JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

JSCDM

Vol. 5 No. 1 (2024) 57-69
https://publisher.uthm.edu.my/ojs/index.php/jscdm

CSWin Transformer-CNN Encoder and Multi-Head Self-Attention Based CNN Decoder for Robust Medical Segmentation

Dr. J. Pandu[1]*, G. Ravi Shankar Reddy[2], Dr. CH. Ashok Babu[3]

[1] *Professor and dean IQAC Department of ECE,*
Sreyas Institute of Engineering and Technology, Nagole, Bandlaguna, Hyderabad, Telangana, INDIA
[2] *Professor, Department of Electronics and Communication Engineering,*
CVR College of Engineering, Hyderabad, Telangana, INDIA
[3] *Professor, Department of Audiology,*
Helen Kellers Institute of Research & Rehabilitation for the Disabled Children, Hyderabad- 500056, INDIA

*Corresponding Author: pandu427@gmail.com
DOI: https://doi.org/10.30880/jscdm.2024.05.01.005

Article Info

Received: 1 December 2023
Accepted: 25 April 2024
Available online: 21 June 2024

Keywords

Medical image segmentation, CSWin transformer, CNN, multi-head self-attention, dilated-uper decoder

Abstract

Convolutional Neural Networks have demonstrated exceptional effectiveness in the field of medical image segmentation by effectively capturing intricate local details such as edges and textures. But still, their limited domain of view often impedes comprehensive representation of global information. Transformers, on the other hand, have shown promise in modeling long-range dependencies, yet, Convolutional Neural Networks occasionally face challenges in effectively capturing high-level spatial features. An ideal segmentation model ought to effectively harness both local and global features to achieve precision and semantic accuracy. This article introduces a novel Cross Shaped Window Transformer framework, employing U-shaped network architecture. This network combines a Convolutional Neural Network encoder with a Multi-Head Self-Attention based CNN decoder. Within the CNN encoder, a transformer path is integrated with a shifted window mechanism, enhancing the representation of both local and global information, thus ensuring robust medical image segmentation. The encoder's skip connections are reinstated using a Multi-Head Self-Attention decoder. To decode a wide range of features and manage distortions in local details, a dilated-Uper decoder is introduced. The Synapse dataset is utilized to assess the effectiveness of the proposed method, revealing that it surpasses existing approaches with an impressive accuracy of approximately 93%.

1. Introduction

Medical image segmentation is a pivotal aspect in computer-aided diagnosis and smart medicine, enhancing diagnostic accuracy and efficiency. It involves delineating structures within images, crucial for tasks like liver and brain tumor segmentation, optic disc delineation, lung and cardiac image segmentation, among others [1]. Different imaging techniques—X-ray, CT, MRI, and ultrasound—have revolutionized medical diagnosis and treatment planning, each with its own advantages and drawbacks [2]. Historically, medical image segmentation relied on techniques like template matching, edge detection, machine learning, statistical models, and active contours. These methods had varying degrees of success but struggled with the unique challenges posed by

medical images, which often suffer from issues like blurriness, noise, and low contrast. As deep learning techniques have advanced, convolutional neural networks have developed as a powerful tool for feature representation and medical image segmentation. CNNs excel at handling image noise, blur, and contrast issues, providing superior segmentation results [3].

Semantic image segmentation, the goal is to classify every pixel in the image, has gained popularity with encoder-decoder structures, such as Fully Convolutional Networks (FCN) [4], U-Net [5], and Deeplab [6]. An encoder is utilized in these structures to extract image features, while a decoder is employed to restore these features back to the original image size, producing segmentation results. The U-Net, unlike typical CNNs used in medical image segmentation, resolves critical issues by adopting a symmetrical structure and integrating skip connections [5]. Unlike regular image segmentation, medical images often consist noise and exhibit unclear boundaries, making object detection solely relying on low-level features a challenging task. Similarly, accurate boundary determination through image semantic features alone is hindered by the shortage of detailed image information. Effectively U-Net overcomes these challenges by merging low-level and high-level image features through skip connections, seamlessly merging low-resolution and high-resolution feature maps [7]. This approach has become the gold standard in most medical image segmentation tasks, sparking numerous significant advancements. However, the localized receptive field of convolutional encoders limits their capacity to handle long-range dependencies and capture global context in medical images.

In response to the triumph of the Vision Transformer (ViT), transformers have lately been tailored for medical imaging to deliver top-tier segmentation [8]. These transformer-based methods analyze the input image or patch by dividing it into sub patches and processing them sequentially, instead of examining the entire input all at once. The primary strength of transformers lies in their capacity to model long-range dependencies using the self-attention mechanism, engaging with every pixel in the image, in contrast to CNNs with their constrained domain of view. This global perspective proves invaluable in medical image segmentation, where contextual information from different parts of the image holds significance. Nevertheless, ViT's computational intensity and its struggle to capture local information, particularly in high-resolution medical data, have led to the proposal of the CSWin Transformer [9]. This transformer minimizes computational redundancy through a shifted window scheme, proving to be robust in high-resolution medical data applications. Beyond preserving global data, the shifted window method also improves the capture of local details. Given the critical significance of accurate segmentation in medical imaging, particularly in tasks such as tumor and multi-organ segmentation, the ability to focus on intricate details is especially advantageous.

This paper introduces a cross-shaped window (CSWin) Transformer based CNN encoder and Multi-Head Self-Attention based CNN decoder for robust medical segmentation. The major contributions of this model are elaborated in the below section.

- The U-shaped CNN employs an enhanced CSWin transformer as its encoder, segmenting images based on multi-scale features extracted from the transformer.
- Within the CSWin transformer, Locally-enhanced Positional Encoding significantly enhances the handling of local positional data's, outperforming existing encoding schemes.
- The inclusion of the multi-head self-attention structure is crucial for reliable extraction of global and local features.
- Additionally, proposed approach utilizes a dilated-Uper decoder, adept at decoding a broad spectrum of features and addressing local detail deformations.

