Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Renal Medical Image Segmentation using Cross-Perceptron Deliberation and Multi-Scale Feature Fusion

Swapna J.[1], Roselin Kiruba R.[2]

Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India.

E-mail: [1]swapnabinesh27@gmail.com, [2]rroselinkiruba@veltech.edu.in

Abstract

Architectures based on Convolutional Neural Networks (CNNs), like U-Net, have demonstrated notable efficiency in the segmentation of renal medical images. However, because convolution processes are limited and have restricted accessible fields, they frequently have trouble capturing long-range dependencies. Recent developments have improved global context modeling by incorporating transformer modules into U-Net variations to address this issue. However, during the global fusion process, these transformers based methods run the risk of losing important local spatial information. This research introduces Multi-Scale MCPA, a unique architecture designed specifically for the segmentation of 2D renal medical images. An encoder, decoder and cross perceptron module are the three main parts of MCPATo provide rich multi-scale feature interaction, the Cross Perceptron primarily uses several Multi-Scale Cross Perceptron (MCP) modules to capture local dependencies. To efficiently model long-range dependencies, these features are spatially unfolded, concatenated, and processed by a Global Perceptron component. A Progressive Dual-Branch Structure (PDBS) is implemented to enhance segmentation performance, particularly for fine-grained structures. During training, this component guides the network to progressively transfer its attention from coarse structural elements to intricate pixel-level representations. The proposed method is specifically designed for 2D medical image segmentation tasks, given the clinical significance of 2D imaging and the high computing demands of 3D models. Experimentation of the proposed approach on multiple publicly accessible datasets from different imaging tasks and modalities, such as OCTA (ROSE), fundus images (DRIVE, CHASE_DB1, HRF), MRI (ACDC), and CT (Synapse), demonstrates that the proposed method reliably outperforms state-of-the-art segmentation methods, accomplishing enhancements of +2.1% Dice score on Synapse CT, +2.6% on ACDC MRI, and up to +3.4% on retinal fundus datasets. The effectiveness and generalizability of MCPA are established by experimental results, which show that it routinely outperforms existing techniques in segmentation accuracy.

Keywords: Medical Image, Multi-Scale, Cross Perceptron Convolutional Neural Networks (CNNs), Segmentation.

1. Introduction

In contemporary Computer Aided Diagnostic (CAD) systems, medical image segmentation is important for accurately locating and defining anatomical features, lesions, or

DOI: https://doi.org/10.36548/jiip.2025.3.025
Received: 24.07.2025, received in revised form: 27.08.2025, accepted: 15.09.2025, published: 30.09.2025

diseased regions [1]. It is comprehensively used in many clinical tasks, such as vascular structure analysis, tumor detection, and organ boundary identification, in imaging modalities like fundus photography, Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Optical Coherence Tomography Angiography (OCTA). Improving computable analysis, treatment planning, and diagnostic accuracy all depend on accurate segmentation [2].

Convolutional Neural Networks (CNNs) have made considerable progress in medical image segmentation during the last ten years [3]. Since of their encoder-decoder architecture and the efficient use of skip connections to preserve spatial information, architectures like U-Net and its derivatives have shown impressive performance [4]. Nevertheless, the intrinsic dependence of CNN-based models on local convolutional kernels limits their capacity to represent distant spatial connections. Performance is frequently subpar as a result of this local receptive field bias, particularly when segmenting complicated structures with global contextual linkages [5].

Transformer-based models have been combined into medical image segmentation workflows in order to overcome these restrictions [6]. Transformers have demonstrated significant promise in modeling global feature dependencies through self-attention processes since their initial introduction in Natural Language Processing (NLP). In an effort to combine the advantages of transformers and CNNs, hybrid models such as TransUNet and Swin-UNet have introduced global feature modeling, which has improved performance [7]. These models do have certain drawbacks, though. In medical imaging scenarios connecting fine structural boundaries or low-contrast regions, local contextual information can be unintentionally extended by global attention mechanisms. Furthermore, transformers typically require substantial training datasets and computational resources, which aren't always feasible in the medical field [8].

Given these complications, we propose a brand new and effective segmentation framework designed especially for 2D medical image segmentation, the Multi-Scale Cross Perceptron Attention Network (MCPA). Contrast to traditional CNN or transformer based architectures, MCPA presents a Cross Perceptron Attention mechanism that uses a single structure to jointly model local and global dependencies [9]. A multi scale encoder, a decoder with skip connections, and a Cross Perceptron module that permits both local feature augmentation and global context aggregation make up the three primary parts of the design [10].

The MCP module, which is at the core of the MCPA network, is designed to extract and integrate data from a diversity of receptive fields. This enables the network to efficiently capture object boundaries and tissue variations at various scales. In order to accurately model long range dependencies, the network uses a Global Perceptron module after local feature extraction. This module aggregates spatially unfolded multi scale structures and learns their global correlations [11].

PDBS improve the segmentation of fine anatomical structures. This element progressively moves training emphasis from coarse structural segmentation to pixel level accuracy, making it easier to detect tiny, thin areas like tissue membranes and blood vessels. The MCPA structure is tailored for 2D renal image data, which continues to be the most common format in many diagnostic workflows because of its efficiency, accessibility, and clinical relevance, in contrast to 3D CNN models, which are computationally costly and frequently impractical in clinical settings [12].

The MCPA network is tested on several publicly available datasets that include a wide range of imaging modalities and anatomical targets, such as fundus photography (DRIVE, CHASE_DB1, HRF), MRI (ACDC), CT (Synapse), and OCTA (ROSE). The experimental results validate the robustness, generalization, and applicability of our approach in real-world clinical scenarios by showing that it achieves state-of-the-art performance across all benchmarks [13].

MCPA, a unique 2D renal medical image segmentation network, is offered in this research. An encoder, decoder, as well as Cross Perceptron component make up architecture. The Cross Perceptron is proposed to further model long-range dependencies within the spatial domain and capture local feature relationships across various scales, making multi-scale feature fusion possible. A PDBS, modeled after the RCE module, is proposed to enhance the segmentation of tiny tissue structures [14]. Beginning with global image-level breakdown and working its way down to fine-grained, pixel-level characteristics, the PDBS progressively directs the network's focus.

The main objectives of the research work are as follows,

- The MCPA framework, which uses a unique Cross Perceptron module to efficiently integrate global attention and local feature extraction.
- A PDBS, which is related to the RCE module and increasingly transfers the network's training attention to pixel-level, more granular data. This arrangement increases segmentation accuracy by including the Cross Perceptron.
- A variety of medical imaging modalities and datasets, such as CT (Synapse), MRI (ACDC), fundus images (DRIVE, CHASE_DB1, HRF), as well as OCTA (ROSE), presentation of MCPA network. Results from experiments show that MCPA consistently performs enhanced than current 2D segmentation methods based on CNN and Transformer. Interestingly, the suggested network preserves a lightweight architecture while achieving greater performance, which lowers overall medical processing costs and computational complexity.

Renal segmentation plays a critical role in medical imaging since the kidneys are affected by a wide range of diseases, including chronic kidney disease (CKD), cystic injuries, tumors, hydronephrosis, and inherited anomalies. Accurate delineation of renal structures in CT, MRI, and ultrasound images is important for diagnosis and screening, treatment planning, surgical guidance, and therapy monitoring.

2. Related Works

Res-UNet [10], UNet++ [27], and Eff-UNet [15] are a few of the improved variants that were developed over the years to enhance performance as well as efficiency. These variants support denser, nested skip connections as well as stronger CNN backbones. The UNet architecture [16] with its encoder-decoder structure connected by skip connections for retention of geographic information has evolved as a central element in renal medical image segmentation. Response Cue Erasing (RCE) was introduced in [17] to solve difficulties in segmenting small tissue structures. RCE improves segmentation by removing confident pixel areas from the input image according to the output of the main branch, prompting the model to concentrate on less confident, often finer areas. One of the limitations of this method is the

potential inconsistency between the segmentation results of the main and assist branches, which could be caused by the abrupt activation of the RCE mechanism.

