Elsevier required licence: © <2025>. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/
The definitive publisher version is available online at 10.1016/j.ultrasmedbio.2025.02.013

# SSAT-Swin: Deep Learning-based Spinal Ultrasound Feature Segmentation for Scoliosis using Self-supervised Swin Transformer

Chen Zhang[a], Yongping Zheng[b], Jeb McAviney[c], Sai Ho Ling[a1]

[a]School of Electrical and Data Engineering University of Technology Sydney NSW Australia
[b]Department of Electronic and Information Engineering Hong Kong Polytechnic University Hong Kong China
[c]ScoliCare Clinic Sydney (South) Kogarah NSW 2217 Australia

## Abstract

**Objective:** Scoliosis, a three-dimensional spinal deformity, requires early detection and intervention. Ultrasound Curve Angle (UCA) measurement using ultrasound images has emerged as a promising diagnostic tool. However, calculating UCA directly from ultrasound images remains challenging due to low contrast, high noise, and irregular target shapes. Accurate segmentation results are therefore crucial to enhance image clarity and precision before UCA calculation.

**Methods:** We propose the SSAT-Swin model, a transformer-based multi-class segmentation framework designed for ultrasound image analysis in scoliosis diagnosis. The model integrates a boundary enhancement module in the decoder and a channel attention module in the skip connections. Additionally, self-supervised proxy tasks are used during pre-training on 1,170 images, followed by fine-tuning on 109 image-label pairs.

**Results:** The SSAT-Swin achieved Dice scores of 85.6±0.8% and Jaccard scores of 74.5±1.2%, with a 92.8% scoliosis bone feature detection rate, outperforming state-of-the-art models.

**Conclusion:** Self-supervised learning enhances the model's ability to capture global context information, making it well-suited for addressing the unique challenges of ultrasound images, ultimately advancing scoliosis assessment through more accurate segmentation.

*Keywords:* Medical image segmentation, Self-Supervised learning, Swin-Transformer, Scoliosis assessment, Ultrasound image

## 1. Introduction

Scoliosis, a deformity of the spine, typically presents as an "S" or "C" shape with a curvature exceeding 10 degrees [1]. It has become a significant health issue globally [2, 3, 4], making early detection and intervention critical to mitigating the associated risks [5, 6].

Radiographic assessment using Cobb's angle remains the gold standard for scoliosis diagnosis [7], supported by research demonstrating high detection accuracy [8, 9]. While measurements based on spinous processes (SP)

and transverse processes (TPs) show good reliability for estimating curve angles, they tend to underestimate the curve magnitude compared to the radiological Cobb method [3, 10]. The center of lamina (COL) method offers high reliability, with variations in estimation accuracy across different curve severities [11]. Beyond curve measurement, ultrasound provides significant advantages in assessing traumatic spinal cord injuries, infant spinal abnormalities, scoliosis treatment, and guiding invasive neuraxial interventions [12, 13, 14]. Therefore, with most methods reliance on multiple X-ray scans introduces significant radiation exposure, urgently necessitating the development of non-invasive, radiation-free diagnostic alternatives [15]. In addition, recent research focuses on the Ultrasound Curve Angle (UCA), a non-invasive, radiation-free method for spinal curvature assessment [16], offering real-time images and ease of use despite challenges of low contrast and irregular target shapes [17, 18].

However, ultrasound image segmentation of spinal structures presents significant challenges due to inherent image complexities and ambiguous anatomical boundaries. Convolutional Neural Networks (CNNs) have demonstrated limitations in accurately detecting Thoracic and Lumbar Bony Features (TBF/LBF, as shown in Figure 1), with previous studies plagued by high feature missing rates and poor detection accuracy [19, 20, 21], which negatively impact the accuracy of algorithms like the UCA calculation [22]. When a spinal feature is missed, the algorithm attempts to compensate by assuming or using previously detected features, which may not be accurate. Therefore, there is a need for a more advanced model that not only enhances segmentation performance but also improves feature detection to minimize such inaccuracies.

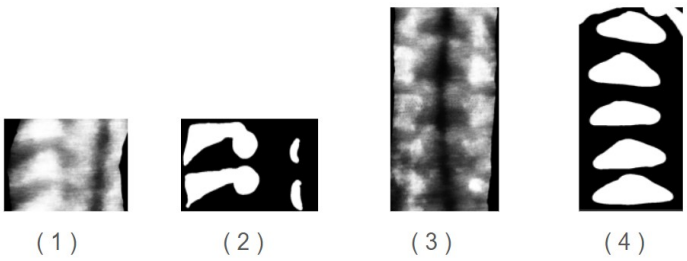


Figure 1: The comparsion of raw and label images: (1) is the TBF raw image (2) is the corresponding label. (3) is the raw image of LBF and (4) is the corresponding label

To address the challenges of ultrasound images in scoliosis diagnosis, we propose a self-supervised learning framework for pre-training the SSAT-Swin model. The pre-training phase leverages tailored proxy tasks, including image inpainting, rotation prediction, and contrastive learning, to help the model learn robust feature representations by capturing anatomical patterns from different spinal regions. This process, conducted on 1170 ultrasound images, allows the model to handle the complex characteristics of ultrasound data, such as low contrast and high noise. After pre-training, the model is fine-tuned on 109 ultrasound image-label pairs for scoliosis diagnosis.

Our main contributions include:

- A novel self-supervised learning framework: We pre-train the SSAT-Swin model on 1170 ultrasound images using tailored proxy tasks to enhance feature extraction, followed by fine-tuning on a dataset of 109 image-label pairs.
- Introduction of a boundary enhancement module: This module, integrated into the decoder side of the Swin Transformer block and the final projection layer, enhances boundary representation and highlights the region of interest (ROI), improving segmentation performance.
- Integration of a channel attention module: This module, applied to the skip connections at each layer, ensures that critical information is emphasized and not overlooked, particularly in the lower-level skip connections.

## 2. Relate works

Scoliosis causes gradual sagging of the spinal cord over time [2], highlighting the importance of a radiation-free automated assessment system. Achieving this necessitates accurate segmentation results.

### *2.1. Landmark Identification for automation*

In the measurement of the UCA, lateral features of the spine are prioritized over the spinous process as the anatomical reference for angle calculation. Before calculating the UCA, it is essential to first identify the exact locations of thoracic and lumbar bony features, particularly those above and below the T12 level (as Figure 2 shows), and determine the most tilted thoracic and lumbar bony feature pairs [23]. Clear identification of TBFs and LBFs plays a critical role in accurate UCA measurement.