Organization. The remaining sections of the paper are structured as follows: Part 2 shows the overview of recent medical image segmentation works. Part 3 provides a comprehensive description of the proposed methodology. Part 4 provides the experimental outcomes and discussion while the article is terminated in part 5.

2. Related Work

In medical image segmentation Unet architecture is utilized as the baseline architecture. Still there is a restriction with Unet-based and traditional Unet networks that prevents them from fully utilizing the result of the convolutional units in the node. To overcome this Tran et al., [10] introduced a novel network model TMD-Unet with three leading enhancements in compression with Unet (1) adjusting (2) dilating (3) integrating with seven datasets. The network's drawback is that, as the number of convolutional filters increases, the network size and calculation time substantially rise. The proposed TMD-UNet is absolutely same like UNet and UNet++. However, the proposed method has better efficiency. For accurate diagnosis, small targets segmentation is crucial in 3D medical image segmentation for that Jiang et al., [11] introduced the 3D medical image segmentation using APAUNet, for axis projection attention networks (APAUNet). By enhancing the boundaries and sizes of APAUNet, it demonstrates the significant advantage it is able to provide in medical image segmentation, particularly for small targets. In this small target segmentation this method was unsuccessful for exploring and designing the fusion strategies.

J. of Soft Computing and Data Mining Vol. 5 No. 1 (2024) p. 57-69

Zhao et al., [12] proposed a brain tumor segmentation network called MM-UNet, which combines multiple modalities, has been developed. This network utilizes a structure with multiple encoders and a single decoder to extract features from various characteristics of the brain tumor, ranging from basic range characteristics to hybrid attention blocks. To produce exact segmentation output these encoders remove the image characteristics. In medical image analysis community, segmentation of hepatic arteries from CT images is essential for surgical planning. Because of the complicated structure and minimal backdrop contrary, automated hepatic vessel segmentation remains extremely difficult. Wu et al., [13] suggested inductive Biased Multi-Head Attention Vessel Net (IBIMHAV-Net) by linking the self-attention and convolution to utilize their advantages and increasing the SWin transformer to three dimensions. Inductive Biased multi-head attention (IB-MSA) is introduced to improve liver vessel segmentation results. This method does not provide accurate results and fails to support multitasking techniques to mitigate the detrimental impact of liver tumors.

In medical imaging, segmentation of complex and low-contrast structures remains challenging. Petit et al., [14] introduced the U-Transformer network connects with a U-shaped architecture with cross- and self-attention Transformers to interact with spatial dependencies and useful for abdominal and complex organ segmentations. This is achieved by fusing two levels of attention mechanisms: In U-Net decoding, self-attention modules take advantage of encoder features interact globally, while non-semantic features are filtered out by skip connections. This approach does not function with medical image tasks like MRI or US images in U-Transformer 3D networks. To address the issue Li et al., [15] introduced Clinical Named Entity Recognition (CNER) technique depend on multi-head self-attention (MHSA) to extract identification efficiency and compensate for the distant dependency loss flaw by extracting multiple levels as well as more thorough text features. To overcome these problems Chinese CNER model multi-head self-attention mechanism merged with BILSTM-CRF was introduced. BILSTM combines the char and word vector to obtain the feature extraction. When applied to different data sets, this method was unable to evaluate the scalability and generalization. A large number of people have died over the past few years because of COVID-19. For early treatment, it is crucial to make an accurate diagnosis.to make an accurate treatment Wang et al., [16] proposed an improved PSSPNN model for identifying COVID-19, secondary pulmonary, pneumonia. Community-captured, tuberculosis, and healthy participants. There are five developments in PSSPNN to exceed max pooling and average pooling. Inspired by VGG-1 a stochastic pooling was introduced. This approach will make diagnosing COVID-19 cases faster and more accurate for radiologists.

The majority of the current approaches mainly depend on a CNN, which is limited in its ability to capture the global context by the localized nature of the convolution process. Urged by the remarkable global modeling powers given by the Swin transformer, He et al., [17] presented the ST-U-shaped network (UNet), a novel semantic segmentation framework for Remote Sensing (RS) images. The ST-UNet is the outcome of this framework's smooth integration of the Swin transformer into the conventional CNN-based UNet. The power of the CNN and Swin transformer are combined in parallel by the novel dual encoder construction introduced by the ST-UNet. In the Swin transformer block, a spatial interaction module (SIM) is utilized to improve the representation capacity of obscured objects by encoding spatial information via pixel-level correlation. To decrease the loss of finer details and diminish smaller-scale features during the patch token downsampling procedure of the Swin transformer, a feature compression module (FCM) is also employed. This improvement greatly raises the accuracy of small-scale ground object segmentation. In order to create a link between the two encoders, overall dependencies from the Swin transformer are hierarchically integrated into the CNN features using a relational aggregation module (RAM).

Overall semantic information can be modeled with the help of the Transformer-based model, which is self-aware. Nevertheless, the spatial information contained in each patch is ignored by the Transformer's present patch computing mechanism. The STransFuse model was presented by Gao et al., [18] as a unique method for semantic segmentation in remote sensing images in order to address these issues. The Transformer and CNN strengths are combined in this model to increase the segmentation perfection for a range of remote sensing images. The coarse-grained and fine-grained representations of features are extracted using a staged model, as opposed to the earlier methods that merged the Transformer model. In order to optimize the features acquired at various phases, an adaptable fusion module is developed. This module uses a self-attentive technique to intelligently combine the semantic information among the features at various scales. On the Vaihingen dataset, the suggested model outperforms the baseline by 1.36% in overall accuracy (OA), while on the Potsdam dataset, it outperforms the baseline by 1.27% in OA.