A new split-based coarse-to-fine vessel segmentation network named OCTA-Net was introduced in [18], especially designed for OCTA images. This structure provides better performance for throughvascular exploration by separately identifying thick and thin arteries. While CNN-based UNet representations have demonstrated promising performance across a range of segmentation tasks, their inherent use of fixed-size convolutional kernels places a ceiling on their ability to process globally informative information and long-range spatial interdependence [18]. As a result, their ability to model more extensive feature correlation remains constrained. Medical image segmentation has improved dramatically over the past few years, with Vision Transformer (ViT) architectures being marked by their capacity to capture global context information.

Hybrid architectures that blend CNNs and Transformers have been the focus of researchers in order to benefit from their strengths. In order to capture local and global features at the same time, models like TransUNet and HiFormer integrate Transformer components into the encoder of CNN-associated UNet models. Likewise, in segmentation tasks, UNETR [19] uses a Transformer-related encoder along with a CNN-related decoder. Despite such advancements, relying on a simple combination of CNN and ViT components to extract and combine both local and global information is still difficult. In order to increase segmentation accuracy, TransUNet, for instance, significantly deepened the model, which can lead to less informative features in the deeper layers [29]. Pure Transformer-based models have become the focus of modern research, which seeks to improve performance by utilizing more complex self-attention mechanisms. The Swin Transformer [20] was initially used as the encoder and decoder in a UNet-like framework by Swin-Unet [21], allowing for categorized feature representation. To further improve segmentation capabilities, MISS Former [6] presents a better transformer context bridge made up of particular blocks that are proposed to capture local as well as global correlations across multi-scale characteristics.

A Cross-scale Global Transformer (CGT) is used by C2FTrans [22] to successfully capture global-scale dependencies with minimal computational complexity. Ductile high kernel consideration, which applies large convolutional kernels to effectively recognize relevant information in volumetric records, is introduced by DLKA-Net [23]. Additionally, a number of Transformer-based UNet versions have shown outstanding performance in a variety of medical image segmentation tasks, including UNETR [7], CoTr [24], nnFormer [25], Scale Former [26], DAE Former [27], Trans Deep Lab [28], PVT-CASCADE [29], and LeViT-UNet-384 [30]. Though the goal of these models is to use self-attention procedures to harness multi-scale feature information, they commonly fail to adequately capture the interplay between local and global dimensions. There is a lot of room for improving segmentation performance because of this limitation. Additionally, many of these networks, like Scale Former [31] and D-LKA Net [32], rely on numerous constraints and higher computational complexity in order to extract features. They therefore necessitate significant computational and medical resources, which might not align with representative financial limitations.

Existing segmentation methods still have a number of drawbacks despite notable advancements. The fixed-size convolutional kernels used by CNN-based U-Net variations limit their capability to capture long-range spatial relationships. Although hybrid CNN-Transformer approaches improve global context modeling, their performance is frequently subpar due to unsuccessful integration of local and global features. Richer global representations are offered

by pure transformer-driven designs, but their applicability in clinical contexts is limited by their potential to overlook specific local information and their high processing requirements. Furthermore, though some approaches (e.g., TransUNet, ScaleFormer, DLKA-Net) enhance model complexity or scale, resulting in increased computing costs and poorer interpretability, others (e.g., RCE) solve small feature segmentation but risk introducing contradictions. Overall, architectures that can efficiently model local as well as global relationships while remaining robust across modalities are still needed.

3. Proposed Methodology

A 2D renal medicinal image segmentation process called the MCPA is illustrated in Figure 1. The network encompasses three main components: an encoder, a decoder, and a Cross Perceptron module depicted in Figure 2. The encoder is designed to extract rich linguistic structures from the input renal image, while the decoder renovates the segmentation mask. Particularly, the MCPA framework supports the use of either a CNN or Transformer backbone for both the encoder and decoder. In contrast to the original UNet architecture, which uses direct skip connections between corresponding encoder and decoder layers, the proposed design presents the Cross Perceptron to replace these independent connections, enabling more effective multi-scale feature interaction. The goal of this proposed design is to allow the network to effectively incorporate multi-scale information by capturing long-range dependencies across several feature scales.

