RotCAtt-TransUNet++: Novel Deep Neural Network for Sophisticated Cardiac Segmentation

Quoc-Bao Nguyen-Le [1,2,3], Tuan-Hy Le [1], Anh-Triet Do [1], and Quoc-Huy Trinh [2, 3]

[1]Le Hong Phong High School for the Gifted, Ho Chi Minh City, Vietnam
[2]Faculty of Information Technology, University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
[3]Viet Nam National University, Ho Chi Minh City, Vietnam

***Abstract*—Cardiovascular disease is a major global health concern, contributing significantly to global mortality. Accurately segmenting cardiac medical imaging data is crucial for reducing fatality rates associated with these conditions. However, current state-of-the-art (SOTA) neural networks, including CNN-based and Transformer-based approaches, face challenges in capturing both inter-slice connections and intra-slice details, especially in datasets featuring intricate, long-range details along the z-axis like coronary arteries. Existing methods also struggle with differentiating non-cardiac components from the myocardium, resulting in segmentation inaccuracies and the "spraying" phenomenon. To address these issues, we introduce RotCAtt-TransUNet++, a novel architecture designed for robust segmentation of intricate cardiac structures. Our approach enhances global context modeling through multiscale feature aggregation and nested skip connections in the encoder. Transformer layers facilitate capturing intra-slice interactions, while a rotatory attention mechanism handles inter-slice connectivity. A channel-wise cross-attention gate integrates multiscale information and decoder features, effectively bridging semantic gaps. Experimental results across multiple datasets demonstrate superior performance over current methods, achieving near-perfect annotation of coronary arteries and myocardium. Ablation studies confirm that our rotatory attention mechanism significantly improves segmentation accuracy by transforming embedded vectorized patches in semantic dimensional space.**

I. Introduction

Medical image segmentation is crucial for disease and tumor detection. While Manual segmentation remains the gold standard in delineating pathological structures, it is labor-intensive, time-consuming, and prone to human error [1]. Automated segmentation is increasingly needed to reduce reliance on expert knowledge and speed up the process. With its intricate structures and fine details, the heart presents significant challenges in this context. Previous studies primarily are binary segmentation tasks using single-labeled dataset [2], [3]. Recent studies opted for multi-class segmentation with two main datasets, MMWHS [4] and ACDC [5], in 2017. Nevertheless, these datasets only annotate basic regions in an unsophisticated way that lack significant details, such as coronary arteries and heart capillaries. More sophisticated-annotated datasets such as ImageCHD [6] with 8 labels (2021) and VHSCDD with 12 labels (2023) challenge SOTA networks. Additionally, SOTA networks, both CNN-based and Transformer-based networks, have not undergone evaluation using the same cardiac datasets, leading to the lack of fair comparison of these networks. In this paper, we conduct experiments with SOTA networks (both CNN-based and Transformer-based approaches) and propose our novel self-designed architecture that proves its superiority.

The content of this paper is organized as follows. In Section II, we briefly review existing methods related to our work. Then, we present our proposed solution in Section III. Experiments and result analysis are discussed in Section IV. Finally, the conclusion and implication are in Section V.

II. Related works

Fully Convolutional Neural Networks (FCNs), has become the de facto standard in medical image segmentation [7], [8]. UNet [9] introduced direct skip connections joining feature maps at the same scale to mitigate detail loss in deeper layers. UNet++ [10] further improved upon UNet by incorporating nested skip connections. ResUNet [11] employs ResNet units with atrous convolutions and pyramid pooling to address the semantic gap. However, CNN-based methods struggle with capturing long-range dependencies and global contexts due to their inherited locality [5]. Attention mechanisms, such as those in U-Net Attention [12], attempt to enhance performance by focusing on relevant details and ignoring distractions. Despite these advancements, CNN-based approaches still yield weak performance, particularly with structures exhibiting significant inter-patient variability [1], [5].

Initially designed for NLP tasks, transformers are known for their Multi-head Self-Attention (MSA) mechanism, which excels at capturing long-range interactions. In computer vision and segmentation, TransUNet [5] utilizes a Transformer encoder for global information learning and CNN decoders for spatial details extraction. Swin-Unet [13], conversely, replaces CNNs with a complete Transformer architecture, employing a shifted window mechanism for detail extraction and patch-expanding layers for upsampling. However, current Transformer-based methods focus self-attention solely on patch interactions and skip connections, processing volumetric data slice by slice and limiting inter-slice information integration. This constraint affects TransUNet's ability to achieve seamless segmentation across adjacent slices.

3D networks like UNet 3D [14] and VNet [15] preserve inter-slice details but face limitations in GPU memory and computational demands. Thus, we introduce RotCAtt-TransUNet++, a lightweight 2.5D network, to overcome these issues.