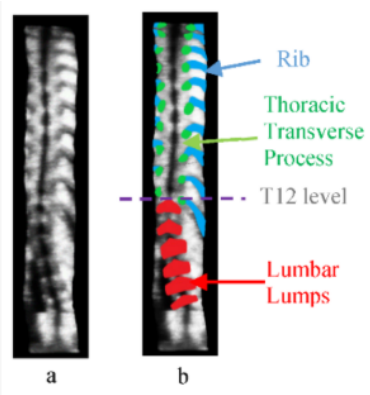


Figure 2: Various regions of interest in an ultrasound spinal image, with the original image (a) and its corresponding marked bone features (b).

Currently, this identification process is manual, heavily reliant on the expertise of doctors, and time-consuming. To automate this step, it is necessary to develop an architecture capable of accurately segregating bony features from ultrasound images, amidst noise and speckles [24]. Once TBFs and LBFs are successfully segmented, they can be used to identify the most skewed regions of the spine for UCA measurement. Therefore, segmentation is a crucial step in automating scoliosis analysis using the UCA method.

### *2.2. Current Segmentation Method*

Numerous models have excelled in medical image segmentation, particularly CNN-based architectures like U-Net and its variants, which have demonstrated commendable results [25, 26]. Specific models, such as Dense-U-Net for general segmentation [27] and H-DenseUNET for liver and tumor segmentation [28], have shown promising outcomes. Attention mechanisms have further enhanced segmentation, with ASCU-Net employing attention gates for skin lesion detection [29].

However, despite these advancements, traditional CNNs face challenges in effectively extracting long-range features, which are essential for accurately segmenting complex medical images [25].

The advent of transformer-based networks has introduced new possibilities for medical image segmentation. Initially developed for natural language processing [30], Transformers have shown significant success in computer vision, particularly with the Vision Transformer [31]. However, challenges remained regarding local information extraction and fixed receptive fields. The Swin Transformer [32] addressed these by implementing a window-based self-attention mechanism, effectively integrating local and global features.

Swin-Unet has emerged as a leading transformer-based model in medical imaging, achieving remarkable results across various competitions and datasets [33]. Additionally, SwinUNETR [34] is a 3D model designed for self-supervised pre-training, achieving state-of-the-art performance on public leaderboards like MSD and BTCV.

In conclusion, while CNN-based networks have historically excelled in computer vision, particularly in medical image segmentation [35], the emergence of transformer-based algorithms has opened up new opportunities for advancement in this field.

## 3. Methods

### *3.1. Architecture overview*

The proposed architecture, illustrated in Figure 3. The input image is divided into 4 patches, and attention scores are computed using a window size of 7 within each patch. This process uses window-based multi-head self-attention (W-MSA) and shifted window-based self-attention (SW-MSA) from Swin-Transformer [32] to capture both local and

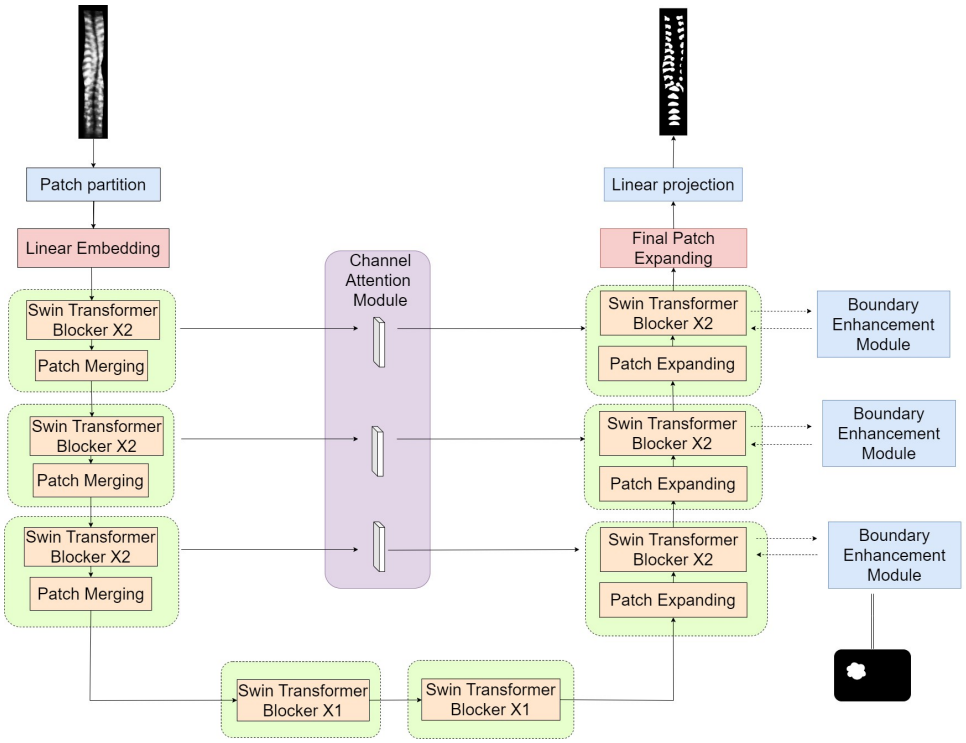


Figure 3: The SSAT-Swin model features two key modules: the boundary enhancement module, which improves boundary detection, and the channel attention module, which emphasizes important channels to ensure crucial information is highlighted during processing.

global features. In the encoder, patch merging reduces dimensionality, while patch expanding occurs in the decoder. Each skip connection incorporates a channel-attention module [36] to emphasize important channels. Additionally, a boundary enhancement module, inspired by the attentive feedback network [37], is integrated into each decoder layer's Swin Transformer block and the final projection, enhancing edge detection.

### *3.2. Boundary Enhancement module*

To augment the feature's integrity, we used a boundary enhancement module, showcased in Figure 4, which effectively separates the background and ROI while accentuating the boundary of the bone feature.

While predicting the shape and structure of targeted bone features, the segmentation result is enhanced by the boundary enhancement component. To achieve that, we get the predicted result P($i$), use a maxpool operation to obtain the feature map P($m$), and subtract P($i$) from P($m$) to get the boundary map P($b$) [37]. In the end, in order to highlight the boundary, we add that with P($i$) to enhance the boundary. The boundary enhancement is used at different levels within the network as well as after the final projection layer.