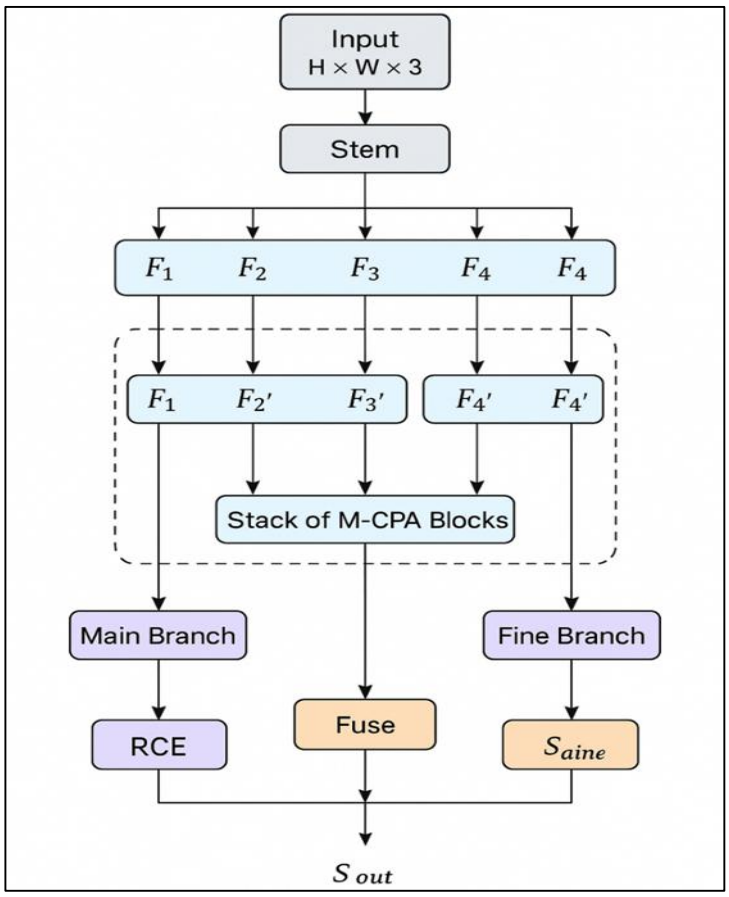

Figure 1. Design of Proposed Methodology

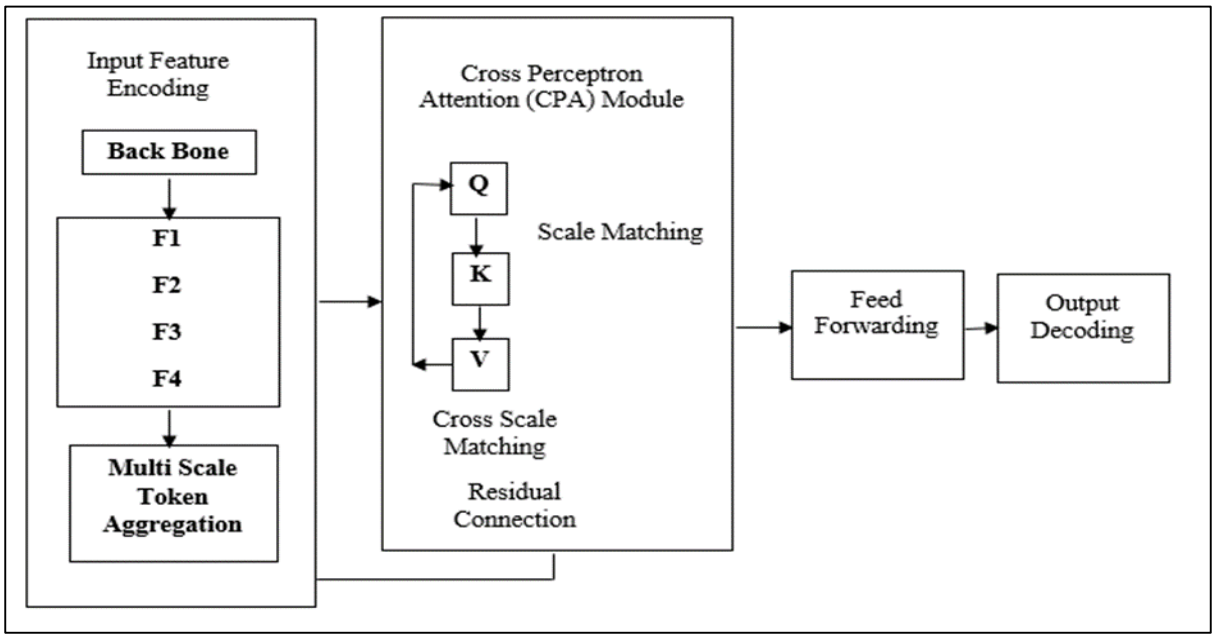

Figure 2. The Architecture of the Cross Perceptron

A PDBS, based on the RCE module, additionally addresses the difficulties of semantic segmentation in renal medical images, particularly those containing fine anatomical components, like retinal vascular segmentation in fundus images. The suggested architecture has dual parallel branches, each related to MCPA context, to improve segmentation accuracy and preserve structural details, as shown in Figure 3. Main branch is designed to perform coarse-grained complete segmentation, while the fine branch targets fine-grained, exhaustive segmentation. To facilitate active interaction among the dual branches, the RCE module is integrated as a communication bridge. Furthermore, during training, a Progressive Regularization Loss (PR Loss, denoted as LP), that progressively shifts the learning focus from the main branch to the fine branch. This progressive approach enhances the model's capability to capture intricate anatomical details, thereby improving complete segmentation performance.

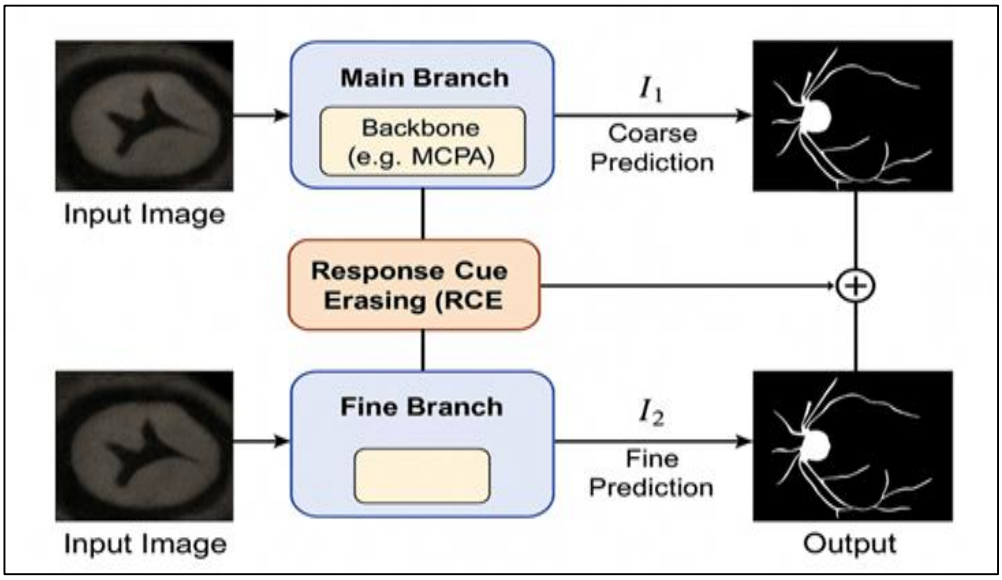

Figure 3. Overall Architecture of Proposed Progressive Dual-Branch Design

3.1 Multi-Granularity Perceptron Attention Framework

The encoder and decoder in the suggested approach are based on the recently extensively used Transformer-based architecture, with the Shunted Self-Attention (SSA) network serving as the backbone. SSA presents a Multi-Scale Token Aggregation (MTA) process in contrast to conventional Transformers, which use self-attention at a similar scale to compute query (Q), key (K), and value (V). This makes it conceivable to extract keys and values at different scales, which helps the model preserve detailed tokens for miniature structures while combining tokens that belong to enormous objects in an adaptive manner. Each of the four stages that make up the encoder and decoder includes a Patch Embedding (for the encoder) or a Patch Expansion (for the decoder), which is monitored by a number of SSA blocks.

Using the Patch Embedding process, the input picture (x \in \mathbb{R}^{H \times W \times C}) is initially downsampled and then molded into a categorization of compacted 2D patches (x_P). Three convolutional layers with kernel dimensions of 7, 3, and 2, and strides of 2, 1, and 2 are used in Stage 1's Patch Embedding process. In comparison to the original renal image, this leads to a four-fold reduction in spatial dimensions. Only a single convolution layer with a kernel dimension of 3 is employed in Stages 2-4, and each stage downsamples feature maps twice.

In a similar manner, the deconvolution layers' patch expand function progressively returns the feature maps' spatial dimensions to the initial input size in the decoder from Stage 1 to Stage 4. Decoder Stages 1 through 3 add an additional processing step, where the Cross Perceptron module features are concatenated with the output of the patch expand procedure at every stage. After that, a linear layer is applied to this combined feature map, halving the number of channels. The SSA block then receives the generated features for additional processing. MTA gathers appropriate information across multiple scales, permitting the network to capture both fine details as well as broad patterns. The SSA Feed Forward Network (SSA FFN) applies non-linear transformations to the generated features, refining and improving the representations for subsequent layers. Equation (1) defines the attention scores, which are then calculated by concatenating the aggregated tensors from several scales.

$$\begin{aligned} Qi &= XAi^{Q} \\ Ki, Vi &= MTA(X,ri)Ai^{K}, MTA(X,ri)Ai^{V} \\ Vi &= Vi + DA(Vi;\theta 1) \end{aligned} \tag{1}$$

Here, MTA(X, ri) stands for the MTA layer in the i^{th} attention head, which applies a convolutional layer to implement the downsampling rate ri. A depth-wise convolution layer parameterized by is represented by DW(·). More tokens in the key (K) and value (V) illustrations are combined as the value of θ1. As ri grows, this shortens their sequence lengths and enhances the model's ability to capture properties of larger-scale objects. Equation (2) then formulates the SSA Feed-Forward Network (SSA FFN), applying the output of the SSA module.

$$\begin{aligned} x &= Linear(a) \\ x &= Linear(\sigma(a + DW(x;\theta 2))) \end{aligned} \tag{2}$$

In this formulation, Linear(·) represents a fully connected (linear) layer, and σ() denotes the GELU activation function. For a comprehensive explanation of the SSA mechanism, readers are discussed to [18]. To optimize segmentation performance, apply a hybrid loss

function that combines Cross-Entropy Loss and Dice Loss, both of which measure the inconsistency between the predicted segmentation and the ground truth labels. Overall loss function is defined as follows:

$$\text{Ltotal} = \alpha\text{LCE} + (1-\alpha)\text{LDice} \quad (3)$$

In this case, the weighting hyperparameter, α, is set to 0.4 as suggested by earlier studies.

