Optimized Breast Lesion Segmentation in Ultrasound Videos Across Varied Resource-Scant Environments

Yunhao Li[1], Zibin Chen[1], Junming Yan[1], Ziyu Ding[1], Jie Li[1], Teng Huang[1], Xiaoqing Pei[2], Zheng Zhang[3], Qiong Wang*[4], and Yan Pang*[1]

[1] Institute of Artificial Intelligence, Guangzhou University, China
[2] Cancer Center, Sun Yat-sen University
[3] Moxibustion and Rehabilitation, Guangzhou University of Chinese Medicine
[4] Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

Abstract. Medical video segmentation plays a crucial role in clinical diagnosis and therapeutic procedures by enabling dynamic tracking of breast lesions across frames in ultrasound videos, thereby improving segmentation accuracy. However, the existing methods struggle to strike a balance between segmentation accuracy and inference speed, which impedes their real-time deployment in resource-limited medical environments. To overcome these challenges, we introduce a rapid breast lesion segmentation framework named RbS. RbS employs the Stem module and RbSBlock to enhance representations through intra-frame analysis of ultrasound videos. Moreover, we have developed two versions of RbS: RbS-S boasts enhanced segmentation accuracy, while RbS-L ensures faster inference speeds. Experimental evidence indicates that RbS surpasses current leading models in both segmentation efficiency and prediction accuracy, particularly on resource-limited devices. Our contribution significantly propels the progress of developing efficient medical video segmentation frameworks suitable for various medical platforms.

Keywords: Breast Lesion Segmentation · Resource-limited · Ultrasound Video.

[0] This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFB2704300, NSFC Key Project U23A20391, NSFC General Project 62072452, Regional Joint Fund of Guangdong (Guangdong–Hong Kong–Macao Research Team Project) under Grant 2021B1515130003, Scientific Research Project for Guangzhou University under Grant YJ2023041, Shenzhen High-tech Zone Development Special Plan Innovation Platform Construction Project, the proof of concept center for high precision and high resolution 4D imaging (*Corresponding authors: Qiong Wang (wangqiong@siat.ac.cn) & Yan Pang (yanpang@gzhu.edu.cn.)

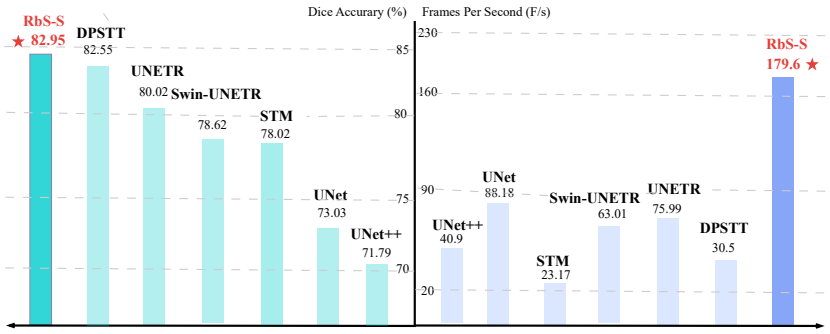

Fig. 1: RbS exhibits remarkable task-level generalization capabilities surpassing other SOTA medical video segmentation approaches on the BUV2022 dataset. With regard to the Dice Coefficient and Inference Speed of video segmentation tasks, we present two models: RbS-S and RbS-L. RbS-S establishes a new benchmark by achieving a Dice Coefficient of 82.95% while maintaining a speed that is six times faster than the runner-up model, DPSTT.

1 Introduction

Over the past few years, the domain of medical imaging has undergone a significant revolution with the introduction of semantic segmentation in Ultrasound Video Analysis (UVA) [2], notably in the field of breast lesion segmentation. These advancements have led to significant improvements in clinical diagnosis and treatment procedures, establishing a new standard in medical practices. Unlike traditional approaches relying solely on single images [4], UVA introduces an additional temporal dimension, enabling the tracking of changes in target areas over time [14]. This enriched temporal data provides an insightful tool for scrutinizing the behavior of breast lesions across consecutive video frames [11]. Furthermore, effectively leveraging this temporal information has the potential to significantly enhance segmentation performance in medical video analysis, thus improving the accuracy of semantic segmentation outcomes.

Despite the advantages, Integrating contemporary temporal data into ultrasound frames presents notable challenges [16]. Exploring an additional temporal dimension often results in heightened complexity and resource demands. Although efforts have been made to reduce computational complexity through temporal independence, current approaches still demand substantial resources [24]. Hence, the primary task is to devise efficient video segmentation algorithms that achieve a balance between minimizing computational expenses and maximizing performance.

