2023 4th International Conference on Communications, Information, Electronic and Energy Systems (CIEES)
23 – 25 November, 2023, Plovdiv, Bulgaria

Hybrid Residual Blocks - Transformers Improving Cardiac MRI Segmentation in Deep Learning

Maysaa Abd Ulkareem Naser
Dept. of Computer Information Systems,
University of Basrah
Basrah, Iraq
Dept. of Computer Science
Kufa University,
Kufa, Iraq
maysaa.naser@uobasrah.edu.iq

Abbas H. Hassin Al-Asadi
Dept. of Computer Information Systems
University of Basrah
Basrah, Iraq
abbas.hasin@uobasrah.edu.iq

***Abstract*—Cardiovascular disease is one of the primary causes of mortality and disability. By assisting in early diagnosis, segmentation techniques based on deep learning might help to lessen their severity, although extremely high levels of accuracy are required. Introduce a novel and flexible architecture that raises the bar for semantic segmentation accuracy in this study. Our contributions focus on the merging of strong self-attention mechanisms obtained from Transformer with traditional convolutional neural network (CNN) architecture. This hybrid model demonstrates improved feature representation, contextual understanding, and pixel-wise predictions, setting new standards in semantic segmentation tasks. the average dice coefficients for ACDC dataset by our model, AVG DICE 92.52%, RV 90.77%, MYO 90.00%, LV 96.28%**

Keywords—image segmentation, vision transformer, residual blocks, deep learning.

I. INTRODUCTION

Individualized modeling of the heart has been employed for non-invasive cardiac rhythm problem diagnosis and therapy, including risk stratification of heart attack patients, estimate of the re-entry site and clinical ablation guidance [1]. The precise production of the individualized model, which is now mostly handled by seasoned professionals, is the key to the therapeutic use of cardiac models [2]. An automated segmentation approach is essential for the medical application of customized cardiac modeling, because simulation takes a long time and manual segmentation is subjective, unpredictable, and time-consuming. Traditional techniques for segmenting medical pictures have been substantially supplanted by convolutional neural networks (CNNs). more specifically, fully convolutional networks (FCNs) [1, 3]. These topologies frequently produce below-average outcomes, especially for target structures that exhibit significant texture, shape, and size inter-patient variation. This is especially true for target structures with substantial size, shape, and texture differences between patients. Existing research suggests developing self-attention strategies based on CNN properties to address this issue. [4]

In contrast, the field of medical image segmentation has shown tremendous promise for the capacity of transformers to recognize distant links by leveraging self-attention. Since the 2015 FCN proposal, Convolution Neural Networks (CNN) in particular have shown to be successful when applying deep learning techniques for medical image semantic segmentation. Convolution operations are geographically localized, which is a weakness of the CNN-based method since it prevents it from describing long-range associations [5]. U-shaped convolution neural networks (CNNs) have gained popularity with the publication of UNet, Ronneberger et al. [2] Self attention mechanisms have been suggested as a solution to this issue Schlemper et al. [6] As a result, structures with varied forms and scales (such as brain lesions with different sizes) are not segmented with enough accuracy. This is because the challenge of obtaining multi-scale contextual information has not yet been resolved. Transformers are a more suitable alternative method for representing global contextual data.

Unlike more recent cutting-edge image recognition methods, by breaking the image up into patches, Vision Transformer (ViT) Dosovitskiy et al. [7] models the relationship between these patches as sequences offered a knowledge distillation strategy. Bakas et al. [8] carried out a thorough analysis to find the most efficient way for segmenting brain tumors. built Vision Transformers with the intention to teach [5]. Due to the three-dimensional nature of medical images from CT and MRI, volumetric segmentation is crucial. However, the most recent research shows that Transformer, a purely self-attention based approach, beats RNN-based approaches with various blocks in the natural language process [9].

The ViT is then proposed as a method to examine Transformer's feature learning capabilities in the context of computer vision [10]. The suggested Swin Transformer performs well on the tasks of semantic segmentation identification, object detection, and picture classification. It outperforms the ViT/DeiT [11] and ResNe(X)t models [12] on all three of the challenges. with comparable latency. Encoder-Decoder type segmentation models, like TransUNet, are at the forefront of other ViT segmentation backbone research. [13], which offers the ViT augment ordinary UNet in the encoder and Swin-UNet [14], which directly shows the Swin-ViT blocks inserted into U-shape backbone network.

