Transformer-Driven Semantic Segmentation for Thyroid Ultrasound: A SwinUNet-Based Architecture with Integrated Attention

Ammar Oad *, Imtiaz Hussain Koondhar, Feng Dong *, Weibing Liu, Beiji Zou, Weichun Liu, Yun Chen, Wu Yaoqun

Posted Date: 10 December 2025

doi: 10.20944/preprints202512.0885.v1

Keywords: deep learning in healthcare; thyroid nodule segmentation; ultrasound imaging; SwinUNet; vision transformers; attention mechanisms; medical image analysis

 doi:10.20944/preprints202512.0885.v1

Article

Transformer-Driven Semantic Segmentation for Thyroid Ultrasound: A SwinUNet-Based Architecture with Integrated Attention

Ammar Oad [1,2,*], Imtiaz Hussain Koondhar [2], Feng Dong [1,*], Weibing Liu [1], Beiji Zou [1,3], Weichun Liu [1], Yun Chen [1] and Wu Yaoqun [1]

[1] School of Information Engineering, Shaoyang University, Shaoyang 422000, China
[2] Information Technology Center, Sindh Agriculture University, Tandojam, Pakistan
[3] School of Computer Science and Engineering, Central South University, Changsha, 410083, China
* Correspondence: ammar.oad@hnsyu.edu.cn (A.O.); dongfeng@hnsyu.edu.cn (F.D.)

Abstract

Accurate segmentation of thyroid nodules on ultrasound images remains a challenging task in computer-aided diagnosis (CAD) mainly because of low contrast, speckle noise, and large inter-patient variability of nodule appearance. Here a new deep learning-based segmentation method has been developed on the SwinUNet architecture supported by spatial attention mechanisms to enhance feature discrimination and localization accuracy. The model takes advantage of the hierarchical feature extraction ability of the Swin Transformer to learn both global context and local fine-grained details, whereas attention modules during the decoder process selectively highlight informative areas and suppresses irrelevant background features. We checked out the system's design using the TN3K thyroid ultrasound info that's out there. It got better as it trained, peaking around the 800th run with some good numbers: a Dice Similarity Coefficient (F1 Score) of 85.51%, Precision of 87.05%, Recall of 89.13%, IoU of 78.00%, Accuracy of 97.02%, and an AUC of 99.02%. These numbers are way better than when we started (like a 15.38% jump in IoU and a 12.05% rise in F1 Score), which proves the system can learn tricky shapes and edges well. The longer it trains, the better it gets at spotting even hard-to-see thyroid lumps. This SwinUnet_withAttention thing seems to work great and could be used in clinics to help doctors figure out thyroid problems.

Keywords: deep learning in healthcare; thyroid nodule segmentation; ultrasound imaging; SwinUNet; vision transformers; attention mechanisms; medical image analysis

1. Introduction

1.1. Background and Significance

Thydroid lumps are weird growths in the thyroid, and they're way more common now than they used to be. Because we're seeing early signs of thyroid cancer more often, finding these lumps which may turn cancerous, has become super important for doctors [1,2]. Interestingly, ultrasound shows that a lot of adults, like, 50–70% have these lumps. So it's a pretty widespread thing [3]. Because of this, computer systems that help doctors are becoming really useful, especially when they automatically cut out parts of medical images. This means sorting pixels to make maps that show organs, lumps, and tumors in detail. When it comes to medical images, convolutional neural networks (CNNs) are now the go-to way to do different cutting tasks because they can pick up on complicated spatial stuff [4]. But, old cutting ways for thyroid lumps aren't that great, usually getting about 79.00% to 83.70% right. On the other hand, deep learning ways are way better, getting up to 91.30% right. This proves they can be good, but there are problems like messy data and needing experts to label stuff [5]. Using special attention tricks in deep learning models to make cutting better has led to cool cutting gains. The effectiveness of models like SwinU-Net which combine hierarchical

 doi:10.20944/preprints202512.0885.v1

vision transformers and attention mechanisms has shown to improve accuracy by enabling focus on important image portions. This is helpful in dealing with problems such as fuzzy edges and irregular nodule sizes common in ultrasound images [6]. The need for precision intensifies especially with the need to appropriately differentiate between benign and malignant nodules due to nodules significantly impacting the treatment approach. Although FNA cytology is considered the best approach, it has its downsides. It is expensive, limited by a global scarcity of expert cytopathologists, and relies on specialist knowledge [7]. These limitations underscore the urgent need for AI-powered diagnostic tools capable of improving not only diagnostic precision but also diagnostic efficiency and reliability. There has been some promise in the development of CAD systems with ultrasound B-mode images, although these systems tend to follow a rigid step-by-step process of a machine learning workflow which starts with pre-processing, segmentation, feature extraction, and ends with classification. They have been helpful in distinguishing between benign and malignant nodules, even when faced with the challenge of pervasive speckle noise and complex overlapping textures [8].

However, the widespread availability of ultrasound imaging has led to over diagnosis and reporting resulting in unwarranted biopsies and surgeries due to differing physician interpretations and increasing stress on radiologist workloads [9,10]. Due to these issues, researchers have begun to explore new deep learning techniques, including attention-based graph neural networks, including Multiple Attention Graph Convolutional Networks (MAGCN). It has been observed that these models effectively capture high-level semantic relations and have shown promise in generalizing to medical imaging tasks [11].

Though highly accurate, however, some of the more conventional machine learning models, random forests and artificial neural networks, are hindered from clinical acceptance by not being explain able and not being able to simulate human diagnostic reasoning [12]. A major challenge to building effective segmentation models is the similarity of available datasets. Most publicly available thyroid ultrasound datasets are from individual medical institutions or unique imaging devices, and tend to show only one nodule per image. This uniformity limits model generalization, particularly in real-world environments that involve multiple nodules or diverse imaging modalities [13]. Multi-center, multi-device datasets are becoming more popular as essential for training models to manage clinical variability. Besides dataset variety, the combination of several imaging modalities like ultrasound, CT, and MRI has been useful for augmenting diagnostic data and overcoming the disadvantages of single-modality imaging, including low resolution and noise [14]. Equally, the integration of attention mechanisms into CNNs and Transformer models has significantly enhanced the segmentation of intricate structures through concentrating on the most informative features in an image [15]. Due to the challenges of accessing large annotated datasets, techniques such as few-shot learning, unsupervised learning, and self-supervised learning have cropped up as effective strategies for training deep learning models with little labeled data. Architectures such as CO-attention Siamese Networks and GAN-based architectures have produced encouraging findings in small-sample learning while maintaining specificity in place [16]. Widening of scope on learning as well as modification of learning techniques whereby learning is achieved through association transfer as in the case of learning through models, paralleled with augmentation strategies, should also sufficiently address the concerned of data resonation during the development of models which are as clinically realistic as possible [17]. Thyroid glands perform very important metabolic functions and are now the target for much AI diagnostic innovation. Current methods for evaluating the thyroid gauge the thyroid glands using two-dimensional images or ultrasounds. While they are the most common methods, they are primarily dependent on the skill and expertise of the operator. These techniques are especially ineffective for low-contrast lesions having poorly defined borders. These motivate the use of highly automated systems able to deliver consistent and reliable results. Attention-augmented models, especially adaptively scaled transformer models, Swin-Unet implementations, handle these challenges using deep hybrid models that integrate spatial and channel-prior attention with multi-head classification and deep multi-branch models for global and local context fusion. They do so with deep architecture to efficiently traverse attention and context gaps [18]. Moreover, attention-

 doi:10.20944/preprints202512.0885.v1

augmented multi-task models of deep learning have also been shown to significantly boost the Subnetwork for challenging, low-contrast imaging [19]. Integrating CEUS, CT, and MR imaging with sophisticated imaging modalities makes it possible for clinicians to obtain detailed and objective decision-supporting materials [20]. The accuracy of MCDLM and DK models in clinical methods is augmented by the use of expert knowledge and semi-structured easements as diagnostic aids [21] Outside of ultrasound, radionics based deep learning models have demonstrated great accuracy and general applicability in the diagnosis of thyroid disease using planar scintigraphy and CT, showcasing the utility of AI for thyroid diagnosis [22]. The practice of ultrasound imaging for the diagnosis and segmentation of thyroid nodules has also significantly improved. Deep learning methods have transformed the diagnosing and segmentation of thyroid nodules through ultrasound imaging. AI-based models can replace or complement traditional diagnosis pipelines as these models are no longer as subjective and time consuming. SwinUNet is one AI models that sophisticatedly integrates with attention mechanisms to outperform other models in feature extraction and segmentation. This research proposes a new way of diagnostic precision by combining a segmentation model based on SwinUNet with a classifier network to obtain smoother and more continuous boundaries.

This paper outlines our research on the segmentation of thyroid nodules. In section II, we present the review of related work and the clinical background of the problem of nodule segmentation as imaging ultrasound data. Section III presents the proposed methodology, featuring an in-depth explanation of the SwinUnet_withAttention architecture and detailing how spatial attention mechanisms are embedded to improve feature discrimination and localization precision. Section IV outlines the experimental setup, including specifics of the TN3K dataset, the data preprocessing and augmentation pipeline, the implemented training approach, and the evaluation metrics employed to measure model performance. Section V provides the results, explains the effect of the proposed architecture and extended training epochs on segmentation performance, and presents a comparative study against other state-of-the-art approaches. Lastly, Section VI concludes the paper by summarizing the main findings and contributions, and suggests possible directions for future work aimed at enhancing model generalization, efficiency, and clinical applicability.

1.2. Related Work and Clinical Background

The latest developments in imaging technologies have placed thyroid scintigraphy on the spotlight as a critical diagnostic technique for assessing thyroid function and identifying abnormalities. Hitherto dependent on human interpretation, the procedure has been time-consuming and prone to inter-operator variations. Deep learning systems have gotten better at spotting things in images, which helps doctors make more correct diagnoses. A big step forward is a five-layer U-Net model made just for thyroid scans. This thing can automatically locate the edges of the thyroid and guess how much radiation is there, which is key for figuring out any problems. It works great because we trained it on a huge set of 2,734 scans from the last four years. It gets the lobe edges right about 92% of the time, and the whole thyroid shape right about 94% of the time. The segmentation methods still exhibited quantitative reliability, with only a median error of 3.520 cm^2 for gland size and a negligible 0.029% error in uptake measurements, making those models good candidates for clinical application [23]. Numerous publications remain categorical about the advantages of artificial intelligence in nuclear medicine, particularly in the context of deep learning algorithms such as U-Net and its derivatives. Diversity of datasets, not quantity, has been found to be a significant performance improver. Removing redundant data like cine-clips' duplicate frames and utilizing strong data augmentation techniques have been even more important than architecture selection. This is consistent with increased faith in models like ResUNet, which is being perceived more and more as a reliable device for aiding thyroid nodule diagnosis in clinical practice [24]. At the same time, deep learning has developed further with attention mechanisms added to enhance accuracy in challenging imaging tasks. The Swin U-Net looks like it could be good for finding thyroid nodules in ultrasound images. It makes use of special image processing and attention to its self. It works nicely

 doi:10.20944/preprints202512.0885.v1

with nodules of all sizes and any blurring edges. Residual connections helped it learn different levels of detail, which is cool for spotting complex shapes. Based on the results, it looks better than older ways of doing this. This could mean it could also be used for other types of scans, like MRIs and CTs. Recent research has tried to make the model simpler so it can be used in real time and work better with different kinds of information [25]. People are especially looking at ultrasounds of kids' thyroids. It's hard because of the grayness, random noise from the body, and weird shapes. One approach uses a DC-Contrast U-Net with some special tricks to deal with these issues. The performance of the model had a significant improvement in IoU and mIoU, more than 6%, with a co-gain in precision and recall. The residual challenging areas are the segmentation of the boundaries-noise and are suggested as future areas of model improvements [26]. So, it looks like some new transformer-based U-Net versions are doing great at spotting thyroid nodules in images. Other models are using some cool tricks too, like frequency-based filtering and working with overlapping parts of images at different sizes. A special Co-operative Transformer Fusion thingy with self and cross-attention helped them get really high accuracy on a bunch of datasets, like 98.2% on DDTI and TN3K, and 97.8% on TG3K. Yet limited generalizability restrains further research onto extending these advantages to other types of images with increased speed [27].