In contrast, Zhang et al. [19] presented a hybrid approach for semantic segmentation of remote sensing images, particularly for very high resolution (VHR) imaging, which incorporates a transformer and CNN. This model uses an encoder-decoder architecture, with the encoder section using a new universal support, the Swin transformer, to efficiently describe large spatial dependencies and extract features. The decoder section applies CNN-based models' proven tactics and useful building blocks to the segmentation of RS images. In order to capture multi-scale context, the architecture uses an atrous spatial pyramid pooling block based on depth wise separable convolution (SASPP). A U-shaped decoder has developed to increase the feature maps' size.

Furthermore, in order to maintain local data's and improve the transfer of multi-scale features, the encoder and decoder feature maps have three skip connections each with the similar size. Squeeze-and-excite (SE) channel attention is added to the segmentation process in order to enhance the characteristics. Moreover, to supply edge restrictions for semantic segmentation, a second boundary detection branch is incorporated. Ablation tests on the Vaihingen and Potsdam benchmarks of the International Society for Photogrammetry and Remote Sensing (ISPRS) were used to assess the efficacy of different network components in great detail.

While these efforts have shown promising outcomes, the emphasis has primarily been on developing a robust encoder through the integration of CNN with the Swin transformer. Insufficient attention has been given to the thoughtful design of the decoder, which is equally critical for effective image segmentation. Additionally, RS images often feature extensive spatial dimensions, whereas the Swin transformer typically requires smaller-scale inputs. To address these challenges, this paper presents an innovative framework known as CSwin-transformer-based bi-decoder transformer for semantic segmentation of RS images.

3. Methods

This article presents a unique approach that merges a CNN encoder with a CNN decoder based on Multi-Head Self-Attention for reliable medical image segmentation. The framework incorporates CSWin Transformer, which utilizes a shifted window mechanism to enhance the representation of both local and global information. Additionally, (LePE) is introduced to enhance the accuracy of local positional data, surpassing existing encoding schemes. To address local detail deformations, the Dilated-Uper Decoder is introduced. Through these methods, the framework guarantees robust medical image segmentation with exceptional accuracy.

3.1 Outline of the Framework

Fig. 1 depicts the CSWin transformer with CNN encoder and multi head self-attention decoder for medical segmentation network. The proposed method consists of two paths a U-shaped CNN encoder path and the shifted window transformer path. Sequentially the data's in the CNN encoder encodes though the down sampling and convolution functions. The input image traverses through the patch partition layer to downscale dimensions, allowing for the visualization of high-level features via a convolution operation, before being directed into the transformer blocks. Subsequently, these processed features are directed into the transformer blocks. At each level, data's from two paths is merged using addition operations, forming a combined set of information that is then conveyed to the CNN-based decoder for the final segmentation prediction. In this proposed method CSWin transformer block contains cross shaped window self-attention layer. Afterwards, self-attention mechanism is used to process the transformer block. As a usual encoder-decoder structure, the proposed CSWin transformer is utilized to encode the depth of the feature extraction, internal correlation mapping and decodes to reconstruct the feature maps. Especially, the images are converted by ITM to provide correct inputs and compute the attention maps to obtain local and global illustrations of images. Hence it takes an output slightly various from the decoder which is up-sampled and concatenated output of CSWin. Dilated-Uper decoder is used to renovate the concealed illustrations of the CSWin encoder.