Despite the fact that a recent study on network design with ViT produced promising outcomes for the discipline of computer vision, one of the major obstacles to implementing cutting-edge methods in clinical medical image processing is still the training approach [13]. Due to the high cost of clinician annotation and the necessity of high expert level ability, medical data typically results in a significant volume

979-8-3503-3691-7/23/$31.00 ©2023 IEEE

of raw data with a tiny fraction of annotations in the community of medical imaging. In this paper, we describe an enhanced UNet architecture that makes use of the attention mechanism and residual learning. In brief, the paper's contributions are as follows:

1) Enhancing Feature Representation in the Encoder

Our key contribution lies in enhancing the encoder component of our architecture using a series of Residual blocks. Unlike conventional approaches, we apply a multi-layered residual design that includes three convolutional layers, each having Parametric Rectified Linear Unit (PReLU) activation and Batch Normalization. The incorporation of such residual structures empowers the model to capture intricate features, while the inclusion of Dropout layers maintains generalization capabilities.

2) Transformer-Infused Decoding

Another pivotal aspect of our contribution is the integration of self-attention mechanisms inspired by Transformers into the decoder's design. The Transformer component consists of two ViT blocks, each with Multi-Head Self-Attention and Multi-Layer Perceptron (MLP) layers. In addition to helping the model capture long-range relationships, this strategic integration improves spatial comprehension and feature refinement, leading to predictions that are more accurate.

3) Novel Up-sampling Strategy

Incorporating a novel up-sampling strategy, we introduce the Up-ResBlock function within the decoder. This function enhances the reconstruction process by applying residual connections to up-sampled feature maps, effectively mitigating information loss during decoding. The strategic concatenation of encoder and decoder features ensures a seamless flow of information and contributes to the model's pixel-wise accuracy.

4) Comprehensive ViT Integration

Our work extends beyond conventional architectures with the introduction of ViT-based decoding. The ViT-block function is at the core of this integration, offering a combination of Multi-Head Self-Attention, MLP, and skip connections. This innovative strategy shows off the adaptability of our architecture in handling various semantic segmentation tasks by enabling the model to recognize intricate relationships and patterns in the data.

5) Efficient Patch-Based Processing

To accommodate the integration of Transformers, we introduce the patch-extract and patch-embedding layers. This approach streamlines the application of self-attention mechanisms and enhances the overall model's computational efficiency.

The organization of this paper is as follows: The first part: contains the introduction. The second part: contains an explanation of previous studies related to transformers and segmentation of medical images, Part Three: Contains the working method and its installation, the fourth part: contains a discussion of the results, and finally the fifth part contains an evaluation of the proposed method.

II. Related Works

A. Vision transformers

The first vision transformer (ViT), created by Dosovitskiy et al. [7], may employ SA to identify distant (global) relationships among the pixels. Recent efforts to enhance ViT have resulted in new SA blocks being created. [15] Tu and others. Reduce the computational expense of ViT by integrating CNNs and taking into account data-efficient training techniques [5]. Wang et al. [16] created a pyramid vision transformer for PVT utilizing a spatial reduction attention technique. PVT is changed by the authors into PVTv2. Positional encoding and Mix-FFN blocks are used in SegFormer, a free hierarchical transformer proposed by Xie et al. [17]. A pyramid vision transformer for PVT has been developed by Wang et al. [16] using a spatial attention reduction method.overlapped patch embedding, a linear complexity attention layer, and a convolutional feed-forward network [18]. A method for creating a hierarchical hybrid CNN transformer in MaxViT employing a multi-axis self-attention methodology was recently described by authors [15]. Despite the fact that multi-scale transformer backbone design is not given much care and multi-scale transformers are only partially capable of processing spatial information, Lin et al. [19] emphasize that vision transformers have shown substantial potential. To get around the same restrictions, this technique employs a multi-scale vision transformer with attention-based decoding.

B. Medical Image Segmentation

The classification of the pixels of lesions or organs in endoscopy, CT, MRI, etc. may be viewed as a dense prediction challenge in the context of medical picture segmentation. Dong [20] is one illustration. Because of their complex encoder-decoder architecture, U-shaped designs, such as those created by Ronneberger et al. [2], Zhou et al. [21], Huang et al. [22], and Lou et al. [23], are often employed in medical picture segmentation. According to Zhou et al. [21], the creators of UNet++, UNet is a hierarchical encoder-decoder design that utilizes sub-networks connected by dense skip links. The full-scale skip connections with intra-connections between the decoder blocks are examined by UNet3Plus. Transformers are now often employed to categorize medical image data, claim [22].

Dong and the others 18 Using PVTv2, the encoder is Wang et al. [18], and the opposite emphasis Attention blockers in the decoder include Fan et al. [24], Chen et al. [13], and Woo et al. [25] authors of PraNet. The designers of CASCADE, Rahman and Marculescu [26], offer a cascaded decoder that enhances features by using attention modules. By combining Transformer and U-Net, Fu et al. [27] are able to segment medical pictures effectively and totally automatically. Various authors Create a brand-new DualNorm-UNet for network normalization that simultaneously integrates regionally specific local statistics with global image-level data [28]. We suggest our model, a system for segmenting medical images that incorporates residual blocks and transformer blocks, when compared to earlier techniques based on pure convolution and pure transformer, our suggested hybrid network performs and is more resilient.