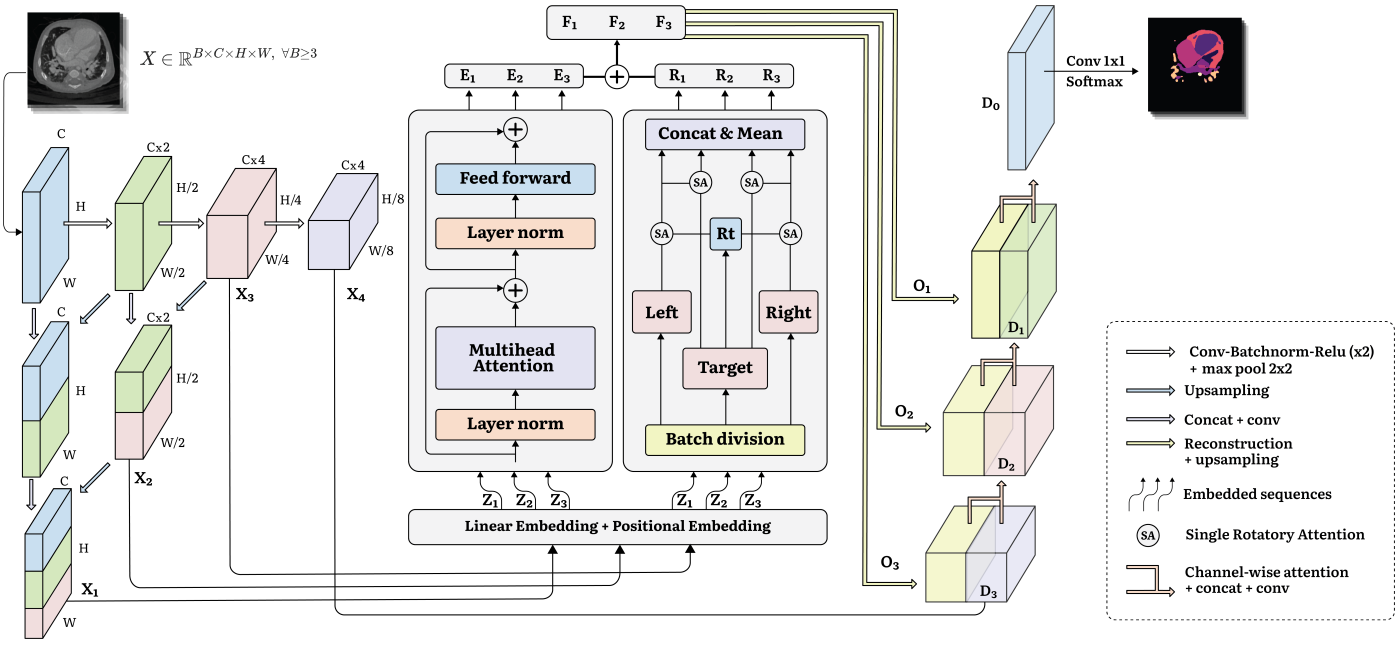

Fig. 1. RotCAtt-TransUNet++ Architecture: combining rotatory attention mechanism with channel-wise attention gates for enhanced feature fusion in the decoder. Leveraging the Transformer-Unet hybrid model with enriched nested skip connections for multiscale feature extraction.

III. Our Proposed Method

A. Architecture Overview

The architecture diagram can be seen in Figure 1. Through meticulous experimentation and ablation studies, we observed the efficacy of the UNet++ [10] architecture coupled with nested skip connections to preserve crucial information in achieving superior segmentation results. We are also inspired by pyramid pooling at different scales of Zhao et al. [16]. Thus, instead of the conventional CNN-based feature extraction approach, such as ResNet-50 in TransUNet [5], we employ dense downsampling alongside nested skip connections, yielding four distinct feature maps X_1, X_2, X_3, X_4 at varying resolutions and depths.

Unlike TransUNet and its variants, which only embed the last lowest-resolution feature maps, we employ linear embedding for multiscale feature maps. Specifically, the first three feature maps X_1, X_2, X_3 undergo linear embedding with a different patch size p to produce different embedded vector $z_i^j \in Z_i | i \in \{1, 2, 3\}$, which simultaneously go through transformer blocks to capture the interactions between patches and rotatory attention mechanisms to aggregate the information from adjacent slices. Within these transformer blocks, comprising N transformer layers, the embedded sequence patches traverse self-attention mechanisms and multilayer perceptrons, facilitating robust intra-slice information capture and yielding new encoded image representations E_1, E_2, E_3.