J. of Soft Computing and Data Mining Vol. 5 No. 1 (2024) p. 57-69

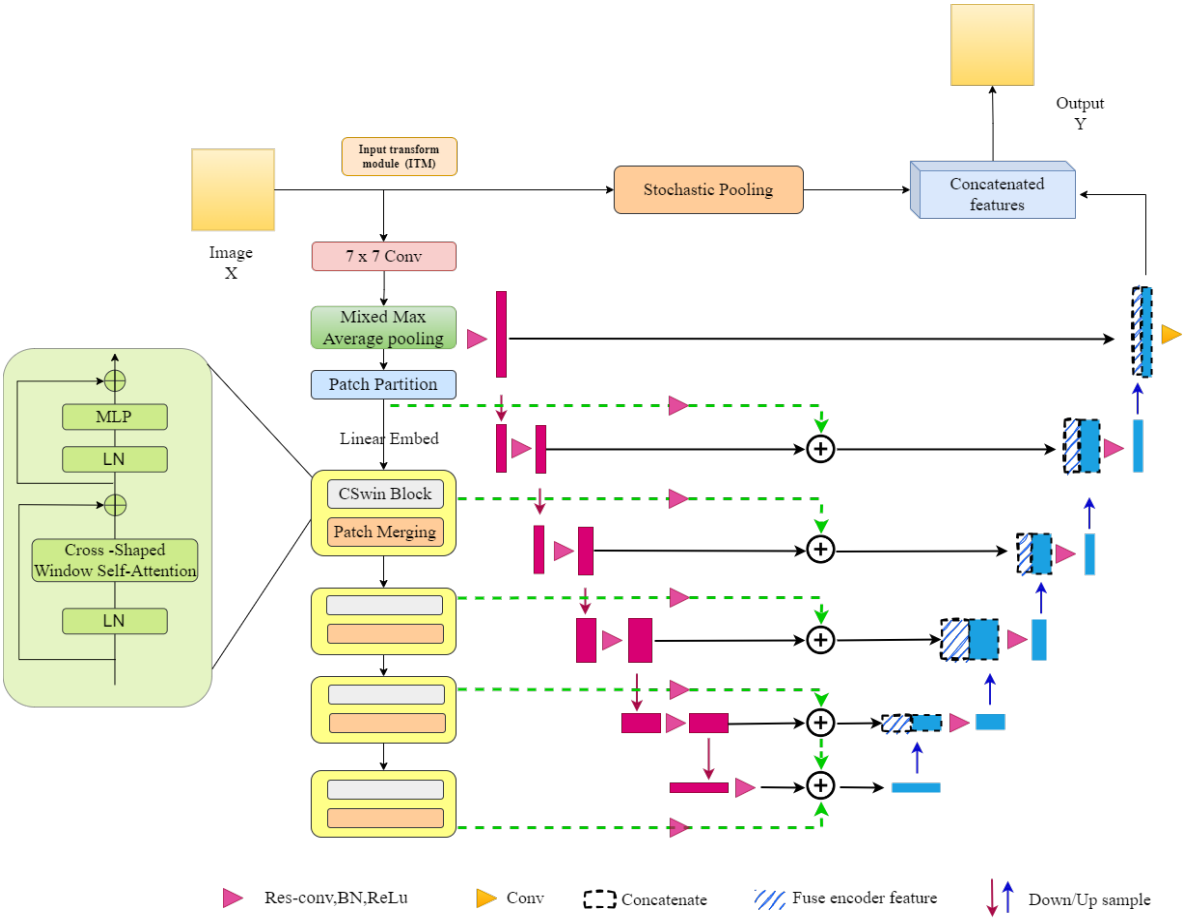

Fig. 1 *Proposed model framework*

3.2 CSWin Transformer Block

In Transformer architecture, a critical challenge lies in computationally intensive nature of global self-attention, while local self-attention frequently restricts the interaction scope of every token. To handle this problem our proposed model utilizes the Cross-Shaped Window self-attention mechanism, operates in both vertical and horizontal stripes concurrently to form the cross-shaped window. This approach involves dividing the input feature into stripes of uniform width, allowing for a thorough mathematical review the impact of stripe width. By adjusting the stripe width across various layers of the Transformer network, we strike a balance, ensuring robust modeling capacity while containing computational costs. Furthermore, a Locally-enhanced Positional Encoding (LePE) is presented which effectively manages local positional data compared to current encoding methods. LePE inherently accommodates various input resolutions, proving particularly adept for downstream tasks. Embedded within a hierarchical structure, the CSWin Transformer showcases competitive performance across prevalent vision tasks.

The CSWin transformer has proper execution for computing images and decreases the computational redundancies hierarchically through shifted window. Based on the improvements of Vision Transformer (ViT) it is difficult to capture high-resolution medical data. Unlike (ViT) a non-overlapping local window are created for effective path interaction modelling and CSWin transformer patches feeds them directly. Afterwards the transformer patches are processed by the Multi-Head Self-Attention mechanism. In addition to preserve global information high-resolution medical data are used to provide robust segmentations in medical field. In certain medical image segmentation tasks, the CSWin approach can help but it cannot match the local specificity of a carefully designed CNN encoder, since the fine details of medical images are often crucial. This suggests combining CNN encoder and CSWin transformer to achieve the best results [20].

3.3 CSWin Transformer Based CNN Encoder

The backbone of vision task is transformer based cross-shaped window (CSWin). CSWin transformer is typically utilized for feature extraction and improves the efficiency of transformer and CNNs. The shifted window block

and window block are merged by the CSWin transformer, which gives the CSWin encoder access to both local and global feature map data. The CSWin encoder integrates the patches at each level for a hierarchical multi-scale framework [9].

For an input image with the size of $h \times w \times 3$, the advantage of convergent overlaid convolutional token embedding (7×7convolutional layer with 4 strides) is used in order to acquire $\frac{h}{4} \times \frac{w}{4}$ patch tokens, and every tokens has C-dimensions. Convolutional stems are utilized for the replacement of patchily stems with CSWin transformer to enhance training efficiency and maintain overall stability. These entire network has four steps to generate a hierarchical representation. Among two adjacent stages the convolutional layer (3 ×3, stride 2) is utilized to increase the channel dimension for double time and decrease the number of tokens. Thus, the framed characteristic maps has $\frac{h}{2^{i+1}} \times \frac{w}{2^{i+1}}$ tokens for the i^{th} step which is same as VGG/ResNet the backbone of conventional CNN. In order to preserve several tokens, each and every stage contains a sequence of N_i CSWin transformer block. The CSWin transformer block features two distinct heads it swaps the self-attention mechanism with cross shaped window self-attention. CSWin Transformer based CNN encoder is included as a parallel component to the self-attention segment as a way to implement the regional inductive bias [21].

3.4 Module for Input Transformation

CSWin transformers require small-scale inputs, so the Input Transform Module (ITM) was developed to manage this issue. Instead of sampling the feature maps directly, it has been demonstrated that splitting the sampling process into numerous steps can produce better results. First two steps if ITM reduces the input size through down-sampling. ITM extracts features from the input images with different scales, and then down-samples and up-samples these segments to maximize the semantic and appearance information. Up-sampling is done by bilinear interpolation and up-sampling is done by transformer block it consist of two groups of batch normalization layer and convolutional layer. A mixed max average pooling and stochastic pooling is used to minimize the size of feature maps. In stochastic pooling, pooled map responses are selected by sampling from a multinomial distribution based on the activations of each pooled region. The pooled activation is then simply a_l,

$$\delta_j = a_l \ \ where\ l \approx P(p_1 \dots \dots, P\left|R_j\right|) \tag{1}$$

From equation 1 the selected samples from the multinomial distribution based on p to pick a location l within the region. Comparing this stochastic pooling method to max-pooling and average-pooling, it exhibits reduced training and testing errors. The mixed max average pooling estimates the spatial information and decreases the dimensions of future map. As a result, stochastic pooling can depict multiple modalities of activation within a region.

For the given input image $X \in R^{h \times w \times 3}$ is initially converted into feature map with $\frac{h}{4} \times \frac{w}{4} \times c$. Where, c is the concatenated layer. The down-sampled feature map will obtain in four different scale. At the end these down-sampled feature maps are extended to $\frac{h}{4} \times \frac{w}{4}$ and then concatenated to make the final output Y used by CSWin transformer encoders. A feature map Y is first divided by two until fed into a CSWin transformer encoder. Every patch is considered as token and the patch size is4×4. Afterwards, the separated patches are predicted into patch embedded module [21].