III. Proposed Method

We give a thorough explanation of our painstakingly designed image segmentation model in this section. The robust and accurate framework for picture segmentation produced by our method, which seamlessly combines the powerful capabilities of Convolutional Neural Networks (CNNs) and Vision Transformers (VITs), is shown in Fig 1.

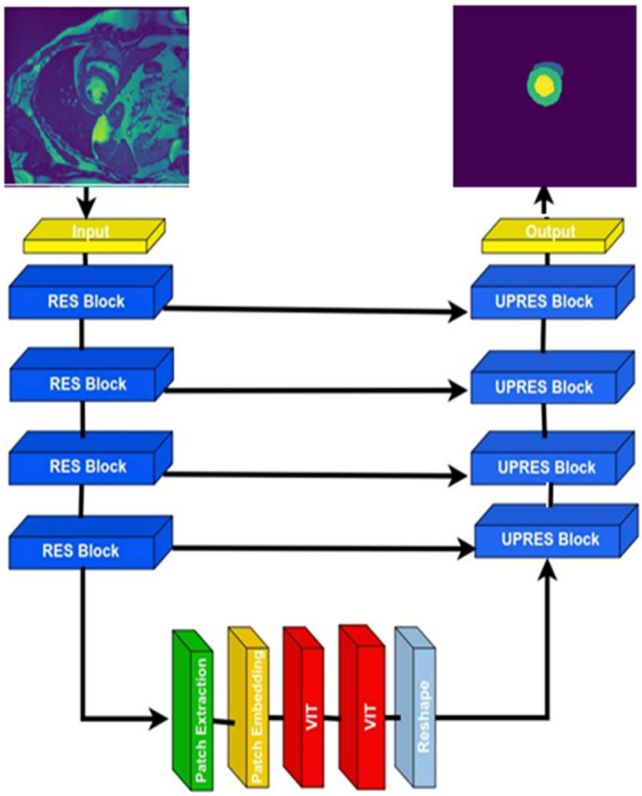

Fig. 1. Capabilities of convolutional neural networks and vision transformers.

The schematic representation above illustrates the core components of our model RT-ResUNet, that exploits residual learning, squeeze and excitation operations, patch Extraction, patch Embedding, two Vision Transformers, Reshape, and the attention mechanism for accurate and offering a visual of how CNNs and VITs collaboratively contribute to the image segmentation process.

Our methodology is driven by a singular goal: precision in image segmentation. This pursuit of accuracy is underpinned by the harmonious synergy of the following potent components:

C. Encoder

The Encoder serves as the foundational pillar of our image segmentation model, designed to meticulously capture hierarchical features from the input image. This component is comprised of four Residual blocks, each of which plays a crucial role in the feature extraction process.

Residual Blocks: Hierarchical Feature Extraction at the core of our Encoder, the Residual blocks are the workhorses responsible for hierarchical feature extraction. Each Residual block consists of N sequential Conv2D blocks, where N corresponds to the number of Conv2D blocks within a Residual block in the encoder. The Conv2D block, in turn, comprises a convolutional layer, PReLU activations, and batch normalization, meticulously crafted to enhance the model's feature learning capabilities (as depicted in Fig. 2).

Skip Connections: Preserving Information to facilitate gradient flow and mitigate the vanishing gradient problem, each Residual block is augmented with a skip connection. This architectural feature allows the input tensor to be added directly to the output tensor of the block. This not only aids in the preservation of critical information but also promotes efficient training of deep networks.

Down sampling Mechanism: However, it is important to note that down sampling occurs within a specific segment of each Residual block, as illustrated in Fig. 3. After the skip connection, the tensor undergoes further processing through a sequence of operations, which includes PReLU activations, dropout, and a Conv2D block. This particular segment is where the down sample operation takes place, reducing the spatial dimensions of the tensor.

The Encoder's primary function is to capture and encode hierarchical features present in the input image. The Residual blocks, each consisting of N Conv2D blocks, ensure that features at varying levels of abstraction are adequately represented. The skip connections retain the vital information flow while the down sampling approach fine-tunes the feature maps' spatial resolution. Our model can comprehend both low-level and high-level visual information due to the hierarchical feature extraction procedure, which eventually improves the precision and accuracy of image segmentation.

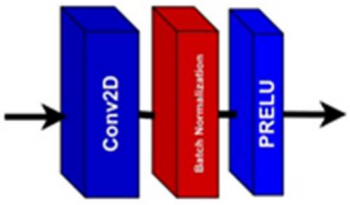

Fig. 2. Schematic representation of Conv2D block.

