Enhancing Medical Image Segmentation with Self-Attention-Embedded LSTM Mechanism

Zexuan Ji, *Member, IEEE*, Shunlong Ye, and Xiao Ma, *Student Member, IEEE*

Abstract*—Vision-based measurement technologies in medical imaging have greatly enhanced automated diagnostics. However, fine-grained medical image segmentation remains challenging due to the complex structures and the need for accurate localization of anatomical regions. In this paper, we propose a novel deep learning model named LSTMSA, which integrates Self-Attention (SA) mechanism within a Long Short-Term Memory (LSTM) module to address two key challenges in medical image segmentation. Firstly, conventional approaches suffer from premature convergence during training, leading to suboptimal segmentation results. LSTMSA achieves a more balanced training process by embedding SA within LSTM, allowing the model to focus on local information early on while fully leveraging LSTM's temporal modeling capabilities throughout training. Secondly, SA mechanism introduces rank collapse issue, potentially resulting in information loss and blurred segmentation in complex medical images. LSTMSA is designed to preserve high-dimensional feature representations, ensuring that details and contextual information are maintained. Experimental results on four medical image datasets and twenty baselines demonstrate the superiority of LSTMSA over existing methods, preserving critical anatomical features and achieving state-of-the-art segmentation accuracy. The code is available at https://github.com/yeshunlong/LSTMSA.

***Index Terms*—medical image segmentation, deep learning, long short-term memory (LSTM), self-attention (SA), rank preservation, feature representation.**

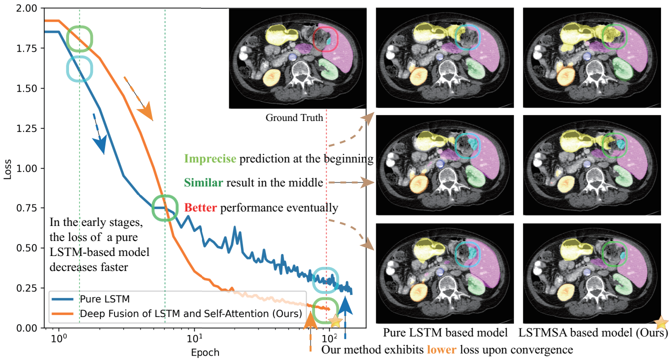


Fig. 1. The comparison of loss curves between LSTM based model and LSTMSA based model (Ours). Accompanying the curves are CT scan images demonstrating the ground truth and the progressive improvement of the deep fusion model's segmentation accuracy, with the final image marked by a star symbolizing the lowest loss point.

I. Introduction

THE rapid advancement of vision-based methods has significantly propelled the progress of computer-aided medical diagnostic technologies [1]. The nature of medical images, characterized by intricate anatomical structures and the need for precise localization [2], [3], requires innovative approaches to achieve accurate and robust segmentation results [4], [5], [6]. Recent years, researchers have been dedicated to enhance model performance through various contextual modules. For example, they have explored the individual implementation of Long Short-Term Memory (LSTM) and Self-Attention (SA) mechanism to enhance the extraction of contextual information. LSTM is adept at capturing temporal and sequential dependencies within the data [7], while SA mechanism excels in modeling global contextual relationships [8]. Each approach independently contributes to the improvement of context-awareness in segmentation models, addressing the critical challenge of capturing long-range dependencies within medical images [9]. These strategies have yielded promising results in medical imaging applications, contributing to more precise segmentation and clinically relevant outcomes [10], [11]. Furthermore, such advanced methodologies enhance the model's ability to interpret complex image sequencesand enables the dynamic adjustment of focus [12], [13], thereby improving the detection of subtle nuances and variations within medical images [14]. This synergy between temporal understanding and attentive contextual analysis significantly boosts the model's diagnostic precision, providing a more detailed and comprehensive understanding of patient-specific conditions [15].