This research aims to improve the effectiveness of medical segmentation frameworks designed for resource-constrained medical facilities. Here, we introduce a novel and resource-efficient model called **R**apid **b**reast Lesion **S**egmentation (RbS). RbS integrates temporal and structural information seamlessly through its unique asymmetric encoder-decoder structure [13], significantly improving

the processing of medical video segmentation tasks. We introduce RbSBlock, a novel component designed to efficiently aggregate the extracted representations. Through a contextual integration process, RbSBlock guides representation learning through three constituent components: Integration, Regional Context Aggregator, and Dynamic Semantic Augmentor. Moreover, we introduce the Stem module. The Stem module efficiently manages the coherent and gradual changes in low-level semantic representations extracted from consecutive frames by mapping these representations onto shared vectors and performing dynamic updates in memory. By striking an optimal balance between high accuracy and high inference efficiency, our proposed RbS models achieve remarkable results, excelling in segmentation accuracy and computational efficiency. Fig. 1 demonstrates exceptional generalization abilities, surpassing other advanced medical video segmentation methods in the BUV2022 dataset [15]. We present two versions of the model: RbS-Solid (RbS-S) and RbS-Lite (RbS-L), each excelling in different aspects.RbS-S establishes a new standard with a Dice Coefficient of 82.95%, while RbS-L sets a new benchmark in speed, achieving 216.3 frames per second.

The main contributions highlighted in this paper can be concisely summarized in the following manner:

- We propose RbS, a highly efficient and rapid medical video segmentation method. RbS effectively leverages temporal data to balance segmentation accuracy and efficiency, rendering it suitable for deployment across diverse resource-limited medical platforms.
- We introduce two innovative components, the Stem module, and the RbS-Block.The Stem module efficiently manages temporal information, while the RbSBlock adeptly aggregates extracted representations through a unique contextual integration process. Together, these components enhance the balance between precision and computational efficiency in medical video segmentation tasks.
- To evaluate our model, we compared it with several resource-constrained platforms and multiple benchmarks for medical video segmentation. We introduced two optimization iterations, RbS-S and RbS-L, focusing on specific performance parameters. Results on the BUV2022 dataset confirm that both RbS variants accelerate inference time and significantly improve segmentation accuracy, underscoring the robust capabilities of our RbS model in ultrasound video segmentation.

2 Related Work

Recent advancements have introduced innovative hybrid transformer-based algorithms that integrate transformer and convolutional layer techniques [9, 25]. These algorithms, such as Swin UNETR [7]. and UNETR [8], effectively handle representations derived from high-definition ultrasound images but face computational challenges due to complexity. Additionally, the direct application of these image segmentation methods may inadvertently neglect crucial temporal

context, resulting in temporal inconsistencies. To address this issue, the innovative Space-Time Memory Networks (STM) method [20] utilizes a memory network to extract crucial information from a time-based buffer containing all previous video sequences. Building upon this approach, DPSTT incorporates a memory bank with decoupled transformers to monitor the temporal movement of lesions in ultrasound videos. However, DPSTT requires substantial data augmentation to prevent overfitting and is characterized by slow processing speed, highlighting potential limitations. Thus, the challenge in semantic segmentation for medical video analysis lies in effectively harnessing the abundance of available temporal data.

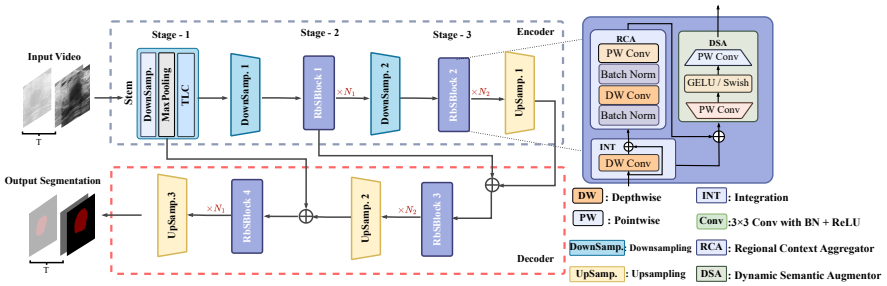

Fig. 2: Overview of the RbS. In addressing medical video segmentation tasks, our methodology adopts an asymmetric encoder-decoder structure to seamlessly integrate structural information. The input data is downsampled through the convolution layer during the initial stage. Following the initial stage, the RbSBlock assumes a critical component in the remaining stages. This lightweight block, acting as a key conduit for effective information transmission, is constructed with RCA, INT, and DSA modules. The specifics of RbSBlock are discussed in Section 3.3.

3 Method

To enhance both video segmentation accuracy and inference speed, we developed an innovative model, RbS, aimed at enhancing video segmentation accuracy and inference speed, as illustrated in Fig. 2. The model is based on an asymmetric encoder-decoder framework, efficiently managing data dimensions. The encoder utilizes convolutional layers for downsampling, while the decoder employs deconvolutional layers for upsampling, with skip connections facilitating integration. The centerpiece of our RbS model is the introduction of two innovative modules: the Stem module and RbSBlock, which are fundamental to its operation. Stem enhances feature extraction efficiency by leveraging inter-frame information, while the RbSBlock is designed to improve the effectiveness of model sampling operations, thereby furthering feature extraction. Further details about the RbS, Stem, and RbSBlock are provided in sections 3.1, 3.2, and 3.3.

3.1 Framework

Our RbS framework draws inspiration from the UNet-based framework [23]. It employs an asymmetric U-shaped structure comprising an encoder and a decoder, depicted in Fig. 2. The input image X has dimensions $W \times H \times C$. Each encoder stage i ($i \in 1, 2, 3$) downsamples the features to dimensions $\frac{W}{2^{i+1}} \times \frac{H}{2^{i+1}} \times C_i$. Conversely, each decoder stage upsamples the feature map to its corresponding dimension. Finally, segmentation probabilities are generated by applying a convolutional layer and an appropriate activation function to the features extracted from the final decoder step.