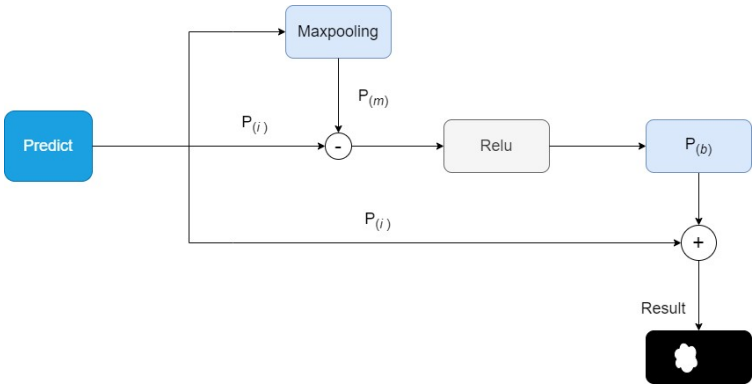


Figure 4: The prediction result $P(i)$ is subtracted from the max pooling result $P(m)$, followed by applying a ReLU function to enhance features. Finally, an addition operation integrates the refined features, contributing to the model's prediction refinement.

### 3.3. Channel Attention Module

This module enhances input feature representation. First, adaptive average pooling reduces the time series dimension, obtaining a global channel description [38]. This channel attention strengthens key information, improving the model's understanding of complex tasks.

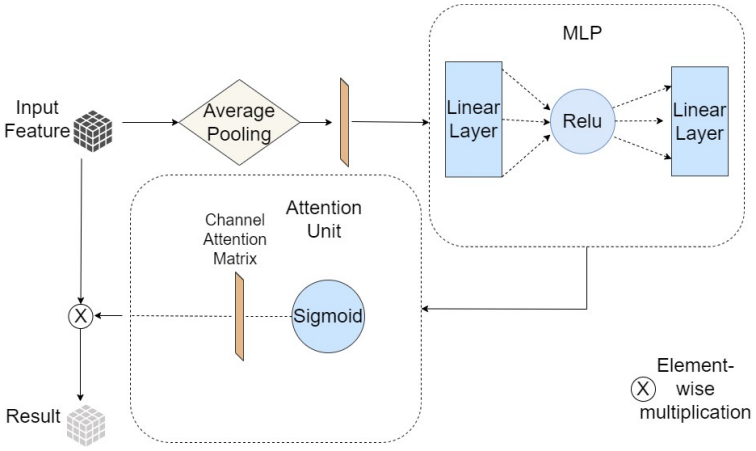


Figure 5: The channel attention mechanism reduces the channel dimension of input features via adaptive average pooling to form a global description. Then, a sequence of two linear layers with ReLU reduces attention to learn channel weights. These weights are normalized using a sigmoid function and applied to the input features, adjusting each channel through element-wise multiplication to produce the final output features.

As shown in Figure 5, by incorporating the channel attention module in the skip connection, the model optimizes channel-wise feature adjustment, enhancing its focus on key information. This process improves the model's ability to express input data, particularly in complex tasks, by emphasizing important features and suppressing irrelevant ones.

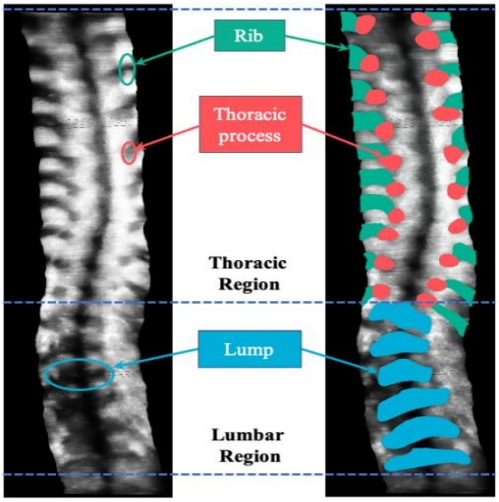


Figure 6: A pair of data, the image on the left side is the original ultrasound image, while on the right side is the ground truth label overlapping the raw data.

## 4. Experiments

### *4.1. Dataset*

As Figure 6 shows, there are two important parts for each case.

1. The LBF is at the bottom of the photos, as shown in blue.
2. The upper part is TBF, as red and green colors.

This research utilizes input images acquired from the Scolioscan system, which employs 3D ultrasound image techniques to generate 3D Volume Projection Images (VPI). To ensure spatial accuracy, a cross-wire calibration procedure was conducted using the Levenberg-Marquardt nonlinear algorithm to align 2D images with positional sensor data. Specifically, ultrasound gel was applied during the scanning process to eliminate gaps between the probe and the skin, ensuring consistent image quality. Technically, a total of 1,170 images were pre-trained using self-supervised proxy tasks, followed by fine-tuning with 109 images selected from 109 patients. The patient cohort included 82 females and 27 males, with an average age of 15.6 ± 2.7 years, all presenting varying degrees of spinal deformity. From the 3D ultrasound voxel data, medical experts selected nine 2D images per patient, captured at different image depths, and a RankNet is used for selecting the best quality slice image for each case [39]. These images, despite variations in size, typically have a resolution of approximately $2600 \times 640$ pixels [1].

All experimental procedures involving human participants were approved by the Institutional Review Board, and informed consent was obtained as required. The study strictly adhered to the principles outlined in the Declaration of Helsinki. Furthermore, human subject consent for the project was granted by the Hong Kong Polytechnic University Ethics Committee (Approval No. HSEARS20180906005).

### 4.2. Data Pre- and Post- process

During the pre-processing stage, the grayscale-converted original images and their corresponding label arrays are stored as NumPy arrays in .npz format. The image pixel values range from 0 to 255, while the labels are constrained to values of 0, 1, or 2 within a 2D array. To further reduce noise and enhance image quality, a Gaussian operator is applied as part of the pre-processing workflow [40]. Each case is then consolidated into a single NumPy array, which is split into training and testing sets at a 0.9:0.1 ratio. The images are subsequently resized to $448 \times 448$ dimensions to improve model performance.

During post-processing, the image is restored to its original size of 640 x 2600. The results display the original class alongside a combination of classes 1 and 2, multiplied by 255 for better visualization.