Despite the notable advantages of SA and LSTM mechanism, these modules still encounter certain challenges, especially in clinical applications. Firstly, the conventional LSTM module embedding often results in a rapid initial reduction of loss, which can lead to **premature convergence**. This issue not only impedes the model's ability to retain long-term memory but also limits further accuracy improvements during later training stages [16]. In medical image segmentation, this premature convergence can severely compromise the model's capacity to capture intricate anatomical details over extended sequences, thus undermining segmentation performance. The root of this issue lies in the model's inefficient propagation of past information due to recurrent connections, coupled with a neglect of local details, which ultimately prevents the LSTM from fully realizing its potential. A balanced approach that fosters the synergy between SA and LSTM is essential for sustained learning progress. Fig. 1 illustrates the loss curves of standard LSTM network compared to our hybrid

---

This work was supported by National Science Foundation of China under Grants No. 62072241. *(Corresponding author: Shunlong Ye.)*

Z. Ji, S. Ye and X. Ma are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: jizexuan@njust.edu.cn; yeshunlong@njust.edu.cn; maxiao@njust.edu.cn).

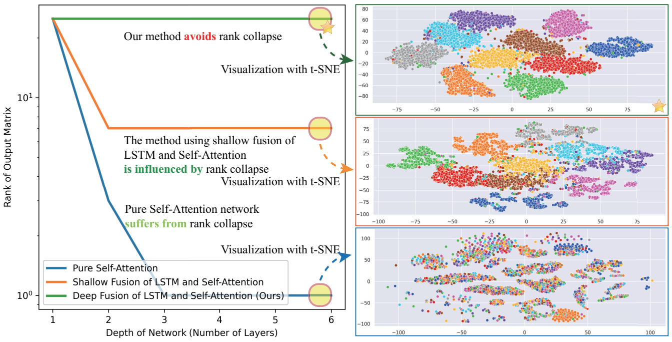


Fig. 2. Comparative analysis of the rank of the output matrix in neural networks of varying depths, contrasting the performance of pure SA, shallow fusion of LSTM and SA, and deep fusion of LSTM and SA (Ours). The three t-SNE visualizations on the right highlight the clustering of features at different layers.

model, where our approach demonstrates a steadier and more uniform reduction in loss, indicative of enhanced learning stability. As training progresses, our method not only achieves a lower loss level in a more stable manner but also effectively integrates LSTM and SA strategies for improved temporal retention and long-range dependency capture. This stability is critical in medical image segmentation, where the ability to accurately segment across varying temporal and spatial contexts directly impacts clinical outcomes. Visual evidence from computed tomography (CT) scan images supports the quantitative improvements, showing enhanced segmentation accuracy that closely mirrors the ground truth over time.

Another significant challenge in medical image segmentation arises from the use of pure SA networks, particularly their susceptibility to a phenomenon where the network converges exponentially (with depth) to a rank-1 matrix, causing all tokens to become indistinguishable [17], a problem we refer to as **rank collapse**. In the context of medical image segmentation, this rank collapse can lead to a loss of critical spatial information, reducing the model's ability to distinguish between different tissue types or pathological features, thereby diminishing segmentation quality. This issue is driven by the SA mechanism functioning as an ensemble of shallow networks, which inherently limits matrix rank. To counteract this, advanced architectural designs are necessary to preserve high-dimensional features. Fig. 2 evaluates rank stability across network depths, demonstrating the effectiveness of our deep fusion method. Our approach consistently maintains a high rank, showcasing resilience against rank collapse, unlike pure SA networks, which exhibit a significant reduction in rank with increasing depth. t-SNE [18] visualizations further validate our model's superior feature representation, leading to smoother loss transitions and enhanced segmentation accuracy by maintaining precise feature delineation.

To address the aforementioned challenges in medical image segmentation, we present a novel deep learning model known as SA-embedded LSTM (LSTMSA). Within the architecture of LSTMSA, we employ a deep embedding of the SA mechanism into the LSTM module, with the overarching goal of achieving a smoother training process and more efficient feature representation. This innovative approach involves integrating the SA mechanism at various time steps within the LSTM module, with the dual objective of maintaining a stronger emphasis on contextual information during the early stages of training, while harnessing the full spatiotemporal modeling capabilities inherent to LSTM throughout the entirety of the training process.

