Traitement du Signal
Vol. 42, No. 5, October, 2025, pp. 2787-2795
Journal homepage: http://iieta.org/journals/ts

APTA-UNET: A Novel Framework for Precision Medical Image Segmentation

Kumaresan Velusamy[1], Nagarajan Rajendran[2*]

[1] Department of ECE, St. Michael College of Engineering & Technology, Kalayarkoil 630551, India
[2] Department of Electrical and Electronics Engineering, Syed Ammal Engineering College, Ramanathapuram 623502, India

Corresponding Author Email: rn@syedengg.ac.in

https://doi.org/10.18280/ts.420529

Received: 10 April 2025
Revised: 22 August 2025
Accepted: 24 September 2025
Available online: 31 October 2025

Keywords:
medical image segmentation, attention-augmented modules, TrioFusion Attention mechanisms, residual connection, segmentation accuracy

ABSTRACT

In modern healthcare, medical image segmentation is a critical component that identifies and localizes the anatomical structures and pathological regions within medical scans accurately. However, image segmentation is a challenging task based on anatomical variation, varying image qualities and complex spatial dependencies. These challenges lead to delayed diagnostic accuracy and treatment planning. The proposed work presented an APTA-UNet architecture with an advanced mechanism of two modules as AP module and the TA module to improve spatial and contextual awareness. In the proposed UNet, the bottleneck structure implemented an attention augmentation with Positional Embedding named as AP module that is used to distinguish positional relationships and contextual dependencies across the image. Furthermore, the proposed architecture introduces a TrioFusion Attention Module named TA module that is used as a residual connection between the encoder-decoder sides. This TriFusion Attention Module is an integration of HaloNet Attention, Axial Attention, and Non-Local Attention. This fusion is used to learn both local and global spatial relationships to enhance the ability to recognize intricate details and broader spatial dependencies. The experimental result used three medical datasets of CVC-ClinicDB, COVID-19 CT, and Breast Ultrasound to validate APTA-UNet performance. It achieves a Dice Similarity Coefficient of 0.949, 0.957, and 0.904 for the CVC-ClinicDB Lesion, COVID-19 CT, and Breast Ultrasound datasets, respectively, compared to 0.939, 0.947, and 0.88 for the best performing baseline model.

1. INTRODUCTION

In modern healthcare technologies, the medical imaging tool plays a most essential part in it [1]. This tool enables the visualization of anatomical structures and pathological conditions with higher precision. Despite important advancements, the effective utilization of these images remains a challenge based on the absolute complexity and volume of data generated. This challenge is overcome by processing a segmentation on it [2]. The segmentation is used to partition the input data into meaningful regions to enable clinical analysis and decision-making. Medical imaging segmentation is used in various applications like brain tumour, breast cancer, kidney stones, organ delineation, and treatment planning to enhance prediction accuracy [3].

In recent decades, there are numerous imaging modalities have been employed across diverse medical domains such as radiology, cardiology, oncology, and neurology. Several Techniques like X-ray, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound and positron emission tomography (PET) have provided detailed pathological and physical behaviour in an image [4, 5]. Each modality has a specific benefit that differs from others X-rays are cost-effective imaging of dense structures like bones, CT scans give internal organs cross-sectional views; MRI provides visualizing the brain's soft tissues and functions, ultrasound is used for real-time imaging capabilities and PET is used for metabolic and functional imaging in oncology. This process supports to process of an exact diagnosis and treatment planning [6].

However, these modalities have several limitations: Manual understanding is hard and consumes more time [7]. The complex in nature has noise, artefacts and overlapping structures that complicate the process of segmentation. To overcome these issues, some traditional methods like thresholding, edge detection and region growing are applied [8]. However, these methods are not good at performance and often struggle to achieve high accuracy in the presence of irregular boundaries. These challenges have to be driven by some reliable approaches to ensure robustness and accuracy in medical imaging [9].

Based on these considerations, deep learning (DL) has emerged for medical imaging segmentation [10]. In recent times, DL has had a better performance in all the sectors which can handle larger data and complex patterns with better accuracy [11]. The convolutional neural network (CNNs) model is a popular DL method that automatically learns hierarchical features from raw data to adapt to complex patterns and variations in medical images. Also, a few Techniques like U-Net, Generative Adversarial Networks (GANs) and so on also have special segmenting behaviour with high accuracy.

In recent times, there are numerous enhancements have also been added to a DL model to achieve an effective segmentation that helps for an earlier disease prediction. In this work, a novel APTA-UNet segmentation model is presented using various datasets. This APTA-UNet method has two main modules, as AP module and the TA Module, that show its uniqueness from other UNets to attain effectiveness. The key enhancement modules are given in the following.

i) **AP Module:** This module improves the capturing of long-range dependencies and spatial relationships, leading to more accurate and detailed segmentations.
ii) **TA Module:** This module enhances feature representation by capturing diverse spatial dependencies, increasing the model's ability to handle complex image patterns.

The remaining part of this work contributed to: Section 2 described related works and section 3 presented the proposed methodology of APTA-UNet. Section 4 described the experiment result and discussion. Finally, a conclusion summary is given in section 5.

2. RELATED WORKS

Azad et al. [12] presented a U-Net model that has received tremendous attention and has backbone, bottleneck or skip connections with a Transformer architecture. It also addressed a probabilistic prediction of the segmentation map.

Chen et al. [13] proposed a novel Dense-Res-Inception Network (DRINet), which comprises a convolutional block with dense connections, a deconvolutional block with residual inception modules, and an unspooling block. This model achieved a higher level of accuracy compared with other existing models.