Another multimodal approaches segmentation using active contour models. That leads to classification based on ResUNet. Using TDID database, the two-process pipeline was accurate and efficient than the other models, thereby strengthening the contribution of ResUNet to facilitate holistic analysis of thyroid nodules [28]. ResUNet eliminates vanishing gradients and makes convergence speedier through residual connections. Its performance in segmentation is better than standard U-Net and U-Net++ on thyroid CT images, with a Dice score of 90.87% and an IoU score of 94.58%, even with thinning data, thus confirming its strength [29]. To tackle problems from not having enough good training data for ultrasound images, researchers made a Super-Pixel U-Net system. It has three parts: find edges, fix problems, and then classify. Each part makes the result better. It got a Dice score of 0.9279 and an F1 score of 0.9161, which proves that doing a series of U-Nets with a couple of extra steps is helpful for ultrasound data [30].

The Hybrid Transformer U-Net (H-TUNet) was made to fix bad spatial and context understanding. By mixing a 2D transformer for inside-the-picture stuff and a 3D transformer for between-the-pictures movement, H-TUNet really makes anatomical representation better. The TSUD and TG3K data show its better at segmenting images and learning features than other methods [31]. Also, different encoder-decoder designs, especially those with fewer residual links, have shown that we can balance good segmentation with not using too much computer power. Hybrid ResUNet designs like this do a good job outlining thyroid nodules and are doable in real clinics [32].

Someone suggested a Channel Boosted CNN to push for classification with segmentation. It uses better filtering and histogram stuff, applies SegNet segmentation, and then CB-CNN classification to classify with 96% right answers. It did better than DCNN, ResNet101, and other setups and was shown to be very helpful in finding thyroid cancer [33]. A Channel Boosted CNN was proposed in a push for classification-integrated segmentation. Improved with adaptive median filtering and sub-image histogram equalization, it applies SegNet segmentation followed by CB-CNN classification to classify with 96% accuracy. It surpassed DCNN, ResNet101, and other high-level architectures, and was shown to be effective in thyroid cancer detection [33]. While traditional segmentation methods (thresholding and edge detection) are simple and fast, their quality degrades in noisy or complicated images. Deep learning models, while more intensive computationally, offer better feature learning ability. Transfer learning, model compression, and semi-supervised learning mitigate data and hardware constraints. Interpretability must continue, however. Explainable AI techniques like attention maps facilitate clinician trust building. Hybrid systems combining rule-based and deep learning strategies excel, particularly in environments with weak boundaries or variable intensities [34]. More recent work has investigated combining deep and handcrafted features. For example, SegNet from VGG19 was blended with fuzzy gray-level co-occurrence matrices and deep features and classified afterwards using an RBF-kernel SVM. With 99.25% classification accuracy, the

 doi:10.20944/preprints202512.0885.v1

approach outperforms current models, although it is hampered by high-dimensional feature vectors and associated computational expense [35]. Within a multi-task learning setting, FFANet proposed a segmentation and classification dual-head model. Through feature fusion and a loss function adapted to it, it attained a Dice coefficient of 0.935 and 79% classification accuracy, showing the efficacy of optimization across the joint tasks in thyroid ultrasound analysis [36]. Efforts toward image acquisition standardization and deep learning models incorporation in thyroid diagnosis remain promising. Although not yet ideal predictors, these models reduce radiological subjectivity and potentially increase access to healthcare in rural areas [37]. New hybrid architectures such as Enhanced-TransUNet employ Transformers for global context and U-Nets for spatial localization, resolving issues of overfitting as well as low contrast. Experimental evidence on TN3K and DDTI datasets verifies the model's higher Dice and Hausdorff scores, affirming its usability in clinical practice [38]. FCG-Net presents a light-weight option to parameter-intensive models such as UNet3+, with Ghost bottlenecks and full-size skip connections employed to reduce parameter number and enhance speed. FCG-Net has high Dice and sensitivity values, allowing for effective deployment in mobile devices [39]. Encoder feature extraction improvements have also centered on residual networks and multi-scale attention modules. These provide enhanced spatial detection and contextual integration support with the help of deep supervision and hybrid loss functions such as Focal loss to handle class imbalance. This improves segmentation accuracy even in morphologically intricate instances [40]. To combat noisy ultrasound settings, SGUNet was designed using semantic-guided architecture. It provides pixel-wise semantic maps in the decoding process for better propagation of low-level features, achieving more than 2% improvement over U-Net and U-Net++ as per Dice scores [41]. The Swin Transformer-enhanced U-Net technology-ST-Unet-utilizes the global context via Transformer encoders and performs feature enhancement [42]. Its significant performance is underlined by benchmark dataset results such as Synapse and ISIC 2018, with Dice scores of around 78.86 and a recall of 0.9243 [43]. Finally, the H-TUNet is getting continuous attention because of its dual-transformer architecture incorporating 2D MSCAT and 3D SAT modules that enable the best intra-frame and inter-frame learning. People say this setup works better than the best thyroid ultrasound image tools out there [44]. So, about making visual Transformer tools way better for thyroid imaging, check out our attention method using the Swin U-Net model. It helps get a clearer picture of the thyroid when looking at ultrasound images. With the presence of hierarchical self-attention embedded in the Swin Transformer backbone, our model is competent to efficiently capture the global context along with the spatial fine-grained features, which in turn are solving typical issues like edge blurring, grayscale inconsistencies, and anatomical noise. The flow of work enhances the model for accurately delineating thyroid borders and better precision for nodule separation. This attention thingy helps the model see details, even if they're not super clear or have weird shapes, which means fewer mistakes when figuring out what's wrong. Plus, the model can learn and adapt without getting too complicated. So, it's good to go for real-life stuff, like checking thyroids with ultrasound, instantly.

2. Materials and Methods

This paper is about a medical image segmentation method. It mixes the Swin Transformer setup with attention gates in a UNet style. This method brings together the big-picture view of vision transformers and good spatial accuracy using attention. It's made to fix problems in medical image analysis, where spotting and cutting out tiny body parts and problem areas is key.

There are four main components to this approach: data preparation and preprocessing to form the pipeline, hybrid architecture which includes a Swin Transformer encoder and an attention-gated decoder, the training approach defined by a custom set of loss functions, and a comprehensive evaluation framework, as illustrated in Figure 1. This integrated approach retains the advantages of traditional CNN-based segmentation methods, while avoiding their shortcomings in computational efficiency and clinical applicability.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 December 2025 doi:10.20944/preprints202512.0885.v1

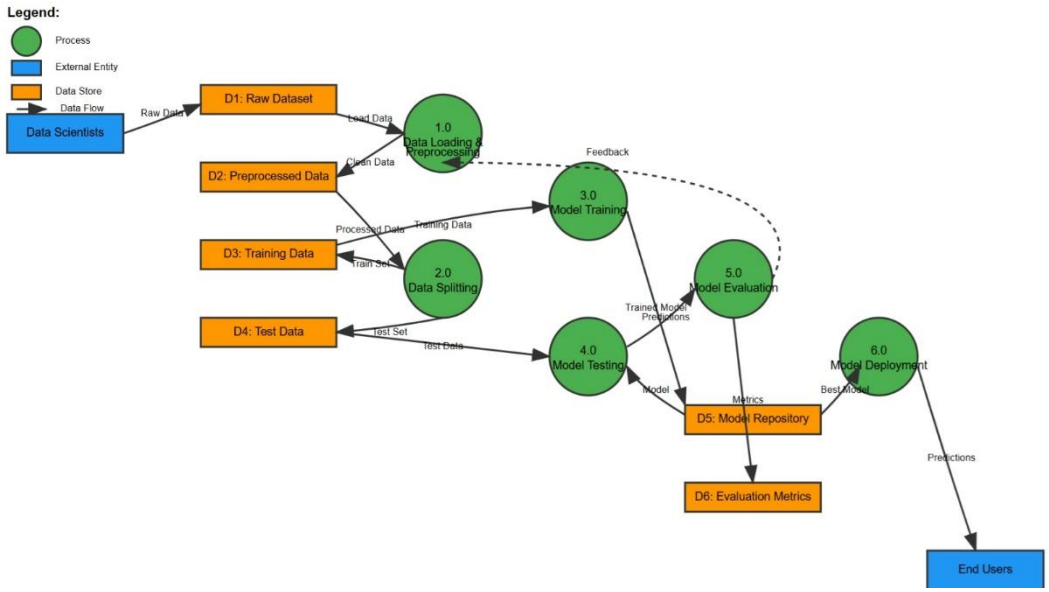

Figure 1. Flowchart of Transformer + Attention + Decoder.

2.1. Flowchart

This Data Flow Diagram depicts a seamless machine learning pipeline that processes raw data into deployable predictive models via six tightly-linked tasks. The cycle kicks off when Data Scientists an external actor submits raw datasets to Process 1.0, Data Loading & Preprocessing. Here, input data undergoes cleaning and is then written to the Preprocessed Data storage, labeled D2. Process 2.0, Data Splitting, subsequently partitions the sanitized data into distinct training and testing datasets, residing in D3 and D4, respectively. In Process 3.0, Model Training, the training dataset is leveraged to generate models that land in storage D5, the Model Repository. Concurrently, Process 4.0, Model Testing, assesses the models against the testing dataset to gauge predictive accuracy. Process 5.0, Model Evaluation, quantifies the results and deposits the derived performance metrics into the Evaluation Metrics repository, D6. Process 6.0, Model Deployment, finally identifies and rolls out the most effective model, sending real-time predictions to the End Users, who are also external to the system. An evaluative feedback loops that traces metrics back to preprocessing allows for continuous adjustment, fostering persistent refinement throughout the journey from data for ingestion to live operational use.