3.2 Inter-Frame Information Transmission: Stem Moduel

To address the need for efficient feature extraction and interframe information transfer, our proposed RbS model integrates the Stem module. This module comprises three essential components: downsampling, max-pooling, and Time Locality-driven Caching (TLC). The downsampling process involves applying a stride-2 convolutional layer to the input data, reducing its dimensions to yield a semantic feature map of size $\frac{W}{2} \times \frac{H}{2} \times C_i$. Subsequently, the MaxPooling module further downsamples the feature map using a stride-2 MaxPooling layer, resulting in a feature map of size $\frac{W}{4} \times \frac{H}{4} \times C_i$, and adjusting the dimension of the final stage feature maps in the decoder. Guided by the principle of temporal locality, TLC anticipates future memory requirements and prefetches shallow semantics from future-neighbor frames, dynamically updating cache memory with breast lesion features. This reduces latencies in pattern retrieval and storage requirements, thus accelerating overall processing speed. Simultaneously, TLC manages the cache, clearing it as new data enters the Stem module, ensuring temporal consistency and optimizing memory resource allocation. This adherence to temporal consistency is crucial for avoiding confusion in shallow semantic representation, thereby bolstering reliability in subsequent data processing stages.

3.3 Intra-Frame Representation Capture: RbS Block

In the context of video image data, shallow feature extraction can effectively utilize temporal information, while deep semantic feature extraction faces challenges that directly impact the overall segmentation performance of the model. Traditional sampling operations are insufficient for effectively extracting features from deep semantic feature maps, resulting in a heavy reliance on transformer-based implementations [12]. Although these implementations achieve efficient feature extraction, they also incur high computational costs. In order to address this issue, we introduce RbSBlock, a lightweight feature extraction module designed for RbS. RbSBlock consists of three components: Integration, Regional Context Aggregator, and Dynamic Semantic Augmentor. This module facilitates efficient feature extraction, overcoming the limitations of traditional sampling methods and offering an effective solution for deep semantic feature extraction.

Integration Given a video frame image volume X with dimensions $W \times H \times C$, where W, H, and C denote the dimensions, including width, height, and number of channels of the initial resolution, respectively, the RbSBlock module accepts this volume as input. In the Integration module, the input volume is partitioned into N patches determined by $N = HW/P^2$, with P denoting the patch size of each token. This partitioning process utilizes a depthwise convolutional layer [10] with a kernel size of 3 and a padding size of 1. Unlike conventional Transformer-based approaches, our method converts these fixed-size patches into a sequence of linear embeddings. Furthermore, we utilize the Regional Context Aggregator (RCA) to capture local representations through the incorporation of image-specific inductive biases.

Regional Context Aggregator We present the Regional Context Aggregator (RCA) module for efficient aggregation of local features in video segmentation tasks. RCA organizes patches into non-overlapping windows of size $W_s \times W_s$, aggregating information from neighboring patches to enhance local context understanding. This is accomplished through a combination of depthwise and pointwise convolutional layers [18], preceded by batch normalization for stability and improved performance. The depthwise convolutional layer performs convolutions on each channel, while the pointwise convolutional layer merges features while maintaining spatial dimensions. Both layers employ a 3x3 filter with a padding of 1 to maintain spatial resolution. Furthermore, a residual connection enhances information flow within RCA, enabling real-time processing and enhancing overall efficiency. This approach effectively balances computational complexity and segmentation performance, improving regional context comprehension and feature integration in video applications.

Dynamic Semantic Augmentor In the dual-layer MLP of the DSA module, the first layer employs a pointwise convolution with an expanded kernel depth from C to $4C$, enhancing the network's capability to capture diverse hierarchical features crucial for medical image segmentation. Subsequently, the output undergoes processing by another PW convolution layer and is then restored to its original size C. This strategic adjustment aims to balance feature richness with computational efficiency. computational efficiency. To accommodate different semantic features at various levels, our method customizes activation functions for different layers. Shallow-level semantics, focusing on low-level features such as texture and edges of breast lesions, are handled with the GELU activation function. Its smooth nonlinearity effectively captures these features, facilitating improved learning and conveying intricacies. For deep-level semantics, involving complex patterns and semantic information, Swish activation is chosen for its stronger nonlinearity. By employing appropriate activation functions such as GELU and Swish for different semantic levels, the network optimizes its ability to effectively learn and represent both shallow and deep-level semantics. Consequently, this leads to an enhancement in the network's expressive capability,

generalization ability, and overall effectiveness and accuracy of medical video segmentation tasks.

