Hybrid CNN-Transformer model for medical image segmentation with pyramid convolution and multi-layer perceptron", Biomedical Signal Processing and Control, 2023 <1 %

Publication

- 35 dokumen.pub <1 %

Internet Source

- 36 orca.cardiff.ac.uk <1 %

Internet Source

- 37 www.frontiersin.org <1 %

Internet Source

- 38 Keles, Ahmet. "Transport Properties of Chiral p-wave Superconductor-Normal Metal Nanostructures.", Proquest, 2014. <1 %

Publication

- 39 Richard Bieck, David Baur, Johann Berger, Tim Stelzner, Anna Völker, Juliane Neumann, Christoph-E. Heyde, Thomas Neumuth. "An interactive system for muscle and fat tissue identification of the lumbar spine using semantic segmentation", Current Directions in Biomedical Engineering, 2021 <1 %

Publication

- 40 Submitted to WHU - Otto Beisheim School of Management <1 %

-
- 41 Wenbo Qi, H. C. Wu, S. C. Chan. "MDF-Net: A Multi-Scale Dynamic Fusion Network for Breast Tumor Segmentation of Ultrasound Images", IEEE Transactions on Image Processing, 2023
Publication <1 %
-
- 42 Xiaosong Li, Fuqiang Zhou, Haishu Tan, Wanning Zhang, Congyang Zhao. "Multimodal medical image fusion based on joint bilateral filter and local gradient energy", Information Sciences, 2021
Publication <1 %
-
- 43 Zihuang Wu, Hua Chen, Xinyu Xiong, Shang Wu, Hongwei Li, Xinyu Zhou. "BMANet: Boundary-guided multi-level attention network for polyp segmentation in colonoscopy images", Biomedical Signal Processing and Control, 2025
Publication <1 %
-
- 44 deepai.org
Internet Source <1 %
-
- 45 ia903409.us.archive.org
Internet Source <1 %
-
- 46 Ji, Wen. "Unsupervised Domain Adaptation for Medical Image Segmentation Using Generative Adversarial Alignment.", Hong Kong University of Science and Technology (Hong Kong)
Publication <1 %
-
- 47 Lei Bi, Ulrich Buehner, Xiaohang Fu, Tom Williamson, Peter Choong, Jinman Kim. "Hybrid CNN-Transformer Network for Interactive Learning of Challenging

Musculoskeletal Images", Computer Methods and Programs in Biomedicine, 2023

Publication

48 Lequan Yu, Hao Chen, Qi Dou, Jing Qin, Pheng Ann Heng. "Integrating Online and Offline Three-Dimensional Deep Learning for Automated Polyp Detection in Colonoscopy Videos", IEEE Journal of Biomedical and Health Informatics, 2017

Publication

49 Mary G. Nam, Suh-Yeon Dong. "Classification of Companion Animals' Ocular Diseases: Domain Adversarial Learning for Imbalanced Data", IEEE Access, 2023

Publication

50 Submitted to University of Nottingham
Student Paper

51 Yiyang Zhao, Kelei He, Yizhe Zhang, Yi Zhou, Tao Zhou, Dong Liang. "Structured prompt-guided knowledge injection for medical image segmentation", Pattern Recognition, 2026

Publication

52 iris.unimo.it
Internet Source

53 m.moam.info
Internet Source

54 nottingham-repository.worktribe.com
Internet Source

55 ppms.cit.cmu.edu
Internet Source

56 uh-ir.tdl.org
Internet Source

57 uu.diva-portal.org
Internet Source

58	www.southampton.ac.uk Internet Source	<1 %
59	www.techscience.com Internet Source	<1 %
60	Celaya, Adrian. "PocketNet: A Smaller Neural Network for Medical Image Analysis", Rice University, 2024 Publication	<1 %
61	Davood Karimi, Septimiu E. Salcudean. "Reducing the Hausdorff Distance in Medical Image Segmentation With Convolutional Neural Networks", IEEE Transactions on Medical Imaging, 2020 Publication	<1 %
62	Jingdian Yang, Xiangxiang Wang. "An SSM enhanced ViT model for aerial imagery based disaster risk inspection", Image and Vision Computing, 2026 Publication	<1 %
63	K. Narmadha, P. Varalakshmi. "Enhanced brain tumour segmentation using a hybrid dual encoder–decoder model in federated learning", Scientific Reports, 2025 Publication	<1 %
64	Lele Xu, Ye Li, Jinzhong Xu, Yue Zhang, Lili Guo. "BCTNet: Bi-Branch Cross-Fusion Transformer for Building Footprint Extraction", IEEE Transactions on Geoscience and Remote Sensing, 2023 Publication	<1 %
65	Qunchao Jin, Hongyu Hou, Guixu Zhang, Zhi Li. "FEGNet: A Feedback Enhancement Gate Network for Automatic Polyp Segmentation", IEEE Journal of Biomedical and Health Informatics, 2023	<1 %