The rotatory attention block, conceived to treat the batch size as multiple continuous slices, selectively processes three consecutive slices—designating the first as the left, the second as the target, and the third as the right—culminating in the production of 3 vectors R_1, R_2, R_3 encapsulating information from adjacent slices in the volumetric data. Integration of interslice and intra-slice information yields F_1, F_2, F_3, which are then reconstructed to their original resolution via upsampling techniques, resulting in O_1, O_2, O_3.

Finally, X_4 undergoes concatenation with O_3, perpetuating this iterative process until the final segmentation map is obtained after 1×1 convolution.

subsectionFeature Extraction with Nested Skip Connections The input is structured as $(B, 1, H, W)$, representing the batch size, channels, height, and width. The batch size also represents the number of adjacent slices aggregated in the rotatory attention block. This input undergoes convolution to yield X_1^1, with shape (B, C, H, W), where $C = 64$. The resulting feature maps are downsampled to X_2^1, with dimensions $(B, C \times 2, \frac{H}{2}, \frac{W}{2})$. Then, X_2^1 is upsampled to $(B, C \times 2, H, W)$.and concatenated with X_1^1 along the C axis, resulting in $(B, C \times 3, H, W)$. This undergoes further convolution to produce X_1^2, which shares the same shape as X_1^1 but includes aggregated information from X_2^1. This process continues through subsequent lower-resolution images.

If we designate the desired number of different-resolution outputs as D, we have $X_i^j \ \forall i \in \{1, \ldots, D-1\}$ and $\forall j \in \{1, \ldots, D-i\}$, where X_i^j has shape $(B, C \times 2^{i-1}, \frac{H}{2^{i-1}}, \frac{W}{2^{i-1}})$. The D-th resolution map has a shape of $(B, C^{D-2}, \frac{H}{2^{D-1}}, \frac{W}{2^{D-1}})$, and bypasses both the Transformer block and Rotatory Attention block but is instead used for the decoder. For $D = 4$, the resulting feature maps are X_1^3, X_2^2, X_3^1. simply denoted as X_i for $i \in 1, 2, 3$. These are linearly embedded via convolution operations E to produce patches represented as embedded vectors $z_j^{p_i} \in Z_i$ where Z_i has shape