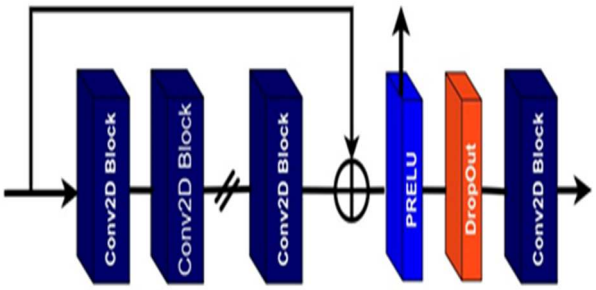

Fig. 3. Schematic representation of residual block.

D. Vision Transformers (VITs)

Following the Encoder's feature extraction phase, as we move from the Encoder to the Decoder, we are easily transitioned to the Transformer component. This segment comprises two Vision Transformer (ViT) blocks, they are essential for capturing long-range relationships and context in the data. Our objective is to transform the feature representation from the Encoder into a format that the Transformer can effectively process [7]. To achieve this, we perform tokenization by reshaping the input feature tensor, denoted as $\boldsymbol{x} \in \boldsymbol{R}^{(\boldsymbol{H}\times\boldsymbol{W}\times\boldsymbol{C})}$, into a sequence of flattened 2D patches $\left\{\boldsymbol{x}_{\boldsymbol{p}}^{\boldsymbol{i}} \in \boldsymbol{R}^{(\boldsymbol{P}^{\boldsymbol{2}}*\boldsymbol{C})} \middle| \boldsymbol{i} = \boldsymbol{1},..,\boldsymbol{N}\right\}$. Each of these patches has a spatial dimension of $\mathbf{P} \times \mathbf{P}$, and N represents the total number of patches, calculated as $\boldsymbol{N} = \frac{\boldsymbol{HW}}{\boldsymbol{P}^{\boldsymbol{2}}}$ (input sequence length).

Patch Embedding, Capturing Patch Information with tokenization complete, we move on to Patch Embedding. Here, our objective is to map the vectorized patches ($\boldsymbol{x_p}$) into Representing a latent D-dimensional embedding space via a trainable linear projection. for these patches to maintain crucial positional information, we introduce specific position embedding. The addition of position embedding ensures that the model retains a sense of the spatial relationships between patches. This operation can be summarized as follows:

$$\boldsymbol{z_0 = [x1_{PE}; x2_{PE}; \cdots; xN_{PE}] + E'_{POS}}, \quad (1)$$

where $\boldsymbol{E}$ represents the patch embedding projection matrix of size $(\boldsymbol{P^2 \cdot C}) \times \boldsymbol{D}$, and $\boldsymbol{E_{pos}}$ represents the position embedding matrix of size$\boldsymbol{NxD}$. These position embeddings are learned during training, allowing the model to adapt to different spatial contexts.

Vision Transformer, we develop a Transformer encoder made up of L levels in the Vision Transformer (VIT), with each layer meticulously designed to capture complex connections in the input data. These layers effortlessly switch between Multi-Head Self-Attention (MSA) and Multi-Layer Perceptron (MLP), two essential building components. The essence of our VIT architecture lies in its hierarchical composition. Each layer - identified by the $L-th$ layer can be expressed as follows:

$$z_l` = \mathrm{MSA}\big(\mathrm{LN}(z_l - 1)\big) + z_l - 1 \quad (1)$$

$$z_l = \mathrm{MLP}\big(\mathrm{LN}(z_l`)\big) + z_l`. \quad (2)$$

These equations unveil the intricate operations within each layer. Here, $\boldsymbol{z_l}$ represents the output state of the layer, while $\boldsymbol{z_l - 1}$ is the preceding layer's output. $\boldsymbol{LN}(.)$ signifies the crucial layer normalization operation, ensuring stable training and consistency throughout the model. For a visual grasp of the architecture and structural layout of a Transformer layer, please refer to Fig. 4, which provides a concise illustration of this integral component.

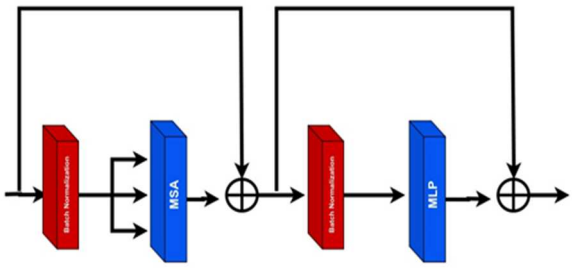

Fig. 4. Structure of a transformer layer.

E. Decoder

The Decoder component is instrumental in our model, tasked with the critical role of reconstructing high-resolution segmentation masks [13]. This is achieved through the utilization of multiple Up-Residual blocks, closely resembling the Residual blocks in the Encoder [12]. However, their primary function diverges significantly; instead of down-sampling, these Up-Residual blocks specialize in up-sampling, enabled by Conv2DTranspose layers.