3.5 Cross-Shaped Window Self-Attention

The primary full self-attention mechanism's computing complication is equal to the size of feature map, in spite of its great long-range context modeling capacity. As such, it will have extremely high compute costs for vision tasks like segmentation and object detection that require maximum outcome in feature maps as input. An easy way to mitigate this problem is to use self-attention in a local attention window as well as haloes or shifted windows to enhance the receptive field. To attain global receptive filed, additional blocks must be stacked, and the token within every transformer block has a limited attention area. A CSWin - Cross-shaped window self-attention mechanism, which expands the attention region more virtually, achieves global self-attention. It works by executing self-attention in parallel vertical and horizontal stripes to create a cross-shaped window [9].

3.6 Multi-Head Self-Attention Based CNN Decoder

Multi-head attention is a multi-way attention mechanism that runs numerous times simultaneously. A linear transformation is performed on the independent attention outputs to obtain the expected dimension by concatenating them. The main component of the transformer is a Multi-Head Self-Attention (MHSA) module that can be accessed using a remote, which combines the representation subspaces, the model can infer attention jointly. The MHSA used in the CNN transformer decoder. In order to segment semantically, the hierarchical

J. of Soft Computing and Data Mining Vol. 5 No. 1 (2024) p. 57-69

feature maps are sent to the CNN decoder by the CSWin transformer encoder. In this article 4 heads are used and the dimension of the multi-head is not demonstrated in the subsequent formation. Let, $P \epsilon J^{c\times h\times w}$ is considered as an input characteristic map.

Where, the height and width are denoted as w and h, c is numerous channels. In equation 2 M, R, K $\in J^{d\times h\times w}$ d denotes the dimension of every head and M, R, K is compressed into segments along with size n×d, where n=h w. the final output is evaluated as dot product.

$$Attention(M,R,K) = \underbrace{softmax(\frac{MR^T}{\sqrt{d}})K}_{X} \tag{2}$$

The weights are then determined by context aggregating matrix, which is utilized to extract context information from the values. For a better characteristic aggregation context aggregation is easily moulded to the input context. The estimation of pair-wise attention is immensely excess and ineffective. In equation 3 the primary concept is to design key and value using two estimation: R, K $\in J^{n\times d}$ into low dimensional embedding: $\overline{R}$, $\overline{K}$ $\in J^{r\times d}$ where R=h w << n, h and w, are decreased after sub-sampling.

$$Attention\left(M,\overline{R},\overline{K}\right) = \underbrace{Softmax\left(\frac{M\overline{K}}{\sqrt{d}}\right)}_{\overline{X}:n\times r} \underbrace{\overline{K}}_{r\times d} \tag{3}$$

The method employs a 7×7 convolution followed by bilinear interpolation to reduce the size of the characteristic map, effectively down sampling it. The up sampled and concatenated output of the CSWin encoder is used as the input for a single multi-head self-attention CNN decoder. This decoder provides enhanced global context information, offering more detailed decoding for the input characteristics. While the concepts of both encoder and decoder are akin, they necessitate two inputs: the low-resolution information sourced from the decoder and the high-resolution characteristics obtained from the encoder's skip connections [5].

3.7 Dilated–Uper Decoder

The Dilated-Uper decoder is proposed to decode a broad spectrum of features while managing local detail distortions. By carefully adjusting dilation rates, it ensures accurate segmentation results in medical imaging.

In semantic segmentation dilated convolution is commonly employed to extend the limit of convolutional neural networks. To grab the wide-range dependencies among the pixels and dilated convolution it accept the ITM and CSWin transformer encoder used in dilated decoder to increase the global context. The mixed max average pooling and stochastic pooling is used to renovate the concatenated feature on the basis of dilated convolution to expand the accessible dilated field. [20] The mixed max average pooling provides the highest feature map inputs which is pooled with the output of other three feature maps. The mixed max average pooling results are linked to create a feature map with similar dimension. These feature maps are expanded with 7×7 convolution for smoothing and dimension matching. In dilated Uper decoder additional convolutions are already summed for connecting the outputs to generate the hierarchical dilated convolution. Dilated convolution fills the convolution kernel with 0's to produce better accessible area. 1 dimension dilated convolution formula is described as,

$$p(i) = \sum_l^L j(i + r.l)r(l) \tag{4}$$

In equation 4 $p(i)$ and $j(i)$ represents the input and output features, x is the 1D vector, r is the rate of dilation and $r(l)$ represents the dilated filter. The significant hybrid dilated convolution (HDC) is used to organize the dilated convolutions with hybrid dilation rates which completely cover the input feature map without holes [21].

4. Result and Discussion

This section provides an overview of the dataset utilized in this introduced method. The experimental setup was performed to assess the effectiveness of the introduced approach.

4.1 Datasets

In this method synapse dataset is used to evaluate the CSWin transformer-CNN encoder and Multi-Head Self-Attention based CNN Decoder model. And it is implemented using python. This model was compared with the performance of several other existing methods [22].

4.1.1 Synapse Multi-organ Dataset

The implemented CSWin transformer -CNN encoder and Multi-Head Self-Attention based CNN Decoder model is evaluated using synapse multi-organ dataset for segmentation. There are 3779 axial abdominal images included in the dataset, which comprises 30 cases.