Algorithm 1: MCPA + PDBS Training

Aim: With two parallel branches (Main and Fine) and a Progressive Regularization Loss (PR Loss) that progressively aligns the Fine branch to the Main branch, the objective is to train MCPA (encoder + Cross-Perceptron + decoder).
Training set inputs D={(I(i),Y(i))}
Model parameters: θMCPA, θfine, and θmain
Batch size B, initial LR $\eta 0$, optimizer, and epochs are hyperparameters.
Emax PR schedule parameters: exponent α, E1 (end of transition), and E0 (end of initial stage).
Loss weights: any weight for LP (implied via f(E)) and λ for CE/Dice hybrid (or α previously).
The following defaults are suggested:
E_max = 200, E0 = 0, E1 = 50 (tuneable), α = 1 (linear ramp), λ = 0.4, optimizer = Adam, optimizer = cosine decay, B = 20, η0 = 1e-4, and scheduler = cosine decay.
Steps for advanced training (numbered)

1. Set the MCPA parameters (backbone weights, CPA/M-CPA, G-Perceptron, and decoders) to their initial values. Set up the cosine LR scheduler and optimizers.
2. To Emax for era E=1:

i. Determine the PR scale factor (E).

$$f(E) = \begin{cases} 0, \text{E} \le \text{E0} \\ (\dfrac{\text{E} - \text{E0}}{\text{E1} - \text{E0}})^{\alpha} \text{E0} < \text{E} < \text{E1} \\ 1, \text{E} \ge \text{E1} \end{cases}$$

ii. For every little batch of pictures I with labels Y: (Main branch employs the whole MCPA flow: encoder → MCP/CPA → G-Perceptron → decoder.) Forward Main branch:1=Pmain(l; θmain).

iii. To create modified input used for Fine branch, applies Response Cue Erasing (RCE) procedure. Top-k confident pixels from I (those with the highest likelihood in 1I) are erased using lmod=R(I,I1).

iv. Moving forward Branch of fineness: I2=Pfine(lmod; θfine)
(Fine branch: CNN-based alternative, identical MCPA architecture, but with an emphasis on fine details.)

v. Determine the segmentation losses:

vi. Dice L$Dice$(I2,Y) and Cross-Entropy LEC(I2,Y). Decrease in hybrid segmentation: Lsc=λLCE(1−λ)L$Dice$

vii. LP=f(E)‖l1−l2‖2 is the progressive regularization loss.

viii. Loss total: L=$Lseg$+LP.

ix. (If desired, calculate the segmentation loss on both branches and combine; L=Lmain+Lfine+LP is a popular choice.)

x. Backward + update: use an optimizer to update after computing gradients with respect to θmain and θfine. Use mixed precision if you'd like
3. Scheduler step: use cosine decay to update LR.
4. Checkpoint and validation: assess using the val set (PA, Dice, SE, SP). Save the best checkpoint.
5. Finish the training.

3.2 Cross Perceptron

The Cross Perceptron module incorporates data from various spatial resolutions to efficiently fuse multi-scale features. Four Cross Perceptron Segments Perceptron 1 through Perceptron 4 are used for simulating the local relationships among feature maps at dissimilar scales. This improves the final representations by adding more local detail and semantic meaning. Inspired by the method [6], it applies a Global Perceptron (G-Perceptron) segment to capture global dependencies. This module concatenates and clarifies multi-scale characteristic vectors along spatial dimensions for thorough context modeling.

$$\text{Attention(Q, K,V)} = \text{softmax(QKT/}\sqrt{}\text{ dk)V} \quad (4)$$

where KT is transpose of K as well as dk is dimensionality of attention heads.

The central element of Perceptrons 1 through 3 is the Cross Perceptron Attention (CPA) module. Two inputs are established by the CPA module: one from the output of the associated encoder stage and the other from a preceding Perceptron or a lower stage. A linear layer processes each input independently to produce the value (V), key (K), and query (Q) projections. The Multi-Head Cross-Attention (MHCA) mechanism then uses these projections to calculate attention values among Q, K, and V that come from various sources. A Typical self-attention formulation is used for the attention computation.

Sequence lengths vary in this research because the inputs to the CPA module are derived from encoder outputs at different stages or from other MCP modules. In order to calculate cross-attention values (as explained in Equation (4)), it is necessary to align the sequence lengths of the output features Q, K, as well as V. According to Table 1, Q, K, and V inside every CPA segment (from Perceptron 1 to Perceptron 3) are given particular sequence lengths in order to strike a balance between computing efficiency and accuracy.

Within the same MCP, it is critical that the length of Q stays the same for every CPA module. On the other hand, K and V can have the similar length (as in Perceptron 1) or different lengths (like in Perceptrons 2 and 3). The outputs of several CPA modules are concatenated within each MCP module, and their dimensionality is further reduced by passing them via a Linear Layer. The output will match the unique feature dimensionality of the appropriate encoder step. The Feed-Forward Network (FFN) module subsequently performs additional processing on the output. The following is a mathematical definition of the entire FFN operation:

$$\text{Xi} = \text{Xi}*+\text{Linear}(\sigma(\text{Linear}(\text{Xi}*))) \quad (5)$$

where X$i*$ indicates the FFN's input and Xi stands for its output. It should be noted that the FFN's output preserves the same dimensions as its input. In particular, the outputs of the

FFN for the second, third, and fourth stages are represented by the symbols X2, X3, and X4, accordingly.

Replace the CPA module in Perceptron 4 with a Modified Cross Perceptron Attention (M-CPA) segment to further enhance performance. The M-CPA uses a MTA component to offer multi-scale K and V representations, just as the SSA design does. The Q, K, and V sequence length combinations of Perceptron 4's M-CPA are also considered. Furthermore, features from two adjacent scales are concatenated to generate each M-CPA input. For example, F1 and F2 are concatenated to produce F12, and F3 and F4 are combined to form F34. The following is a formal definition of the operation:

$$F12=\text{Remodel}(F1[-1,C1])\ \text{cRemodel}(F2[-1,C1])$$

$$F34=\text{Remodel}(F3[-1,C1])\ \text{cRemodel}(F4[-1,C1]) \quad (6)$$

Here, the channel depth is denoted as $C1$. Proportions of Q, K, and V are modified appropriately since every input to the M-CPA is created by adding two feature maps. As shown in Table 1, the channel diameters of Q, K, and V are significantly compacted to 32 in order to diminish computing overhead. The channel complexities of F12 and output F34 from Perceptron 4 are the same. The G-Perceptron module receives these two features concatenated. SSA is applied to the combined feature first. An SSA-FFN is used to expand feature representation after paying attention to capture dependencies. Four features F1, F2, F3, and F4 are produced when the feature encoding of global dependencies is finally divided according to the original channel order.

The scales F1, F2, F3, and F4 are consistent with these characteristics, as are G-Perceptron's Q, K, and V sequence lengths. Through the integration of global features and efficient global dependency modeling, the G-Perceptron enhances model performance by implementing multi-head attention throughout the global characteristic space.

3.3 Progressive Dual Branch Design

A PDBS related to MCPA aims to improve segmentation performance for medical images by utilizing delicate tissue features, such as retinal blood vessels in fundus images. The Main Branch and Fine Branch are two components of PDBS, which uses MCPA as the backbone network, as shown in Figure 3. It should be noted that the majority of current methods for segmenting these complex structures are typically CNN-based. This inclination most likely results from the comparatively little investigation of Transformers in this field, which has been encouraged by concerns about how well they can capture fine-grained local features.