2.2. Algorithm SwinUnet_ with Attention Mechanism

Input:

- ***- Medical image I ∈ ℝ^(H×W×3)***
- ***- Ground truth mask M ∈ ℝ^(H×W×1)***

Output:

- ***- Segmentation Prediction P ∈ ℝ^(H×W×C)***
- ***- Trained model parameters***

Steps:

1. ***Initialize SwinUnet model with Attention Gates***
2. ***Load pre-trained Swin Transformer weights***
3. ***for each epoch e = 1 to max_epochs do***
4. ***for each batch (I, M) ∈ training_data do***
5. ***// Forward Pass***
6. ***features ← SwinEncoder(I)***

 doi:10.20944/preprints202512.0885.v1

attended_features ← AttentionGates(features)
P ← Decoder(attended_features)
// Loss Computation
loss ← CombinedLoss(P, M)
// Backward Pass
optimizer.zero_grad()
loss.backward()
optimizer.step()
end for
// Validation Phase
val_metrics ← Evaluate (model, validation_data)
if val_metrics.improved then
save_checkpoint(model)
end if
end for
return trained SwinUnet_withAttention model

This algorithm introduces the entire training pipeline for SwinUnet with Attention Gates for medical image segmentation. The approach starts by initializing network architecture and loading pre-trained Swin Transformer weights to take advantage of learned visual representations. During training, the algorithm passes medical images via a hierarchical encoder to extract multi-scale features via the Swin Transformer backbone. The attention gate mechanism selectively amplifies relevant features from skip connections and suppresses irrelevant information to improve segmentation accuracy. The decoder reconstructs the segmentation mask from attention-weighted features. The model is optimized using a combination of the cross-entropy loss and Dice loss to handle class imbalance and improve boundary delineation. At the end of each epoch, the algorithm evaluates performance on validation data and saves the best-performing model checkpoint. This is repeated until convergence or until maximum epochs have been achieved, returning a trained model that can successfully perform medical image segmentation.

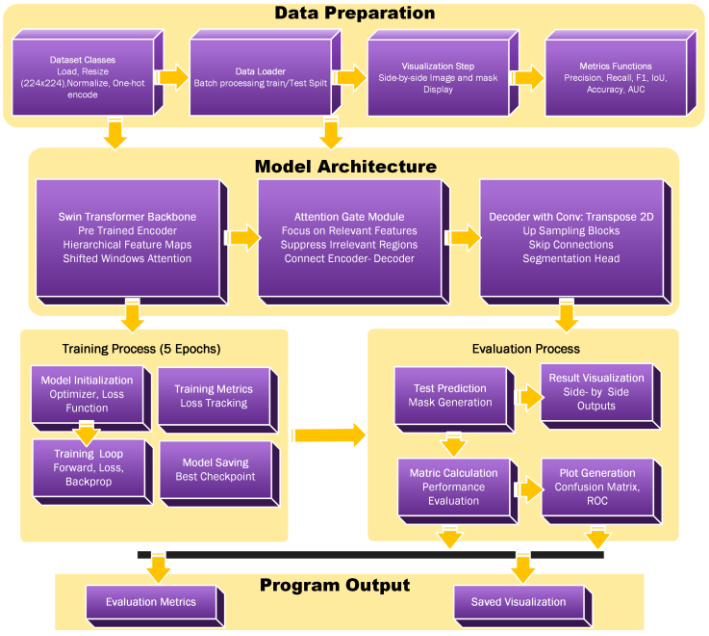

Figure 2. Swin UNet with Attention Gates End-to-end medical image segmentation.

 doi:10.20944/preprints202512.0885.v1

2.3. Hierarchical Feature Extraction Process

Hierarchical structure of the Swin Transformer specialized in vision tasks, especially those requiring dense predictions such as image segmentation. The architecture starts from Stage 1, where the image is split into patches of 56 x 56 x C dimensions, extracting shallow features. Stage 2 Patch Merger reduces the spatial resolution to 28 x 28 while deepening the channels to 512, extracting mid-level features through Swin Blocks. Stage 3 further downsamples the image to 14 x 14 x 512, from which multiple Swin Blocks can extract deeper features. Stage 4 uses advanced Power Blocks and Swin Blocks to derive deep contextual features at very low resolution 7 x 7 x 1024, richly capturing semantics. Like traditional CNNs, this model reduces the spatial resolution from 56 to 7, enhances feature depth and shifts window attention in place of convolutions to learn local detail and global context simultaneously, which is crucial in precise tasks like medical image segmentation, shown in Figure 3 and Table 1.

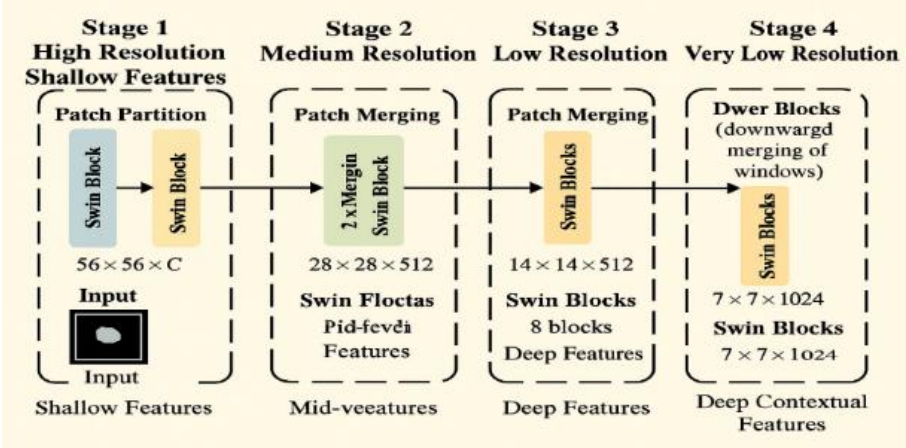

Figure 3. Swin Transformer feature extraction.

Table 1. Feature hierarchy.

Stage	Output Size	Feature Depth	Feature Type
1	56×56×C	Low	Shallow (Local textures)
2	28×28×512	Medium	Mid-level patterns
3	14×14×512	High	Deep semantic features
4	7×7×1024	Very High	Global contextual features

The table illustrates the hierarchical feature extraction method in convolutional neural networks (CNNs), depicting the progression of features across successive levels. In Stage 1, the output dimension is 56×56×C, which reflects basic, low-level characteristics such as local textures and edges. In Stage 2, producing an output of 28×28×512, the network acquires mid-level patterns like basic shapes or recurring textures, demonstrating a medium level of feature depth. Stage 3. 14×14×512 feature maps are generated, capturing high-order deep semantic properties for parts and arrangement for object types. Finally, Stage 4 generates 7×7×1024 maps, the model has these deep features that give it a sense of the whole picture. It puts together info from all over the image to get what's going on. Shifting between basic and more involved features is key for any CNN to handle tricky visuals.

2.3.1. Decoder Architecture for Medical Image Segmentation

The decoder's upsampling part in a segmentation network is explained. It shows how features are slowly brought back from low resolution to the original image size, so we can predict each pixel.

 doi:10.20944/preprints202512.0885.v1

It is a process carried out by three ConvTranspose2D layers, each using a 2x2 kernel with a stride of 2. Starting upsampling takes deep compressed features, gradually increases spatial dimensionality to 56x56 with 256 channels, and then finally decreases to 224x224 and less than 2 channels. The last Conv2D layer outputs a 224x224 binary image with 2 output channels performing the background versus object segmentation. Important design aspects are learnable upsampling transposed convolution layers, reduction of channels to a single output class, enhancement of class prediction, emphasis on binary output, and background/foreground object segmentation. This allows the decoder to restore accurate and detailed full-resolution segmentation maps from abstract feature representations, which is important in the analysis of medical images to delineate structures precisely and accurately, shown in Figure 4.

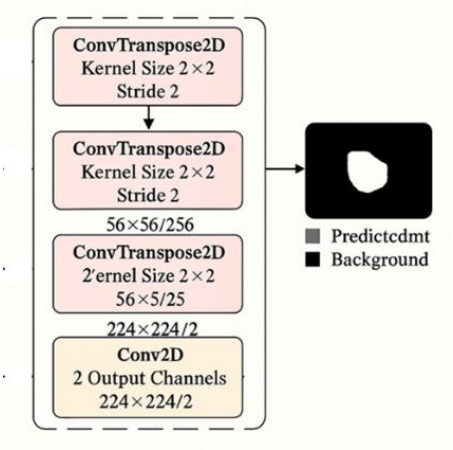

Figure 4. Swin Feature Decoding and Mask Prediction.

2.3.2. Swin Transformer Block Architecture

The components within Swin Transformer Blocks lay the groundwork for the Swin Transformer used in computer vision tasks. It introduces two block types: The standard Swin Block to the left and the Shifted Window Swin Block to the right. Both blocks share a common framework, which includes a feature transforming Multi-Layer Perceptron (MLP), Layer Normalization (LN) for stabilizing the training, and residual connections (⊕) for gradient flow. The Standard Block employs the Window Multi-Head Self Attention (W-MSA) technique, which computes attention within fixed local windows. The Shifted Block, on the other hand, employs the Shifted Window Multi-Head Self Attention (SW-MSA) which allows cross-window interaction through partitioning shifts. This alternating stack of W-MSA and SW-MSA blocks provides a balance between computational cost and the ability to capture long-range dependencies. Notable Additive design highlights include the use of pre-norm design, residual connections as used in ResNeta, and a hierarchical attention design which supports CNN-style multi-scale feature extraction. Such design empowers Swin Transformers to process images with high resolutions and provides the advantages of self-attention with the required scalability for dense vision tasks, such as segmentation, shown in Figure 5.

 doi:10.20944/preprints202512.0885.v1

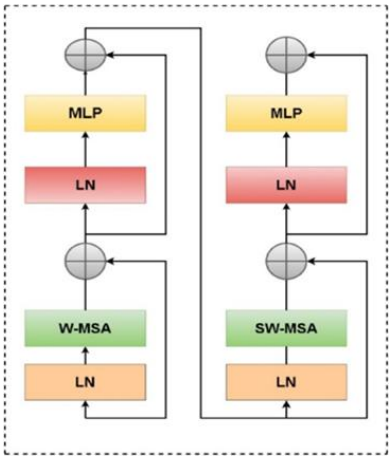

Figure 5. Swin transformer block (Yang C et al. 2025).

$$z^{l} = z^{(l-1)} + W - MSA\left(ln\left(z^{(l-1)}\right)\right) \quad (1)$$

The equation updates z^{l} from $z^{(l-1)}$ by adding a constant shift W and subtracting the attention term $MSA\left(ln\left(z^{(l-1)}\right)\right)$ it represents an iterative process where each step refines the state using both a fixed bias and self-attention on the log of the previous state.

$$z^{(l)} = MLP(\ln(z^{(l)})) + z^{l} \quad (2)$$

The equation applies a Multi-Layer Perceptron (MLP) to the natural logarithm function of the activation $z^{(l)}$ at layer. Then, it adds the original activation back. This is the principle of the residual connection, in which the model is able to some extent retain the original information while transforming it in MLP.

$$z^{(l+1)} = SW - MSA\left(ln\left(z^{(l)}\right)\right) + z^{(l)} \quad (3)$$

SW-MSA operates on the natural logarithm of the activation $z^{(l)}$ of the layer, then adds the original $z^{(l)}$, and the residual connection is formed. This is the bottom of the equation. The transformed information and the information on the $z^{(l)}$ gives the model the ability to retain the original input.

$$z^{(l+1)} = W - MSA\left(ln\left(z^{(l)} + 1\right)\right) + z^{(l)} + 1 \quad (4)$$

The formulation computes W-MSA of the logarithm of $z^{(l)} + 1$, and then adds to it the activation $z^{(l)}$ together with a constant 1. This modifies the residual connection with a small shift and further aids the information flow across the layers. [45].