4.2 Evaluation Metrics

Numerous evaluation metrics applied in this model is Dice Similarity Coefficient (DSC), the 95th percentile Hausdroff Distance (HD95), Mean accuracy (mAcc), Intersection over union (IoU), Mean intersection over union (mIoU), precision, accuracy, sensitivity, and F1 score.

i) Dice Similarity Coefficient (DSC)

The most used similarity evaluation function is dice coefficient. It is typically employed to determine how similar or comparable two samples are to one another. The range of its value spans from 0 to 1.To improve the segmentation its value must be nearer to 1. For the given two set A & B the metric is defined as,

$$DSC = \frac{2*|A \cap B|}{|A|+|B|} \tag{5}$$

Where, ground truth image pixel is represented as A and detected pixel image is denoted as B. the detected pixel image must be high.

ii) The 95th percentile Hausdroff distance (HD95)

Hausdroff distance is determined as the distance among the detected ground truth images and pixel images. Highly segmented accuracy is represented by the smaller value of Hausdroff distance.

$$HD(S,L) = \max\left\{\left\{k_{s\in S}^{th} min_{g\in G\|S-L\|}\right\}, \left\{k_{g\in G}^{th} min_{S\in S\|S-S\|}\right\},\right\} \tag{6}$$

Where, the ground truth image is represented as G and segmented image is denoted in the form of s.

iii) Intersection over union (IoU)

It is a region-based metric it also obtains false alarms and ignored cases. IoU is measured by dividing overlap among the identified and ground truth annotation by the union of these. The mathematical equation for the IOU is given as

$$IOU = \frac{Tr_p}{Tr_p+Fa_{pv}+Fa_n} \tag{7}$$

iv) Mean interaction over union (mIoU)

It defines the ratio of the total of true positives, false positives, and true negatives to true positives.

$$mIoU = \frac{1}{N}\sum_i^N \frac{Tr_p}{Tr_p+Fa_p+Fa_n} \tag{8}$$

Where, N is the numerous classes.

v) Mean accuracy (mAcc)

The ratio between the sum of true positive and the sum of false negative and true positive is equivalent to the accuracy in classification tasks.

$$mAcc = \frac{1}{N}\sum_i^N \frac{Tr_p}{Tr_p+Fa_n} \tag{9}$$

vi) Precision

It is the ratio among the true positive and sum of true positive and false positive.

$$precision = \frac{Tr_p}{Tr_p+Fa_p} \tag{10}$$

vii) Accuracy

The accuracy value is determined by calculating the ratio between the sum of true positive and true negative, and the sum of false negative, false positive, and true positive.

$$Accuracy = \frac{Tr_p+Tr_n}{Tr_P+Tr_n+Fa_p+Fa_n} \tag{11}$$

J. of Soft Computing and Data Mining Vol. 5 No. 1 (2024) p. 57-69

viii) Recall
This is the ratio of true positives to the sum of true positives and false positives.

$$Recall = \frac{Tr_p}{Tr_p + Fa_n} \tag{12}$$

ix) Sensitivity
It is the ratio between the true positives and the sum of true positives and false negatives.

$$\frac{Tr_p}{Tr_p + Fa_n} \tag{13}$$

4.3 Experimental Results

The overall outcome of this model is analyzed between the proposed CSWin transformer -CNN encoder and Multi-Head Self-Attention based CNN Decoder model with previous SOTA transformer-based segmentation methods, such as SAS-UNet [23], U-Net [24], Trans-Unet [25] and APAUnet [11] with the performance metrics namely, Intersection over Union (IoU), 95% Hausdorff Distance (HD), Dice-Similarity Coefficient (DSC), mean accuracy (mAcc), mean intersection over union (mIoU), recall, accuracy, precision and sensitivity. The Synapse dataset, a publicly accessible multi-organ segmentation dataset which contains 30 cases encompassing 3779 axial abdominal CT clinical images is utilized, for the evaluation process.

In Fig. 2, the experimental outcomes of multi-organ synapse dataset segmentation are compared with various existing techniques. From the figure it is found that the suggested approach adeptly segments complicated and fine structures, delivering highly accurate segmentation outcomes that demonstrate more robust to complex backgrounds.

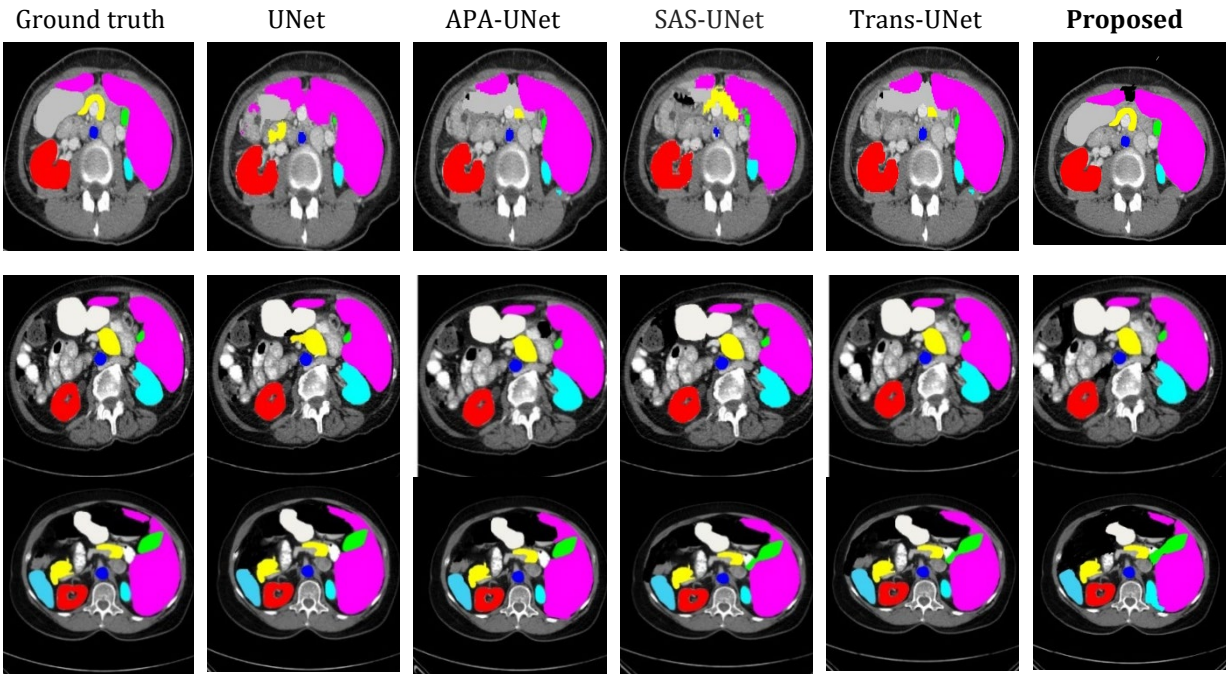

Fig. 2 *Segmentation results for synapse dataset*

Here the above table represents the image result for synapse dataset compared with existing techniques and proposed method. From the obtained data set it carries three various input images based on ground truth.The segmented images for the four existing models such as SAS-UNet [23], U-Net [24], Trans-Unet [25] and APAUnet [1] are implemented in CSWin transformer. From the above image outcomes this introduced method gives better segmentation outcomes than others.