3.4 Loss Funcition

To accomplish the task of ultrasound video segmentation, we employ the soft dice loss function and the BCEWithLogitsLoss to ensure numerical stability. Additionally, we utilize the focal loss technique to address potential data imbalance issues. These calculations can be represented using Equations 1, 2, and 3.

$$L_{dice}(G,Y) = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I} G_{i,j}Y_{i,j}}{\sum_{i=1}^{I} G_{i,j} + \sum_{i=1}^{1} Y_{i,j}} \tag{1}$$

$$\begin{aligned} L_{BCE}(G,Y) = -\sum_{j=1}^{J}\sum_{i=1}^{I}(G_{i,j}log(\frac{1}{1-e^{Y_{i,j}}})) \\ +(1-G_{i,j})log(1-\frac{1}{1-e^{Y_{i,j}}})) \end{aligned} \tag{2}$$

$$\begin{aligned} L_{focal}(G,Y) = -\sum_{j=1}^{J}\sum_{i=1}^{I}(Y_{i,j}G_{i,j}^{2}log(1-G_{i,j}) \\ +(1-Y_{i,j})(1-G_{i,j})^{2}log(G_{i,j})) \end{aligned} \tag{3}$$

where I and J represent the cumulative number of data samples and categories, respectively. Correspondingly, $Y_{i,j}$ and $G_{i,j}$ stand for the predicted probability and the actual label of class j for the i-th data point, respectively. In a more official sense, our goal is to reduce the comprehensive loss function by educating the combined loss function that integrates these three elements, as depicted in the equation 4.

$$L_{tot} = \lambda_1 L_{focal} + \lambda_2 L_{dice} + \lambda_3 L_{BCE} \tag{4}$$

where the optimal values of $\lambda_1 = \lambda_2 = \lambda_3 = 1$ is selected using a hyper-parameter grid search optimization.

4 Results and discussion

4.1 Dataset

The BUV2022 dataset [15] is a meticulously curated collection of data specifically designed for the purpose of breast lesion segmentation in ultrasound video sequences. It consists of 63 distinct video sequences, each representing a different patient and contributing a total of 4619 frames. These frames have been meticulously annotated at the pixel level by domain experts, guaranteeing highly precise ground truth labels. The dataset encompasses videos originating from various ultrasound machines, resulting in diverse spatial resolutions ranging from 580x600 to 600x800 pixels. This dataset serves as a valuable resource for researchers and practitioners in the field, enabling the development of advanced algorithms and techniques for breast lesion segmentation in ultrasound imaging.

4.2 Evaluation Metrics

In this study, we use four important metrics to evaluate the performance and efficiency of models:**1**. *Dice Similarity Coefficient (Dice)* [3] is widely used in evaluating segmentation performance, quantifying the similarity between predicted and ground truth data. **2**. *Jaccard similarity coefficient (Jaccard)* [3] is a statistical measure used to quantify the similarity or dissimilarity between two finite sets of samples. **3**. *Precision* [22] serves as a crucial metric for evaluating the segmentation quality of the model, as it measures the accuracy of positive sample predictions. **4**. *Recall* assesses the accuracy of identifying pixels belonging to the true targets [22].

Table 1: Quantitative comparison with divergent approaches on the BUV2022 and execution speed comparison on the RTX3090 device. In this table, the unit of the metric (Dice, Jaccard, and Recall), which symbolizes the model segmentation performance are expressed in percentage % (the highest values are marked in **red**), while the unit of the metric FPS, which symbolizes the execution speed, is conveyed in terms of images per second (the highest values are marked in **green**). Our performance metrics are marked in gray.

Methods	Dice	Jaccard	precision	Recall	FPS
UNet [23]	73.03	62.47	79.46	72.72	88.18
AttUNet [21]	62.10	48.19	73.90	70.74	113.73
UNet++ [27]	71.79	61.24	82.80	68.84	40.9
TransUNet [5]	65.47	53.58	71.67	66.82	65.1
SETR [26]	66.49	54.80	75.33	66.43	21.61
STM [20]	78.02	68.58	82.01	79.10	23.17
AFB-URR [17]	80.18	70.34	80.08	85.91	11.84
ViViT [1]	67.39	54.46	75.54	66.83	24.33
UNETR [8]	74.44	62.82	84.97	66.06	75.99
Swin UNETR [6]	77.28	66.47	**85.87**	70.16	63.01
DPSTT	82.55	73.64	83.89	84.55	30.5
RbS-L	82.29	73.33	83.03	86.25	**216.3**
RbS-S	**82.95**	**74.03**	85.52	**87.01**	179.6

4.3 Segmentation Performance Evaluations

We compare our model's accuracy and speed with eleven existing models on an NVIDIA RTX 3090, as summarized in Table 1. Transformer-based methods typically exhibit superior video segmentation accuracy due to their intricate global attention mechanism computations. However, they often operate at slower speeds compared to CNN-based approaches. For instance, DPSTT, a transformer-based model, achieved the highest Dice coefficient among video segmentation models at 82.55%, outperforming the best CNN-based method, UNet, by 9.52%. Nonetheless, the fastest transformer-based method, UNETR, only achieves an inference

speed of 75.99 FPS, which is merely 26.82% of the speed attained by the leading CNN-based approach, AttUNet. Hence, achieving a balance between segmentation accuracy and inference speed continues to pose a significant challenge in medical video analysis.