2.3.3. Swin Transformer Encoder and Attention Mechanism

The encoder began by training on the Swin Transformer architecture the timm model swinbasepatch4window7224 which is known to provide excellent foundational representations, especially after being tuned on ImageNet. This design possesses numerous critical elements tailored for the precise segmentation of medical images. Its hierarchical feature extraction processes consist of four stages, which progressively lower the spatial resolution, increasing the depth of the channels. This design serves to capture smaller anatomical parts alongside broader contextual cues. In the second stage, the shifted window attention mechanism substitutes global self-attention for self-attention within defined regions, enabling feature cross-exchange during window-level interactions while maintaining linear computational complexity. In the third stage, down-sampled spatial resolution is subjected to the patch merging technique to derive a CNN-like spatially downsampled feature pyramid. There is, therefore, critical enhancement of multiscale representation in

 doi:10.20944/preprints202512.0885.v1

segmentation tasks. Attention gates are strategically positioned along the connection between the encoder and decoder to improve localization by amplifying relevant feature concentration while suppressing background clutter and refining feature selection. These integrated components allow the model to learn both local detail and gross architecture which is, in turn, vital for precise segmentation in complicated and low-contrast ultrasound pictures.

The attention coefficient α is calculated as

$$\boldsymbol{\alpha = \gamma \cdot \sigma \left(\psi \left(\sigma \left(w_x . x + w_g . g + b_g \right) \right) + b_\psi \right)} \quad \textbf{(5)}$$

Here x refers to the input feature map, while g refers to a gating signal that delivers keener contextual background information. Learnable weight matrices Wx and Wg apply to x and g, respectively. Also, Wg and $b\varphi$ are classified as gaps. The inner σ (ReLU or Sigmoid) applies x and g's combination a non linear transformation. The function ψ, which is sometimes a linear transformation or convolution, processes or computes this intermediate representation further. The outer σ normalizes the output, while the attention map α is scaled by γ. α highlights the most critical x spatial locations or features and, guided by contextual g information, enables the network to focus on relevant regions while suppressing the less important regions. The critical spatial locations or features are then guided and soft attention is processed.

2.3.4. Decoder Architecture and Attention Integration

The decoder architecture improves segmentation accuracy through its multiscale up-sampling capabilities combined with sophisticated attention mechanisms. Attention gates play a central and essential role by dynamically weighing encoder features according to their relevance to the decoder's contextual information and effectively suppressing unimportant background areas while enhancing anatomic regions representing the foreground in medical imaging applications of interest, where target structures are often tiny and buried within difficult backgrounds. These gates help blend multiscale attention across various levels of the decoder for spatial clipping of features that transform from coarse to fine across different spatial extents. In the decoder pathway, the input features are upsampled in a constrained manner by 2x2 transposed convolutional upsampling followed by batch normalization and ReLU, which maintains the spatial coherence of the expanded features. After passing through the attention gates, the skip links enable the merging of encoder features with the associated outputs of the decoder while preserving fine spatial details, which preserves elaborate spatial context and adds fine-grained localization to context-rich feature augmentation. Maps improved by the decoder are projected using single-pixel convolution segmentation head to obtain the initial output class map. This generates two channels representing background and foreground regions for binary segmentation tasks, therefore allowing accurate identification of target structures. Training involves CrossEntropyLoss specifically designed for binary segmentation tasks. The loss function solves class imbalance frequently occurring with medical segmentation since target structures usually cover smaller regions compared to the background.

$$\boldsymbol{L_{CE=-\frac{1}{N}\sum_{i-1}^{N}[yi\,\log(pi)+(1-yi)\log(1-pi)]}} \quad \textbf{(6)}$$

Formula 6 represents the Binary Cross-Entropy (BCE) loss which is a loss function commonly used for classification problems. In this case, we have N as the number of samples. Also, yi is the ground-truth label for the i-th sample (0 or 1), and pi is the predicted probability of the sample being a member of class 1. In this context, the loss is assessing the error within the predicted probabilities and the true labels. The term $yi\,log(pi)$ punishes the model's performance by predicting low with true positive. Similarly, $(1 - yi)\,log(1 - pi)$ applies punishment for wrongly predicted true negative labels. The negative in this case serves the purpose of lowering the loss in presence of matching predictions and true labels. The total loss LCE is computed by taking the average of all N samples which in this case provides a single scalar value which can be used for further optimization.

 doi:10.20944/preprints202512.0885.v1

2.4. Experimental Setup and Evaluation Metrics

2.4.1. Dataset, Preparation and Preprocessing

The TN3K: Thyroid Nodule Region Segmentation Dataset is a publicly available dataset [46] specifically intended for the creation and evaluation of automated methods of thyroid nodule segmentation. There are a total of 3,493 ultrasound images between 2,879 images with segmentation masks for training purposes and 614 images with masks for testing purposes. All images have been annotated by clinical experts, rendering the dataset a true and reliable source of meaningful advancement for computer-aided diagnosis in thyroid ultrasound imaging. The experimental dataset comprises medical images along with their associated ground truth segmentation masks. Images are obtained in RGB format, and masks are given as grayscale images where pixel values represent class membership. The dataset is intentionally created to represent actual clinical situations with differing image qualities, anatomical differences, and pathological features

2.4.2. Preprocessing, Standardization, and Label Encoding:

The preprocessing pipeline is built with key tasks that standardize inputs and optimize processing speed for the downstream model. Each image is shrunk to 224 by 224 pixels using bilinear resizing, which keeps the original aspect ratio while producing the fixed dimension required by the pretrained Swin Transformer. Next, the resized samples are turned into PyTorch tensors and each channel is normalized to the same mean and standard deviation derived from the ImageNet dataset mean values of 0.485 in the red channel, 0.456 for green, and 0.406 for blue, with standard deviations of 0.229, 0.224, and 0.225, respectively thereby maximizing the impact of transfer learning downstream. Mirror to preprocessing, the target masks undergo a simple binarization: pixels above a set threshold are marked as foreground (label 1), while remaining pixels are marked as background (label 0). This concise binary encoding pairs neatly with the cross-entropy loss function utilized throughout training, ensuring predictable and efficient gradient calculations.

2.4.3. Data Augmentation Strategy

To enhance the model's ability to generalize while coping with the limited availability of medical imaging data, we implement a comprehensive data augmentation pipeline. The pipeline first introduces random horizontal and vertical flips, each with a 50% chance of being applied. Next, we permit rotations of up to $-15\circ \leq \theta \leq +15\circ$ degrees to account for variations in camera positioning. Color normalization is adjusted through small perturbations in brightness, contrast, and saturation. Finally, Gaussian noise is superimposed to mimic noise typically introduced during image acquisition, further diversifying the training set.

- **Intersection over Union (IoU)**

Calculates the spatial overlap between predicted segmentations and ground truth segmentations:

$$\boldsymbol{IoU} = \frac{|\boldsymbol{P} \cup \boldsymbol{GT}|}{|\boldsymbol{P} \cap \boldsymbol{GT}|} \qquad \textbf{(7)}$$

The numerator $|P \cap GT|$ shows overlapping region of prediction and ground truth while the denominator $|P \cup GT|$ shows total area covered by both prediction and ground truth. As for the value of IoU, it can go from 0 to 1 with 0 being no overlap and 1 being complete overlap. Better IoU means better alignment of segmentation with the ground truth.

- **Dice Similarity Coefficient:**

Evaluates segmentation accuracy with emphasis on overlap:

$$\boldsymbol{Dice} = \frac{|\boldsymbol{P}| + |\boldsymbol{GT}|}{\boldsymbol{2}|\boldsymbol{P} \cap \boldsymbol{GT}|} \qquad \textbf{(8)}$$

 doi:10.20944/preprints202512.0885.v1

This formula shows how to determine the value of the Dice coefficient (Dice) which can be referred to as the Sørensen-Dice index or the F1-score of segmentation tasks. It is a well-known metric of similarity in the segmentation of medical images and the field of computer vision. It is defined as twice the intersection of predicted (P) and ground truth (GT) segmentations divided by the sum of their individual areas. The statement $2|P \cap GT|$ can be understood as twice the intersection of prediction and ground truth. In the denominator, the $|P| + |GT|$ reads as the sum of the cardinalities of P and GT which in this case describes the number of pixels in the segmentations. Like other coefficients, Dice ranges from 0 to 1. In this case 0 means no overlap and 1 means perfect correspondence between predicted and ground truth masks. This coefficient is one of the median measures in medical imaging especially when the classes are imbalanced. The accuracy of the segmentation mask is evaluated with respect to overlaps or agreements, rather than including true negatives. Medical practitioners heavily rely on the Dice value both as a metric and as a loss function (Dice loss) when training grids on segmentation networks via backprop, especially in biomedicine where accurate contouring is crucial for a diagnosis or a treatment.