Up-Residual blocks serve as the vital bridge between the abstract, the encoder's high-level feature extraction and the output of the segmentation. They meticulously apply Convolutional layers, Parametric Rectified Linear Unit (PReLU) activations, and Batch Normalization to enhance the richness and context of the encoded features.

The distinctive feature of these Up-Residual blocks is their proficiency in up-sampling, effectively expanding the spatial dimensions of the data. Conv2DTranspose layers are the key enablers of this up-sampling process, allowing the model to regain any fine-grained spatial information that could have been lost during the first down-sampling steps.

Final Layer, At the pinnacle of the Decoder resides the final layer, a critical element for multi-class segmentation. This layer is meticulously designed, incorporating a Convolutional layer that operates on the up-sampled features. Its paramount function lies in the application of the Soft-Max activation function. The Soft-Max activation function is a cornerstone of multi-class segmentation tasks. It expertly transforms the raw model output into a probability distribution over multiple class. Each pixel in the output corresponds to a specific class, and the Soft-Max activation ensures that the pixel values are normalized, summing up to 1 across all classes. This normalization process provides a clear and intuitive delineation of class membership for each pixel in the final segmentation map. For a visual grasp of structure and organization of the Decoder component, please refer to Fig. 5, which offers a schematic representation of this pivotal segment of our model.

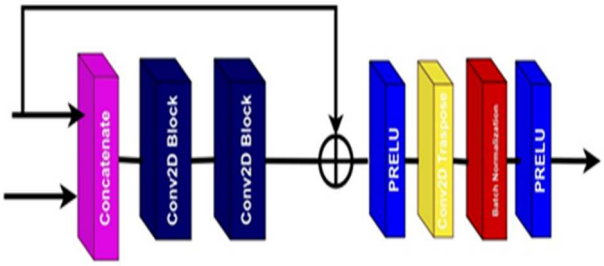

Fig. 5. Schematic representation of up-residual block.

IV. Implementation Details

F. Data Preprocessing and Data Enhancement

ACDC Dataset .100 cardiac MRI images from various sources are part of the ACDC collection. From each MRI scan, we slice three organs into two-dimensional (2D) slices, such as the myocardium (MYO), left and right ventricles (LV and RV).

Synapse multi-organ dataset We employed a subset of the labelled training set for testing (training sample = 14, validation sample = 7, test sample = 9) and just the second scenario from the ACDC. In order to assess the model's effectiveness, we employed the mean dice similarity coefficient (DSC) for eight abdominal organs: the aorta, gall bladder, spleen, left kidney, right kidney, liver, pancreas, and stomach.

Images from the same dataset are all cropped to the same size after being resampled to the same target spacing. The training approach uses a number of data augmentation techniques, including as rotation, scaling, Gaussian blur,

Gaussian noise, brightness correction, and contrast enhancement, because there aren't enough training examples.

G. Evaluation Metrics

Our loss function offers a harmonic fusion of two essential elements, Binary Cross-Entropy (BCE) Loss and Dice Loss, to enhance image segmentation tasks. Pixel-wise class distribution discrepancies and segmentation mask similarities are efficiently balanced by this hybrid loss function.

1) Binary Cross-Entropy (BCE) Loss:

The BCE Loss serves as a foundational element, quantifying the disparity between predicted and true pixel-wise class distributions. It is expressed mathematically as:

$$\mathrm{BCE}(y_{\mathrm{true}}, y_{\mathrm{pred}}) = -\frac{1}{N\sum_{i=1}^{N}\left[y_{\mathrm{true}}^{(i)}\log\left(y_{\mathrm{pred}}^{(i)}\right)+\left(1-y_{\mathrm{true}}^{(i)}\right)\log\left(1-y_{\mathrm{pred}}^{(i)}\right)\right]}. \quad (3)$$

Here, $y_{true}^{(i)}$ represents the true label for pixel i, $y_{pred}^{(i)}$ denotes the corresponding predicted probability, where N is the segmentation's overall pixel count.

2) Dice Loss:

The Dice Loss assesses the degree of similarity between the expected and actual segmentation masks as a complement to the BCE Loss. It employs the Dice coefficient, mathematically defined as:

$$\mathrm{Dice}(y_{\mathrm{true}}, y_{\mathrm{pred}}) = 2.|\, y_{\mathrm{true}} \cap y_{\mathrm{pred}} \,|/|\, y_{\mathrm{true}} + |\, y_{\mathrm{pred}} \,| \quad (4)$$

Here, signifies the intersection of pixels correctly classified as belonging to the target class, corresponds to the total amount of pixels in the predicted mask and the total amount of pixels in the real mask. To ensure numerical stability, a small constant term, denoted as 'smooth,' is added to the denominator.