### 4.3. Implementation

The network is designed in Python 3.7 and Pytorch 1.13.1. Data augmentations like flips and rotations are used for all the training images. As mentioned above, the input image is 448x448 and the patch size is 4. The window size is 7. We used an NVIDIA Quadro RTX 8000 GPU with a total memory of 32GB.

#### 4.3.1. Pre-training

We pre-train the SSAT-Swin model encoder using multiple proxy tasks and adopt a multi-objective loss function for self-supervised representation learning. We utilize three proxy tasks: masked image inpainting, rotation prediction, and contrastive learning. During pre-training, three projection heads are attached to the encoder, which are later removed for the downstream segmentation task.

- Masked Image Inpainting: In the masked image inpainting task, a portion of the ultrasound image's region of interest (ROI) is randomly masked, and the model reconstructs the missing areas. Using a transpose convolution layer during pre-training, the output $\hat{X}_M$ is optimized with the L1 loss:

$$L_{\text{inpaint}} = \|X - \hat{X}_M\|_1 \tag{1}$$

- Image Rotation: The image rotation task predicts the rotation angle of an input image from four categories (0°, 90°, 180°, and 270°). An MLP head predicts softmax probabilities $\hat{y}_r$ with a cross-entropy loss:

$$L_{\text{rot}} = -\sum_{r=1}^{R} y_r \log(\hat{y}_r) \tag{2}$$

- Contrastive Learning: Self-supervised contrastive learning improves representation by maximizing mutual information between positive pairs (same image views) and minimizing it for negatives. The contrastive loss for representations $v_i$ and $v_j$is:

$$L_{\text{contrast}} = -\log\left(\frac{\exp(\text{sim}(v_i, v_j)/t)}{\sum_{k=1,k\neq i}^{2N}\exp(\text{sim}(v_i, v_k)/t)}\right) \tag{3}$$

#### *4.3.2. Fine-Tune*

During training, we used the popular SGD optimizer with a momentum of 0.9 and a weight decay of 1e-4 for backpropagation. Cross-validation was performed multiple times, maintaining a 90% training and 10% test split. The total loss was computed by combining weighted entropy, dice loss, and cross-entropy loss, as described in Equations (4) and (5), where $y_{\text{true}}$ represents the true labels and $y_{\text{pred}}$ the predicted output.

- We employed dice loss, which is formulated as :

$$Dice\_Loss(y_{\text{true}}, y_{\text{pred}}) = 1 - \frac{2 \cdot \sum(y_{\text{true}} \cdot y_{\text{pred}})}{\sum y_{\text{true}}^2 + \sum y_{\text{pred}}^2} \tag{4}$$

- Additionally, we utilized cross-entropy loss, which is formulated as:

$$Cross\ Entropy\ Loss(y_{\text{true}}, y_{\text{pred}}) = -\sum y_{\text{true}} \cdot \log(y_{\text{pred}}) \tag{5}$$

For quantitative performance evaluation, we utilized three key metrics: Jaccard scores, Dice scores, and detection rate. For detection rate assessment, we manually calculated values, prioritizing continuous and distinct boundaries for predicted bone structures. Our focus is on ensuring that the predicted bone boundaries demonstrate sufficient clarity and continuity.

### *4.4. Result*

We conduct a quantitative comparative evaluation of our SSAT-Swin against seven other networks: Swin-Unet, TransUnet, TransDeepLab, Unet, Unet++, Feature Pyramid Network (FPN), and H2Former [26, 33, 41, 42, 43, 44, 45]. For the subsequent qualitative analysis, we further compare the different models by directly assessing the predicted segmentation images, providing a more nuanced understanding of their respective representations.

#### 4.4.1. Quantitative Evaluation

Table 1 presents eight records corresponding to different models with the corresponding dice score and standard deviation.

Table 1: Comparison of Dice Scores for Segmentation Performance

| | SSAT-Swin | Swin-Unet | Trans-Unet | Trans DeepLab | H2Former | Unet | Unet++ | FPN |
|---|---|---|---|---|---|---|---|---|
| 1 | **85.8%** | 84.7% | 84.3% | 83.5% | 84.9% | 81.4% | 82.2% | 80.0% |
| 2 | **86.7%** | 85.5% | 86.6% | 85.3% | 85.7% | 83.3% | 82.6% | 81.4% |
| 3 | **85.6%** | 84.5% | 84.9% | 83.0% | 84.7% | 80.6% | 82.4% | 80.5% |
| 4 | **86.8%** | 85.6% | 86.2% | 85.0% | 85.8% | 84.0% | 82.9% | 81.3% |
| 5 | **85.5%** | 84.4% | 83.5% | 83.1% | 84.6% | 81.7% | 81.9% | 81.0% |
| 6 | **84.4%** | 83.3% | 84.1% | 81.4% | 83.5% | 79.5% | 80.1% | 78.1% |
| 7 | 84.7% | 83.6% | **84.9%** | 82.9% | 83.8% | 83.0% | 83.0% | 80.9% |
| 8 | **85.4%** | 84.5% | 85.2% | 84.1% | 84.7% | 81.9% | 80.6% | 79.9% |
| Average | **85.6%** | 84.5% | 85.0% | 83.5% | 84.7% | 81.9% | 82.0% | 80.4% |
| Std Dev | 0.8 | 0.8 | 1.0 | 1.3 | 0.8 | 1.5 | 1.1 | 1.1 |

Our proposed SSAT-Swin consistently outperforms its counterparts, achieving an average Dice score of 85.6%, closely followed by TransUnet at 85.0% and Swin-Unet at 84.5%. Notably, SSAT-Swin demonstrates a lower standard deviation of 0.8%, indicating greater consistency in segmentation performance.