4.4 Performance Analysis

Here an analysis has been done between performance metrics with existing and proposed model. The methods such as U-Net, APAUnet, SAS-UNet, and Trans-Unet are implemented in CSWin transformer with the evaluation

metrics namely DSC, HD95, IoU, mIoU, mAcc, precision accuracy, recall and sensitivity. This analysis has been conducted using the synapse dataset.

The following table 1 illustrates the comparative evaluation of numerous segmentation techniques based on their performance metrics in segmenting specific anatomical structures. These methods include established architectures like U-Net, APA-UNet, SAS-UNet, Trans-UNet, and a proposed CSWin-transformer. The evaluation metrics encompass Intersection over Union (IoU) for distinct anatomical structures which includes Aorta, Gallbladder, Right and Left Kidneys, Liver, and Spleen. The U-Net demonstrates robust performance across multiple structures, notably excelling in Liver segmentation with a high IoU of 91.3% and an overall Mean IoU (mIoU) of 73%. The APA-UNet, while generally effective, exhibits lower performance in the Spleen segmentation with an IoU of 14.98%. On the other hand, the SAS-UNet shows competitive performance, particularly strong in the Spleen segment with an IoU of 33.73%. The Trans-UNet portrays superior results for various structures, especially in the Aorta and Spleen. However, the CSWin-transformer proposed outperforms all other methods, showing the highest IoU across all anatomical structures, particularly excelling in Aorta, Gallbladder, Liver, and Spleen segments. It showcases significant improvements in mean accuracy and IoU metrics, indicating its superiority in accurately segmenting various anatomical structures compared to the other models evaluated. As a result of such improvements, it is possible to accurately identify boundaries of large organs as well as to identify the evolution of anatomical features within an image. The findings also indicate that dual attention is more effective for performance gain than the base structure.

Table 1 *Comparison with some existing technique on the synapse dataset*

Methods	**IoU (%)**						**Evaluation Metrics**	
	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Spleen	mIoU (%)	mAcc (%)
U-Net	80.45	72.93	67.54	72.79	91.3	28.46	73	82.51
APA-UNet	77.67	68.23	74.32	67.41	82.28	14.98	62.98	73.45
SAS-UNet	81.24	70.4	66.13	69.55	87.58	33.73	73.04	81.01
Trans-UNet	86.23	74.11	75.43	73.51	75.41	34.76	72.02	80.67
Proposed	90.10	83.23	75.43	75.44	88.11	39.26	76.08	83.55

Table 2 illustrates the evaluation of the suggested approach. From the performance analysis the suDSC metric have reached a high value. The overall value demonstrated that the accuracy rate is high. Based on the above analysis it concluded that this approach generalized well on the segmentation process among different types of medical data. The proposed model outperforms all other methods, showcasing the highest DSC and HD95 values, indicating superior overlap and reduced distances between contours.

Table 2 *Performance analysis of proposed method*

Methods	DSC	HD95	Precision	Accuracy	Recall	Sensitivity
U-Net	0.8654	0.7654	0.8122	0.7954	0.6675	0.8232
APA-UNet	0.8577	0.7845	0.8302	0.8032	0.7123	0.8367
SAS-UNet	0.8677	0.8566	0.8872	0.8122	0.7654	0.8465
Trans-UNet	0.8254	0.8108	0.8723	0.8512	0.8076	0.8698
Proposed	0.9043	0.8893	0.8967	0.9345	0.8365	0.8832

The above table offers an understanding of various segmentation models, including U-Net, APA-UNet, SAS-UNet, Trans-UNet, and a proposed model. Each model's performance is evaluated based on crucial segmentation metrics. The proposed model demonstrates a superior Dice Similarity Coefficient (DSC), indicating better spatial overlap among the predicted and ground truth masks. Its lower 95% Hausdorff Distance (HD95) suggests reduced distances between predicted and true boundaries, underscoring enhanced segmentation precision. Moreover, across metrics like precision, accuracy, recall, and sensitivity, the alternative methods are consistently outperformed by the proposed approach. These findings emphasize the pivotal role modifications within the proposed model, contributing to improved segmentation accuracy, boundary delineation, and fine detail capture crucial for precise medical image analysis and diagnosis.

5. Conclusion

This research has introduced a novel CSWin Transformer framework, which combines the strengths of Convolutional Neural Networks (CNNs) and Transformers for robust medical image segmentation. By effectively harnessing both global and local features, the model overcomes the limitations of traditional Transformers, and CNNs achieving remarkable accuracy in segmenting medical images. The proposed CSWin Transformer, a fusion

J. of Soft Computing and Data Mining Vol. 5 No. 1 (2024) p. 57-69

of CNN encoder and Multi-Head Self-Attention-based CNN decoder, emerges as a promising solution. This innovative approach integrates a transformer path with shifted window mechanism, effectively enhancing both global and local information representation, essential for robust medical image segmentation. Additionally, the introduction of a dilated-Uper decoder further enriches the model's capability to handle a wide spectrum of features and local details. From the experimental results multi-organ Synapse dataset is utilized to compare with various existing techniques for segmentation. The experimental outcomes demonstrate that the proposed model has better results in the metrics such as DSC of 0.9043%, HD95 of 0.8893%, Precision of 0.8967%, Accuracy of 0.9345%, Recall of 0.8365% and Sensitivity of 0.8832%. The future work will focus on designing a lightweight segmentation model for medical images in 3D.