The Solid Version, RbS-S, as shown in Table 1, is our proposed solution to this challenge. The RbS-S outperforms DPSTT in video segmentation accuracy by 0.40% in the Dice coefficient, 0.39% in the Jaccard index, and 2.46% in Recall. In addition, our precision is only slightly lower by 0.35% compared to the highest 85.87%. Furthermore, it boasts an inference speed of 179.6 FPS, which is six times that of DPSTT. Even when compared to the fastest model, AttUNet, our RbS-S displays a 1.58 times faster performance, while significantly outperforming it in accuracy metrics, including improvements of 20.85%, 25.84%, 11.62%, and 16.27% in Dice coefficient, Jaccard index, precision, and Recall, respectively.

Table 2: Compared to other methods on both desktop GPU and mobile phone. For easy description, We footnote the RTX 3060 on the desktop with ‡. Both for FPS and milliseconds per image, we indicate the associated performance fame rate in **green** and fame duration in **red**. The performance of our model is marked in gray.

Model	Desktop GPUs						Mobile phone			
	GTX 1660		RTX 3060‡		Telsa V100s		Apple A15 Bionic		Apple A14 Bionic	
	FPS↑	ms↓	FPS↑	ms↓	FPS↑	ms↓	FPS↑	ms↓	FPS↑	ms↓
AttUNet	26.9	37.1	30.9	32.3	104.1	9.6	57.1	17.5	45.8	21.8
UNETR	46.9	21.3	78.7	12.7	89.2	11.2	56.7	17.6	48.5	20.6
Swin-UNETR	70.4	14.2	68.4	14.6	71.4	14.0	4.4	223.6	3.1	319.3
RbS-S	67.5	14.8	156.0	7.1	158.2	6.7	48.2	20.03	41.3	24.2
RbS-L	**80.5**	**12.9**	**162.5**	**6.3**	**187.9**	**5.9**	**58.6**	**17.1**	**48.9**	**20.3**

Moreover, in our relentless pursuit to enhance processing speed, we developed a faster version called RbS-L. The speed of RbS-L is increased to 216.3 FPS, which is 7.09 times faster than DPSTT, with only a minor decline in accuracy by 0.26% in the Dice coefficient, 0.86% in the precision, and 0.31% in the Jaccard index. Notably, the Recall metric actually improved by 1.70%.

4.4 Speed Evaluation across multiple source-limited Platforms

To comprehensively assess our medical video image segmentation model's adaptability and versatility, we conducted speed performance evaluations across five hardware configurations. These configurations include desktops and mobile platforms, representing diverse real-world scenarios. The results of these performance evaluations are summarized in Table 2, demonstrating the efficiency and adaptability of our model across various environments. Our model consistently performs well across various platform specifications and hardware constraints, showcasing its versatility.

Speed Evaluation on Desktop GPUs We comprehensively evaluated our model's inference speed by testing it on various setups with different computational capabilities, including desktop GPUs such as the GTX 1660 SUPER, RTX

3060 12G, and Telsa V100s 32G. The GTX 1660 SUPER, equipped with 6GB of memory and 1408 CUDA cores, offers modest computational performance, making it suitable for less demanding tasks. Conversely, the RTX 3060, with 12GB of memory and 3584 CUDA cores, provides substantial computational power, making it suitable for larger deep-learning tasks. Finally, the Telsa V100s, with 32GB of memory and 5120 CUDA cores, offers exceptional computational performance, making it ideal for demanding medical video segmentation tasks. We performed a comparative analysis of our model's inference speed against AttUNet, UNETR, and Swin UNETR on identical hardware, as summarized in Table 2. The results demonstrate that RbS-L consistently outperforms the competition on all tested devices. Additionally, RbS-S maintains a speed advantage over the others, except for RbS-L. For instance, on the RTX 3060 12G, RbS-S achieves 156.0 FPS, nearly doubling the speed of UNETR at 78.7 FPS. On the Telsa V100s 32G, RbS-S even surpasses AttUNet by 54.1 FPS. Moreover, RbS-L outpaces RbS-S's inference speed on all devices, notably by 29.7 FPS on the Telsa V100s. This improvement can be attributed to the efficient design of RbSBlock, which effectively reduces the computational load while maintaining high precision.

Speed Evaluation on Mobile Chips Mobile platforms present a unique environment for the implementation of medical video and image segmentation, given their ubiquitous use and distinct performance characteristics compared to desktop systems. For our evaluation, we elected to utilize the Apple A15 Bionic and Apple A14 Bionic as our testing platforms. We employ the Tera Operations Per Second (Tops) metric to assess the computational prowess of these processors, with the Apple A15 Bionic and Apple A14 Bionic exhibiting Tops values of 15.8 and 11, respectively.

We benchmarked the inference speed of our RbS model against competitors on these two mobile devices. The results in Table 4 highlight RbS-L's superior speed on both mobile platforms. For instance, on the Apple A15 Bionic, RbS-L's inference time outperforms AttUNet and UNETR by 0.4 ms and 0.5 ms, respectively, for single-image inference. This advantage extends to the Apple A14 Bionic, where RbS-L's inference times surpass AttUNet and UNETR by 1.5 ms and 0.3 ms. Particularly when compared with Swin-UNETR, RbS-L's speed advantage becomes more pronounced.RbS-L's inference times were 17.1 ms and 20.3 ms on the Apple A15 Bionic and Apple A14 Bionic, respectively. In contrast, Swin UNETR's inference times are substantially longer, registering 223.6 ms and 319.3 ms on the respective devices. This shows that RbS-L is 13.08 and 15.73 times faster than Swin UNETR on these platforms.