Table 2: Comparison of Jaccard Scores for Segmentation Performance

| | SSAT-Swin | Swin-Unet | Trans-Unet | Trans DeepLab | H2Former | Unet | Unet++ | FPN |
|---|---|---|---|---|---|---|---|---|
| 1 | **74.6%** | 73.5% | 73.0% | 71.8% | 73.2% | 68.9% | 68.4% | 69.9% |
| 2 | **76.5%** | 74.8% | 76.4% | 74.4% | 75.2% | 71.5% | 70.7% | 70.4% |
| 3 | **74.3%** | 73.2% | 73.8% | 71.1% | 73.8% | 67.6% | 69.0% | 70.1% |
| 4 | **76.2%** | 75.0% | 75.8% | 74.0% | 75.0% | 72.5% | 71.5% | 70.9% |
| 5 | **74.2%** | 73.1% | 71.7% | 71.1% | 73.6% | 69.1% | 70.7% | 69.4% |
| 6 | **72.7%** | 71.4% | 72.6% | 68.7% | 70.5% | 66.0% | 67.3% | 66.8% |
| 7 | 73.4% | 71.9% | **73.8%** | 70.9% | 74.4% | 71.0% | 70.1% | 71.0% |
| 8 | **74.4%** | 73.3% | 74.3% | 72.6% | 73.0% | 69.4% | 71.0% | 67.6% |
| Average | **74.5%** | 73.3% | 73.9% | 71.8% | 73.6% | 69.5% | 69.8% | 69.5% |
| Std Dev | 1.2 | 1.2 | 1.6 | 1.8 | 1.5 | 2.1 | 1.5 | 1.5 |

Table 2 displays the records corresponding to different models, presenting their respective Jaccard scores and standard deviations. Our proposed SSAT-Swin model particularly stands out, achieving an enhancement of around 1.2% over its baseline, Swin-Unet. Remarkably, SSAT-Swin emerges as the leading model, attaining the highest Jaccard score of 74.5% while exhibiting a low standard deviation of 1.2%, indicating its robustness and consistency in segmentation.

The detection rate result is shown as Table 3:

Table 3: Detection Rates for Segmentation Performance Comparison

| | SSAT-Swin | Swin-Unet | Trans-Unet | Trans DeepLab | H2-Former | Unet | Unet++ | FPN |
|---|---|---|---|---|---|---|---|---|
| Average | **92.8%** | 90.8% | 91.2% | 86.4% | 91.0% | 77.8% | 78.9% | 79.1% |

As it shows, the average result for each model, SSAT-Swin leads the pack with a notable score of 92.8%, showcasing its effectiveness in the given task. Following closely is the Swin model, achieving a commendable score of 90.8%, indicating a strong performance but slightly trailing behind SSAT-Swin. TransUnet, with a score of 91.2%, demonstrates a decent level of competence, albeit with a lower score compared to the SSAT-Swin.

We calculate the UCA results based on manual calculations, following the procedure outlined in [22]. Specifically, we identify the T12 Rib pair in ultrasound images to determine the vertebral hierarchy for spinal measurement. Next, we select the most inclined thoracic and lumbar features, draw central lines, and compute key spinal angles, including the Main Thoracic Angle and Lumbar Angle, for scoliosis assessment.

Table 4: Comparison of Range Values and Mean Error between each model with its ground truth for Thoracic and Lumbar Regions

| | Thoracic Range (°) | Lumbar Range (°) | Mean error with Std (°) |
|---|---|---|---|
| Ground Truth | 2.04 – 15.43 | 4.01 – 20.24 | - |
| TransDeepLab | 2.30 – 18.41 | 2.01 – 17.40 | 2.81 ± 1.45 |
| TransUnet | 2.20 – 16.10 | 3.50 – 22.08 | 1.95 ± 1.10 |
| Swin-Unet | 2.10 – 14.80 | 3.11 – 18.40 | 2.02 ± 1.18 |
| SSAT-Swin | 2.10 – 14.90 | 3.98 – 21.98 | **1.76 ± 0.95** |

Table 4 compares the predicted angle ranges for the thoracic and lumbar regions across models and their mean errors relative to the ground truth. SSAT-Swin achieves the most stable performance, with thoracic (2.10°–14.90°) and lumbar (3.98°–21.98°) ranges closely matching the ground truth (2.04°–15.43° and 4.01°–20.24°) while also having the lowest mean error (1.76° ± 0.95°), indicating more precise landmark detection. TransUnet also performs well with a mean error of 1.95° ± 1.10°, but its lumbar range deviates slightly. Swin-Unet and TransDeepLab exhibit higher mean errors (2.02° ± 1.18° and 2.81° ± 1.45°, respectively), with TransDeepLab showing the most deviation, particularly in lumbar predictions. Overall, SSAT-Swin demonstrates the best balance between accuracy and robustness, making it the most reliable model for UCA estimation.

We also compared the dice score for two classes (LBF, TBF) for SSAT-Swin, Swin-Unet, and TransUnet. The result is shown as Table 5. In detail, SSAT-Swin achieves the highest average performance for Class 1 (82.2%) with the lowest standard deviation, indicating more consistent results. For Class 2, SSAT-Swin achieves an average score (83.3%) comparable to TransUnet (83.3%) but with slightly better consistency. Swin performs slightly lower in both

Table 5: Validation Performance Comparison for Two Classes (%)

| | Class 1 | | | Class 2 | | |
|---|---|---|---|---|---|---|
| Validation time | SSAT-Swin | Swin | TransUnet | SSAT-Swin | Swin | TransUnet |
| 1 | 81.8 | 81.4 | 81.4 | 83.7 | 84.3 | 77.1 |
| 2 | 83.3 | 82.8 | 84.1 | 85.8 | 85.6 | 86.7 |
| 3 | 80.5 | 80.0 | 79.8 | 87.0 | 87.2 | 86.8 |
| 4 | 83.4 | 83.6 | 84.9 | 83.7 | 83.5 | 85.0 |
| 5 | 82.6 | 81.4 | 81.0 | 76.6 | 76.4 | 82.5 |
| 6 | 81.5 | 81.3 | 81.9 | 82.1 | 81.3 | 82.3 |
| 7 | 81.2 | 80.4 | 82.5 | 82.2 | 82.0 | 83.1 |
| 8 | 83.0 | 82.1 | 83.1 | 85.5 | 85.0 | 82.6 |
| Average | **82.2** | 81.6 | 82.3 | **83.3** | 83.2 | 83.3 |
| Std | **1.1** | 1.2 | 1.7 | **3.2** | 3.3 | 3.1 |

classes.