Acknowledgement

I confirm that all authors listed on the title page have contributed significantly to the work, have read the manuscript, attested to the validity and legitimacy of the data and its interpretation, and agree to its submission.

Conflict of Interest

The authors declare that they have no conflict of interest.

Author Contribution

We confirm that all authors listed on the title page have contributed significantly to the work, have read the manuscript, attest to the validity and legitimacy of the data and its interpretation, and agree to its submission.

References

[1] Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., & Nandi, A. K. (2022). Medical image segmentation using deep learning: A survey, *IET Image Processing*, *16(5),* 1243-1267. https://doi.org/10.1049/ipr2.12419

[2] Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., & Rueckert, D. (2020). Deep learning for cardiac image segmentation: a review, *Frontiers in cardiovascular medicine*, *7*, 25. https://doi.org/10.3389/fcvm.2020.00025

[3] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, *25*.

[4] Tragakis, A., Kaul, C., Murray-Smith, R., & Husmeier, D. (2023). The fully convolutional transformer for medical image segmentation, In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3660-3669.

[5] Gao, Y., Zhou, M., & Metaxas, D. N. (2021, September 27–October 1). UTNet: a hybrid transformer architecture for medical image segmentation, In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference,* Strasbourg, France, Proceedings, Part III 24, pp. 61-71.

[6] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE transactions on pattern analysis and machine intelligence*, *40(4),* 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

[7] Zhang, J., Zhang, Y., Jin, Y., Xu, J., & Xu, X. (2023). Mdu-net: Multi-scale densely connected u-net for biomedical image segmentation, *Health Information Science and Systems*, *11(1),* 13. https://doi.org/10.1007/s13755-022-00204-9

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. & Uszkoreit, J., (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[9] Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D. & Guo, B. (2022). Cswin transformer: A general vision transformer backbone with cross-shaped windows, In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12124-12134.

[10] Tran, S. T., Cheng, C. H., Nguyen, T. T., Le, M. H., & Liu, D. G. (2021). Tmd-unet: Triple-unet with multi-scale input features and dense skip connection for medical image segmentation, *Healthcare, 9(1),* 54. https://doi.org/10.3390/healthcare9010054

[11] Jiang, Y., Zhang, Z., Qin, S., Guo, Y., Li, Z., & Cui, S. (2022). APAUNet: axis projection attention UNet for small target in 3D medical segmentation, In *Proceedings of the Asian Conference on Computer Vision*, pp. 283-298.

[12] Zhao, L., Ma, J., Shao, Y., Jia, C., Zhao, J., & Yuan, H. (2022). MM-UNet: A multimodality brain tumor segmentation network in MRI images, *Frontiers in oncology*, *12*, 950706. https://doi.org/10.3389/fonc.2022.950706

[13] Wu, M., Qian, Y., Liao, X., Wang, Q., & Heng, P. A. (2023). Hepatic vessel segmentation based on 3D swin-transformer with inductive biased multi-head self-attention, *BMC Medical Imaging*, *23(1)*, 91. https://doi.org/10.1186/s12880-023-01045-y

[14] Petit, O., Thome, N., Rambour, C., Themyr, L., Collins, T., & Soler, L. (2021). U-net transformer: Self and cross attention for medical image segmentation. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pp. 267-276.

[15] Li, C., & Ma, K. (2022). Entity recognition of Chinese medical text based on multi-head self-attention combined with BILSTM-CRF, *Mathematical Biosciences and Engineering*, *19(3)*, 2206-2218. https://doi.org/10.3934/mbe.2022103

[16] Wang, S. H., Zhang, Y., Cheng, X., Zhang, X., & Zhang, Y. D. (2021). PSSPNN: PatchShuffle stochastic pooling neural network for an explainable diagnosis of COVID-19 with multiple-way data augmentation, *Computational and Mathematical Methods in Medicine*, *2021*, 1-18. https://doi.org/10.1155/2021/6633755

[17] He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., & Xue, Y. (2022). Swin transformer embedding UNet for remote sensing image semantic segmentation, *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-15. https://doi.org/10.1109/TGRS.2022.3144165

[18] Gao, L., Liu, H., Yang, M., Chen, L., Wan, Y., Xiao, Z., & Qian, Y. (2021). STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation, *IEEE journal of selected topics in applied earth observations and remote sensing*, *14*, 10990-11003. https://doi.org/10.1109/JSTARS.2021.3119654

J. of Soft Computing and Data Mining Vol. 5 No. 1 (2024) p. 57-69

[19] Zhang, C., Jiang, W., Zhang, Y., Wang, W., Zhao, Q., & Wang, C. (2022). Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery, *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-20. https://doi.org/10.1109/TGRS.2022.3144894

[20] Li, H., Liu, H., Hu, D., Yao, X., Wang, J., & Oguz, I. (2023). CATS v2: Hybrid encoders for robust medical segmentation. *arXiv preprint arXiv:2308.06377*.

[21] Liu, Y., Zhang, Y., Wang, Y., & Mei, S. (2023). BiTSRS: A bi-decoder transformer segmentor for high-spatial-resolution remote sensing images, *Remote Sensing*, *15(3),* 840. https://doi.org/10.3390/rs15030840

[22] The synapse dataset was taken from https://www.synapse.org/#!Synapse:syn3193805/wiki/217789 assessed on October, 2023.

[23] Samal, S., Gadekellu, T. R., Rajput, P., Zhang, Y. D., & Balabantaray, B. K. (2023, May). SAS-UNet: Modified encoder-decoder network for the segmentation of obscenity in images, In *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW)*, pp. 45-51.

[24] Yin, X. X., Sun, L., Fu, Y., Lu, R., & Zhang, Y. (2022). U-Net-Based medical image segmentation. *Journal of healthcare engineering*, *2022*. https://doi.org/10.1155/2022/4189781

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