The contributions of this paper can be summarized as follows:

- Deep Fusion of LSTM and SA: Our first innovation lies in the deep integration of the LSTM and SA mechanism within a single module. This integration enhances the model's capability to capture both global and local contextual information effectively, allowing for superior medical image segmentation results.
- Resolving Premature Convergence and Rank Collapse Challenges: The second noteworthy innovation of this work is the resolution of two prominent challenges in contextual models. Firstly, LSTMSA effectively addresses the issue of premature loss convergence, ensuring a more balanced and stable training process. Secondly, it tackles the problem of rank collapse inherent in pure SA networks, preserving high-dimensional feature representations, and averting information loss, particularly in complex medical image structures.
- State-of-the-art (SOTA) Performance on Four Public Datasets and Improvements on Twenty Baselines: Our third innovation manifests in the validation of LSTMSA's performance on diverse medical image datasets. Through extensive experimentation, LSTMSA consistently outperforms existing methodologies, achieving SOTA results in terms of segmentation accuracy and robustness. Moreover, LSTMSA demonstrates significant improvements over twenty baseline models across four widely used medical image datasets. These superior outcomes underscore the practical significance and clinical utility of our proposed model.

II. BACKGROUND

A. *Literature Review*

Separated LSTM and SA for Embedded Contextual Information in Medical Image Segmentation: Recent advancements in medical image segmentation have focused on embedding contextual information into Convolutional Neural Networks (CNNs), particularly within encoder-decoder architectures such as UNet [19]. These models capture both local semantic and global contextual features, improving segmentation of fine-grained structures [5]. Moreover, combining CNNs with RNNs, such as LSTM, has been effective for capturing temporal dependencies in 3D medical images, aiding in the segmentation of complex anatomical structures [6]. The inclusion of SA mechanism has further enriched these networks from the other side, enhancing their ability to capture contextual dependencies and improving performance across diverse medical imaging datasets [20], [21], [22]. Collectively, these advancements have significantly improved the accuracy and reliability of medical image segmentation.

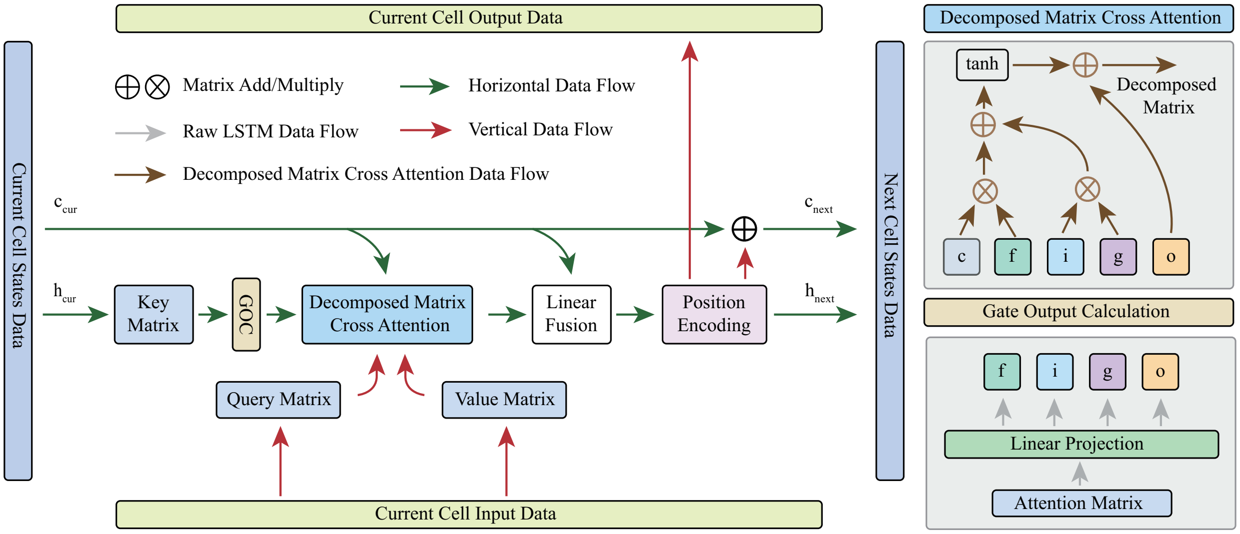


Fig. 3. Overview of LSTMSA module. The vertical data flow represents the processing over time in the input sequence, and the horizontal data flow represents the state updates within each cell.