### *4.4.2. Qualitative Evaluation*

We have presented segmentation results for various models alongside their respective raw images and labels, as depicted in Figure 7:

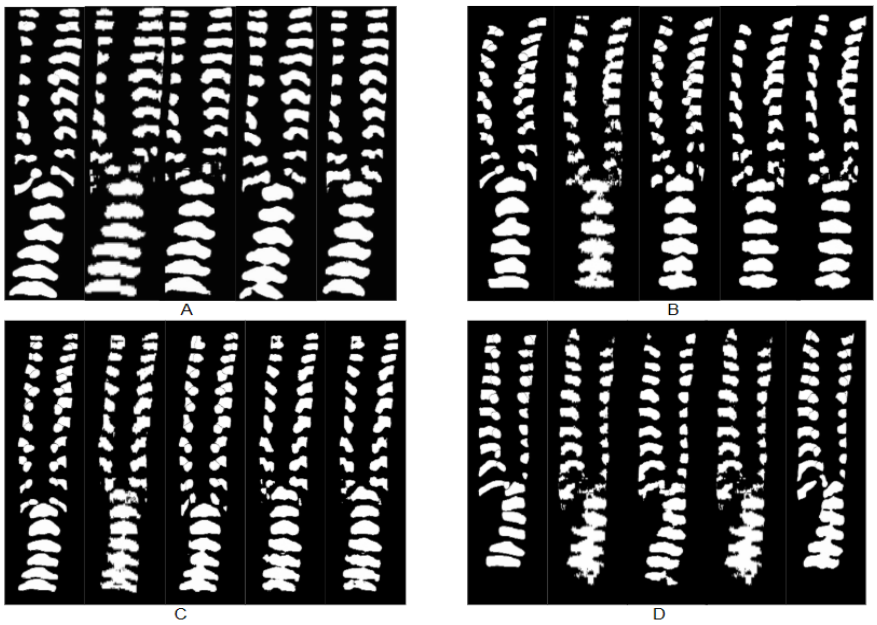


Figure 7: Qualitative comparison of segmentation performance across different models. From left to right: Raw Label, TransDeepLab, TransUnet, Swin-Unet, and SSAT-Swin. Results are shown for four different cases (A, B, C, and D)

Our proposed model, SSAT-Swin, consistently delivers superior detection capabilities, particularly in delineating fine-grained spinal features. TransUnet and TransDeepLab, however, struggle with structural continuity and exhibit incomplete segmentation, particularly in regions with irregular or noisy patterns. Notably, our proposed model, SSAT-

Swin, excels in capturing the intricate boundaries and shapes of vertebral structures, achieving clearer delineation even in challenging cases, especially for Case D, with its extreme spinal curvature, which poses significant segmentation challenges, where TransUnet and TransDeepLab produce fragmented results, and Swin-Unet shows partial improvement but struggles with complex boundaries, while SSAT-Swin demonstrates superior accuracy and robustness by capturing intricate details and reducing errors.

#### *4.4.3. Ablation Studies*

We conducted two supplementary trials using the basic Swin architecture, enhanced separately with boundary enhancement and channel attention, alongside our novel SSAT-Swin model as well as without the self-supervised version.

Table 6: Dice and Jaccard Score Comparison for Swin-Unet, Swin-1 (Swin + Boundary Enhancement), Swin-2 (Swin + Channel Attention), and our model SSAT-Swin-1(without Self-Supervised) and SSAT-Swin-2(with Self-Supervised)

| Model | Dice Average | Dice Std | Jaccard Average | Jaccard Std |
|---|---|---|---|---|
| Swin-Unet | 84.5% | 0.5 | 73.3% | 0.7 |
| Swin-1 | 84.7% | 0.6 | 73.1% | 1.3 |
| Swin-2 | 84.6% | 0.7 | 73.3% | 0.9 |
| SSAT-Swin-1 | 85.0% | 0.7 | 74.0% | 1.1 |
| SSAT-Swin-2 | **85.6%** | 0.8 | **74.5%** | 1.3 |

Based on the Table 6, we observe that Swin-1 and Swin-2 show improved performance compared to the base model, Swin-Unet, in terms of both Dice and Jaccard scores. Swin-1 demonstrates slightly better performance, which emphasizes the importance of adding additional modules, particularly in enhancing the consistency of model predictions. The inclusion of self-supervised learning in SSAT-Swin further emphasizes the significance of incorporating both channel attention and boundary enhancement. SSAT-Swin with self-supervised learning achieved a substantial increase in both Dice and Jaccard scores, reaching average values of 85.6% and 74.5%, respectively. This indicates that combining the advantages of Swin-1 and Swin-2 with self-supervised learning results in a more robust model, capable of achieving higher prediction accuracy and robustness.

In addition, we visualize the difference between images processed with and without the boundary enhancement module, as shown in Figure 8. Boundary enhancement algorithms are employed to increase the contrast of pixel values near edges. Both the TBF and LBF methods exhibit significant boundary differences. The edges of each bone feature are prominently highlighted.

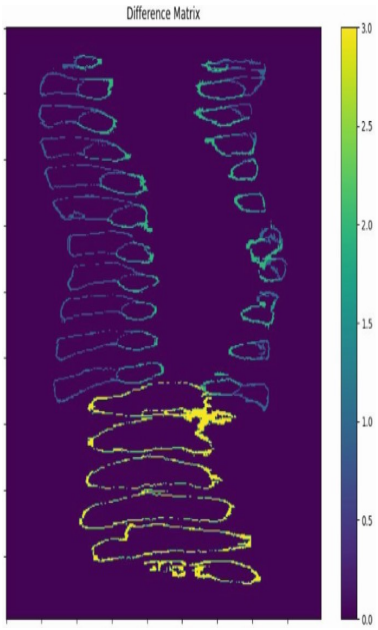


Figure 8: Difference matrix, the class number is highlighted as a difference color on the right side

## 5. Discussion

Unlike CNNs, Transformers-based networks leverage self-attention mechanisms to capture global information, effectively modeling long-range dependencies, which is crucial for tasks like ultrasound image segmentation. Ultrasound images often suffer from noise, low contrast, and other challenges, making it difficult to rely solely on local information to segment the target regions accurately.

### *5.1. Superiority of Self-supervised*

As demonstrated in Table 6, self-supervised learning tasks significantly enhance the proposed model's performance, leading to notable improvements in Dice and Jaccard scores compared to the non-self-supervised baseline. Tasks such as masked image inpainting, rotation prediction, and contrastive learning provide the model with robust, generalized feature-learning capabilities to address the challenges of ultrasound imaging, including high noise, low contrast, and irregular shapes. As illustrated in Figure 7, even in the worst-case scenario (case D), the SSAT model achieves superior performance. Masked image inpainting enhances spatial understanding by reconstructing missing or occluded regions, improving resilience to incomplete or noisy data common in clinical scenarios. Rotation prediction enables the model to adapt to varied orientations, compensating for irregularities caused by scoliosis or diverse acquisition angles. Contrastive learning further strengthens the model's ability to discern subtle differences in similar features, aiding segmentation of complex structures such as bone features in low-contrast images. These

self-supervised strategies not only improve robustness to unseen datasets but also enhance adaptability for clinical applications, validating their superiority over traditional supervised methods and highlighting their potential to drive more accurate and reliable diagnoses in real-world settings.