Therefore, a CNN-based MCPA is created in addition to using the Transformer-based MCPA architecture outlined in Section 3.1 to replace the Transformer-based modules. While the Fine Branch is intended to capture fine-grained features in medical images, the Main Branch is responsible for segmenting large-scale regions. Both branches follow the design principles commonly used in architectures such as UNet by employing CNN-based architectures for their encoders and decoders. An RCE module is provided to facilitate resourceful cooperation between the two branches. By connecting the Main Branch output to the Fine Branch input, this module allows the Main Branch segmentation cues from training to directly inform the Fine Branch. The input renal image and the accompanying ground truth label are designated. A segmentation map I1 is generated following processing by the Main Branch's MCPA module.

Large blood vessels and other perceptible tissue features in retinal images are disregarded from the original renal image using the RCE module. The Fine Branch then uses the resulting image to focus its training on capturing finer tissue details, such as retinal vessels. A detailed portrayal of this entire technique may be found in:

$$\begin{aligned} I1&=PMain(I,\theta1)\\ I2&=PFine(R(I,I1),\theta2) \end{aligned} \tag{7}$$

In this case, () stands for the RCE module, and PMain and PFine consistently refer to the Main Branch and Fine Branch, respectively. θ1 and θ2 characterize the parameters of these two networks. In particular, from the Main Branch's output, the top k pixels with the extreme confidence ratings are selected. Large-scale tissue features that are relatively simple to distinguish throughout the renal image are characterized by these high-confidence pixels. Input I for the Fine Branch is created using the remaining pixels. A revised feature map I2 is then generated by passing this transformed renal image to the Fine Branch network, which emphasizes segmenting larger tissue structures with more complicated characteristics. In order to execute similarity among the outputs of the Main Branch and the Fine Branch, we adopt a stability loss, denoted as $\|I1-I2\|2$, based on the regularization loss suggested in [26]. This promotes alignment between the two branches, establishing the network's overall resilience. However, the Main Branch's segmentation output might not be at its best in the early training phases when the model is still not fully tuned. As a result, employing this output to guide the Fine Branch may not only be ineffective but also hinder training, which would prevent the network from achieving optimal performance.

To solve this problem, the training procedure is divided into three separate stages, each of which is governed by the regularization loss: the initial, evolution, and final stages. A Progressive Regularization Loss (PRLoss, LP) is employed to help the Main Branch and the Fine Branch align seamlessly and efficiently. Throughout training, this loss function progressively reduces the difference between the outputs of the two branches, encouraging consistency and avoiding negative transmission in the early stages.

$$Lp=f(E)\cdot(\ I1-I2\ 2) \tag{8}$$

Let (E) be a transformation function that increases monotonically and is reliant on the training epoch E. The beginning and ending epochs of the transition stage are characterized by the two hyperparameters, (E0) and (E1), respectively. The countenance for the function (E) is:

$$f(E)=\begin{cases} 0, \mathrm{E}\le \mathrm{E0} \\ (\frac{\mathrm{E}-\mathrm{E0}}{\mathrm{E1}-\mathrm{E0}})^{\alpha} \mathrm{E0}<\mathrm{E}<\mathrm{E1} \\ 1, \mathrm{E}\ge \mathrm{E1} \end{cases} \tag{9}$$

In particular, only the Main Branch is engaged throughout the first training phase (E≤E0), when (E0)=0. For large-scale tissue structures, this permits the model to learn coarse segmentation results quickly. The regularization loss is not applied at this time, and the Fine Branch is still dormant. In order to ensure stable and efficient meeting, the transition stage (E0<E<E1) should only necessitate modest feature weight adjustments rather than important updates due to the non-convex structure of the loss function.

In the meantime, the function (E) steadily and smoothly rises from 0 to 1, signifying the Fine Branch's slow activation and the associated loss of regularization. This design facilitates a gradual shift in training focus to the trickier fine tissue segmentation task. A progressive

hyperparameter α, controlled by power law function is added to further stabilize training process. Interestingly, f(E) exhibits linear development when α=1.

To optimize segmentation performance, full Progressive Dual branch Arrangement as well as the full regularization cost are used in the last training stage (E≥E1), where (E1)=1. Consequently, Equation (3) is replaced with the updated total loss *L* as follows:

$$Ltotal = \lambda LCE + (1-\lambda)LDice + LP \quad (10)$$

In this case, λ is a weighting hyperparameter that was premeditated using the rules.

4. Experimental Results and Discussion

4.1 Experimental Setup

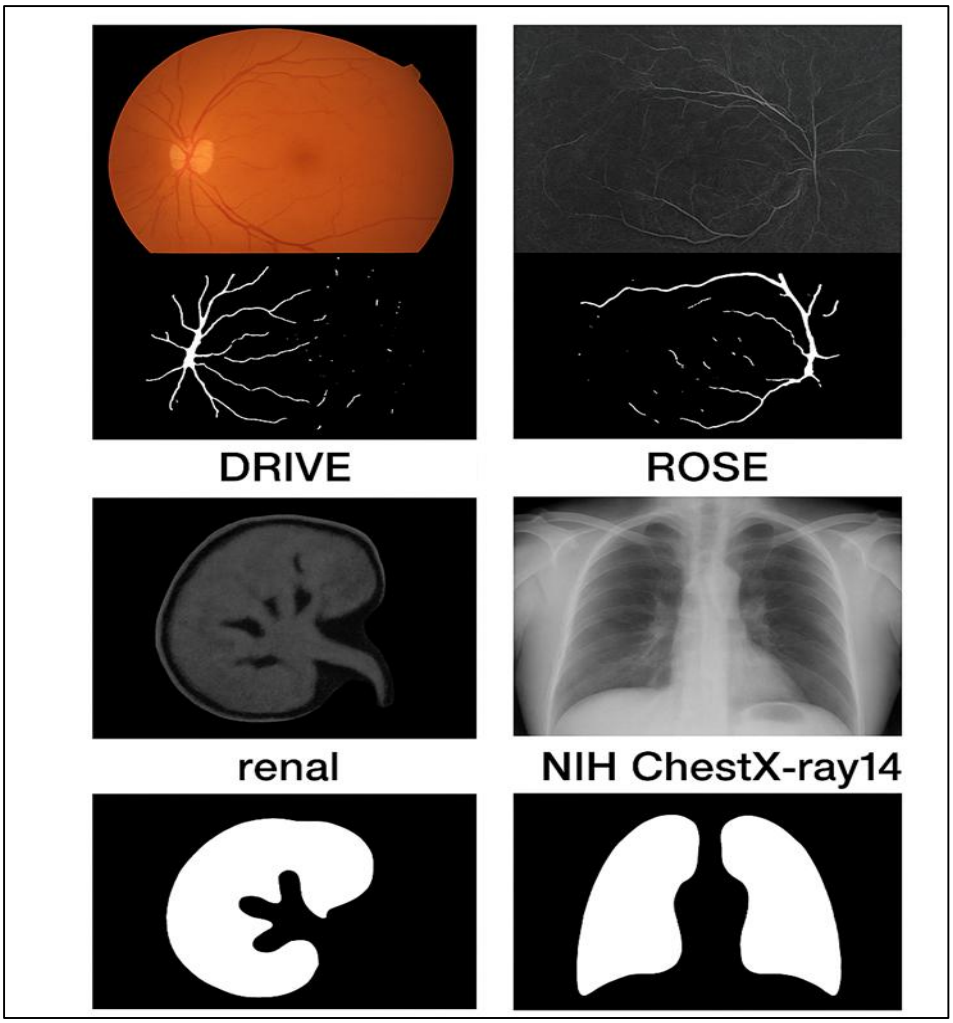

Figure 4. Different Dataset Classification

The effectiveness of the proposed Transformer-related segmentation framework employing the PDBS was assessed through experiments conducted on the ROSE dataset. The dataset comprises 1,077 high-resolution medical images with annotations of characteristics of

fragile tissue, such as renal vessels. The dataset is divided into 70% of the data for training, 15% for validation, and 15% for testing. Figure 4 lists the photos from the varied dataset.