Hybrid LSTM and SA Mechanism in Medical Image Segmentation: Several studies have explored hybrid approaches that combine LSTM and SA mechanisms for medical image segmentation [23], [24], [25]. Typically, these methods use shallow fusion strategies, stacking the two modules where one's output serves as the other's input [26], [27]. This approach, however, fails to fully leverage the complementary strengths of LSTM and SA. Alternatively, deeper integrations have been proposed, such as SwinLSTM [28], and RWKV [29], achieving more intricate fusions. Nevertheless, these models frequently fail to adequately tackle key challenges mentioned above, namely, premature convergence and rank collapse, and they do not deeply explore the individual contributions of LSTM and SA. Consequently, they may miss opportunities to optimize the interaction between these components, which in turn limits their effectiveness in medical image segmentation.

B. Loss Convergence Analysis

Lemma: *A rapid reduction in loss at the early stages of training suggests an aggressive weight adjustment strategy.*

Intuitive Proof: The initial phase of training a neural network is critical for setting the trajectory towards optimal performance. During this period, a sharp decrease in the loss function, L, signifies that the network's weights, W, are being adjusted significantly in response to the computed gradients. This rapid reduction can be intuitively understood as the network quickly moving towards a more favorable region within the loss landscape.

The crux of the proof involves demonstrating that such an aggressive weight adjustment mechanism, directly contributes to the network's inability to preserve information over long periods, thereby affecting its long-term dependency modeling capabilities [16]. This is because an aggressive decay of weights W inherently biases the network towards recent data points, diminishing its capacity to integrate and remember older information. Thus, a faster initial decrease in L not only signifies efficient early learning but also implies a propensity towards aggressive weight decay, leading to compromised long-term memory within the network. This relationship underscores the intricate balance between rapid learning and the preservation of historical data in the context of neural network training dynamics.

Based on the aforementioned information, we observe a logical chain that encapsulates the dynamics of network training and performance. Rapid convergence is characterized by an initial swift decline in loss, which can be associated with the network weights decaying at an aggressive rate. This aggressive decay of network weights is, in fact, indicative of poorer long-term memory capabilities within the network [16]. These insights collectively underscore a pivotal aspect of neural network behavior, particularly in the context of training efficacy and memory retention capabilities.

C. Rank Collapse Analysis

The limitation of the pure SA mechanism lies in its potential to introduce the issue of ”token uniformity” when using deep network to process input sequences, which can subsequently impact the network's performance and generalization capabilities [17]. Specifically, SA networks tend to assign equal importance to each token when processing input sequences, resulting in an output matrix with a lower rank. This problem is known as the ”token uniformity” problem, and its fundamental cause can be attributed to the decomposition and induction bias inherent in the SA mechanism.

The ”token uniformity” problem can have significant repercussions for sequence processing tasks. Since the network cannot efficiently distinguish the importance of individual tokens in the input sequence, it may result in the model overlooking crucial information, affecting its ability to accurately model the internal structure of the sequence. In tasks such as medical image segmentation, this uniformity may lead to inefficient attention to specific regions within the image, thereby impacting the accuracy of segmenting target structures.

III. METHOD

A. Module Design Overview

As illustrated in Fig. 3, the LSTMSA module is designed to address two critical challenges: premature convergence in LSTM based models and rank collapse in pure SA mechanism. By deeply embedding SA into the LSTM architecture, LSTMSA effectively balances the strengths and mitigates the limitations of both components, enabling effective capture of sequential temporal dependencies typical of LSTM and the intricate internal dependencies facilitated by SA for medical image segmentation. The module's design is guided by the following principles:

Firstly, to counteract rapid convergence in LSTM models, which often leads to the loss of effective long-term memory modeling early in training, LSTMSA integrates the SA matrices calculation into LSTM's basic units. This integration enhances the model's capacity to capture long-range dependencies by dynamically adjusting based on global sequence information, thereby mitigating premature convergence. SA allows the model to maintain a better understanding of long-term dependencies by leveraging global contextual information.