### *5.2. Better object shape*

Our proposed model, SSAT-Swin, surpasses baseline models like Swin-Unet and TransUnet in prediction accuracy, primarily due to our boundary enhancement module. This module refines boundary representations in target regions, which is vital for accurate segmentation in low-contrast ultrasound images. By integrating it into the decoder of the Swin Transformer block and the final projection layer, the model effectively focuses on boundary details while maintaining global context. However, challenges persist, as some bone features appear smaller or irregular compared to ground truth labels, indicating potential limitations in boundary continuity during window shifting. Addressing this issue is crucial for improving segmentation precision, and leveraging the Swin Transformer's ability to capture local information may help enhance performance further.

### *5.3. Information loss*

The lower detection rates of CNN-based networks compared to Transformer-based networks can be attributed to the occurrence of discontinuous shapes, with CNN predictions often showing irregularities where bone structures may only display head and tail portions while missing the middle segment. This issue may arise from information loss or neglect in lower-level skip connections within the architecture. To address this, we propose adding a channel attention unit to ensure that valuable information in the U-Net-style architecture's skip connections is not overlooked; this module automatically assigns weights to different feature channels, enabling the model to focus on the most informative features. In the Swin Transformer-based architecture, the window-based self-attention mechanism effectively models local and global relationships while reducing computational complexity. By integrating both the channel attention unit and a boundary enhancement module, our SSAT-Swin model can leverage the Swin Transformer's efficient global modeling capability alongside local detail features, allowing it to accurately capture the overall structure of ultrasound images and concentrate on crucial boundary regions. This leads to precise segmentation of irregular bone structures, resulting in high-quality masks for UCA calculations and theoretically enhancing overall segmentation performance.

## 6. Conclusions and future works

In this paper, we introduce SSAT-Swin, a Transformer-based model designed for segmenting bony structures in spine ultrasound images. Our model incorporates self-supervised pre-training, which enhances its ability to learn from unlabeled data, contributing to its superior performance compared to the baseline Swin-Unet.

Future work will focus on two directions. Firstly, expanding and diversifying the dataset, including different sources of inputs to improve robustness and ensure the model's generalizability across different populations and conditions. Secondly, automating UCA calculations by integrating segmentation with the automatic identification of the most skewed spinal regions.


## Acknowledgements

This study was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Nos. R5017-18 and B-Q86J).


## Conflict of Interest

Y.P. Zheng serves as a consultant for Telefield Medical Imaging Limited in the development of Scolioscan and holds multiple patents related to 3D ultrasound imaging for scoliosis, licensed to the company through Hong Kong Polytechnic University. He is also a director and shareholder of Telefield Medical Imaging Limited. The other authors declare no conflicts of interest relevant to this article.

## Data availability statement

The data that support the findings of this study are openly available in GitHub at https://github.com/suni88/HongKong-PolyU-Ultrasound-spine-dataset.

## References


[1] C.-W. J. Cheung, G.-Q. Zhou, S.-Y. Law, T.-M. Mak, K.-L. Lai, Y.-P. Zheng, Ultrasound volume projection imaging for assessment of scoliosis, IEEE transactions on medical imaging 34 (8) (2015) 1760–1768.

[2] H. Kim, H. S. Kim, E. S. Moon, C.-S. Yoon, T.-S. Chung, H.-T. Song, J.-S. Suh, Y. H. Lee, S. Kim, Scoliosis imaging: what radiologists should know, Radiographics 30 (7) (2010) 1823–1842.

[3] Y.-P. Zheng, T. T.-Y. Lee, K. K.-L. Lai, B. H.-K. Yip, G.-Q. Zhou, W.-W. Jiang, J. C.-W. Cheung, M.-S. Wong, B. K.-W. Ng, J. C.-Y. Cheng, et al., A reliability and validity study for scolioscan: a radiation-free scoliosis assessment system using 3d ultrasound imaging, Scoliosis and spinal disorders 11 (2016) 1–15.

[4] H.-D. Wu, W. Liu, M.-S. Wong, Reliability and validity of lateral curvature assessments using clinical ultrasound for the patients with scoliosis: a systematic review, European Spine Journal 29 (2020) 717–725.

[5] S. L. Weinstein, L. A. Dolan, J. C. Cheng, A. Danielsson, J. A. Morcuende, Adolescent idiopathic scoliosis, The lancet 371 (9623) (2008) 1527–1537.

[6] J. C. Cheng, R. M. Castelein, W. C. Chu, A. J. Danielsson, M. B. Dobbs, T. B. Grivas, C. A. Gurnett, K. D. Luk, A. Moreau, P. O. Newton, et al., Adolescent idiopathic scoliosis, Nature reviews disease primers 1 (1) (2015) 1–21.

[7] J. Cobb, Outline for the study of scoliosis, Instructional course lecture (1948).

[8] Y. Yao, W. Yu, Y. Gao, J. Dong, Q. Xiao, B. Huang, Z. Shi, W-transformer: Accurate cobb angles estimation by using a transformer-based hybrid structure, Medical Physics 49 (5) (2022) 3246–3262.

[9] M. Zhao, N. Meng, J. P. Y. Cheung, C. Yu, P. Lu, T. Zhang, Spinehrformer: A transformer-based deep learning model for automatic spine deformity assessment with prospective validation, Bioengineering 10 (11) (2023) 1333.

[10] R. C. Brink, S. P. Wijdicks, I. N. Tromp, T. P. Schlösser, M. C. Kruyt, F. J. Beek, R. M. Castelein, A reliability and validity study for different coronal angles using ultrasound imaging in adolescent idiopathic scoliosis, The Spine Journal 18 (6) (2018) 979–985.

[11] R. Zheng, D. Hill, D. Hedden, J. Mahood, M. Moreau, S. Southon, E. Lou, Factors influencing spinal curvature measurements on ultrasound images for children with adolescent idiopathic scoliosis (ais), PLoS One 13 (6) (2018) e0198792.