66 Yang Yu, Rongrong Ni, Yao Zhao, Siyuan Yang, Fen Xia, Ning Jiang, Guoqing Zhao. "MSVT: Multiple Spatiotemporal Views Transformer for DeepFake Video Detection", IEEE Transactions on Circuits and Systems for Video Technology, 2023

Publication

67 Yunchu Zhang, Jianfei Dong. "MAEF-Net: MLP Attention for Feature Enhancement in U-Net based Medical Image Segmentation Networks", IEEE Journal of Biomedical and Health Informatics, 2023

Publication

68 atrium.lib.uoguelph.ca
Internet Source

69 cdn.techscience.cn
Internet Source

70 download.bibis.ir
Internet Source

71 eprints.usm.my
Internet Source

72 export.arxiv.org
Internet Source

73 ijsrem.com
Internet Source

74 jeeemi.org
Internet Source

75 pmc.ncbi.nlm.nih.gov
Internet Source

76 thesai.org
Internet Source

77 theses.lib.polyu.edu.hk

<1 %

78

w3.uqo.ca

Internet Source

<1 %

79

www.sciencecast.org

Internet Source

<1 %

80

www.spandidos-publications.com

Internet Source

<1 %

81

"Algorithms and Architectures for Parallel Processing", Springer Science and Business Media LLC, 2026

Publication

<1 %

82

"Artificial Intelligence and Soft Computing", Springer Science and Business Media LLC, 2026

Publication

<1 %

83

"Medical Image Computing and Computer Assisted Intervention – MICCAI 2021", Springer Science and Business Media LLC, 2021

Publication

<1 %

84

Zhaojun Liu, Shuaijie Xing, Xiaoli Zhang, Yuncong Feng, Xiaohan Hu. "P2TNet: Using Pyramid Pooling Transformer and Reverse Attention for Polyp Segmentation", IET Image Processing, 2026

Publication

<1 %

85

Zhenyu Yao, Yuping Su, Honghong Yang, Yumei Zhang, Xiaojun Wu. "TFSWA-ResUNet: music source separation with time-frequency sequence and shifted window attention-based ResUNet", EURASIP Journal on Advances in Signal Processing, 2025

Publication

<1 %

86 chao liu, Liang Yuan, wendong xiao, Cainian Li, Kai Lv, ran teng. "Trans-EWA-MCKMMD: Robust Unsupervised Domain Adaptation via Entropy-Enhanced Transformer and Multi-Layer Cauchy MMD", Measurement Science and Technology, 2026

Publication

<1 %

87 "Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries", Springer Science and Business Media LLC, 2022

Publication

<1 %

88 "Computer Vision – ECCV 2016", Springer Nature, 2016

Publication

<1 %

89 Xiaojuan Liu, Yilong Liu, Yongjun Zhu, Fang Huang, Jie Li, Shanxiong Chen. "Structure-aware context-enhanced and dual-path synergistic decoding network for atrophic gastritis segmentation", Engineering Applications of Artificial Intelligence, 2026

Publication

<1 %

Exclude quotes Off

Exclude matches Off

Exclude bibliography On

FINAL GRADE

/0

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40

PAGE 41

PAGE 42

PAGE 43

PAGE 44

PAGE 45

PAGE 46

PAGE 47

PAGE 48

PAGE 49

PAGE 50

Kartik K

HD-MixNet



Quick Submit



Quick Submit



Delhi Technological University

Document Details

Submission ID

trn:oid:::1:3581115002

Submission Date

May 28, 2026, 8:02 PM GMT+5:30

Download Date

May 28, 2026, 8:11 PM GMT+5:30

File Name

Theses_HD-MixNet.pdf

File Size

1.2 MB

40 Pages**9,471 Words****52,859 Characters**

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.