3) Weighted Sum:

The final loss function is an intelligently weighted sum of these two integral components, formulated as follows:

$$\mathrm{Loss}_{\mathrm{final}} = w_{\mathrm{BCE}}\mathrm{BCE}(y_{\mathrm{true}}, y_{\mathrm{pred}}) + w_{\mathrm{Dice}}\mathrm{Dice}(y_{\mathrm{true}}, y_{\mathrm{pred}}) \quad . \quad (5)$$

The weights for the BCE Loss and Dice Loss are shown here as w_{BCE}, w_{Dice}, and correspondingly. A deliberate decision was made to give Dice Loss a larger weight because of its increased sensitivity to minute segmentation mistakes, which encourages accuracy.

V. Discussion and Results

In this part, we go into great depth on the experimental outcomes produced by our algorithm and examine how various factors affect the model's performance, which we contrasted using the appropriate ACDC and Synapse datasets. We explicitly examine the effects of different learning analyses from a quantitative perspective. The model is Trans CASCADE [24], the best transformer-based model from Table 1 with an average dice coefficient of 91.63%. Despite the fact that R50-UNet, the best convolution-based model, with an average dice coefficient of 87.55%, our recommended RT-ResUNet outperforms Trans CASCADE and R50-U Net. Even if these networks' accuracy in their current form is already fairly good, the network update we suggest is still quite efficient. Indicating that our approach may produce superior prediction analysis show Fig. 6. Layer by layer comparison of our technique's results shows that they are very close to the genuine value and that excellent outcomes are still feasible for the right ventricle. This is difficult to divide. We measured our Synapse experiments and compared our RT-ResUNet to a range of transformer- and U-net-based baselines, as shown in Table 2. The dice factor serves as the primary assessment metric. The residual block-transformer based technique with the highest average score, Swin-Unet, earns a score of 79.13. Swin Unet has significantly lower results than DualNorm-UNet, which boasts the best CNN-based results with an average of 80.3 7 and TF-Unet 85.64. Our RT-ResUNet outperforms the average performance of Swin Unet and DualNorm UNet and TF-Unet, which is a significant improvement over Synapse. Qualitatively. In this study, we show that superior global and distant semantic information interactions may be learnt by integrating hybrid residual block-transforms procedures, leading to improved segmentation outcomes.

A. Results ACDC Cardiac Segmentation

Table 1 lists the outcomes of three distinct cardiac segmentation techniques on the ACDC dataset using MRI data modality. Compared to previous approaches, our method has higher Dice ratings. demonstrates segmentation with the highest Dice scores for RV (90.77%) MYO (90.0%)and LV (96. 28%).We created two experimental scenarios for running experiments on the ACDC dataset. By employing 80 training data as the training set and 40 test data as the test set, the first scenario utilizes the whole dataset. We divided the 100 labeled training data into 70 training sets, 20 validation sets, and 10 test sets in order to statistically analyze our results. The real labels for the 20 test circumstances were not included in the instruction. We may infer from these results that our technique performs best when used with different types of medical imaging data.

Results based on data from the ACDC are shown on Table 1, for each organ, DICE scores (%) are presented. We display the outcomes of OURS averaging. The dark color represents the results of our method.

TABLE I. Results based on data from the ACDC

Architectures	AVGDICE	RV	MYO	LV
R50+UNet [11]	87.55	87.10	80.63	94.92
ViT+CUP [11]	81.45	81.46	70.71	92.18
VIT[4]	81.45	81.46	70.71	92.18
TransUNet [11]	89.71	88.86	84.53	95.73
Swin-UNet [12]	90.00	88.55	85.62	95.83
R50-VIT[4]	87.57	86.07	81.88	94.75
MISS Former [20]	90.86	89.55	88.04	94.99
Trans-CASCADE [24]	91.63	89.14	90.25	95.50
(Ours)	**92.52**	**90.77**	**90.00**	**96.28**

TABLE II. RESULTS FROM THE MULTI-ORGAN SYNAPSE DATASET. FOR EACH ORGAN, DICE SCORES (%) ARE PRESENTED. THE DARK COLOR REPRESENTS THE RESULTS OF OUR METHOD