Summary of Speed Evaluation Our experiments demonstrate the impressive speed performance of our proposed models, RbS-L and RbS-S, across various platforms. The RbS-L model is characterized by its high speed, making it highly effective for swift and efficient segmentation tasks. Conversely, although RbS-S operates at a marginally slower speed compared to RbS-L, it is particularly notable for its exceptional accuracy, a feature of paramount importance in the realm of medical imaging. We recommend RbS-S for tasks prioritizing ac-

curacy over speed. By ensuring both speed and accuracy, our suggested models provide versatile solutions for diverse computing environments.

4.5 Ablation Study

During our ablation study, we delve into the intricate effects of block configuration, block depth, and stage count on the performance outcomes of the RbS model. These insightful findings guide us in determining the optimal configurations for the RbS model, allowing us to strike a careful balance between speed and accuracy tailored to the specific demands of medical image segmentation tasks.

Table 3: Quantitative comparison with different Block Settings of RbS. Employing the RbS-L as a representative, we adjusted the settings solely in RbSBlock and evaluated the performance by comparing it with the BUV2022 dataset. In this table, "INT" represents Integration while "RCA" stands for Regional Context Aggregator, also we footnote the FPS on the Tesla V100 with ⊙ and the FPS on the RTX 3090 with ⊖. The modules definition: ①: Backbone ②: Integration(INT) ③: Regional Context Aggregator(RCA) ④: Dynamic Semantic Augmentor(DSA); Both for segmentation and FPS, we indicate the associated performance loss in **green** and the improvement in **red**. The performance metrics of the S version are marked in gray.

Modules	Dice	Jaccard	F1	Precision	Recall	$FPS_{⊙}$ (V100)	$FPS_{⊖}$ (3090)
①	64.28 -18.01	50.24 -23.09	67.55 -15.42	57.34 -25.69	84.77 -1.48	1165.7 +976.3	1233.5 +1017.2
①+②	70.22 -12.07	56.80 -16.53	72.88 -10.09	63.56 -19.47	85.18 -1.07	867.3 +677.9	900.7 +684.4
①+②+③	79.59 -2.70	70.58 -2.75	78.89 -4.08	75.77 -7.26	85.16 -1.09	370.5 +181.1	429.8 +213.5
①+②+③+④	**82.29**	**73.33**	**82.97**	**83.03**	**86.25**	**189.4**	**216.3**

Block Setting This section assesses the effectiveness of the RbSBlock in image feature extraction, with RbS-L serving as a representative example. As shown in Table 3, experimental results demonstrate that RbS-L achieves FPS rates of 1165.7 and 1233.5 on the V100 GPU and 3090 GPU, respectively, but with only a Dice coefficient of 64.28%. However, after integrating the INT module into our model, we observed a 5.94% improvement in the Dice coefficient. This improvement is attributed to the initial feature extraction facilitated by the 1x1 convolution inherent in the INT module. Furthermore, incorporating the RCA module which utilizes 3x3 depth convolution for global feature extraction, enhances the Dice coefficient by 15.31%. To further augment feature extraction efficiency, we introduce a class-linear convolution DSA on top of the INT and RCA modules, which utilizes distinct activation functions for shallow and deep features, achieving a maximum accuracy of 82.29%. It is noteworthy that at this stage, the FPS on the V100 and 3090 GPUs stands at 189.4 and 216.3, respectively, notably surpassing the conventional FPS of 90 demonstrated by typical ultrasound visual sensor devices [19].

Block Depth We analyze how the depth of RbSBlocks within a stage influences the effectiveness of the RbS model in medical image segmentation. This

analysis is visualized in Fig. 2, where the encoder and decoder incorporate varying quantities of RbSBlocks (labeled N_1 and N_2). Table 4 presents the evaluation results across five unique combinations of N_1 and N_2.When these values are designated as 2 and 3, the resulting Dice score is 82.29%, the Jaccard index is 73.33%, the F1 score is 82.97%, Precision is at 83.03%, and Recall is 86.25%. FPS on V100s and 3090 desktop GPUs are 189.4 and 216.3, respectively. Interestingly, reducing N_1 and N_2 results in a shallower model, leading to increased FPS but also leading to a decline in various video segmentation metrics. For example, with N_1 and N_2 set at 1 and 2, the model's inference speed improves by 32.05 FPS and 68.89 FPS on Tesla V100s and 3090, respectively. However, this improvement is countered by reductions in Dice, Jaccard, F1, and Precision scores by 1.23%, 3.10%, 2.82%, 6.93%, and 0.54%, respectively. In contrast, increasing both N_1 and N_2 values, such as setting them to 3, results in a decrease in FPS (by 34.10 and 36.70 on the Tesla V100s and 3090, respectively), but an enhancement in video segmentation accuracy is observed, with the Dice score peaking at 82.95%. However, further increments in N_1 and N_2 (to 3 and 4, respectively) lead to a decline in all model metrics, accompanied by a significant drop in inference speed. Considering these factors, we conclude that the third and fourth configurations highlighted in Table 4 (RbS-Lite and RbS-Solid, respectively) are the most suitable as the final versions of our model.