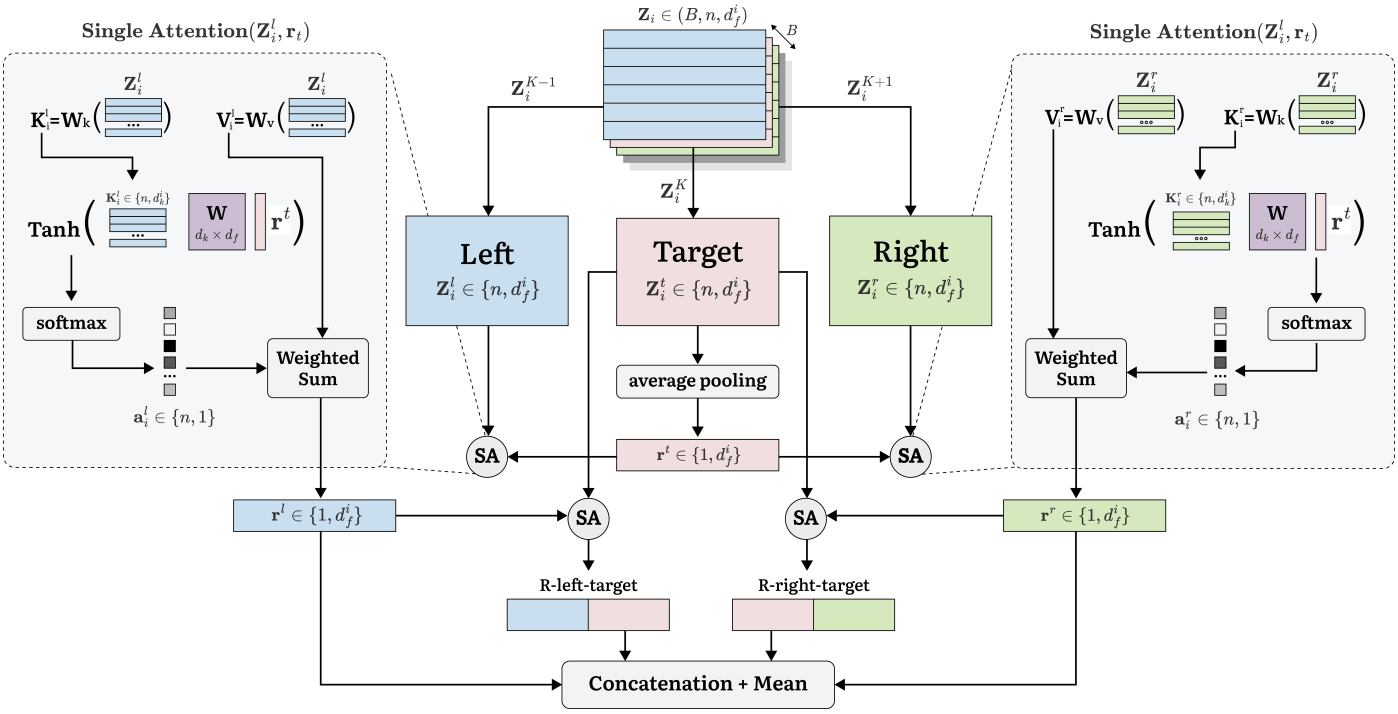

Fig. 2. The rotatory attention first uses the target slice to compute new representations for the left (previous slice) and right (next slice) context with a single attention module to capture the crucial inter-connectivity information from adjacent slices. Then, it uses these left and right representations to calculate the new representations for the target slice integrating essential information into the current slice.

(B, n_i, d_f^i) and $1 \leq j \leq n_i$. The sequence length and feature dimension of Z_i are $n_i = \frac{H_i \times W_i}{p_i^2}$ and d_i^f, respectively. Ensuring uniformity across n_i for all i, we establish $D-1$ patch sizes $p_i = 2^{D-i+1}$, where i ranges from 1 to $D-1$, implying that $p = \{2^4, 2^3, 2^2\}$ and the smallest patch size is $2^2 = 4$, given $D = 4$

B. Linear Embedding and Positional Embedding

Patch Embedding involves transforming vectorized patches $\hat{z}_j^{p_i} \in Z_i$ into a latent space of d_i dimensions using a trainable linear projection. To preserve the spatial information of the patches, we incorporate position embedding specific to each patch, which is then combined with the patch embeddings.

$$Z_i = E_i(X_i) + E_{\text{pos}}^i$$
$$Z_i = \hat{Z}_i + E_{\text{pos}}^i$$
$$[z_1^{p_i}, \ldots, z_n^{p_i}] = [\hat{z}_1^{p_i}, \ldots, \hat{z}_n^{p_i}] + [e_1^i, \ldots, e_n^i]$$

where E_i is the convolution operation to perform patch embbeding on X_i and produce $\hat{Z}^i$, while $E_{\text{pos}}^i \in (B, n, d_f^i)$ denotes the position embedding, Z_i is the linear embedding projection after adding vectors $\hat{z}_j^{p_i} \in (B, 1, d_f^i)$ with positional vectors $e_j^i \in (B, 1, d_f^i)$.

C. Transformer Block

The Transformer encoder consists of N layers of Multi-head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks. Therefore the output of the l-th $\in N$ layer can be formulated as follows:

$$\bar{Z}_i^{l'} = \text{MSA}(\text{LN}(Z_i^l)) + Z_i^l$$
$$Z_i^{l+1} = \text{MLP}(\text{LN}(\bar{Z}_i^{l'})) + \bar{Z}_i^l$$
$$\ldots$$
$$\bar{Z}_i^{N-1} = \text{MSA}(\text{LN}(Z_i^{N-1})) + Z_i^{N-1}$$
$$Z_i^N = \text{MLP}(\text{LN}(\bar{Z}_i^{N-1})) + \bar{Z}_i^{N-1}$$

where $LN(\cdot)$ denotes the layer normalization operator and Z_i^l is the encoded image representation at scale i. In each layer l-th, the encoded image representation Z_i undergoes a self-attention mechanism, enabling encoded patches to learn how to attend to each other. Mathematically, the attention scores $A_i = \text{Attention}(Q_i, K_i, V_i)$ for Z_i are computed using scaled dot product as follows:

$$A_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_f^i}}\right) V_i$$

where $Q_i = W_q(Z_i), K_i = W_k(Z_i), V_i = W_v(Z_i)$ and $Q_i, K_i, V_i \in (B, n, d_f^i)$. The Multi-Layer Perceptron (MLP) also contains a fully connected layer of size $d_i \times 4$ in the middle. The resulting E_i maintains the same shape as Z_i, which learns the intra-slice information or the relationship between patches in one 2D image slice.