Methods	DSC(avg)	Aorta	Gallbladder	Kidney (L)	Kidney -(R)	Liver	Pancreas	Spleen	Stomach
VIT [4]	67.86	70.19	45.10	74.70	67.40	91.32	42.00	81.75	70.44
R50-VIT [4]	71.29	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
DualNorm-Net [26]	80.37	86.52	55.51	88.64	86.29	95.64	55.91	94.62	79.80
SQNet [27]	73.76	83.55	61.17	76.87	69.40	91.53	56.55	85.82	65.24
TransUNet[11]	77.48	87.23	63.16	81.87	77.02	94.08	55.86	85.08	75.62
Swin-Unet[12]	79.13	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
TF-Unet[25]	85.46	87.45	63.10	92.44	93.05	96.21	79.06	88.80	83.57
R50-net[11]	74.68	87.74	63.66	80.60	78.19	93.74	56.90	85.87	74.16
Ours	**87.75**	**88.23**	**85.85**	**87.45**	**84.85**	**96.20**	**70.75**	**89.01**	**85.43**

B. Maintaining the Integrity of the Specifications

To evaluate the model's efficacy for Synapse data, we used the mean dice similarity coefficient (DSC) for eight abdominal organs, including the aorta, gall bladder, spleen, left kidney, right kidney, liver, pancreas, and stomach. For ACDC, we just applied the second scenario. As can be shown, for 2D medical picture segmentation, both of our versions outperform every CNN and transformer-based method. See Table 2**,** our approach yields the highest average DICE score (87.75%) of all the approaches.

In Fig. 6 the original image in the left, predicted image in middle, actual image in the right, Yellow represents the left ventricle, green the myo, and blue the right ventricle. We chose a number of patients' findings at random to be visualized.

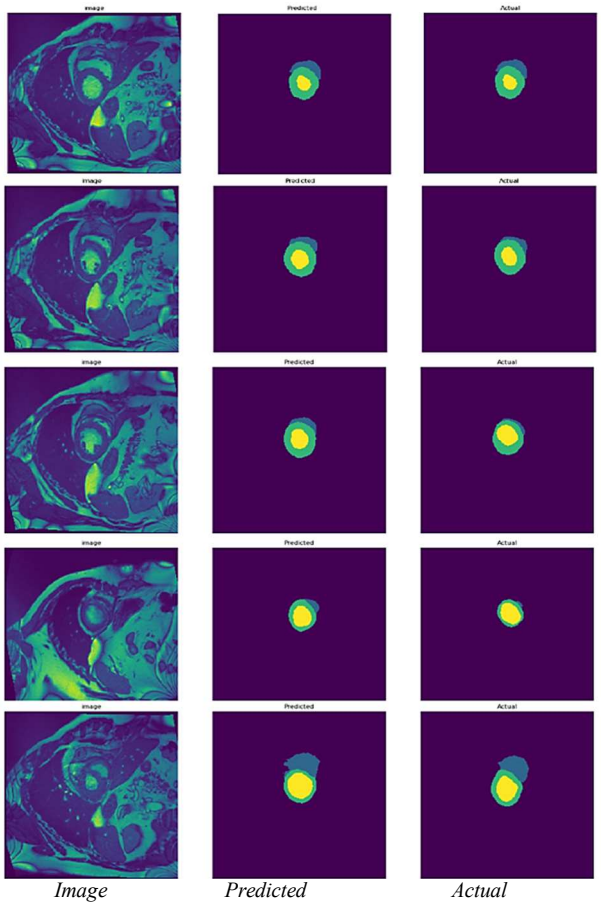

Fig. 6. Example of segmentation results.

VI. CONCLUSIONS

In this research, we suggest the RT-ResUNet, a brand-new network for segmenting medical images. Convolution and self-attention from the interconnected backbone of RT-UNet, high effectively utilizes the fundamental CNN properties to construct hierarchical object conceptions at various sizes using a U-shaped hybrid architectural design nusing a series of Residual blocks. The distinctive feature of these Up-Residual blocks is their proficiency in up-sampling, effectively expanding the spatial dimensions of the data. Conv2DTranspose layers are the key enablers of this up-sampling process, allowing the model to regain any fine-grained spatial information that could have been lost during the first down-sampling steps. Our model can comprehend both low-level and high-level visual information due to the hierarchical feature extraction procedure, which eventually improves the precision and accuracy of image segmentation., we develop a Transformer encoder made up of L levels in the Vision Transformer (VIT), with each layer meticulously designed to capture complex connections in the input data. These layers effortlessly switch between Multi-Head Self-Attention (MSA) and Multi-Layer Perceptron (MLP), two essential building components. Our objective is to transform the feature representation from the Encoder into a format that the Transformer can effectively process, Play Transformer's potent self-attention mechanism as well, which ties global context to long-term dependencies and characteristics retrieved by convolution. RT-ResUNet hybrid structure is the basis for context. Based on this hybrid structure, RT-ResUNet has advanced significantly over earlier segmentation techniques based on Transformer. We anticipate that RT-ResUNet will soon be able to replace manual segmentation techniques, greatly increasing the efficacy of specialized models.

REFERENCES

[1] X. Wang, R. Girshick, A. Gupta and K. He, "Non-local Neural Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7794-7803, doi: 10.1109/CVPR.2018.00813.