[12] B. Y. Hwang, D. Mampre, A. K. Ahmed, I. Suk, W. S. Anderson, A. Manbachi, N. Theodore, Ultrasound in traumatic spinal cord injury: a wide-open field, Neurosurgery 89 (3) (2021) 372–382.

[13] N. A. Tawfik, A. T. Ahmed, T. E. El-Shafei, M. R. Habba, Diagnostic value of spinal ultrasound compared to mri for diagnosis of spinal anomalies in pediatrics, Egyptian Journal of Radiology and Nuclear Medicine 51 (2020) 1–11.

[14] J. Zhang, X. Cui, S. Chen, Y. Dai, Y. Huang, S. Zhang, Ultrasound-guided nusinersen administration for spinal muscular atrophy patients with severe scoliosis: an observational study, Orphanet Journal of Rare Diseases 16 (2021) 1–8.

[15] A. R. Levy, M. S. Goldberg, N. E. Mayo, J. A. Hanley, B. Poitras, Reducing the lifetime risk of cancer from spinal radiographs among people with adolescent idiopathic scoliosis, Spine 21 (13) (1996) 1540–1547.

[16] D. Yang, T. T.-Y. Lee, K. K.-L. Lai, T.-P. Lam, W. C.-W. Chu, R. M. Castelein, J. C.-Y. Cheng, Y.-P. Zheng, Semi-automatic ultrasound curve angle measurement for adolescent idiopathic scoliosis, Spine Deformity (2022) 1–9.

[17] S. Trac, R. Zheng, D. L. Hill, E. Lou, Intra-and interrater reliability of cobb angle measurements on the plane of maximum curvature using ultrasound imaging method, Spine deformity 7 (2019) 18–26.

[18] J. A. Noble, D. Boukerroui, Ultrasound image segmentation: a survey, IEEE Transactions on medical imaging 25 (8) (2006) 987–1010.

[19] S. Banerjee, J. Lyu, Z. Huang, H. F. F. Leung, T. T.-Y. Lee, D. Yang, S. Su, Y. Zheng, S.-H. Ling, Light-convolution dense selection u-net (lds u-net) for ultrasound lateral bony feature segmentation, Applied Sciences 11 (21) (2021) 10180.

[20] S. Banerjee, J. Lyu, Z. Huang, F. H. Leung, T. Lee, D. Yang, S. Su, Y. Zheng, S. H. Ling, Ultrasound spine image segmentation using multi-scale feature fusion skip-inception u-net (siu-net), Biocybernetics and Biomedical Engineering 42 (1) (2022) 341–361.

[21] G. Kossoff, Basic physics and imaging characteristics of ultrasound, World journal of surgery 24 (2) (2000) 134–142.

[22] S. Banerjee, Z. Huang, J. Lyu, F. H. Leung, T. Lee, D. Yang, Y. Zheng, J. McAviney, S. H. Ling, Automatic assessment of ultrasound curvature angle for scoliosis detection using 3-d ultrasound volume projection imaging, Ultrasound in Medicine & Biology (2024).

[23] T. T.-Y. Lee, K. K.-L. Lai, J. C.-Y. Cheng, R. M. Castelein, T.-P. Lam, Y.-P. Zheng, 3d ultrasound imaging provides reliable angle measurement with validity comparable to x-ray in patients with adolescent idiopathic scoliosis, Journal of orthopaedic translation 29 (2021) 51–59.

[24] P. U. Pandey, N. Quader, P. Guy, R. Garbi, A. J. Hodgson, Ultrasound bone segmentation: A scoping review of techniques and validation practices, Ultrasound in Medicine & Biology 46 (4) (2020) 921–935.

[25] N. Siddique, S. Paheding, C. P. Elkin, V. Devabhaktuni, U-net and its variants for medical image segmentation: A review of theory and applications, Ieee Access 9 (2021) 82031–82057.

[26] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[27] S. Cai, Y. Tian, H. Lui, H. Zeng, Y. Wu, G. Chen, Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a

convolutional neural network, Quantitative imaging in medicine and surgery 10 (6) (2020) 1275.

[28] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, P.-A. Heng, H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes, IEEE transactions on medical imaging 37 (12) (2018) 2663–2674.

[29] X. Tong, J. Wei, B. Sun, S. Su, Z. Zuo, P. Wu, Ascu-net: attention gate, spatial and channel attention u-net for skin lesion segmentation, Diagnostics 11 (3) (2021) 501.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

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

[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.

[33] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: European conference on computer vision, Springer, 2022, pp. 205–218.

[34] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, A. Hatamizadeh, Self-supervised pre-training of swin transformers for 3d medical image analysis, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 20730–20740.

[35] C. Tian, Y. Xu, L. Fei, J. Wang, J. Wen, N. Luo, Enhanced cnn for image denoising, CAAI Transactions on Intelligence Technology 4 (1) (2019) 17–23.

[36] C. Li, Y. Tan, W. Chen, X. Luo, Y. Gao, X. Jia, Z. Wang, Attention unet++: A nested attention-aware u-net for liver ct image segmentation, in: 2020 IEEE international conference on image processing (ICIP), IEEE, 2020, pp. 345–349.

[37] M. Feng, H. Lu, E. Ding, Attentive feedback network for boundary-aware salient object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1623–1632.

[38] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, Eca-net: Efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11534–11542.

[39] J. Lyu, S. H. Ling, S. Banerjee, J. Zheng, K.-L. Lai, D. Yang, Y.-P. Zheng, X. Bi, S. Su, U. Chamoli, Ultrasound volume projection image quality selection by ranking from convolutional ranknet, Computerized Medical Imaging and Graphics 89 (2021) 101847.

[40] J. F. Corney, P. D. Drummond, Gaussian quantum operator representation for bosons, Physical Review A 68 (6) (2003) 063822.

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

[42] R. Azad, M. Heidari, M. Shariatnia, E. K. Aghdam, S. Karimijafarbigloo, E. Adeli, D. Merhof, Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation, in: International Workshop on PRedictive Intelligence In MEdicine, Springer, 2022, pp. 91–102.

[43] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, Springer, 2018, pp. 3–11.

[44] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.

[45] A. He, K. Wang, T. Li, C. Du, S. Xia, H. Fu, H2former: An efficient hierarchical hybrid transformer for medical image segmentation, IEEE Transactions on Medical Imaging (2023).