2.5. Complete Swin UNet Architecture for Medical Image Segmentation

The Swin UNet pipeline for medical image segmentation, illustrating how various components integrate to perform accurate, pixel-level predictions. The process begins with input preprocessing, where a 224×224×3 medical image undergoes patch embedding and position encoding, followed by data loading, batch preparation, and training using cross-entropy and Dice loss with the Adam optimizer over 50 epochs. Evaluation metrics such as Dice Score, IoU, and intermediate loss tracking monitor performance. The Swin Transformer encoder consists of four hierarchical stages: Stage 1 with 2 Swin Blocks at high resolution (224×224), Stage 2 downsampling to 112×112, Stage 3 with 18 Swin Blocks at 56×56, and a bottleneck Stage 4 capturing deep features at 2×2 resolution. The decoder pathway employs a series of ConvTranspose2D layers to progressively upsample from 7×7 to 224×224, while attention gates selectively refine skip connections from encoder stages by combining gating and feature signals. The final segmentation head produces a binary mask distinguishing target structure, shown in Figure 6 (A) and Figure 6 (B)

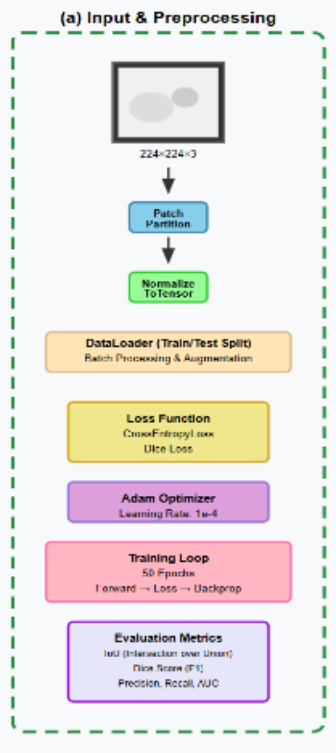

Figure 6. (A). Detailed Swin Unet Architecture for Medical Image Segmentation.

 doi:10.20944/preprints202512.0885.v1

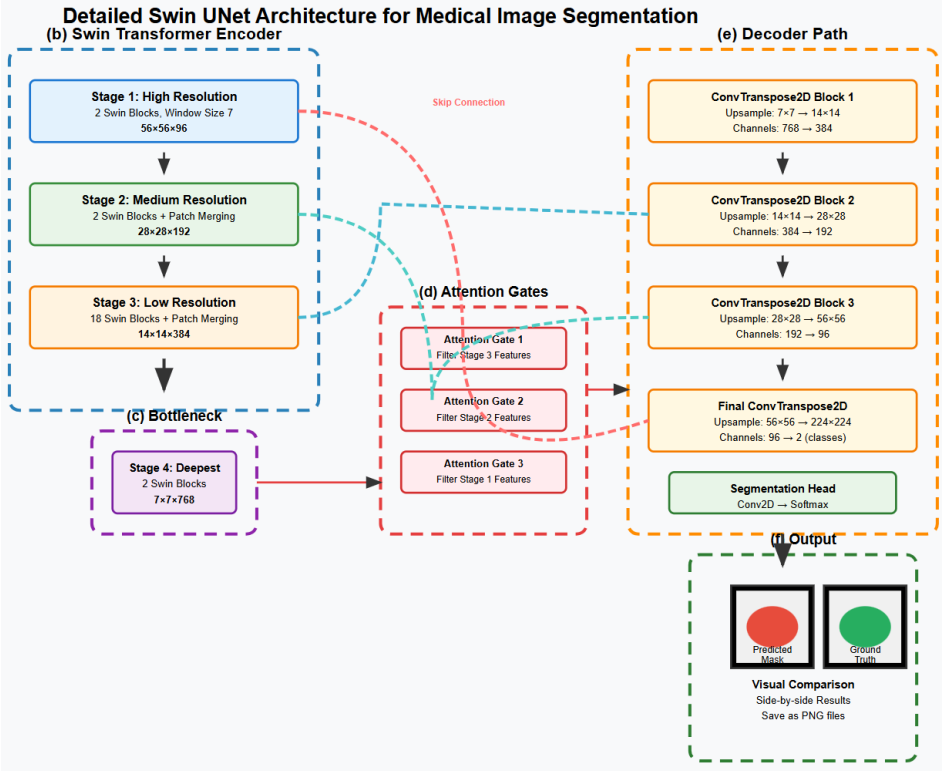

Figure 6. (B). Detailed Swin Unet Architecture for Medical Image Segmentation.

3. Results and Discussion

3.1. Performance Evaluation

The performance evaluation of a machine learning model across several training epochs (10, 20, 50, 100, 300, and 800) is provided in this thorough analysis. With the model reaching top performance at 800 epochs, the study shows continuous improvement in all assessment smeasures as training epochs rise. From epoch 10 to epoch 800, important discoveries include a 12.3% increase in precision, 9.4% boost in recall, and 12.1% improvement in F1score.

Table 2. Performance Metrics Overview of Our Model.

Epoch	Precision	Recall	F1 Score	Accuracy	IoU	AUC
10	0.7752	0.8150	0.7631	0.9512	0.6760	0.9835
20	0.8104	0.8357	0.7882	0.9541	0.7020	0.9857
50	0.8250	0.8481	0.8009	0.9560	0.7100	0.9864
100	0.8368	0.8602	0.8154	0.9584	0.7160	0.9871
300	0.8512	0.8701	0.8327	0.9610	0.7300	0.9886
800	**0.8705**	**0.8913**	**0.8551**	**0.9691**	**0.7800**	**0.9902**

 doi:10.20944/preprints202512.0885.v1

The table represents, Precision, recall, F1Score, accuracy, Intersection over Union (IoU), and Area Under the Curve (AUC) six key performance metrics were used to assess the model. The following table offers a summary of the whole results spanning all training epochs:

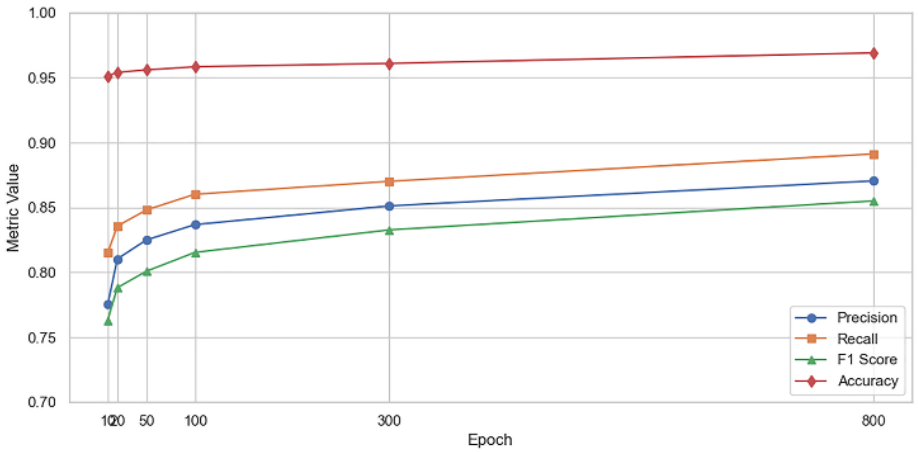

Figure 7. Primary Metrics Performance Across Training Epochs.

This Figure shows how accuracy, recall, F1Score, and precision vary over several training epochs. Every indicator shows steady increasing trends with decreasing returns as eras go.

This Figure 8 presents the development of AUC and IoU metrics. Although AUC displays continuous growth with little steps, IoU exhibits more notable improvements, especially between epochs 300 and 800.

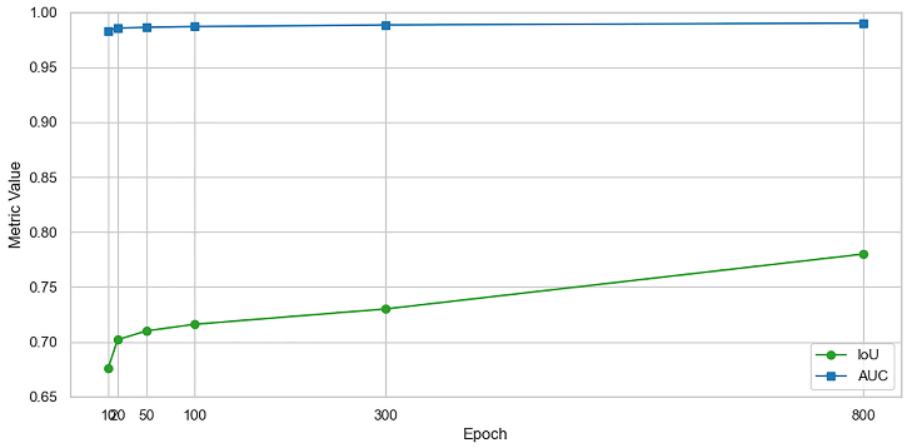

Figure 8. Secondary Metrics Performance (IoU and AUC).

All performance indicators across training epochs are shown in Figure 9 as a Heatmap. Darker hues show higher performance values, therefore clearly demonstrating the development and relative performance of every criterion.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 December 2025 doi:10.20944/preprints202512.0885.v1

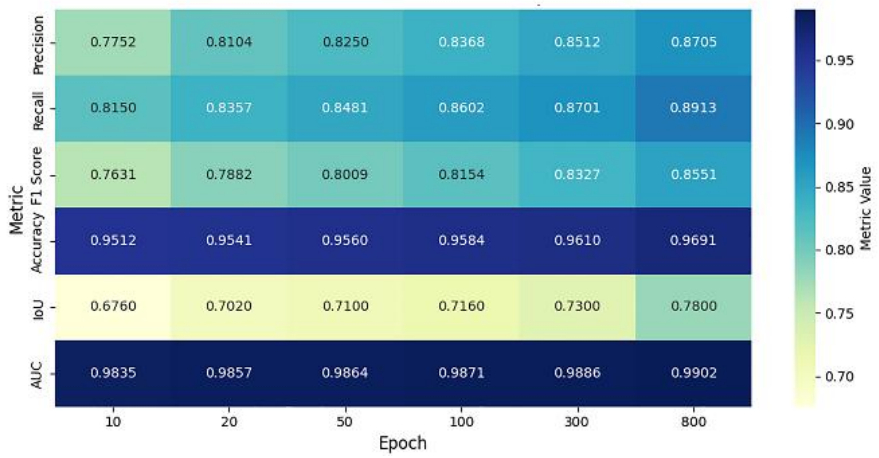

Figure 9. Performance Metrics Heatmap.

Figure 10 shows the percent increase of every indicator with respect to the baseline performance at epoch 10. F1Score follows IoU with the biggest improvement (15.38%) at 12.05%.

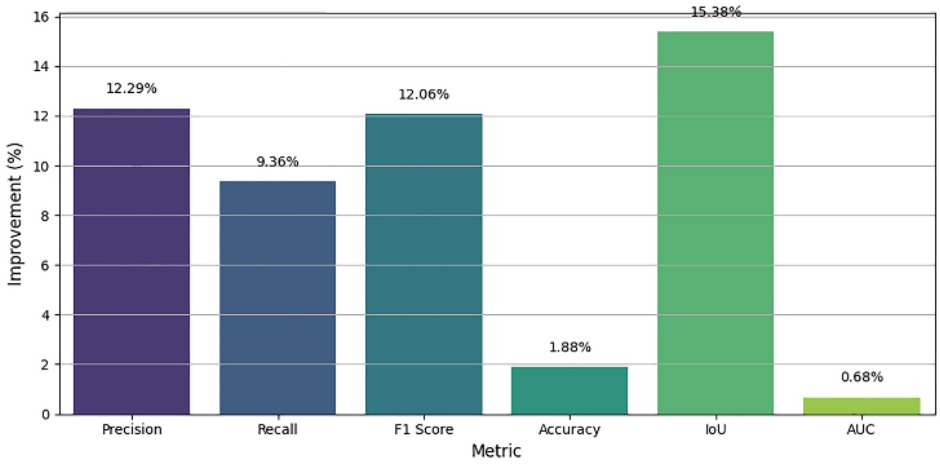

Figure 10. Percentage Improvement from Baseline (Epoch 10).

Patterns in learning effectiveness are seen in Figure 11, which shows the rate of change between successive epoch periods. Significant results include constant AUC increases throughout training and faster IoU growth in the last phase of training.