Following its PyTorch implementation, the model was proficient on an NVIDIA RTX 3090 GPU using the Adam optimizer, a batch size of 20, and an initial learning rate of 1×10^{-4}. A cosine decay schedule was used to dynamically alter the learning rate during training. The network was trained over 200 epochs to ensure convergence. Xavier (Glorot) activation was utilized to weight layers permitting robust gradient transmission, and regularized dropouts were incorporated into the network to enhance applicability as well as consistency. A data augmentation pipeline was also utilized to improve resilience to retinal structural heterogeneity and overfitting. In order to simulate various imaging situations, enhancement process encompassed random flips in the vertical and horizontal directions, varying contrast and brightness modifications, and elastic distortions.

4.2 Quantitative Results

The conditions used to estimate the analysis of the suggested models were Pixel Accuracy (PA), Dice Coefficient, Sensitivity (SE), and Specificity (SP). The proposed method using several segmentation models. Table 1 presents a numerical assessment of the segmentation performance of different approaches.

Table 1. Quantitative Comparison of Segmentation Performance on ROSE Dataset

Method	Pixel Accuracy (%)	Dice Coefficient (%)	Sensitivity (%)	Specificity (%)
CNN	87.7	80.4	83.4	86.7
3D-CNN	89.4	83.8	85.4	81.1
SCNN-ResNet	92.1	79.6	82.2	83.4
GAN	95.7	85.9	86.2	89.4
Ours (PDBS)	96.8	88.6	89.1	91.2

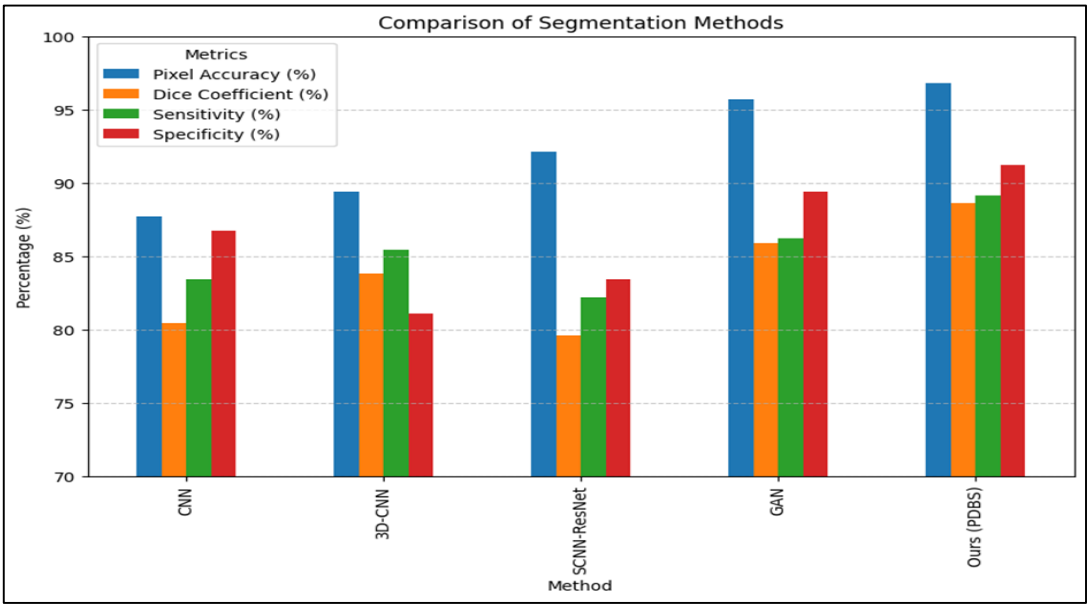

Figure 5. Segmentation Performance on the ROSE Dataset

i. Accuracy of Pixels (PA)

Pixel accuracy is the proportion of appropriately classified pixels in the entire image. The proposed PDBS method achieves 96.8% accuracy, which is better than both traditional CNN (87.7%) and even GAN established methods (95.7%). Figure 6 illustrates that PDBS has exceptional segmentation competence overall.

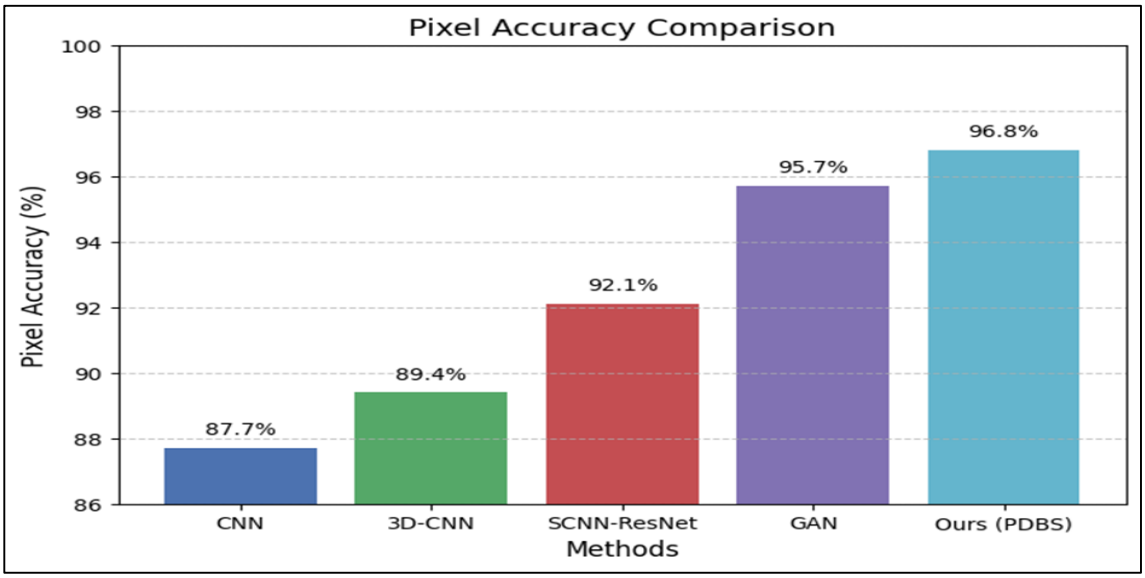

Figure 6. Pixel Accuracy of Compared Methodologies

Pixel accuracy of compared methodologies is shown in figure 6. The Pixel Accuracy statistic appraises the proportion of correctly identified pixels in the entire image. Permitting outcomes, CNN's 87.7% baseline accuracy is rather low due to its limited capacity to capture complex contextual relationships. CNN-ResNet recognises superior contextual responsiveness at 92.1% by using spatial connection; 3D-CNN improves by 89.4% by adding volumetric information, but it still struggles with fine structures. GAN-based separation influences 95.7% through use of adversarial learning, which makes separation outputs more accurate. Proposed method PDBS accomplishes maximum pixel accuracy of 96.8%, demonstrating the efficiency of combining transformer related context modeling, cross attention, and an enlightened dual branch design.

ii. Coefficient of Dice

The Dice score reflects the overlap among predicted and ground-truth segmentations. A higher Dice value indicates improved performance. PDBS again wins with 88.6% when compared to CNN at 80.4% and SCNN-ResNet at 79.6%, indicating superior segmentation quality, particularly in capturing insignificant tissue characteristics.

iii. Recall Sensitivity

Sensitivity quantifies a model’s capability to accurately distinguish separated zones, or true positives. PDBS can accumulate the most relevant features with a high sensitivity of 89.1% without missing any important sites. In medical imaging, false negatives can be quite damaging, subsequently this is crucial.

iv. Particularity

The specificity of a model limits its competence to avoid false positives. Here, the PDBS technique also scores the highest at 91.2%, indicating that it effectively avoids incorrectly classifying inappropriate background regions as target structures. The proposed PDBS architecture constantly outperforms conventional and deep learning related methodologies across all criteria.