Wang et al. [14] presented CNN segmentation for both 2-D segmentation of multiple organs from fetal MRI and brain tumour core for training with higher accuracy. Also, Feng et al. [15] presented a novel Context Pyramid Fusion Network (CPFNet) designed for multiple global pyramid guidance (GPG) among the encoder and the decoder initially. It also provides various data from the global context to construct a skip-connection.

Huang et al. [16] explored a novel UNet 3+ that has full-scale skip connections which were minimum details with maximum semantics from feature maps. It also has deep supervision to learn the hierarchical representations. This method minimised the parameters to improve the efficiency of computation.

Cao et al. [17] developed a Swin-Unet that has a Transformer for the local-global semantic feature. It used shifted windows as the encoder and symmetric decoder for global context and patch expanding layer to perform the up-sampling operation.

Weng et al. [18] implemented a Neural architecture search (NAS)-Unet that is stacked by the DownSC and UpSC process. It is updated by a differential architecture strategy simultaneously. It attained a good segmentation result on various datasets.

Yan et al. [19] implemented an Axial Fusion Transformer UNet (AFTer-UNet) that has convolutional layers and transformers used for a long-range signal. This model has minimum parameters with less GPU memory for training models.

Park et al. [20] presented a Unicorn model that has multiple time series images which have a bottleneck layer with convolution operations to capture the spatiotemporal variables. This method attained a higher MAE which was 12% better than previous models.

Huang et al. [21] compared various models of UNet, Res-UNet, Attention Res-UNet, and nnUNet by assessing their results in brain tumour, polyp and multi-class heart segmentation tasks. The nnUNet shows superior overall performance across the experiments consistently.

Alrfou et al. [22] developed parallel CNN and Transformer encoders to extract the higher transfer-learning benefits. It is pre-trained on materials microscopy images with an accuracy of 79.9% in Dice Similarity Coefficient (DSC), respectively.

Liao et al. [23] explored Lightweight Mamba UNet (LightM-UNet) that has Residual Vision Mamba to capture deep semantic features. This model processed a long-range spatial dependency effectively with a linear computational complexity.

Al Qurri et al. [24] designed a Three-Level Attention (TLA) model that has Attention Gate (AG), channel attention and spatial normalization in it. The AG presented structural information and attention used for interdependencies between channels.

Khan et al. [25] experimented with a Hybrid Attention-Based Residual Unet (HA-RUnet) that has a residual block to capture low- and high-level features from MRI volumes. This method was trained on the BraTS-2020 dataset and achieved a result of DSC of 0.867, 0.813, and 0.787 and also the sensitivity of 0.93, 0.88, and 0.83, respectively.

3. PROPOSED METHODOLOGY

In the proposed system, the medical image segmentation is done by a novel APTA-UNet architecture. The proposed UNet Architecture is given in Figure 1, which has four sections such as encoder path, bottleneck layer, decoder path and Residual connections. These layers are used to capture both low-level and high-level features effectively.

3.1 Encoder path

In the Encoder Path, the initial input image is processed by a convolutional layer. These layers are used to increase the filter sizes and decrease the spatial dimensions. These layers are used to extract low-level features like edges and textures. The Max pooling layers are used to downsample the feature maps and reduce computational cost.

3.2 Bottleneck layer

In the UNet Bottleneck layer, the proposed AP Module is used in it. The AP module is an integration of both the Attention-Augmentation and positioning embedding process. The Attention-Augmented Convolutional Module is used to capture both local and global dependencies [26]. After the third downsampling block, it operates on the activation maps obtained where the feature map dimensions are 32×32 with 128 channels. It effectively learns complex spatial relationships using four attention heads, a kernel size of 3×3 and depths of queries (dk=40) and values (dv=4) respectively. This module has a feature map that is concatenated with regular convolutional outputs from the last downsampling block to improve the data passed to the decoder.