Table 4: Quantitative comparison of different Block Depth settings of RbS. We assess the reasonableness of the allocation strategy adopted by RbSBlock in the shallow and deep layers of RbS by varying the Block Depth setting and monitoring the consequent fluctuations in both its task segregation performance and execution speed. Based on the observation results, we divide two model versions Lite(L) and Solid(S) of RbS by applying different Block Depth settings. These two versions of RbS are marked in gray, while the other indicator colors applied in this table are the same as in Table 3.

N_1, N_2	Dice	Jaccard	F1	Precision	Recall	FPS(V100)	FPS(3090)
1, 2	81.06 -1.23	70.23 -3.10	80.15 -2.82	76.10 -6.93	85.71 -0.54	221.45 +32.05	285.19 +68.89
2, 2	81.76 -0.53	71.52 -1.81	81.72 -1.25	76.63 -6.40	85.98 -0.27	206.61 +17.21	240.58 +24.28
L: 2, 3	**82.29**	**73.33**	**82.97**	**83.03**	**86.25**	**189.4**	**216.3**
S: 3, 3	**82.95** +0.66	**74.03** +0.70	**84.03** +1.06	**85.52** +2.49	**87.01** +0.76	**155.3** -34.10	**179.6** -36.70
3, 4	81.38 -0.91	72.65 -0.68	82.68 -0.29	80.45 -2.58	84.57 -1.68	145.83 -43.57	175.26 -41.04

Stage Count Upon defining the architecture and depth of the RbSBlock, our attention shifts to investigating the impact of stage count within the RbS model on medical image segmentation tasks. We observe notable outcomes when the stage count is set to two. As detailed in Table 5, the RbS model achieves high inference speeds, recording 472.94 FPS on the Tesla V100 and 532.86 FPS on the 3090 GPU. However, this acceleration comes at the expense of segmentation performance, with Dice, Jaccard, Precision, and Recall scores measured at 78.57%, 67.20%, 76.37%, and 78.94%, respectively. It becomes clear that while reducing the number of stages enhances inference speed, it simultaneously leads to insufficient feature learning, which in turn compromises segmentation accu-

racy in medical video applications. When we increase the stage count to three, the RbS model experiences a reduction in inference throughput, dropping to 189.4 FPS on the Tesla V100 and 216.3 FPS on the 3090 GPU. Nevertheless, this decrease in speed is offset by significant improvements in segmentation performance, with Dice, Jaccard, and Precision scores rising by 3.72%, 6.13%, and 6.66%, respectively. Interestingly, when the stage count is further increased to four, we observe a reversal of this trend. Dice, Jaccard, Precision, and Recall scores drop by 2.42%, 3.21%, 3.75%, 7.45%, and 0.57%, respectively. This decline can be attributed to the over-extraction of deep semantic features, which becomes particularly problematic when dealing with small and variable medical video datasets. Taking these observations into account, we conclude that the optimal configuration for the RbS model is a stage count of three. This setup strikes a balanced compromise between inference speed and segmentation accuracy, making it the most effective configuration for achieving robust performance in medical video segmentation.

Table 5: Quantitative comparison with different Stage Count settings of RbS. Through a comparative analysis of the performance and execution speeds associated with different settings of Satge Count employed in task segmentation, we determine the optimal Stage settings of RbS is 3.

Stage Count	Dice	Jaccard	Precision	Recall	FPS(V100)	FPS(3090)
2	78.57 -3.72	67.20 -6.13	76.37 -6.66	78.94 -7.31	472.94 +283.54	532.86 +316.56
3	**82.29**	**73.33**	**83.03**	**86.25**	**189.4**	**216.3**
4	79.87 -2.42	70.12 -3.21	75.58 -7.45	85.68 -0.57	219.54 +30.14	225.82 +9.52

4.6 Visualization

Fig. 3 exhibits comparative results from diverse medical video segmentation models applied to the BUV2022 dataset.

The figure begins with Ground Truth instances in the first row, representing accurately annotated data. In the row below, segmentation outcomes from our RbS-S model are depicted. Notably, the RbS-S model yields finer segmentation in areas that generally pose significant challenges, particularly along the edge regions of the frames. Contrasting with results from alternative SOTA models like UNETR, Swin UNETR, and AttUNet in the subsequent rows, RbS-S consistently outperforms them in segmentation quality and accuracy, particularly in challenging regions.