TABLE I
BENCHMARK 9 MODELS ACROSS 4 MEDICAL DATASETS

Architecture	Params	MMWHS			Synapse			ImageCHD			VHSCDD			VHSCDD*		
		DSC	IOU	HD	DSC	IOU	HD	DSC	IOU	HD	DSC	IOU	HD	DSC	IOU	HD
UNet [9]	124.2M	0.78	0.61	28.3	0.61	0.43	30.5	0.72	0.52	26.1	0.50	0.29	39.4	0.449	0.26	89.5
Att-UNet [12]	32.54M	0.84	0.78	15.6	0.51	0.33	44.9	0.86	0.75	20.2	0.40	0.23	42.9	0.51	0.34	92.1
UNet++ [10]	36.64M	0.96	0.9	13.9	0.54	0.38	30.6	0.85	0.71	21.7	0.79	0.62	28.4	0.72	0.68	68.9
Att-UNet++ [17]	38.50M	0.84	0.78	15.6	0.68	0.51	21.5	0.81	0.65	23.7	0.80	0.64	22.6	0.68	0.62	64.7
ResUNet [11]	52.17M	0.76	0.64	17.6	0.47	0.31	40.6	0.68	0.56	34.2	0.56	0.35	41.9	0.61	0.56	40.9
Swin-unet [13]	165.4M	0.87	0.79	17.3	0.77	0.65	**23.9**	0.78	0.64	23.6	0.84	0.73	23.5	0.81	0.73	45.1
Att Swin-UNet [18]	165.4M	0.84	0.73	20.4	**0.79**	0.67	24.5	0.89	0.78	18.7	0.82	0.71	25.6	0.79	0.65	43.1
TransUNet [5]	420.5M	0.91	0.84	15.6	0.76	**0.78**	32.2	0.86	0.72	22.6	0.85	0.71	22.3	0.76	0.75	41.2
RotCAtt-TransUNet++	51.51M	**0.97**	**0.92**	**15.9**	0.68	0.61	25.6	**0.96**	**0.89**	**15.67**	**0.93**	**0.91**	**20.3**	**0.95**	**0.92**	**32.4**

D. Rotatory Attention Block

This technique is commonly applied in natural language processing [19], [20], which involves three main inputs: the target phrase, the previous phrase (left context), and the next phrase (right context). This method assumes that adjacent elements contribute significantly to understanding the central/target phrase. In our scenario, if we denote the current encoded input representation as $Z_i \in (B, n, d_f^i)$, we can treat this as a collection of images $\{Z_i^1, \ldots, Z_i^k, \ldots, Z_i^B\}$ where each $Z_i^k \in (n, d_f^i)$. Thus, three consecutive encoded slices/images can be selected as $\{Z_i^{k-1}, Z_i^K, Z_i^{K+1}\}$ or $\{Z^l, Z^t, Z^r\}$. For simplicity in notation, we temporarily omit the scale index i:

$$Z^l = [z_1^l, \ldots, z_j^l, \ldots, Z_n^l] \in \mathbb{R}^{n \times d_f}$$
$$Z^t = [z_1^t, \ldots, z_j^t, \ldots, z_n^t] \in \mathbb{R}^{n \times d_f}$$
$$Z^r = [z_1^r, \ldots, z_j^r, \ldots, z_n^r] \in \mathbb{R}^{n \times d_f}$$

The goal is to derive a single vector $r \in d_f$ and integrate it with Z^t to adjust the hidden states or transform the position of each embedded patch z_j^t in semantic dimensional space. Z^t is represented as a single vector r^t, incorporating necessary information from the left and right contexts by attention mechanism to filter noise and redundant information. Firstly, a single target representation is formed by:

$$r^t = \text{pooling}(z_1^t, z_2^t, \ldots, z_n^t) = \frac{1}{n}\Sigma_{j=1}^n z_j^t$$

Similar to the self-attention mechanism in Transformer layers, the key and value are extracted from the left context:

$$K^l = W_k^l(Z^l) = [k_1^l, \ldots, k_n^l] \in \mathbb{R}^{n \times d_f}$$
$$V^l = W_v^l(Z^l) = [v_1^l, \ldots, v_n^l] \in \mathbb{R}^{n \times d_f}$$

The r_t is now used as a query to create the context vector out of the left context. The scores are calculated with the activated general score function with tanh activation function, and the attention scores are calculated with the softmax function:

$$S^l = [s_1^l, \ldots, s_j^l, \ldots, s_n^l] = \tanh(K^l \cdot r^t + b^l)$$
$$a_j^l = \frac{\exp(e_j^l)}{\Sigma_{j=1}^n \exp(e_j^l)}$$

A weighted combination of patch embedding is considered as the component representation for left contexts:

$$r^l = \Sigma_{i=1}^n a_i^l \cdot v_i^l$$

In Figure 2, we denote the above process as Single Attention (SA), which is represented as:

$$\text{SA}(Z, r) = \begin{cases} K = W_k(Z), \quad V = W_v(Z) \\ a = \text{softmax}(\tanh(K \cdot r + b)) \\ r = \sum_n a \cdot V \end{cases}$$