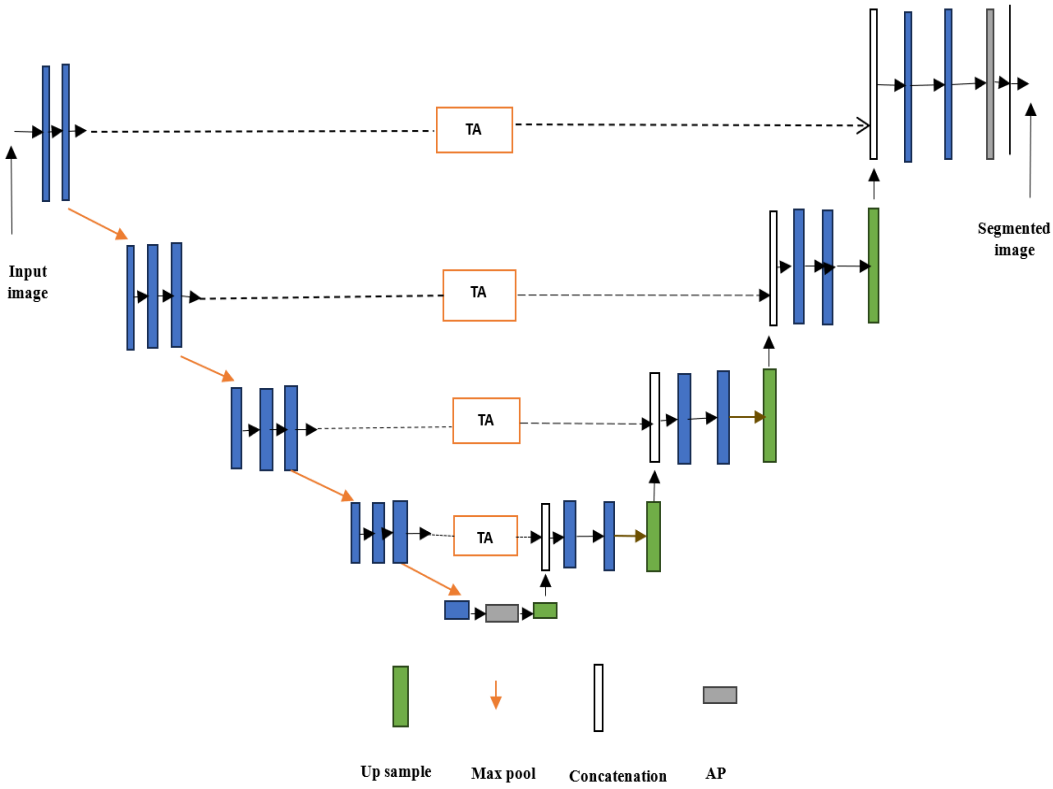

Figure 1. APTA-UNet architecture

To enhance spatial awareness, a Positional Embedding Model is supported which is used in bottleneck. These embeddings are learned using an unsupervised objective function that is inspired by the skip-gram model. It is used to reduce an error in first- and second-order proximities predicting between nodes in the feature graph. The objective function is expressed as in the following.

$$LU(P_v, G) = \sum_{v \in V} \sum_{u \in N(v)} \left[-log\sigma(p_v^T p_u) - Q.E_{u' \sim P_n(v)} \log\left(\sigma(-p_u^T p_{u'})\right) \right] \quad (1)$$

where, p_v and p_u denotes the positional embedding of node u and v, $P_n(v)$ denotes the negative sampling distribution, Q represents the number of negative samples per edge and σ indicates a nonlinear activation function.

The embeddings are computed through multiple fully connected layers, where the t-th layer is defined as:

$$p_v^t = \sigma(W_{emb}^t p_v^{t-1}) \quad (2)$$

where, W_{emb}^t indicates a weight matrix for layer t.

It is added to the feature maps before applying the attention mechanism, enabling the model to retain and utilize spatial and structural context.

To address the class imbalance inherent in medical datasets, an Inverse Class-Weighted Cross-Entropy Loss is employed that is expressed as:

$$CE(z, y) = w_y . - \log\left(\frac{exp(z_y)}{\sum_{j=1}^{C} exp(z_j)}\right) \quad (3)$$

where, $w_y = \frac{1}{\sqrt{n_y}}$, n_y denotes sample frequency of class y and C indicates the total number of classes. This weighting is used to lessen the frequency of classes that contribute more learning process. The Separate weighting is computed to train and validate datasets to account for distributional differences.

In this layer, each pixel is considered as a node in a grid. The grid structure acts as the graph for processing. Each node (pixel) in the grid is connected to its spatial neighbors which form an implicit graph where the edges represent the spatial proximity between pixels. The adjacency matrix for this grid-based graph is constructed based on the local neighborhood of each pixel. Specifically, each node is connected to pixels within a fixed neighborhood. This connection is used to capture the spatial relationships between pixels.

The objective function used for positional embedding aims to reduce the error in predicting the relative positional relationships between neighboring nodes within this grid structure. This is used for the model to capture spatial dependencies in medical images. By adapting the skip-gram model in this manner, the spatial relationships are effectively learned in grid-based image data. It supports the models to segment complex anatomical structures with varying shapes and sizes.

After the bottleneck completion, the enriched feature maps are attained using attention augmentation. The positional embeddings and adaptive loss functions ensure precise segmentation by achieving spatial awareness.

3.3 Decoder path

Here, the bottleneck feature maps are upsampled using transposed convolutions or bilinear interpolation. These layers increased the spatial dimensions of the feature maps gradually.

3.4 Residual connections

The Residual Connection is used to allow gradients to flow easily which helps for training in deeper. For a Residual

Connection among encoder and decoder channels, the novel TP module is used which is a Trio Fusion of HaloNet Attention, Axial Attention and Non-Local Attention models. Each attention mechanism-based residual connection improves local and global context learning.

In the medical image segmentation process, the TP module contributes as:

- **HaloNet** is used to ensure sharp boundary detection and fine-grained features for small lesion segmentation.
- **Axial Attention** models are used for global context efficiently to enable the network to capture large structural patterns with lower computational costs.
- **Non-local Attention** is used to refine long-range dependencies by enabling better segmentation of complex regions.

3.5 HaloNet attention

This attention model is used to localize self-attention to attain an efficient feature extraction in smaller spatial neighborhoods [27]. Defining haloed neighborhoods, it limits attention to a subset of spatial features to minimise computational complexity while preserving local details. The mechanism consists of query block size b and halo size h to define the region of interest for attention. For example, a query block of b=8 and halo size h=3 ensures a 14×14 receptive field. Also, it modified a bottleneck width multiplier and a 1×1 convolution before global average pooling for better feature representation.

The attention operation is mathematically expressed as:

$$HaloNet_{attention}\,(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \qquad (4)$$

where, Q, K and V indicate queries, keys, and values from local spatial neighborhood.

In segmentation tasks, HaloNet preserves fine-grained features used to identify boundaries and small lesions by integrating its output into U-Net's encoder layers.

3.6 Axial attention

This Attention [28] is used to simplify the global attention computation by dividing it across image axes. It is used to apply attention along one axis (rows or columns) at a time instead of attending to all pixels simultaneously. For instance, row attention (k=1) computes:

$$Axial_{attention\_row}(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \qquad (5)$$

While keeping column information independent.

Similarly, column attention (k=2) processes information along columns. This two-pass approach reduces computational complexity from $O(N^2)$ in traditional attention to $O\left(N\sqrt{N}\right)$, where N is the number of pixels.

3.7 Non-local attention

This attention model is used to capture long-range dependencies by enabling every feature map pixel to interact with all others [29]. Unlike HaloNet or Axial Attention, this method aggregates a global context directly which is expressed as:

$$Non-local\,(x) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V + x \qquad (6)$$

where, x indicates the input feature map and the residual connection $+x$ stabilizes training.