[2] O. Ronneberger, P. Fischer, and T.s Brox, "U-net: Convolutional networks for biomedical image segmentation", In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 234–241, Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28, 2015.

[3] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 9992-10002, doi: 10.1109/ICCV48922.2021.00986,2021.

[4] M. Treml, J. Arjona-Medina, Entertainer, et al., “Speeding up semantic segmentation for autonomous driving”,2016.

[5] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou,"Training data-efficient image transformers & distillation through attention”, arXiv preprint arXiv:2012.12877, 2020.

[6] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker and D. Rueckert," Attention gated networks: Learning to leverage salient regions in medical images," Medical image analysis, 53:197–207,2019.

[7] A. Dosovitskiy, L. Beyer, A. Kalashnikov, et al., " An image is worth16x16 words: Transformers for image recognition at scale", arXiv preprint arXiv:2010.11929, 2020.

[8] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara,C. Berger, S. M. Ha, M. Rozycki, et al.,” Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overallsurvival prediction in the brats challenge”. arXiv preprint arXiv:1811.02629,2018.

[9] A. Vaswani, N.Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I.Polosukhin.," Attention is all you need, "In Advances in neural information processing systems, pp. 5998–6008, 2017.

[10] B.Cordonnier, A.Loukas and M. Jaggi," On the relationship between self-attention and convolutional layers, "In ICLR, 2020.

[11] K. Chen et al., "Hybrid Task Cascade for Instance Segmentation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 4969-4978, doi: 10.1109/CVPR.2019.00511.

[12] S. Xie, R. Girshick, P. Dollár, Z. Tu and K. He, "Aggregated Residual Transformations for Deep Neural Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 5987-5995, doi: 10.1109/CVPR.2017.634.

[13] J.Chen, Y.Lu, Q.Yu, X.Luo,E.Adeli,Y.Wang, et al.," Transunet: Transformers make strong encoders for medical image segmentation", arXiv preprint arXiv:2102.04306, 2021.

[14] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang," Swin-unet: Unet-like pure transformer for medical image segmentation", arXiv preprint arXiv:2105.05537,2021.

[15] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. L, "Maxvit: Multi-axis vision transformer", In Proceedings of the European https://doi.org/10.1007/978-3-031-20053-3_27, 2022

[16] W. Wang, E. Xie, X. Li, D-P. Fan, K. Song, D. Liang, T. Lu, P. Luo and L.g Shao,"Pvt v2: Improved baselines with pyramid vision transformer", Computational Visual Media, 8(3):415–424, DOI: 10.1007/s41095-022-0274-8,Springer,2022.

[17] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez and P. Luo," Segformer: Simple and efficient design for semantic segmentation with transformers", Advances in Neural Information Processing Systems, 34:12077–12090, 2021.

[18] W. Wang et al., "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 548-558, doi: 10.1109/ICCV48922.2021.00061.

[19] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, D. Zhang,"Dstransunet: Dual swin transformer u-net for medical image segmentation," IEEE Transactions on Instrumentation and Measurement, 2022.

[20] B. Dong, W. Wang, D. Fan, J. Li, H. Fu and L. Shao," Polyp-pvt: Polyp segmentation with pyramid vision transformers", arXiv preprint arXiv:2108.06932 ,2021.

[21] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh and J. Liang,"Unet++: A nested u-net architecture for medical image segmentation", In Deep learningin Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018.

[22] H. Huang et al., "UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 1055-1059, doi: 10.1109/ICASSP40776.2020.9053405.

[23] A. Lou, S. Guan and M. Loew,"Dc-unet: rethinking the u-net architecture with dual channel efficient Cnn for medical image segmentation", In Medical Imaging 2021: Image Processing, vol. 11596, pp. 758–768, SPIE, 2021.

[24] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen and L. Shao," Pranet: Parallel reverse attention network for polyp segmentation", In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 263–273, https://doi.org/10.1007/978-3-030-01234-2_1 ,Springer, 2020.

[25] S. Woo, J. Park, Joon-Young Lee and I. S. Kweon," Cbam: Convolutional block attention module", In Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19, 2018.

[26] M. M. Rahman and R. Marculescu, "Medical Image Segmentation via Cascaded Attention Decoding," 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2023, pp. 6211-6220, doi: 10.1109/WACV56688.2023.00616.

[27] S. A. Sokolov, T. B. Iliev and I. S. Stoyanov, "Analysis of Cybersecurity Threats in Cloud Applications Using Deep Learning Techniques," 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019, pp. 441-446, doi: 10.23919/MIPRO.2019.8756755.

[28] J. Xiao, L. Yu, L. Xing and A. Yuille,” DualNorm-UNet: Incorporating global and local statistics for robust medical image segmentation”, preprint, arXiv:2103.15858.