Secondly, LSTMSA addresses the rank collapse issue associated with pure SA networks, where SA tends to produce low-rank weight matrices, potentially compromising performance and generalization. To combat this, LSTMSA introduces a novel combination of query, key, and value matrices, decomposed and integrated into the LSTM unit states. This design enhances the model's ability to capture various aspects of the input sequence, thereby maintaining higher rank in the output matrix and improving the network's representational capacity.

B. Module Design Details

Alg. 1 outlines the calculation process of the LSTMSA module. The decision to use the hidden states as the key matrix is based on the necessity of incorporating contextual information that spans across different time steps (patches or slices in medical image). In medical imaging, such temporal and historical context is crucial for accurately segmenting structures that evolve or appear differently over time. This design is supported by our experimental analysis of time-step dependency IV-D, which demonstrates that the key matrix, derived from LSTM's hidden states, provides the most comprehensive representation of this temporal context. On the other hand, the input data, while serving as query and value matrices, allow the model to integrate immediate information with historical context during decision-making. This balance between using historical context (as keys) and immediate input data (as queries and values) is further validated by our ablation study IV-E1, which show that this configuration optimally enhances the model's performance in medical image segmentation.

```
Algorithm 1 LSTMSA Calculation Process
Require: Current cell input $x_{cur}$, current hidden state $h_{cur}$, weight matrices $W_i, W_f, W_o, W_g$, bias vectors $b_i, b_f, b_o, b_g$, sequence length $l$, model dimensionality $d_{\text{model}}$
Ensure: Next cell state $c_{next}$, next hidden state $h_{next}$
Step 1: Gate Output Calculation
    $xh \leftarrow [x_{cur}; h_{cur}]$
    $i \leftarrow \sigma(W_i \odot xh + b_i)$
    $f \leftarrow \sigma(W_f \odot xh + b_f)$
    $o \leftarrow \sigma(W_o \odot xh + b_o)$
    $g \leftarrow \tanh(W_g \odot xh + b_g)$
Step 2: Decomposed Matrix Attention
    $D \leftarrow o + \tanh(c \cdot f + i \cdot g)$
Step 3: Cross Attention
    $CA \leftarrow \text{Softmax}\left(\frac{D \cdot h_{cur}^T}{\sqrt{d_{h_{cur}}}}\right) \cdot D$
Step 4: Feature Fusion
    $x \leftarrow [c_{cur}; CA]$
    $F \leftarrow \text{ReLU}(W(x))$
Step 5: Dynamic Positional Encoding
    $P_{\text{dyn}} \leftarrow \text{Variable}(\text{shape} = [l, d_{\text{model}}], \text{trainable} = True)$
    $PF_{\text{dyn}} \leftarrow F + P_{\text{dyn}}$
Step 6: Computing the Next Cell State
    $c_{next} \leftarrow c_{cur} + PF_{\text{dyn}}$
    $h_{next} \leftarrow PF_{\text{dyn}}$
return $c_{next}, h_{next}$
```

By decomposing and reintegrating query, key, and value matrices within the LSTM framework, the model explores input data nuances more effectively, significantly boosting its ability to represent complex dependencies. The cross attention mechanism excels at uncovering dependencies across input segments, leading to a more granular understanding of sequence relationships. The deep fusion of LSTM states with cross attention outcomes preserves data integrity, effectively addressing rank collapse issues common in pure SA networks. Additionally, the incorporation of dynamic positional encoding differentiates sequence element positions, enhancing the model's understanding of order and structure within sequences, making LSTMSA a robust solution for complex, sequential data tasks.

IV. EXPERIMENTS AND RESULTS

A. Implementation Details

For the assessment of our proposed modules on 2D inputs, we conduct experiments on the following datasets:

- Synapse dataset: This dataset encompasses various organ segmentation tasks[1].
- ISIC2018 dataset: Focusing on skin lesion segmentation, this dataset has been employed in previous studies by Codella *et al.* [30] and Tschandl *et al.* [31].