3.8 Fusion in TP model

The fusion strategy combines the outputs of HaloNet, Axial Attention, and Non-Local Attention to influence their complementary strengths. Outputs from each attention module are concatenated along the channel axis:

$$F_{fused} = Concat(F_{HaloNet}, F_{Axial}, F_{NonLocal}) \qquad (7)$$

where, $F_{HaloNet}, F_{Axial}, F_{NonLocal}$ denotes attention-augmented feature maps.

3.9 Transformation

The concatenated features are passed through a 1×1 convolution to reduce the dimensionality:

$$F_{output} = Conv_{1\times1}(F_{fused}) + F_{residual} \qquad (8)$$

where, $F_{residual}$ denotes skip connection from the encoder.

The fused features are passed through the decoder, where their rich local (HaloNet), axis-wise global (Axial) and full global (Non-Local) context improve segmentation accuracy. At last, the 1x1 convolutional layer produces the final segmentation map with each pixel classified into one of the desired classes.

By integrating these attentions, the APTA-UNet model robustly addresses class imbalance issues and captures lesion progression effectively.

4. EXPERIMENTAL RESULTS

4.1 Dataset description

- **The CVC-ClinicDB Dataset:** It is collected from Barcelona Hospital and includes 612 polyp images extracted from 31 colonoscopy videos. Each image has expert-annotated ground truth of it. The original image resolution is 383×288 pixels [30].
- **COVID-19 CT Dataset:** It was gathered on Kaggle in 2019 and contains CT lung scan images along with corresponding label data. The original image size is 512×512 pixels, but images can be resized to 256×256 as needed [30].
- **Breast Ultrasound Dataset:** It was gathered on the Kaggle dataset that contains 780 PNG images collected in 2018 from 600 women aged 25-75, with an average size of 500×500 pixels. The images are categorized into three classes: normal, benign, and malignant, with ground truth annotations included. It supports research in breast abnormality detection and classification.

For each dataset, the data is split into 80% training and 20% test sets. To augment the data, the techniques of random rotation, flipping, zooming, and translation are applied. In

terms of class distribution, the CVC-ClinicDB dataset contained 450 polyps and 162 non-polyps. The COVID-19 CT dataset had 120 positive and 380 negative cases and the Breast Ultrasound dataset included 260 normal, 290 benign, and 230 malignant images. To mitigate class imbalance, an Inverse Class-Weighted Cross-Entropy Loss is used to balance performance across all classes.

The proposed APTA-UNET was evaluated across three datasets, focusing on segmentation tasks in medical imaging in comparsion with Ground Truth (GT). Metrics used include DSC, Intersection over Union (IoU), Precision, Sensitivity (Recall), Time, and Parameters. Below is a detailed discussion of the results.

DSC: It is defined as the intersection between segmented results to the GT.

IoU: It measures the region of intersection between the predicted and original portions.

Precision: It is the proportion of accurately segmented positive pixels out of all segmented positive pixels.

Sensitivity (Recall): Sensitivity defines the capability of the model to correctly detect positive pixels.

Time (Inference Speed): The computational time was measured for inference on a single image.

Parameters: The number of trainable parameters determines the model's complexity and memory footprint.

Figure 2(a) shows the segmentation results on CVC-ClinicDB Dataset. The proposed model can absolutely differentiate lesion regions with blurred boundaries. It effectively addresses the challenge of segmenting polyps with colors similar to the background. In addition, accurately detecting polyp tissues of varying shapes, sizes, and colors. The regions and boundaries are identified clearly.

The segmentation results on the COVID-19 CT Dataset are given in Figure 2(b). The proposed model preserves more image details and produces segmentation outputs that align closely with the GT images. The segmentation outputs on the Breast Ultrasound Dataset are given in Figure 2(c).

The Figure 3 shows the training and validation accuracy of a model over 250 epochs. Initially, the curves show a steady increase which denotes that the model is learning and improving its ability to generalize. The validation accuracy closely follows the training accuracy and shows minimal overfitting. Overall, the convergence of both curves suggests that the model is performing consistently well on both training and validation data and achieves high accuracy levels by the end of the training process.