The vector r^l is then used as a query to create context out of the target context to integrate information back into the center encoded slice/image to produce $r^{l/r} = SA(Z^t, r^l)$. An analogous procedure can be performed to obtain the right-aware target representation $r^r = SA(Z^r, r^t)$ and $r^{r/t} = SA(Z^t, r^r)$. Finally, to obtain the full representation vector r, we perform concatenation: $r^k = \text{concat}([r^l, r^r, r^{l/t}, r^{r/t}])$ with $r^k \in \mathbb{R}^{1 \times d_f \times 4}$. This r vector contains the aggregated information between 3 consecutive slices, thus we have $B-2$ vectors r^k with $1 < k < B$. The final vector R is achieved as: $R = W_r(\text{mean}(r^k | 1 < k < B))$. But this is only one i-th level output; thus, we have R_i output. This interslice-informational vector is added to encoded intra-slice-informational E_i to retrieve more optimized vectorized patch embeddings F_i.

E. Channel-wise Attention Gate for Feature Fusion

To fuse features with varied semantics between the Channel Transformer and U-Net decoder effectively, we employ a channel-wise cross-attention module, guiding channel and information filtration of Transformer features, resolving ambiguities with decoder features. Mathematically, we take the i-th level output F_i after Transformer and Rotatory blocks to reconstruct or decode the encoded image representations to get $O_i \in \mathbb{R}^{C \times H \times W}$. The reconstructed O_i are taken with i-th level decoder feature map $D_i \in \mathbb{R}^{C \times H \times W}$ as the inputs of Channel-wise Cross Attention. Spatial squeeze is performed by a global average pooling (GAP) layer, producing vector $\mathcal{G}(X) \in \mathbb{R}^{C \times 1 \times 1}$ with its k^{th} channel $\mathcal{G}(X) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W X^k(i, j)$. We use this operation to embed the global spatial information and generate the attention mask:

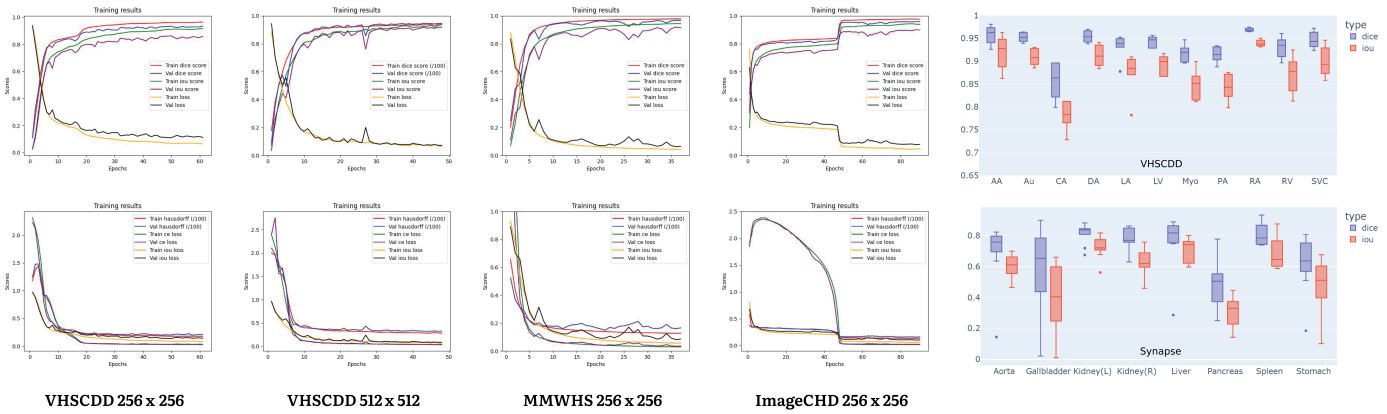

Fig. 3. Training graphs depict the performance of the RotCAtt-TransUNet++ model across datasets, and the box plots depict Dice/IoU scores across classes.

TABLE II
ABLATION STUDY OF ROTATORY ATTENTION ON VHSCDD 512

Type	DSC	IOU	HD	CE
w RotAtt	0.946±0.052	0.918±0.067	32.380±2.59	0.035±0.058
w/o RotAtt	0.904±0.078	0.864±0.037	37.019±2.89	0.048±0.087