4.3 Ablation Study

Figure 7 displays the dice coefficient as well as the sensitivity of ablation studies that gradually delete modules in order to investigate the contributions of various components in the proposed system.

Table 2. Ablation Research Demonstrating Each Suggested Modules Input

Model Variant	Dice (%)	Sensitivity (%)
Baseline SSA + Transformer only	84.5	85.2
+ Cross Perceptron Attention (CPA)	86.0	86.8
+ Modified CPA (M-CPA)	87.1	88.0
+ G-Perceptron	87.8	88.5
+ Progressive Dual-Branch (PDBS)	88.6	89.1

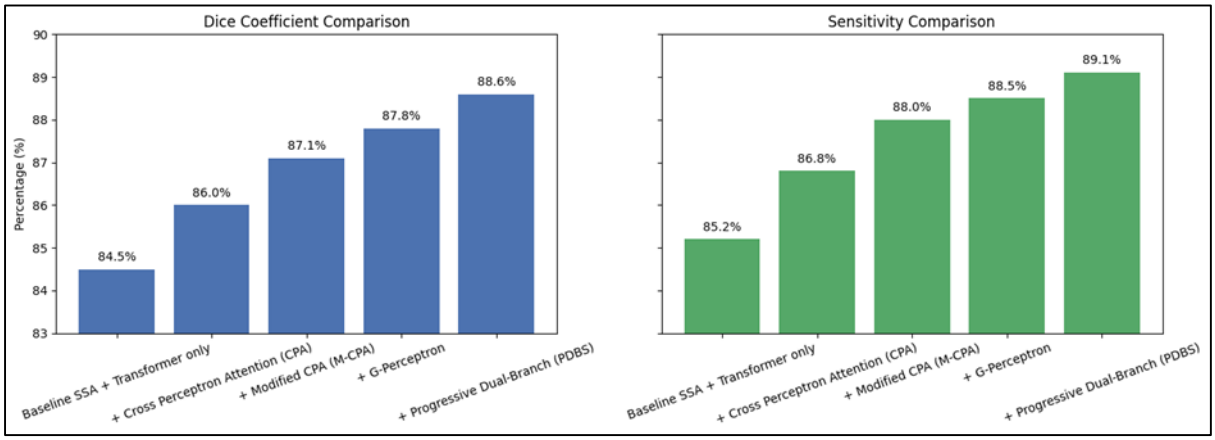

Figure 7. Dice Coefficient and Sensitivity

4.4 Qualitative Findings

Figure 5 shows sample qualitative assessments of segmentation masks created by different models. The suggested technique produces more determined and transparent vascular structures, especially for insignificant capillaries.

4.5 The Efficacy of Progressive Loss

To verify the efficiency of the proposed Progressive Regularization Loss (PR Loss), models trained with and without it were compared. Figure 8 illustrates how training curves exhibit more trustworthy and seamless convergence with PR Loss.

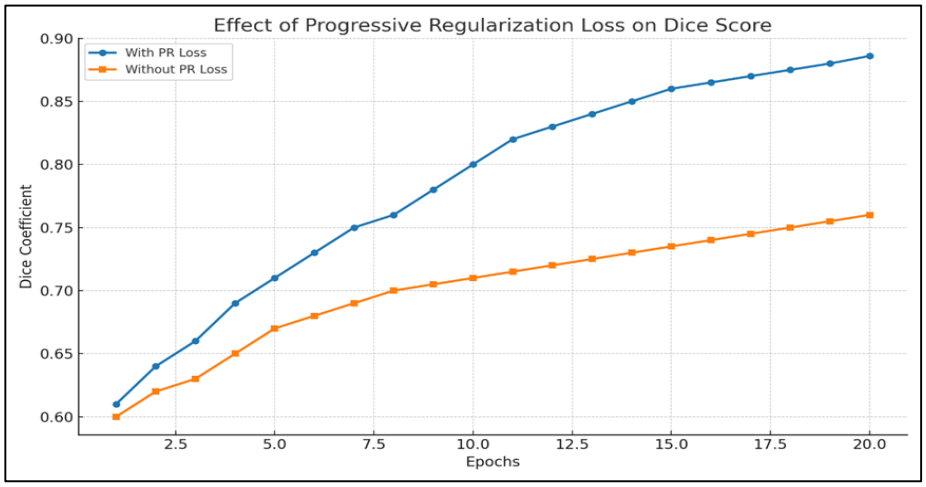

Figure 8. Effect of Progressive Regularization

- **With PR Loss:** The curve demonstrates a steadier and smoother rise in performance, indicating dependability brought about by the Fine Branch's regular alignment.
- **Without PR Loss:** The performance plateaus earlier and develops more deliberately, suggesting that the Fine Branch struggles when it is impulsively focused by raw Main Branch outputs. The effectiveness of the enlightened training approach in improving fine-grained segmentation is amply demonstrated in Figure 9.

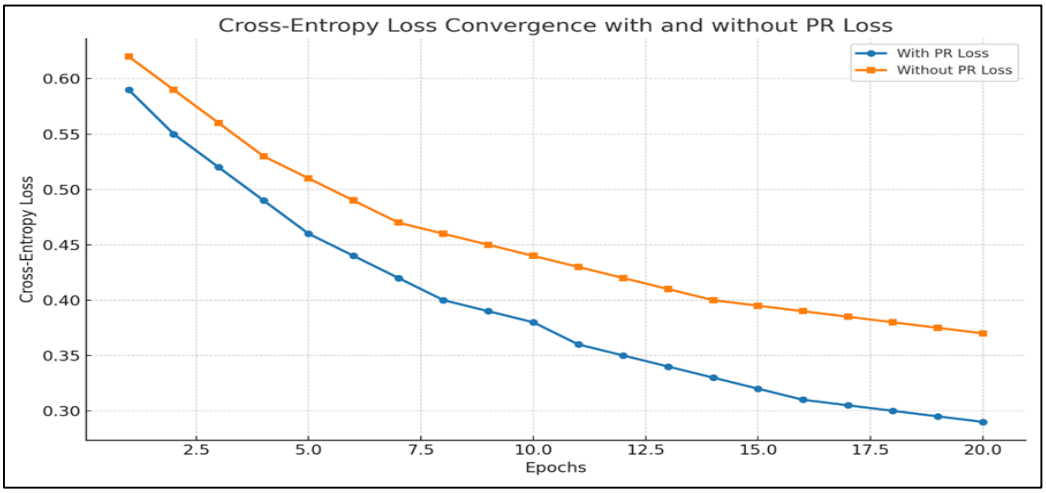

Figure 9. Ceoss-Entropy Convergence

- **With PR Loss:** The loss drops more promptly and steadily, suggesting that the model benefits from more continuous regulation and avoids initial misdirection.
- **In the absence of PR Loss:** The slower and less effective convergence is affected by an initial requirement on the Main Branch. This further illustrates how the Progressive

Regularization Loss enhances training stability, convergence speed, and segmentation accuracy.