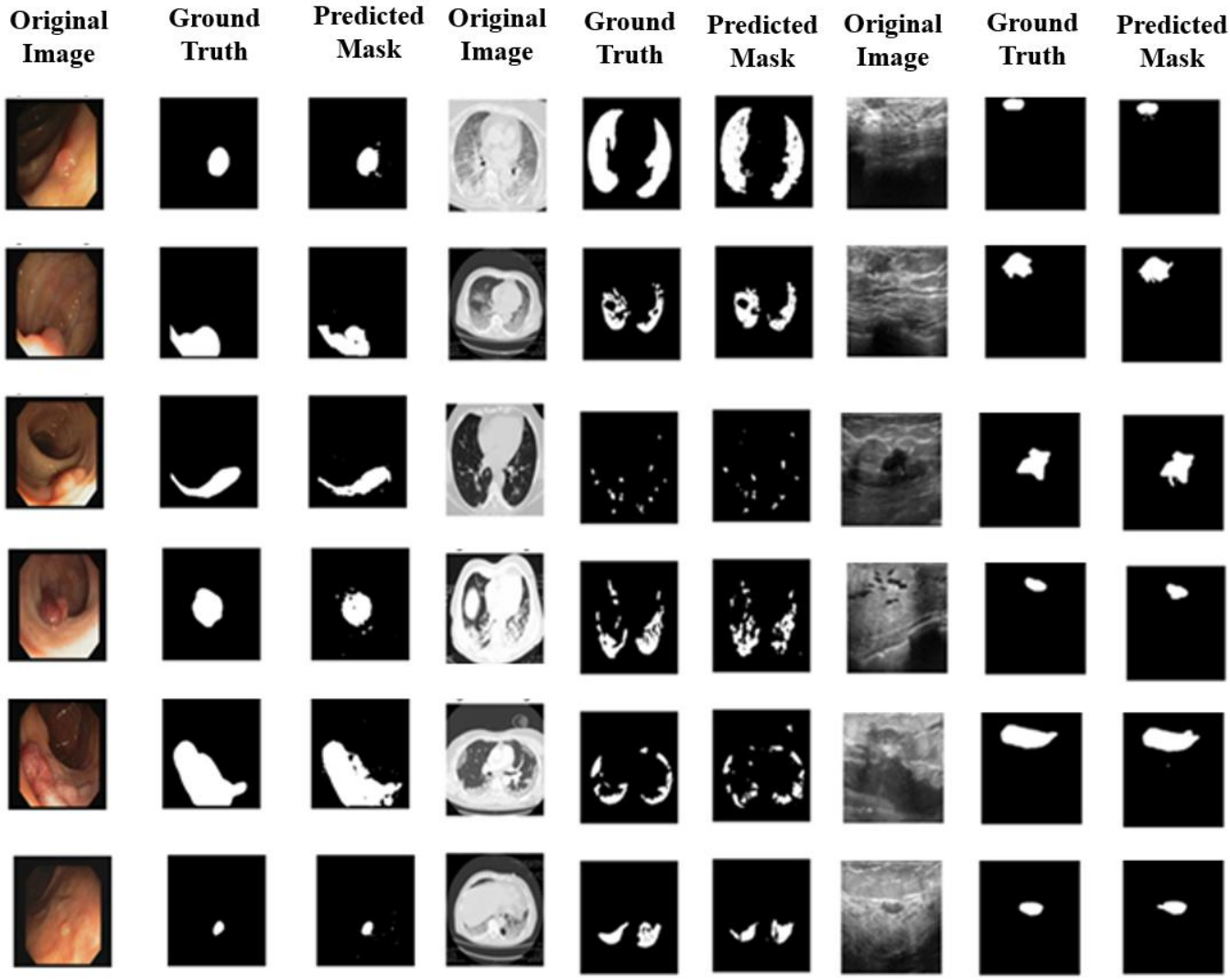

Figure 2. Segmentation results (a) CVC-ClinicDB Dataset (b) COVID-19 CT Dataset (c) Breast Ultrasound Dataset

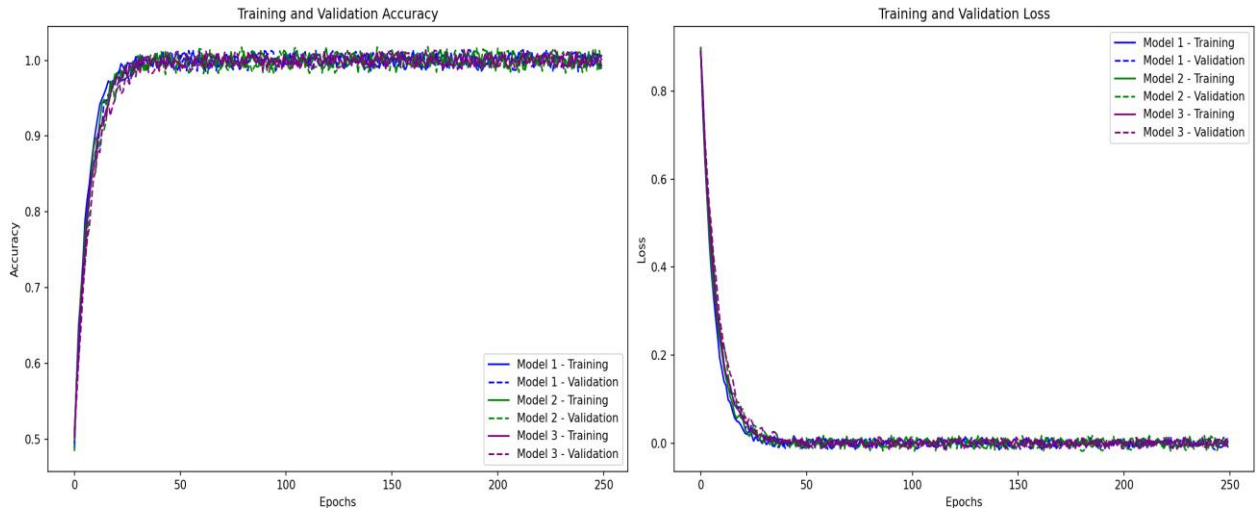

Figure 3. Model accuracy and loss validation result for Model 1-CVC-ClinicDB Dataset, Model 2-COVID-19 CT Dataset and Model 3-Breast Ultrasound Dataset

Table 1. CVC-ClinicDB Lesion segmentation (Dataset 1)

Method	DSC	IoU	Precision	Sensitivity	Time (h)	Parameters (M)
APTA-UNet	0.949	0.942	0.963	0.974	1.3	24.2
AFTer-UNet [19]	0.921	0.865	0.923	0.952	2.7	35.1
H-DenseUNet [31]	0.913	0.848	0.901	0.924	3.8	40.5
DenseRes-UNet [32]	0.902	0.819	0.892	0.915	3.4	38.7
U-Net 3+ [16]	0.892	0.808	0.834	0.893	5.2	37.5
Swin-UNet	0.939	0.935	0.957	0.962	1.5	40.0
MISSFormer	0.935	0.928	0.952	0.957	1.4	38.5

Table 2. COVID-19 CT segmentation (Dataset 2)