In scenarios involving 3D inputs, we perform experiments on the following datasets:

- ACDC dataset: Dataset for heart segmentation[2].
- CVC-ClinicDB dataset: Addressing polyp segmentation in colonoscopy videos, this dataset has been used for benchmarking segmentation methods in the work by Bernal *et al.* [32].

---

[1] https://www.synapse.org/#!Synapse:syn3193805/wiki/217789

[2] https://www.creatis.insa-lyon.fr/Challenge/acdc/

TABLE I
QUANTITATIVE RESULTS ON SYNAPSE MULTI-ORGAN DATASET AND ISIC2018 SEGMENTATION DATASET. DICE SCORES (%), HD95 ARE REPORTED. THE BEST RESULTS ARE BOLDED. THE SECOND BEST RESULTS ARE UNDERLINED. ↑ DENOTES HIGHER VALUE INDICATING BETTER PERFORMANCE, ↓ DENOTES LOWER VALUE INDICATING BETTER PERFORMANCE. BASELINES ARE STARRED.

Synapse											ISIC2018		
Method	DICE ↑	HD95 ↓	Aorta	GB	KL	KR	Liver	PC	SP	SM	Method	DICE ↑	HD95 ↓
UNet [19]	70.11	44.69	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96	UNet [19]	87.41	4.03
TransUNet [4]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62	DWUNet [35]	87.47	4.55
MT-UNet [9]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81	ResUNet [36]	87.91	3.49
SwinUNet [37]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60	UNet++ [38]	88.32	3.83
MISSFormer [39]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	**91.92**	80.81	R2UNet [40]	90.13	3.62
TransCASCADE [41]	82.68	17.34	86.63	68.48	**87.66**	84.56	94.43	65.33	90.79	83.52	DCSAU-Net [33]*	90.41	**2.21**
MERIT [10]*	84.22	16.51	88.38	73.48	87.21	84.31	95.06	**69.97**	91.21	84.15	MSCA-Net [42]	90.52	2.79
LSTMSA (Ours)	**84.85**	**15.83**	**89.51**	73.32	85.88	**84.88**	**95.44**	69.20	90.99	**84.49**	LSTMSA (Ours)	**91.42**	2.53

We assess the performance of our proposed modules on these datasets using the DICE score and Hausdorff distance 95% (HD95) as evaluation metrics. The DICE score quantifies the overlap between predicted and ground truth masks, defined as DICE $= \frac{2\times|X\cap Y|}{|X|+|Y|}$, where X and Y are the predicted and ground truth masks. The HD95 measures the 95th percentile of the maximum distance between the two masks, offering a comprehensive evaluation of segmentation quality and edge accuracy.

Across all the aforementioned datasets, our approach integrate the U-shaped network with our proposed modules as a baseline model. Importantly, we introduce the proposed modules into the baseline models' architecture without any additional alterations to the original network structure and employ the same training strategies as the baseline models, indicating the seamless integration of our proposed modules into existing network structures for enhanced model performance, without necessitating extra training pipeline configuration or hyperparameter adjustments.

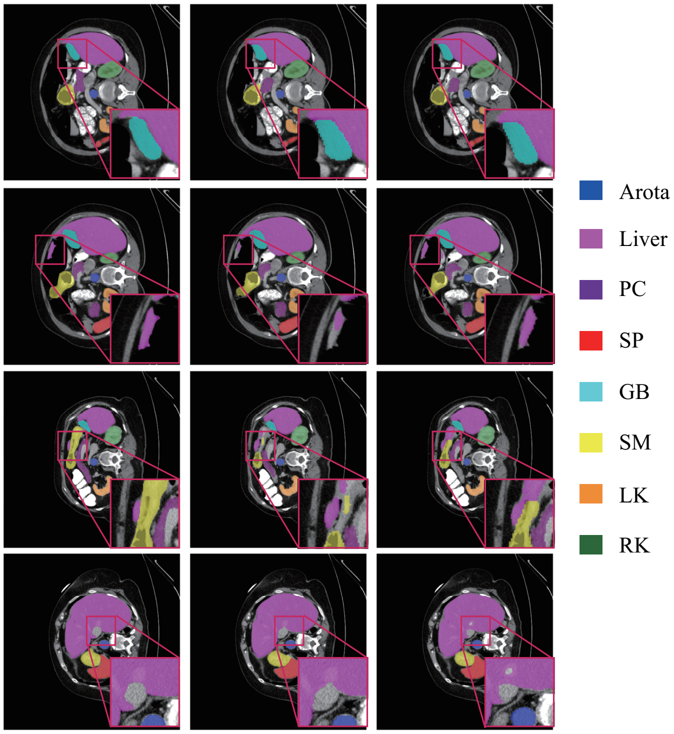
Notably, our experiments encompass twenty baseline models across the four datasets. In the subsequent sections, we will present the results obtained from the best-performing baseline model on each dataset, with results from the remaining baseline models available in subsequent analyses.