5. Conclusion

We showed an MCPA with a PDBS to achieve accurate and reliable 2D medical image segmentation. In order to accurately capture fine-grained anatomical structures and long-range dependencies, the framework proposed here combines global context modeling and multi-scale local feature extraction. PDBS significantly improves segmentation quality by iteratively refining the accuracy of structural representations from coarse to pixel-level. Experiments on publicly accessible datasets on various imaging modalities, such as MRI, CT, fundus photography, and OCTA, consistently show that our algorithm outperforms state-of-the-art methods in terms of segmentation accuracy and structural coherence. Because of its wide use and high clinical employability, it is the best choice for a variety of clinical imaging applications. Light-weight implementations that can be applied in real-time clinical use will be explored, and this system will be extended to 3D volumetric segmentation in the future.

References

[1] Shamshad, Fahad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. "Transformers in medical imaging: A survey." Medical image analysis 88 (2023): 102802.

[2] Gibson, Eli, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P. Pereira, Matthew J. Clarkson, and Dean C. Barratt. "Automatic multi-organ segmentation on abdominal CT with dense V-networks." IEEE transactions on medical imaging 37, no. 8 (2018): 1822-1834.

[3] Azad, Reza, Amirhossein Kazerouni, Moein Heidari, Ehsan Khodapanah Aghdam, Amirali Molaei, Yiwei Jia, Abin Jose, Rijo Roy, and Dorit Merhof. "Advances in medical image analysis with vision transformers: a comprehensive review." Medical Image Analysis 91 (2024): 103000.

[4] Chen, Jieneng, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. "Transunet: Transformers make strong encoders for medical image segmentation." arXiv preprint arXiv:2102.04306 (2021).

[5] Nanni, Loris, Carlo Fantozzi, Andrea Loreggia, and Alessandra Lumini. "Ensembles of convolutional neural networks and transformers for polyp segmentation." Sensors 23, no. 10 (2023): 4688.

[6] Ghazouani, Fethi, Pierre Vera, and Su Ruan. "Efficient brain tumor segmentation using Swin transformer and enhanced local self-attention." International Journal of Computer Assisted Radiology and Surgery 19, no. 2 (2024): 273-281.

[7] Ali, Hazrat, Farida Mohsen, and Zubair Shah. "Improving diagnosis and prognosis of lung cancer using vision transformers: a scoping review." BMC Medical Imaging 23, no. 1 (2023): 129.

[8] Zhou, Hong-Yu, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. "nnformer: Interleaved transformer for volumetric segmentation." arXiv preprint

arXiv:2109.03201 (2021).

[9] Antonelli, Michela, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens et al. "The medical segmentation decathlon." Nature communications 13, no. 1 (2022): 4128.

[10] Hatamizadeh, Ali, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images." In International MICCAI brainlesion workshop, Cham: Springer International Publishing, (2021): 272-284.

[11] Gu, Zaiwang, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. "Ce-net: Context encoder network for 2d medical image segmentation." IEEE transactions on medical imaging 38, no. 10 (2019): 2281-2292.

[12] Li, Yuanyuan, Ziyu Wang, Li Yin, Zhiqin Zhu, Guanqiu Qi, and Yu Liu. "X-net: a dual encoding–decoding method in medical image segmentation." The Visual Computer 39, no. 6 (2023): 2223-2233.

[13] Chen, Xuming, Shanlin Sun, Narisu Bai, Kun Han, Qianqian Liu, Shengyu Yao, Hao Tang et al. "A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy." Radiotherapy and Oncology 160 (2021): 175-184.

[14] Yuan, Feiniu, Zhengxiao Zhang, and Zhijun Fang. "An effective CNN and Transformer complementary network for medical image segmentation." Pattern Recognition 136 (2023): 109228.

[15] Ibtehaz, Nabil, and M. Sohel Rahman. "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation." Neural networks 121 (2020): 74-87.

[16] Sirinukunwattana, Korsuk, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang et al. "Gland segmentation in colon histology images: The glas challenge contest." Medical image analysis 35 (2017): 489-502.

[17] Wang, Wenhai, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. "Pvt v2: Improved baselines with pyramid vision transformer." Computational visual media 8, no. 3 (2022): 415-424.

[18] Kumar, Neeraj, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. "A dataset and a technique for generalized nuclear segmentation for computational pathology." IEEE transactions on medical imaging 36, no. 7 (2017): 1550-1560.

[19] Yun, Boxiang, Yan Wang, Jieneng Chen, Huiyu Wang, Wei Shen, and Qingli Li. "Spectr: Spectral transformer for hyperspectral pathology image segmentation." arXiv preprint arXiv:2103.03604 (2021).

[20] Lin, X., Yan, Z., Yu, L., and Cheng, K.-T. "C2FTrans: Coarse-to-fine transformers for medical image segmentation." arXiv preprint arXiv:2206.14409 (2022).

[21] Li, Di, and Susanto Rahardja. "BSEResU-Net: An attention-based before-activation residual U-Net for retinal vessel segmentation." Computer Methods and Programs in Biomedicine 205 (2021): 106070.

[22] Imran, Azhar, Jianqiang Li, Yan Pei, Ji-Jiang Yang, and Qing Wang. "Comparative analysis of vessel segmentation techniques in retinal images." IEEE Access 7 (2019): 114862-114887.

[23] Aras, Rezty Amalia, Tri Lestari, Hanung Adi Nugroho, and Igi Ardiyanto. "Segmentation of retinal blood vessels for detection of diabetic retinopathy: A review." Communications in Science and Technology 1, no. 1 (2016).

[24] Dong, Fangfang, Dengyang Wu, Chenying Guo, Shuting Zhang, Bailin Yang, and Xiangyang Gong. "CRAUNet: A cascaded residual attention U-Net for retinal vessel segmentation." Computers in Biology and Medicine 147 (2022): 105651.

[25] Ma, Yuhui, Huaying Hao, Jianyang Xie, Huazhu Fu, Jiong Zhang, Jianlong Yang, Zhen Wang, Jiang Liu, Yalin Zheng, and Yitian Zhao. "ROSE: a retinal OCT-angiography vessel segmentation dataset and new model." IEEE transactions on medical imaging 40, no. 3 (2020): 928-939.

[26] Lin, Ailiang, Bingzhi Chen, Jiayu Xu, Zheng Zhang, Guangming Lu, and David Zhang. "Ds-transunet: Dual swin transformer u-net for medical image segmentation." IEEE Transactions on Instrumentation and Measurement 71 (2022): 1-15.

[27] Budai, Attila, Rüdiger Bock, Andreas Maier, Joachim Hornegger, and Georg Michelson. "Robust vessel segmentation in fundus images." International journal of biomedical imaging 2013, no. 1 (2013): 154860.

[28] Zhang, Zhuangzhuang, and Weixiong Zhang. "Pyramid medical transformer for medical image segmentation." arXiv preprint arXiv:2104.14702 (2021).

[29] Wang, Peihao, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. "Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice." arXiv preprint arXiv:2203.05962 (2022).

[30] Azad, Reza, Leon Niggemeier, Michael Hüttemann, Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Yury Velichko, Ulas Bagci, and Dorit Merhof. "Beyond self-attention: Deformable large kernel attention for medical image segmentation." In Proceedings of the IEEE/CVF winter conference on applications of computer vision, (2024): 1287-1297.

[31] Huang, Huimin, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Ruofeng Tong. "ScaleFormer: revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation." arXiv preprint arXiv:2207.14552 (2022).

[32] Azad, Reza, René Arimond, Ehsan Khodapanah Aghdam, Amirhossein Kazerouni, and Dorit Merhof. "Dae-former: Dual attention-guided efficient transformer for medical image segmentation." In International workshop on predictive intelligence in medicine, Cham: Springer Nature Switzerland, (2023): 83-95.