Method	DSC	IoU	Precision	Sensitivity	Time (h)	Parameters (M)
APTA-UNet	0.957	0.950	0.968	0.978	1.5	25.4
AFTer-UNet	0.924	0.872	0.927	0.958	2.9	34.7
H-DenseUNet	0.910	0.843	0.899	0.921	3.6	39.9
DenseRes-UNet	0.899	0.827	0.887	0.913	3.2	38.1
U-Net 3+	0.886	0.812	0.840	0.895	5.0	37.2
Swin-UNet	0.947	0.941	0.960	0.970	1.6	41.0
MISSFormer	0.940	0.932	0.955	0.970	1.7	36.8

Table 1 gives the CVC-ClinicDB Lesion Segmentation performance result where the proposed APTA-UNet attained a better performance when compared to previously proposed U-Net models. The APTA-UNet achieves the highest DSC (0.949) and IoU (0.942) values. Also, it shows the highest precision and sensitivity rate of 0.963 & 0.974, respectively. It proves the robustness of the model for detecting the region of interest portion with reduced false positives. It consumes only a runtime of 1.3 hours and a lightweight architecture of 24.2M parameters which is computationally efficient. The reported "Time (h)" refers to the inference time per image, not the total training time or inference time per epoch. It reflects the average time taken for the model to perform segmentation on a single image after the model has been trained. Also, other models like AFTer-UNet, H-DenseUNet, Swin-UNet, MISSFormer and DenseRes-UNet exhibit lower accuracy (DSC ranging from 0.902 to 0.935) and require significantly more time and parameters. The U-Net 3+ achieves the lowest DSC (0.892), reflecting its limitations. Overall, in the COVID-19 Lesion dataset, the APTA-UNet sets a new benchmark by combining high accuracy with computational efficiency.

Table 2 provides the comparative performance of various UNet methods for the COVID-19 CT liver dataset segmentation. Here, the proposed APTA-UNet achieves the highest effectiveness in all metrics, such as DSC of 0.957, IoU of 0.950, precision of 0.968 and sensitivity of 0.978, respectively. It shows its ability to accurately identify true positives while minimizing false negatives. Also, it achieves this performance with the shortest processing time (1.5 hours) and a relatively low parameter count of 25.4M which shows its efficiency. Also, existing models like AFTer-UNet, H-DenseUNet, and DenseRes-UNet have lower than proposed in all aspects. This shows all other methods are less effective for lung segmentation than the proposed APTA-UNet respectively.

Table 3 represents the performance of different methods for Breast Ultrasound Dataset segmentation. The proposed APTA-UNet outperforms with a DSC of 0.904 and an IoU of 0.827. It also has a high precision (0.892) and sensitivity (0.889) which shows its accuracy and its ability to correctly identify tumor regions. Additionally, it has the lowest processing time of 0.45 hours with a moderate number of parameters (31.2M). Also, other traditional methods attained lower accuracy and also required more time and computational resources with its less effective performance in this segmentation task.

Table 3. Breast Ultrasound Dataset segmentation (Dataset 3)

Method	DSC	IoU	Precision	Sensitivity	Time (h)	Parameters (M)
APTA-UNet	0.904	0.827	0.892	0.889	0.45	31.2
AFTer-UNet	0.876	0.792	0.864	0.875	0.50	35.1
H-DenseUNet	0.858	0.774	0.852	0.874	0.53	40.5
DenseRes-UNet	0.852	0.769	0.848	0.877	0.52	38.7
U-Net 3+	0.835	0.750	0.831	0.879	0.48	37.5
Swin-UNet	0.880	0.805	0.875	0.881	0.5	41.0
MISSFormer	0.880	0.800	0.872	0.879	0.55	38.0

Table 4. Ablation study of TA module

Model Configuration	DSC	IoU	Precision	Sensitivity
HaloNet-only	0.865	0.780	0.845	0.850
Axial-only	0.890	0.810	0.890	0.920
Non-Local-only	0.890	0.825	0.888	0.915
TA	0.949	0.942	0.963	0.974

Table 5. Statistical analysis of APTA-UNet for breast ultrasound, CVC-ClinicDB Lesion, and COVID-19 datasets

Metric	APTA-UNet	AFTer-UNet	H-DenseUNet	U-Net 3+	p-value (t-test)	95% Confidence Interval for DSC
Breast Ultrasound						
DSC	0.904 ± 0.005	0.876 ± 0.007	0.858 ± 0.008	0.835 ± 0.009	0.001	[0.899, 0.909]
IoU	0.827 ± 0.006	0.792 ± 0.007	0.774 ± 0.008	0.750 ± 0.009	0.003	[0.819, 0.835]
Precision	0.892 ± 0.004	0.864 ± 0.006	0.852 ± 0.007	0.831 ± 0.008	0.002	[0.884, 0.900]
Sensitivity	0.889 ± 0.005	0.885 ± 0.006	0.894 ± 0.007	0.879 ± 0.007	0.004	[0.875, 0.894]
CVC-ClinicDB						
DSC	0.949 ± 0.004	0.921 ± 0.005	0.913 ± 0.006	0.892 ± 0.005	0.002	[0.944, 0.954]
IoU	0.942 ± 0.003	0.865 ± 0.004	0.848 ± 0.005	0.808 ± 0.005	0.001	[0.938, 0.946]
Precision	0.963 ± 0.002	0.923 ± 0.003	0.901 ± 0.004	0.834 ± 0.004	0.000	[0.960, 0.967]
Sensitivity	0.974 ± 0.003	0.952 ± 0.004	0.924 ± 0.005	0.893 ± 0.005	0.001	[0.970, 0.979]
COVID-19 CT						
DSC	0.957 ± 0.004	0.924 ± 0.006	0.910 ± 0.007	0.886 ± 0.005	0.001	[0.952, 0.962]
IoU	0.950 ± 0.003	0.872 ± 0.004	0.843 ± 0.005	0.812 ± 0.005	0.002	[0.945, 0.955]
Precision	0.968 ± 0.002	0.927 ± 0.003	0.899 ± 0.004	0.840 ± 0.004	0.000	[0.965, 0.972]
Sensitivity	0.978 ± 0.002	0.958 ± 0.003	0.921 ± 0.004	0.895 ± 0.004	0.001	[0.974, 0.981]

To validate the effectiveness of the TA module, an ablation study is conducted on the CVC-ClinicDB Lesion dataset to evaluate the performance of each attention component independently and in combination. The obtained results are given in Table 4.