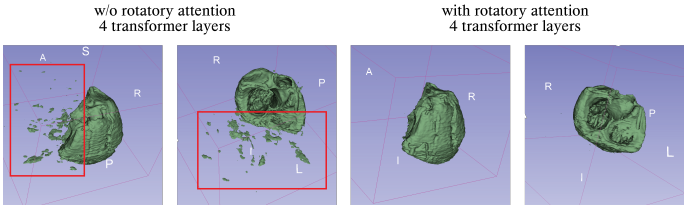

Fig. 4. 3D visualization of ablation study on rotatory attention

$$M_i = L_1 \cdot \mathcal{G}(O_i) + L_2 \cdot \mathcal{G}(D_i)$$

where $L_1 \in \mathbb{R}^{C\times C}$ and $L_2 \in \mathbb{R}^{C\times C}$ and weights of two Linear layers and the ReLU operator $\delta(\cdot)$, encoding channel-wise dependencies. Following ECA-Net [21], which emphasizes avoiding dimensionality reduction for effective channel attention, we use a single Linear layer and sigmoid function to build the channel attention map, then used to excite O_i to $\hat{O}_i = \sigma(M_i)\cdot O_i$. The activation $\sigma(M_i)$ indicates the importance of channels. Finally, the masked $\hat{O}_i$ is concatenated with the up-sampled features of the i-th level decoder.

IV. Experiments and Results

A. Datasets and Evaluation

We experimented with 5 CNN-based and 3 Transformer-based networks with our own network across 4 datasets: Multi-Modality Whole Heart Segmentation (MMWHS), Synapse multi-organ segmentation dataset, The Image Congenital Heart Diseases (ImageCHD) dataset, and Vietnamese Heart Segmentation and Cardiac Disease Detection (VHSCDD). We used an NVIDIA RTX 4090 1X GPU with 24GB memory and 81.4 TFLOPS for training and testing. We reported three metrics: Dice Coefficient Score (DSC), Intersection over Union (IoU) scores, and Hausdorff Distance (HD).

B. Implementation Details

We implement the Dice score differently: it operates on logits before $arg\max$ to maximize confidence scores of predicted pixels per class. At the same time, IoU compares segmentation accuracy between ground truth G and prediction P after $arg\max$ predictions. The loss is the reverse of those metrics:

$$\text{Dice Loss} = 1 - \frac{1}{c}\sum_{c}\frac{2\sum_{ij} P_{ij}^{c} \times G_{ij}^{c}}{\sum_{ij} P_{ij}^{c} + \sum_{ij} G_{ij}^{c} + \epsilon} \quad \forall c \neq 0$$

$$\text{IoU Loss} = 1 - \frac{\sum_{ij} P_{ij} \times G_{ij}}{\sum_{ij}(P_{ij} + G_{ij} - P_{ij} \times G_{ij})}$$

The exclusion of $c \neq 0$ ensures the avoidance of unreal DSC and IoU scores from dominant background pixels. Our combined loss function is defined as:

$$L = \alpha \times \text{IoU Loss} + (1 - \alpha) \times \text{Dice Loss}$$

In our implementation, we set α to 0.6 because we observed that the IoU loss consistently exceeds the Dice loss. Therefore, we opt to increase the penalty on the model.

C. Results and Discussion

The results are summarized in Table I. The VHSCDD* dataset contains images of size 512x512, while VHSCDD and other datasets have images of size 256x256. Our model's lightweight nature is due to having only four transformer layers. The robustness of rotatory attention allows encoded vectorized patches to be effectively transformed in the semantic space, reducing the need for numerous transformer layers. As shown in Figure 3, our network demonstrates the fastest convergence time when applied to cardiac data, thanks to robust long-range interslice connectivity. However, despite RotCAtt-TransUNet++ outperforming other methods across various datasets and metrics, it is less effective on the Synapses dataset.