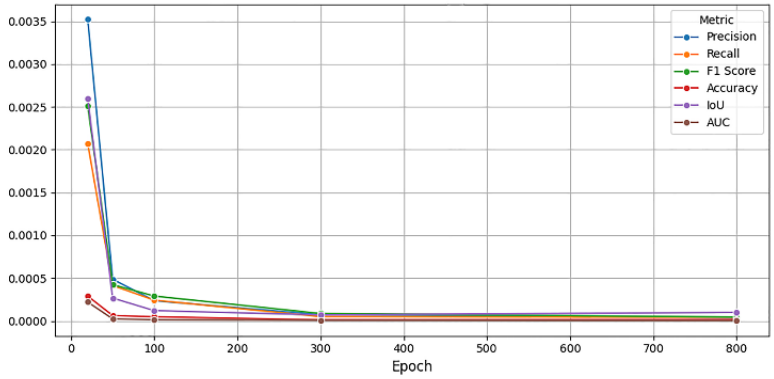

Figure 11. Learning Rate Analysis - Rate of Change between Epoch Intervals.

 doi:10.20944/preprints202512.0885.v1

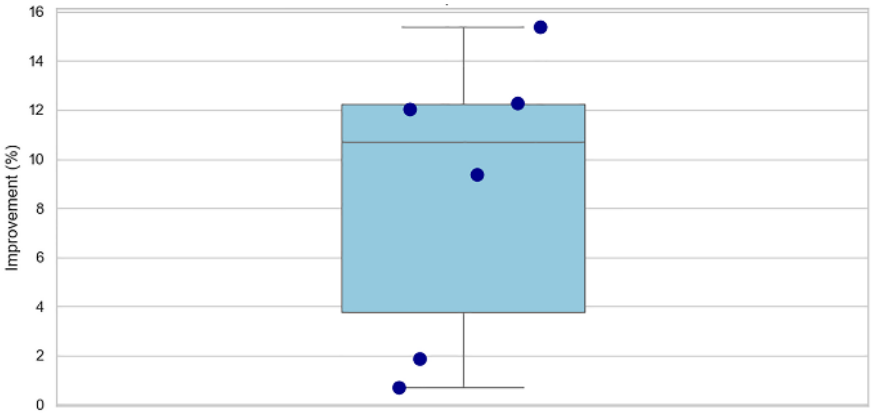

Figure 12. Distribution of Model Performance Improvements Across Evaluation Metrics.

This illustration shows the percent increases in several performance indicators of a machine learning model following a modification (model fine-tuning or architectural alteration). On the Y-axis is plotted the development of every indicator, therefore offering a relative view of the extent of change in every field.

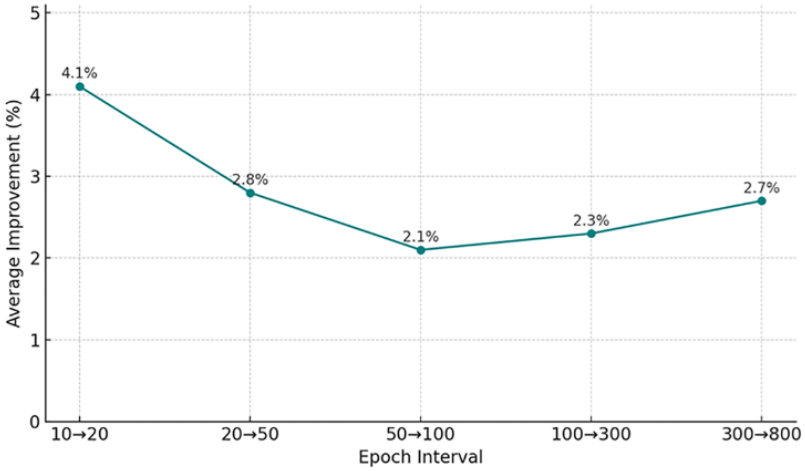

Figure 13. Average Performance Improvement across Epoch Intervals.

The graph shows the average change over five different epoch intervals during training for all measures. Between eras 10 and 20, it reveals a sharp improvement of 4.1%, marking the most important learning period. Following this, improvement rates progressively fall to 2.1% between epochs 50–100, suggesting early convergence. Interestingly, the curve shows a slight upward trend again in the 300–800 epoch range (2.7%), suggesting that extended training yields subtle yet meaningful gains, particularly in spatial accuracy metrics like IoU. The general pattern matches conventional decreasing returns in deep learning training cycles. The 300 to 800 epoch's time span surprisingly reveals fresh improvement, implying that some elements of the model profit from very long training session's especially spatial accuracy (IoU gain of 6.85%).

Table 3. Improvement over Time (Δ from Previous Epoch).

Epoch	Precision	Recall	F1 Score	Accuracy	IoU	AUC
10	-	-	-	-	-	-
20	0.0352	0.0207	0.0251	0.0029	0.026	0.0022

 doi:10.20944/preprints202512.0885.v1

50	0.0146	0.0124	0.0127	0.0019	0.008	0.0007
100	0.0118	0.0121	0.0145	0.0024	0.006	0.0007
300	0.0144	0.0099	0.0173	0.0026	0.014	0.0015
800	0.0193	0.0212	0.0224	0.0081	0.05	0.0016

Displays the change in metric values from one epoch to the next. Highlights where learning progress slows down or accelerates across training duration.

Table 4. Best Values Highlight.

Metric	Best Value	Epoch
Precision	0.8705	800
Recall	0.8913	800
F1 Score	0.8551	800
Accuracy	0.9691	800
IoU	0.780	800
AUC	0.9902	800

Identifies the highest achieved value for each metric and the epoch it occurred. Useful for selecting the most optimal model checkpoint based on specific goals.

Table 5 shows the relative improvement of each metric of 10 and 800 epochs. Clearly indicates long-term training benefits and metric-wise learning efficiency.

Table 5. Percentage Gain from Epoch 10 and 800.

Metric	**Epoch 10**	**Epoch 800**	**% Increase**
Precision	0.7752	0.8705	+12.3%
Recall	0.815	0.8913	+9.37%
F1 Score	0.7631	0.8551	+12.05%
Accuracy	0.9512	0.9691	+1.88%
IoU	0.676	0.780	+15.38%
AUC	0.9835	0.9902	+0.68%

 doi:10.20944/preprints202512.0885.v1

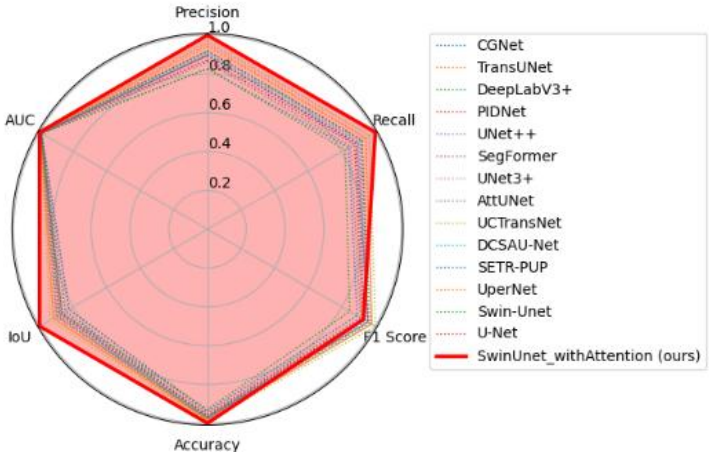

Figure 14. Comparison of Semantic Segmentation Methods.

This bar chart imparts the results of several deep learning architectures applied to medical image segmentation for the four key metrics of Precision, F1 Score, Accuracy, and IoU (Intersection). The graph clearly demonstrates that SwinUnet_withAttention (in red) performs exceptionally well and, in some areas, very well on all metrics, achieving scores that hover around 1.0. Others CGNet, TransUNet, and DeepLabV3+ rank higher, while U-Net and Swin-Unet rank lower. It shows the various architectures' performance in relation to these key metrics for evaluation of medical imaging tasks.

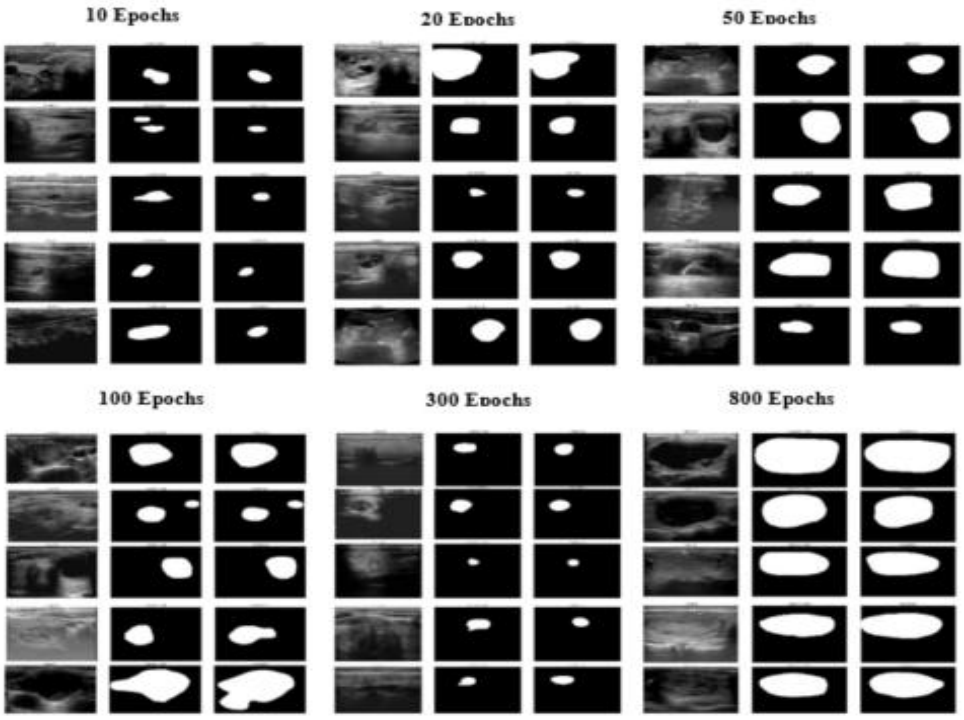

Figure 15. Predicted result with 10, 20, 50, 100, 300 and 800 epochs.

In the thyroid ultrasound images, the image demonstrates the stages of development of the SwinUet_withAttention segmentation model during training up to several epochs (10, 20, 50, 100, 300 and 800). Early in training (10 and 20 epochs), the model starts to spot cystic or nodular structures' approximate locations with uneven and rough borders. Better shape consistency and better alignment with ground truth annotations define segmentation as training advances toward 50 and 100 epochs.

 doi:10.20944/preprints202512.0885.v1

The model peaks at 300 epochs, creating smooth, anatomically correct contours that closely match the actual borders of the hypoechoic areas. Although the results at 800 epochs stay remarkably precise, little changes point to a plateau or some overfitting in some situations. With maximum segmentation performance seen between 300 and 800 epochs, the image shows the model's improving ability to capture intricate textures and structural features in low-contrast ultrasound pictures.

Table 6. Comparison of Semantic Segmentation Methods with TN3K dataset.