The model using only HaloNet attention achieves the lowest performance in all metrics. It shows that local spatial relationships alone are not sufficient for accurate medical image segmentation. The model using Axial attention alone improves the performance compared to HaloNet-only. However, it still does not perform as well as the full TA approach. The non-local attention captures long-range dependencies across the entire image; it still underperforms compared to the full TrioFusion model. Overall, it is observed that combining multiple attention mechanisms supports refining the segmentation by using both local and global spatial relationships.

In addition to the performance metrics, statistical significance tests are conducted to validate the improvements achieved by the APTA-UNet model over competing methods. The paired t-tests are used to compare the segmentation performance between APTA-UNet and other baseline models. In addition, 95% confidence intervals (CIs) for the DSC and IoU values are computed to quantify the uncertainty in the performance estimates. From Table 5, it is observed that the improvements observed in APTA-UNet over baseline models are not only substantial but also statistically significant. The 95% CI for APTA-UNet indicate a high level of consistency in performance with a narrower interval compared to the other models. The p-values of less than 0.05 indicate that the performance of APTA-UNet is statistically significantly higher than the other models.

5. CONCLUSION

The proposed novel APTA-UNet architecture introduces an advanced AP module and TA modules to attain an accurate segmentation to improve spatial context learning in medical image segmentation. This APTA-UNet architecture is validated using three datasets, such as COVID-19 CT segmentation, CVC-ClinicDB lesion segmentation and Breast Ultrasound segmentation. The performance is evaluated using metrics such as Dice, IoU, Precision, Sensitivity, Time, and Model Parameters. The results show that the proposed APTA-UNet achieved Dice scores of 94.9%, 95.7%, and 90.4% across the respective datasets, demonstrating superior segmentation accuracy compared to other advanced UNet-based models. The superior accuracy is attributed to the effective fusion of local and global contextual information enabled by the TrioFusion Attention mechanisms. This demonstrates that the proposed architecture is both efficient and scalable and achieves state-of-the-art performance in medical image segmentation tasks. Future research will focus on extending the model to multi-modal MRI scans to address challenges such as inter-modality data fusion and computational efficiency. This will be achieved by using multi-task learning and optimized attention mechanisms for

improved diagnostic performance.

REFERENCES

[1] Qureshi, I., Yan, J.H., Abbas, Q., Shaheed, K., Riaz, A.B., Wahid, A., Khan, M.W.J., Szczuko, P. (2023). Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. Information Fusion, 90: 316-352. https://doi.org/10.1016/j.inffus.2022.09.031

[2] Liu, X.B., Song, L.P., Liu, S., Zhang, Y.D. (2021). A review of deep-learning-based medical image segmentation methods. Sustainability, 13(3): 1224. https://doi.org/10.3390/su13031224

[3] Chi, W.C., Ma, L., Wu, J.J., Chen, M.L., Lu, W.G., Gu, X.J. (2020). Deep learning-based medical image segmentation with limited labels. Physics in Medicine & Biology, 65(23): 235001. https://doi.org/10.1088/1361-6560/abc363

[4] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7): 3523-3542. https://doi.org/10.1109/TPAMI.2021.3059968

[5] Du, G.T., Cao, X., Liang, J.M., Chen, X.L., Zhan, Y.H. (2020). Medical image segmentation based on U-Net: A review. Journal of Imaging Science & Technology, 64(2): 020508. https://doi.org/10.2352/J.ImagingSci.Technol.2020.64.2.020508

[6] Rayed, M.E., Islam, S.M.S., Niha, S.I., Jim, J.R., Kabir, M.M., Mridha, M.F. (2024). Deep learning for medical image segmentation: State-of-the-art advancements and challenges. Informatics in Medicine Unlocked, 47: 101504. https://doi.org/10.1016/j.imu.2024.101504

[7] Guo, Z., Li, X., Huang, H., Guo, N., Li, Q.Z. (2019). Deep learning-based image segmentation on multimodal medical imaging. IEEE Transactions on Radiation and Plasma Medical Sciences, 3(2): 162-169. https://doi.org/10.1109/TRPMS.2018.2890359

[8] Fu, Y.B., Lei, Y., Wang, T.H., Curran, W.J., Liu, T., Yang, X.F. (2021). A review of deep learning based methods for medical image multi-organ segmentation. Physica Medica, 85: 107-122. https://doi.org/10.1016/j.ejmp.2021.05.003

[9] Wang, R.S., Lei, T., Cui, R.X., Zhang, B.T., Meng, H.Y., 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

[10] Aljabri, M., AlGhamdi, M. (2022). A review on the use of deep learning for medical images segmentation. Neurocomputing, 506: 311-335. https://doi.org/10.1016/j.neucom.2022.07.070

[11] Renard, F., Guedria, S., Palma, N.D., Vuillerme, N. (2020). Variability and reproducibility in deep learning for medical image segmentation. Scientific Reports, 10: 13724. https://doi.org/10.1038/s41598-020-69920-0