This collective visualization encapsulates the superior performance of our RbS-S model. Specifically, it emphasizes its enhanced capability in achieving detailed boundary segmentation, setting it apart in accomplishing precise semantic segmentation in medical video image segmentation tasks. Thus, the figure effectively demonstrates the superior competence of our RbS model in handling

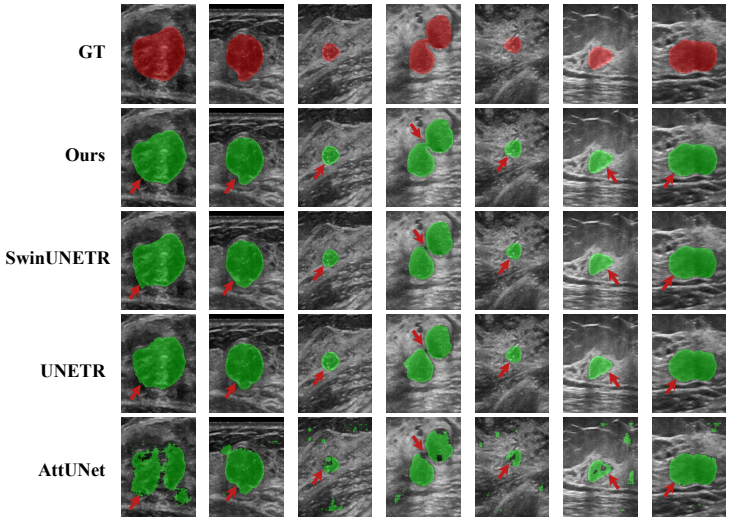

Fig. 3: Comparative display of segmentation results on the BUV2022 dataset. The topmost row is the ground truth. The rows beneath, from second to last, individually exhibit the segmentation outcomes of RbS-S, Swin UNETR, UNETR, and AttUNet. GT: Ground Truth.

complex medical video segmentation tasks, exhibiting promising potential for future applications in this domain.

5 Conclusion

This paper introduces RbS, a novel methodology aimed at streamlining medical video segmentation tasks, particularly focusing on the rapid and accurate segmentation of breast lesions in ultrasound videos. By adeptly integrating an asymmetric encoder-decoder structure with the Stem module and RbSBlock, RbS facilitates the swift and precise extraction of semantic features from video frames, a necessity for high-speed inference in resource-limited environments. Experimental results validate RbS's effectiveness in achieving an exceptional balance between accuracy and speed across multiple platforms with limited resources. In conclusion, RbS represents a significant advancement in the field of medical video image segmentation, establishing a robust foundation for future progress in this crucial domain.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)

2. Asgari Taghanaki, S., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G.: Deep semantic segmentation of natural and medical images: a review. Artificial Intelligence Review **54**, 137–178 (2021)
3. Chang, H.H., Zhuang, A.H., Valentino, D.J., Chu, W.C.: Performance measure characterization for evaluating neuroimage segmentation algorithms. Neuroimage **47**(1), 122–135 (2009)
4. Chen, G.P., Zhao, Y., Dai, Y., Zhang, J.X., Yin, X.T., Cui, L., Qian, J.: Asymmetric u-shaped network with hybrid attention mechanism for kidney ultrasound images segmentation. Expert Systems with Applications **212**, 118847 (2023)
5. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
6. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. pp. 272–284. Springer (2022)
8. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
9. He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., Shen, D.: Transformers in medical image analysis: A review. Intelligent Medicine (2022)
10. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
11. Huang, R., Lin, M., Dou, H., Lin, Z., Ying, Q., Jia, X., Xu, W., Mei, Z., Yang, X., Dong, Y., et al.: Boundary-rendering network for breast lesion segmentation in ultrasound images. Medical Image Analysis **80**, 102478 (2022)
12. Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: Missformer: An effective transformer for 2d medical image segmentation. IEEE Transactions on Medical Imaging (2022)
13. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
14. Lei, T., Zhang, D., Du, X., Wang, X., Wan, Y., Nandi, A.K.: Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network. IEEE Transactions on Medical Imaging (2022)
15. Li, J., Zheng, Q., Li, M., Liu, P., Wang, Q., Sun, L., Zhu, L.: Rethinking breast lesion segmentation in ultrasound: A new video dataset and a baseline network. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV. pp. 391–400. Springer (2022)
16. Li, L., Hu, Z., Huang, Y., Zhu, W., Wang, Y., Chen, M., Yu, J.: Automatic multiplaque tracking and segmentation in ultrasonic videos. Medical Image Analysis **74**, 102201 (2021)
17. Liang, Y., Li, X., Jafari, N., Chen, J.: Video object segmentation with adaptive feature bank and uncertain-region refinement. Advances in Neural Information Processing Systems **33**, 3430–3441 (2020)
18. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)

19. Mejia-Trujillo, J.D., Castaño-Pino, Y.J., Navarro, A., Arango-Paredes, J.D., Rincón, D., Valderrama, J., Muñoz, B., Orozco, J.L.: Kinect™ and intel realsense™ d435 comparison: a preliminary study for motion analysis. In: 2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom). pp. 1–4 (2019). https://doi.org/10.1109/HealthCom46333.2019.9009433
20. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9226–9235 (2019)
21. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
22. Powers, D.M.W.: Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation (2020)
23. Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241. Springer (2015)
24. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K.: Medical image segmentation using deep learning: A survey. IET Image Processing **16**(5), 1243–1267 (2022)
25. Yang, H., Yang, D.: Cswin-pnet: A cnn-swin transformer combined pyramid network for breast lesion segmentation in ultrasound images. Expert Systems with Applications **213**, 119024 (2023)
26. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6881–6890 (2021)
27. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: 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. pp. 3–11. Springer (2018)