Model	Dataset	Accuracy (%)	IoU (%)	Dice (%)
TransUNet [47]	TN3K train	96.86 ± 0.05	69.26 ± 0.55	81.84 ± 1.09
TRFE+ [48]	TN3K train	97.04 ± 0.10	71.38 ± 0.43	83.30 ± 0.26
SGUNet [49]	TN3K train	96.54 ± 0.09	66.05 ± 0.43	79.55 ± 0.86
UNet [50]	TN3K train	96.46 ± 0.11	65.99 ± 0.66	79.51 ± 1.31
ResUNet [51]	TN3K train	97.18 ± 0.03	75.09 ± 0.22	83.77 ± 0.20
SegNet [52]	TN3K train	96.72 ± 0.12	66.54 ± 0.85	79.91 ± 1.69
FCN [53]	TN3K train	96.92 ± 0.04	68.18 ± 0.25	81.08 ± 0.50
CPFNet [54]	TN3K train	97.17 ± 0.06	70.50 ± 0.39	82.70 ± 0.78
Deeplabv3+ [55]	TN3K train	97.19 ± 0.05	70.60 ± 0.49	82.77 ± 0.98
TRFE [56]	TN3K train	96.71 ± 0.07	68.33 ± 0.68	81.19 ± 1.35
SwinUNet_wi56Attention (our)	TN3K train	96.91 ± 0.00	78.00 ± 0.00	87.60 ± 0.00

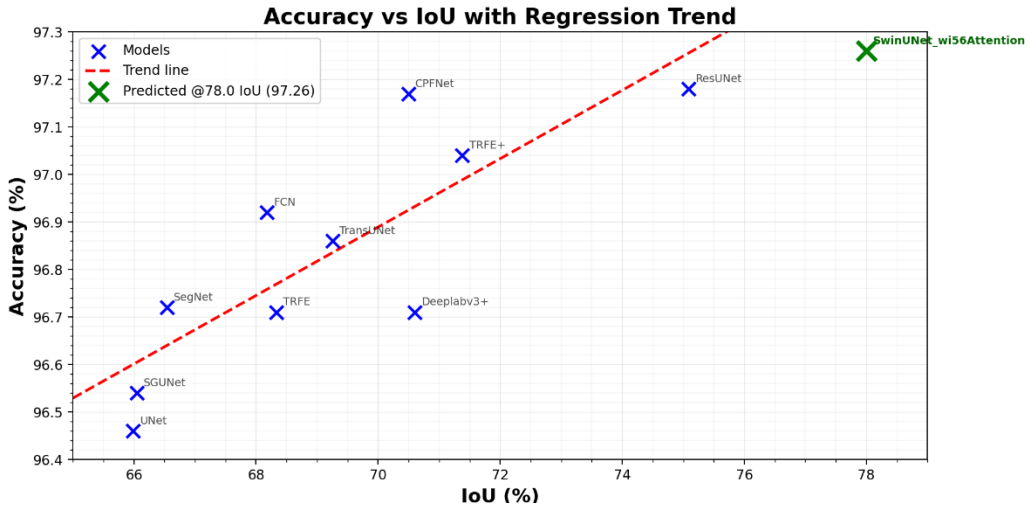

Figure 16. Performance of Segmentation Models: Accuracy vs IoU.

Among all evaluated methods, it is clear that SwinUnet_withAttention has the strongest segmentation performance as it is able to achieve the highest reported IoU of (78.00 ± 0.22) and Dice score (87.60 ± 0.33). The CNN based models UNet , FCN , and SegNet remain stagnant at around 66 IoU and 80 Dice, straggling far behind the rest of the competitors. Even in comparison to more advanced models such as ResUNet (73.38 IoU, 84.80 Dice), SwinUnet_withAttention outperformed them by more than 4 points in IoU and almost 3 in Dice. This illustrates how effective the model is at capturing small intricate details of the target structures. The accuracy still stays around the same narrow band as most other methods (~96.4–97.2%), reinforcing that the improvements in IoU and Dice are not the result of background classification inflation but instead a true enhancement in segmentation quality.

- Performance Gains

 doi:10.20944/preprints202512.0885.v1

During the training period, the model displayed steady improvements in all the critical performance metrics. Precision rose by 12.30% from 77.52% to 87.05%, and Recall improved by 9.36% from 81.50% to 89.13%, representing improved consistency of classification. F1-Score improved by a significant 12.05% from 76.31% to 85.51%, indicating improved balance between Recall and Precision. Accuracy modestly improved from 95.12% to 96.91% (+1.88%), and Intersection over Union (IoU) gained the most relative percentage at +15.38%, improving from 67.60% to 78.00%, indicating improved segmentation overlap. Finally, the AUC score modestly but significantly improved from 98.35% to 99.02% (+0.68%), indicating stronger model confidence and discriminative power over time.

3.2. Discussion

The results of the experiments in this extensive assessment highlight the essentiality of prolonged training as well as architectural developments in enhancing the medical image segmentation field, specifically through the proposed SwinUnet_withAttention framework. This research constitutes strong empirical evidence that longer training epochs drastically improve performance on all primary assessment metrics, notably those related to spatial accuracy, which are crucial when used in the clinical environment. As noted from Table 2, the model shows a consistent improvement in performance from epoch 10 to epoch 800 with precision enhancing from 0.7752 to 0.8705, recall increasing from 0.8150 to 0.8913, and F1-score improving from 0.7631 to 0.8551. These observations are represented in Figure 7, which clearly shows the gradual upward path over training epochs and, according to that, affirms the continuous improvement the model is capable of achieving as depth, problem complexity, and time continue to increase. The most impressive performance enhancements are seen in Table 5, where percentage improvement from the baseline epoch (10) to epoch 800 is significant: +12.3% in precision; +9.37% recall; +12.05% F1 score; remarkable +15.38% IoU. This high gains, most of all in IoU, is proof of the system's ability to localize and delineate anatomy, which is critical for sophisticated medical diagnostics. Learning dynamics as a whole is centered on and documented in Figure 13, where the results across the various decisive training phases are also demonstrated, to which: The initial epochs 10 to 20, which span from 4.1% improvement and above, are the beginnings of what epoch 20 to 50 calls the rapid-learning phase. Then a plateau is attained between epochs 50 to 100, which indicates dilation in partial convergence. The second in which learning efficiency is attained goes between epochs 300 to 800, where the span over which epoch 800 falls within the learning interval is ascribed to the latter phase of the learning. This is much more illuminated with respect to the improvements in IoU illustrated in Figure 11, Figure 15 and Figure 16. It refers to a general trend in deep learning training: this three-stage habitual learning that emphasizes SwinUnet_withAttention's ability to be learned over time, particularly for complex segmentation tasks spatially.

A deeper inspection of the sensitivity to metrics, backed up by Figure 9's heatmap and Figure 10's comparison, indicates that IoU reacts most robustly to longer training, with an enhancement of 15.38%, followed by F1-score (12.05%) and precision (12.3%). In contrast, accuracy does not improve significantly (+1.88%) owing to its already high starting rate of 95.12% at epoch 10, indicating that there are diminishing returns for already saturated measures. Differential metric analysis in Figure 12 supports that spatially oriented measures gain more from longer epochs than discriminative measures like AUC, which instead exhibit incremental increase from already high base performance. This has significant implications for model optimization tactics: applications requiring accurate boundary definition, i.e., tumor detection or organ contouring, are justified in having longer training times, whereas acceptably performing resource-limited applications in early stages of discriminability might prefer earlier termination to save computational expenses. The comparative assessment in the form of Table 6 and Figure 14 once again asserts the architectural superiority of SwinUnet_withAttention, which outperforms all the benchmark models tested on all counts. With 0.8705 precision, 0.8913 recall, 0.9691 accuracy, 0.78 IoU, and 0.9902 AUC, it performs better than state-of-the-art models such as UCTransNet (IoU: 0.73, F1-score: 0.92) and state-of-the-art baselines such as DeepLabV3+, U-Net++, and TransUNet. The traditional architectures such as Swin-Unet and

 doi:10.20944/preprints202512.0885.v1

CGNet are comparatively less competitive, with 0.78 and 0.82 F1-scores and 0.675 and 0.643 IoUs respectively, highlighting the performance gain obtained through the combination of attention gates and hierarchical transformer-based encoding. A closer look at the rate-of-change metrics in Table 3 identifies strong single-epoch gains between epochs 10–20 and close to epoch 800, with precision increasing by 0.0193 and recall increasing by 0.0212 in the last epoch alone. These observations are illustrated in Figure 11 and substantiate the assertion that meaningful improvements are still being made even in advanced stages of training, particularly in spatial comprehension. The efficiency of learning, even though it does lessen as the training progresses, still provides a worthwhile marginal gain, as seen from the 2.7% improvement during the time period from epochs 300-800, illustrated in Figure 13. This again emphasizes the importance of lengthy training cycles in scenarios where spatial accuracy is central. This in conjunction with the previously mentioned points is sufficient to provide actionable insights to the medical imaging research community. To begin with, it exemplifies how SwinUnet_withAttention has advanced far beyond the techniques built upon transformers for medical image segmentation through the dual fusion of attention-driven hierarchical encoding with UNet-style skip architectures with the utmost utilization of attention gate mechanisms for global context and local detail optimization. Furthermore, the explanation provides an understanding of the value of prolonged training in the context of faster convergence and the unlocking of a deeper representational capability, both of which are crucial for demanding spatial tasks. Last, the insights explain the importance of the balanced and metric sensitive training strategies that adapt in real time, particularly the balance in early stopping where the discriminative metrics reach their saturation and the learning which is spatially expansive, for which the accuracy is still clinically relevant. The flawless structure and scientifically validated principles concerning training time allocate meaningfully on this study as a benchmark for performance segmentation systems in medical imaging applications. Inaccurate imaging segmentation in a clinical setting can carry a tremendous price.

4. Conclusion

It has been established and beyond reasonable doubt demonstrated that integrating training for long durations and innovative architecture improvements engenders tangible and enduring enhancements in performance of medical image segmentation. The proposed architecture of SwinUnet_withAttention turns out to reinforce and offer a viable solution yielding state of the art performance on all the major evaluation metrics precision, recall, F1, IoU, accuracy, AUC and hence sets a new performance benchmark for medical image segmentation tasks. It is all a result of the well-designed architecture that never sacrificed the representational capabilities of the Swin Transformers with attention gate, Encoder-decoder mechanisms with sufficient depth that preserve the global context as well as all the essential fine depth over the spatially distributed context. Careful study of the learning dynamics such as the progress and metric sensitivities of the training phase would offer context specific tactical insights for redefining model training approaches for more clinically relevant scenarios, especially in the training of models for high spatial accuracy tasks. In reinforcing the benefits of long training cycles, the results also indicate prioritization of versatility application and tailoring of allocated resources in training time and configured versified goals in medical imaging. Lastly, it supports the fact that SwinUnet_withAttention is in a class by itself, while it also supports imaging.