[12] Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., et al. (2024). Medical image segmentation review: The success of U-Net. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12): 10076-10095. https://doi.org/10.1109/TPAMI.2024.3435571

[13] Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D. (2018). DRINet for medical image segmentation. IEEE Transactions on Medical Imaging, 37(11): 2453-2462. https://doi.org/10.1109/TMI.2018.2835303

[14] Wang, G.T., Li, W.Q., Zuluaga, M.A., Pratt, R., et al. (2018). Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Transactions on Medical Imaging, 37(7): 1562-1573. https://doi.org/10.1109/TMI.2018.2791721

[15] Feng, S.L., Zhao, H.M., Shi, F., Cheng, X.N., Wang, M., Ma, Y., Xiang, D.H., Zhu, W.F., Chen, X.J. (2020). CPFNet: Context pyramid fusion network for medical image segmentation. IEEE Transactions on Medical Imaging, 39(10): 3008-3018. https://doi.org/10.1109/TMI.2020.2983721

[16] Huang, H.M., Lin, L.F., Tong, R.F., Hu, H.J., Zhang, Q.W., Iwamoto, Y., Han, X.H., Chen, Y.W., Wu, J. (2020). UNet 3+: A full-scale connected UNet for medical image segmentation. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 1055-1059. https://doi.org/10.1109/ICASSP40776.2020.9053405

[17] Cao, H., Wang, Y.Y., Chen, J., Jiang, D.S., Zhang, X.P., Tian, Q., Wang, M. (2022). Swin-Unet: Unet-like pure transformer for medical image segmentation. In Lecture Notes in Computer Science, pp. 205-218. https://doi.org/10.1007/978-3-031-25066-8_9

[18] Weng, Y., Zhou, T.B., Li, Y.J., Qiu, X.Y. (2019). NAS-Unet: Neural architecture search for medical image segmentation. IEEE Access, 7: 44247-44257. https://doi.org/10.1109/ACCESS.2019.2908991

[19] Yan, X.Y., Tang, H., Sun, S.L., Ma, H.Y., Kong, D.Y., Xie, X.H. (2022). AfTer-UNet: Axial fusion transformer UNet for medical image segmentation. In 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, USA, pp. 3971-3981. https://doi.org/10.1109/WACV51458.2022.00333

[20] Park, J., Hong, S., Cho, Y., Jeon, J.J. (2024). Unicorn: U-Net for sea ice forecasting with convolutional neural ordinary differential equations. arXiv preprint arXiv:2405.03929. https://doi.org/10.48550/arXiv.2405.03929

[21] Huang, L., Miron, A., Hone, K., Li, Y. (2024). Segmenting medical images: From UNet to Res-UNet and nnUNet. In 2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS), Guadalajara, Mexico, pp. 483-489. https://doi.org/10.1109/CBMS61543.2024.00086

[22] Alrfou, K., Zhao, T., Kordijazi, A. (2025). CS-UNet: A generalizable and flexible segmentation algorithm. Multimedia Tools and Applications, 84: 7807-7834. https://doi.org/10.1007/s11042-024-19242-4

[23] Liao, W.B., Zhu, Y.H., Wang, X.Y., Pan, C.W., Wang, Y.S., Ma, L.T. (2024). LightM-UNet: Mamba assists in lightweight UNet for medical image segmentation. arXiv preprint arXiv:2403.05246. https://doi.org/10.48550/arXiv.2403.05246

[24] Al Qurri, A., Almekkawy, M. (2023). Improved UNet with attention for medical image segmentation. Sensors, 23(20): 8589. https://doi.org/10.3390/s23208589

[25] Khan, W.R., Madni, T.M., Janjua, U.I., Javed, U., Khan, M.A., Alhaisoni, M., Tariq, U., Cha, J.H. (2023). A

hybrid attention-based residual Unet for semantic segmentation of brain tumor. Computers, Materials and Continua, 76(1): 647-664. https://doi.org/10.32604/cmc.2023.039188

[26] Rajamani, K.T., Rani, P., Siebert, H., ElagiriRamalingam, R., Heinrich, M.P. (2023). Attention-augmented U-Net (AA-U-Net) for semantic segmentation. Signal, Image And Video Processing, 17: 981-989. https://doi.org/10.1007/s11760-022-02302-3

[27] Vaswani, A., Ramachandran, P., Srinivas, A., Parmar, N., Hechtman, B., Shlens, J. (2021). Scaling local self-attention for parameter efficient visual backbones. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, USA, pp. 12889-12899. https://doi.org/10.1109/CVPR46437.2021.01270

[28] Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T. (2019). Axial attention in multidimensional transformers. arXiv preprint arXiv:1912.12180. https://doi.org/10.48550/arXiv.1912.12180

[29] Wang, Z.Y., Zou, N., Shen, D.G., Ji, S.W. (2020). Non-local U-Nets for biomedical image segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 34(4): 6315-6322. https://doi.org/10.1609/aaai.v34i04.6100

[30] Shen, T.P., Xu, H.Q. (2023). Medical image segmentation based on transformer and HarDNet structures. IEEE Access, 11: 16621-16630. https://doi.org/10.1109/ACCESS.2023.3244197

[31] Li, X.M., Chen, H., Qi, X.J., Dou, Q., Fu, C.W., Heng, P.A. (2018). H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Transactions on Medical Imaging, 37(12): 2663-2674. https://doi.org/10.1109/TMI.2018.2845918

[32] Kiran, I., Raza, B., Ijaz, A., Khan, M.A. (2022). DenseRes-Unet: Segmentation of overlapped/clustered nuclei from multi organ histopathology images. Computers in Biology and Medicine, 143: 105267. https://doi.org/10.1016/j.compbiomed.2022.105267