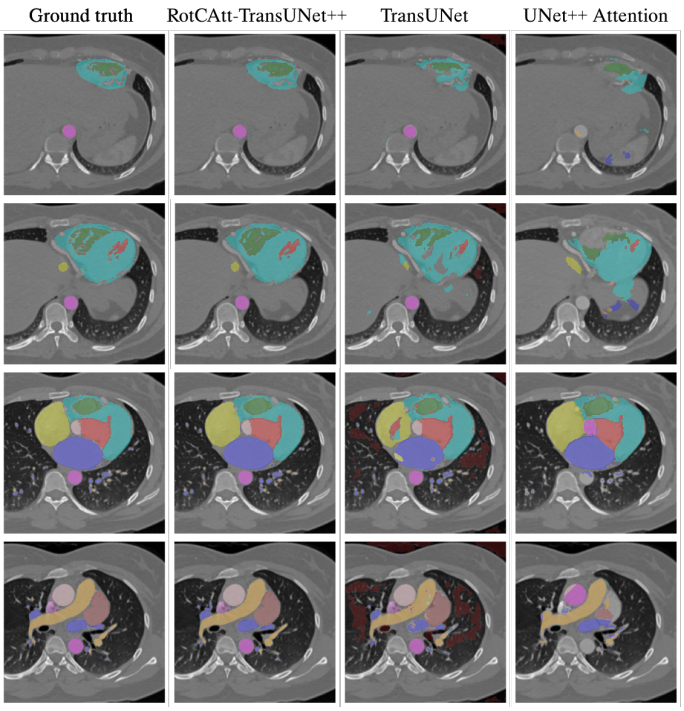

Fig. 5. Segmentation comparison between our method with different ones

After dataset analysis, we conclude this may be due to the discontinuous nature of organ structures in this dataset, where the model struggles to aggregate adjacent slice information or the necessary information is distant (exceeds batch size) along the z-axis. Increasing the number of transformer layers, as in TransUNet, would provide only marginal improvement while significantly increasing model parameters/complexity. Therefore, this area remains open for future improvement. We conducted an ablation study on the VHSCDD dataset to compare results with and without the attention mechanism. As shown in Table II, the DSC and IoU scores drop significantly, and in Figure 4, the "spraying phenomenon" occurs when no rotatory attention is applied. Our attention map analysis revealed that non-cardiac regions outside the heart exhibit high similarity to patches of the myocardium. Additionally, as illustrated in Figure 5, our method achieves near-perfect segmentation across all classes. In contrast, TransUNet (a Transformer-based approach) and UNet++ Attention (a CNN-based approach) did not perform as well. The "spraying phenomenon" is also noticeable in the TransUNet segmentation results.

V. Conclusion and Implication

In conclusion, Transformer-based methods excel in self-attention, while CNN-based methods are strong in localization. Our study introduces RotCAtt-TransUNet++, featuring nested skip connections for multiscale feature extraction in the encoder, followed by transformer layers, and rotatory attention blocks. This architecture enhances image representation and segmentation accuracy, particularly in complex cardiac datasets. Experimental results show near-perfect annotation of critical structures like coronary arteries and myocardium, with the ablation study confirming the effectiveness of rotatory attention. Future research aims to refine the architecture, and integrate advanced techniques to improve segmentation efficiency and clinical outcomes in cardiovascular diseases.

References

[1] R. Azad, M. T. AL-Antary, M. Heidari, and D. Merhof, "Transnorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model," 2022.

[2] A. Suinesiaputra, B. Cowan, J. Finn, C. Fonseca, A. Kadish, D. Lee, P. Medrano-Gracia, S. Warfield, W. Tao, and A. Young, "Left ventricular segmentation challenge from cardiac mri: A collation study," vol. 7085, 09 2011, pp. 88–97.

[3] C. Petitjean, M. Zuluaga, W. Bai, J.-N. Dacher, D. Grosgeorge, J. Caudron, S. Ruan, I. Ben Ayed, M. J. Cardoso, H.-C. Chen, D. Jimenez-Carretero, M. Ledesma-Carbayo, C. Davatzikos, J. Doshi, G. Erus, O. Maier, C. Nambakhsh, Y. Ou, S. Ourselin, and J. Yuan, "Right ventricle segmentation from cardiac mri: A collation study," *Medical Image Analysis*, 10 2014.

[4] S. Park and M. Chung, "Cardiac segmentation on ct images through shape-aware contour attentions," 2021.

[5] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," 2021.

[6] X. Xu, T. Wang, J. Zhuang, H. Yuan, M. Huang, J. Cen, Q. Jia, Y. Dong, and Y. Shi, "Imagechd: A 3d computed tomography image dataset for classification of congenital heart disease," 2021.

[7] P. V. Tran, "A fully convolutional neural network for cardiac segmentation in short-axis mri," 2017.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," 2015.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[10] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," 2018.

[11] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, p. 94–114, Apr. 2020. [Online]. Available: http://dx.doi.org/10.1016/j.isprsjprs.2020.01.013

[12] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018.

[13] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," 2021.

[14] Özgün Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," 2016.

[15] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," 2016.

[16] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," 2017.

[17] C. Li, Y. Tan, W. Chen, X. Luo, Y. Gao, X. Jia, and Z. Wang, "Attention unet++: A nested attention-aware u-net for liver ct image segmentation," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 345–349.

[18] E. K. Aghdam, R. Azad, M. Zarvani, and D. Merhof, "Attention swin u-net: Cross-contextual attention mechanism for skin lesion segmentation," 2022.

[19] S. Zheng and R. Xia, "Left-center-right separated neural network for aspect-based sentiment analysis with rotatory attention," 2018.

[20] G. Brauwers and F. Frasincar, "A general survey on attention mechanisms in deep learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, p. 3279–3298, Apr. 2023. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2021.3126456

[21] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and 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 (CVPR)*, June 2020.