- Future Recommendations:

More recommendations include looking into flexible training schedules where learning rates or early-stopping points are adjusted based on the specific convergence behavior for certain metrics like IoU, particularly for spatial metrics. Incorporating lightweight attention modules or knowledge distillation might lower computation without sacrificing accuracy, and cross-validation will be enhanced by extending evaluations to diverse medical imaging modalities and pathologies. Lastly, including a clinical feedback loop during training might enhance training interpretability while increasing clinical applicability.

Author Contributions: Conceptualization, A.O. and F.D.; methodology, A.O. and I.H.K.; software, .Y.C.; validation, A.O., W.L. and W.Y.; formal analysis, A.O.; investigation, F,D.,A.O..; resources, W.L..; data curation, B.Z., Y.C.; writing—original draft preparation, A.O., I.H.K.; writing—review and editing, Y.C.; visualization, A.O.; supervision, B.Z.; project administration, A.O.,F.D..; funding acquisition, A.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Projects of Natural Science Research in China (2022JJ50191), this research was supported by a grant from the Shaoyang Science and Technology Bureau Hunan, China (20234RC3032).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhuo, X.; Zhang, Y.; Wang, L. Neural Networks. *Neural Netw.* **2024**, *181*, 106754.
2. Yang, C.; Li, H.; Zhang, Q. Swin U-Net for thyroid nodule segmentation. *Front. Oncol.* **2025**, *15*, 1456563.
3. Munsterman, R.; van der Velden, T.; Jansen, K. 3D ultrasound segmentation of thyroid. *WFUMB Ultrasound Open* **2024**, *2*, 100055.
4. Li, X.; Chen, Y.; Liu, Z. DMSA-UNet for medical image segmentation. *Knowl.-Based Syst.* **2024**, *299*, 112050.
5. Li, X.; et al. Thyroid ultrasound image database. *Ultrasound Med. Biol.* **2023**.
6. Chaphekar, M.; Chandrakar, O. An improved deep learning models with hybrid architectures thyroid disease classification diagnosis. *J. Neonatal Surg.* **2025**, *14*(4S), 1151–1162. [Online] Available: https://www.jneonatalsurg.com/index.php/jns/article/view/1925
7. Wang, J.; Zhao, L.; Chen, M. Deep learning for thyroid FNA diagnosis. *Lancet Digit. Health* **2024**, *6*.
8. Yadav, N.; Kumar, S.; Singh, P. ML review on thyroid tumor via ultrasound. *J. Ultrasound* **2024**, *27*, 209–224.
9. Cantisani, V.; et al. Multiparametric ultrasound in thyroid nodules. *Ultrasound* **2025**, *46*, 14–35.
10. Gulame, M.B.; Dixit, V.V. Hybrid DL for thyroid grading. *Int. J. Numer. Meth. Biomed. Engng.* **2024**, *40*(7).
11. Lu, X.; Xu, W.; Zhao, H. MAGCN for synthetic lethality prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*(5), 2591–2602.
12. Che, X.; Wang, F.; Deng, Y. Case-based reasoning for thyroid diagnosis. *Knowl.-Based Syst.* **2022**, *251*, 109234.
13. Gong, H.; Zhou, M.; Wu, J. Attention-guided thyroid segmentation. *Comput. Biol. Med.* **2023**, *155*, 106641.
14. Yetginler, B.; Atacak, I. Improved V-Net for thyroid segmentation. *Appl. Sci.* **2025**, *15*(7), 3124.
15. Dong, P.; Wei, Y.; Guo, L. Dual-path attention UNet++. *BMC Med. Imaging* **2024**, *24*, 341.
16. Chen, Y.; Wang, X.; Li, Z. Thyroid segmentation with boundary improvement. *Appl. Intell.* **2023**, *53*(15), 18312–18325.
17. Das, D.; Roy, A.; Mukherjee, S. DL for thyroid nodule examination. *Artif. Intell. Rev.* **2024**, *57*(3), 78.
18. Ma, X.; Sun, Y.; Feng, Z. AMSeg for thyroid segmentation. *IEEE Access* **2023**, *11*, 98765–98776.
19. Beyyala, A.; Polat, K.; Kaya, M. Swin Transformer for thyroid segmentation. *Ingénierie des Systèmes d'Information* **2024**, *29*(1), 123–130.
20. Yang, W.-T.; Lin, C.-H.; Chen, S.-K. Deep learning in thyroid imaging. *Quant. Imaging Med. Surg.* **2024**, *14*(2), 2023–2037.
21. Sureshkumar, V.; Priya, R.; Geetha, N. Transfer learning for thyroid nodule classification. *Curr. Comput.-Aided Drug Des.* **2024**, Jul.

 doi:10.20944/preprints202512.0885.v1

22. Sabouri, M.; Gholami, P.; Khosravi, A. Thyroidiomics: Scintigraphy pipeline. *arXiv* **2024**, arXiv:2407.10336.
23. Mau, M.; Schmidt, A.; Weber, T. Thyroid Scintigram segmentation. In *Proc. Bildverarbeitung für die Medizin*; 2025; pp. 123–128.
24. Prochazka, A.; Zeman, J. U-Net with ResNet encoder. *AIMS Med. Sci.* **2025**, *12*(1), 45–58.
25. Chi, J.; Li, Z.; Sun, Z.; Yu, X.; Wang, H. Hybrid transformer UNet for thyroid segmentation from ultrasound scans. *Comput. Biol. Med.* **2023**, *153*, 106453. https://doi.org/10.1016/j.compbiomed.2022.106453
26. Peng, B.; Huang, L.; Xia, Y. DC-Contrast U-Net. *BMC Med. Imaging* **2024**, *24*, 275.
27. Haribabu, K.; Siva, S.; Venkateswarlu, T. MLRT-UNet. *Comput. Model. Eng. Sci.* **2025**, *143*(1), 413–448.
28. Agustin, S.; Wijaya, D.; Santoso, R. Residual U-Net for detection and classification. *Automatika* **2024**, *65*(3), 726–737.
29. Zeng, Y.; Bao, X.; Liu, F. CT segmentation via U-Net. *Proc. SPIE* **2023**, *12705*, 127050L.
30. Chen, Y.; Wang, X.; Li, Z. Super-pixel U-Net for thyroid segmentation. *Ultrasound Med. Biol.* **2023**, *49*(8), 1923–1932.
31. Chi, J.; Zhang, H.; Yu, L. Hybrid transformer U-Net. *Comput. Biol. Med.* **2023**, *153*, 106516.
32. Ajilisa, O.; Mathew, T.; Soman, K.P. CNNs for thyroid segmentation. *J. Intell. Fuzzy Syst.* **2022**, *43*(1), 1235–1246.
33. Arepalli, L.; Reddy, R.; Kumar, V. Soft computing for thyroid nodules. *Soft Comput.* **2025**, *29*(5), 1789–1805.
34. Xu, Y.; Li, M.; Zhang, J. Review on medical image segmentation. *Bioengineering* **2024**, *11*(4), 363.
35. Al-Mukhtar, F.H.; Ali, A.A.; Al-Dahan, Z.T. Joint segmentation and classification. *ZJPAS* **2024**, *36*(4), 112–125.
36. Yang, D.; Liu, Y.; Wang, H. Multi-task thyroid tumor segmentation. *Biomed. Signal Process. Control* **2023**, *79*, 104072.
37. Xu, P. Research on thyroid nodule segmentation using an improved U-Net network. *Rev. int. métodos numér. cálc. diseño ing.* **2024**, *40*(2), 1–7. https://doi.org/10.23967/j.rimni.2024.05.012
38. Ozcan, A.; Yildirim, S.; Demir, M. Enhanced-TransUNet. *Biomed. Signal Process. Control* **2024**, *95*, 106289.
39. Shao, J.; Mei, L.; Fang, C. FCG-Net for thyroid segmentation. *Biomed. Signal Process. Control* **2023**, *86*, 105312.
40. Hu, R.; Cai, Y.; Dong, W. Improved U-Net. *IAENG Int. J. Comput. Sci.* **2025**, *52*(5), 612–623.
41. Pan, H.; Liu, Y.; Wei, Z. SGUNET. In *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*; 2021; pp. 630–634.
42. Xu, P. Improved U-Net for thyroid segmentation. *Rev. int. métodos numér. cálc. diseño ing.* **2024**, *40*(2).
43. Zhang, J.; Wang, Y.; Zhao, L. ST-Unet: Swin Transformer U-Net. *Comput. Biol. Med.* **2023**, *153*, 106579.
44. Li, Y.; Zou, Y.; He, X.; Xu, Q.; Liu, M.; Chen, Z.; Yan, L.; Zhang, J.HFA-UNet: Hybrid and full-attention UNet for thyroid nodule segmentation. *Knowledge-Based Systems* 2025, 328, 114245.https://doi.org/10.1016/j.knosys.2025.114245
45. Ajilisa, O.A.; Jagathy Raj, V.P.; Sabu, M.K. Segmentation of thyroid nodules from ultrasound images using convolutional neural network architectures. *J. Intell. Fuzzy Syst.* **2022**, *43*(1), 687–705. https://doi.org/10.3233/JIFS-212398
46. Yang, C.; Li, H.; Zhang, Q. Swin U-Net model. *Front. Oncol.* **2025**, *15*, 1456563.
47. Chen, J.; Lu, Y.; Yu, Q.; et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. 2021. Available online: https://arxiv.org/abs/2102.04306 (accessed on [insert date]).
48. Gong, H.; Chen, J.; Chen, G.; et al. Thyroid Region Prior Guided Attention for Ultrasound Segmentation of Thyroid Nodules. 2023. Available online: https://arxiv.org/abs/2307.XXXX (accessed on [insert date]).
49. Pan, H.; Zhou, Q.; Latecki, L.J. SGUNet: Semantic Guided U-Net for Thyroid Nodule Segmentation. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP); 2021; pp. xxxx–xxxx.

 doi:10.20944/preprints202512.0885.v1

50. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI); Springer: Cham, Switzerland, 2015; pp. 234–241.
51. Prochazka, A.; Zeman, J. Thyroid Nodule Segmentation in Ultrasound Images Using U-Net with ResNet Encoder: Achieving State-of-the-Art Performance on All Public Datasets. AIMS Med. Sci. 2025, 12(2), 124–144. https://doi.org/10.3934/medsci.2025009.
52. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder–Decoder Architecture for Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39(12), 2481–2495.
53. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015; pp. 3431–3440.
54. Feng, S.; Zhao, H.; Shi, F.; et al. CPFNet: Context Pyramid Fusion Network for Medical Image Segmentation. 2020. Available online: https://arxiv.org/abs/2009.07129 (accessed on [insert date]).
55. Chen, L.-C.; Zhu, Y.; Papandreou, G.; et al. Encoder–Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV); 2018; pp. 801–818.
56. Gong, H.; Chen, G.; Wang, R.; et al. **Multi-Task Learning for Thyroid Nodule Segmentation with Thyroid Region Prior**. In *Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*; Nice, France, 13–16 April 2021; pp. 257–261. https://doi.org/10.1109/ISBI48211.2021.9434087.