The specific baseline models utilized on the Synapse, ISIC2018, ACDC, and CVC-ClinicDB datasets are MERIT [10], DCSAU-Net [33], MT-UNet [9], and ESFPNet [34], respectively.

It is important to emphasize that the compared methods in our experiments represent the most advanced approaches for each dataset in recent years. The results achieved by the baseline models signify the SOTA performances attained by these models on their respective datasets.

B. Results on 2D Input

The Synapse Multi-Organ Segmentation dataset is composed of abdominal CT scans consists of eight distinct abdominal organs: the aorta, gallbladder (GB), spleen (SP), left kidney (KL), right kidney (KR), liver, pancreas (PC), and stomach (SM). This dataset comprises a total of 30 abdominal CT scans, each typically spanning 85 to 198 slices, with a pixel resolution of 512 × 512 per slice. The primary objective revolves around the precise segmentation of these eight abdominal organs within the dataset.



(a) Ground Truth (b) MERIT* (c) LSTMSA

Fig. 4. 2D visualization results on the Synapse dataset. (a) Ground truths, (b) MERIT (baseline are starred), (c) LSTMSA (Ours). The red rectangular box indicates the zoomed-in region.

The ISIC2018 dataset constitutes a collection of skin lesion images, encompassing both dermoscopic and clinical variants. The central challenge lies in the development and evaluation of automated segmentation methodologies tailored for accurately delineating skin lesions present within these diverse image types.

The data in Table I shows that networks incorporating the LSTMSA module consistently achieve SOTA results across both 2D datasets. For instance, in the ISIC2018 dataset, LSTMSA achieves a precision score of 91.42, significantly outperforming other baseline methods, indicating superior segmentation performance on 2D inputs. This performance improvement is linked to the LSTMSA's impact on loss attenuation (Fig. 1) and rank preservation. The module stabilizes loss reduction during early training, reducing the risk of rapid loss decline, which facilitates better convergence. Moreover, the

TABLE II

QUANTITATIVE RESULTS ON ACDC AND CVC-CLINICDB DATASET. DICE SCORES (%) AND HD95 ARE REPORTED. THE BEST RESULTS ARE BOLDED. THE SECOND BEST RESULTS ARE UNDERLINED. ↑ DENOTES HIGHER VALUE INDICATING BETTER PERFORMANCE, ↓ DENOTES LOWER VALUE INDICATING BETTER PERFORMANCE. BASELINES ARE STARRED.

ACDC						CVC-ClinicDB		
Method	DICE ↑	HD95 ↓	RV	Myo	LV	Method	DICE ↑	HD95 ↓
TransUNet [4]	89.71	2.54	88.86	84.53	95.73	ColonSegNet [43]	88.62	4.56
SwinUNet [37]	90.00	4.52	88.55	85.62	95.83	FCBFormer [14]	92.53	3.21
MT-UNet [9]*	90.43	2.23	86.64	89.04	95.62	SSFormer-S [44]	92.68	1.45
MISSFormer [39]	90.86	2.13	89.55	88.04	94.99	HarDNet-DFUS [45]	93.32	1.29
PVT-CASCADE [41]	91.46	1.09	88.9	89.97	95.50	FANet [46]	93.55	1.15
TransCASCADE [41]	91.63	1.09	89.14	90.25	95.50	TGANet [47]	94.57	1.47
MERIT [10]	92.32	1.08	90.87	90.00	96.08	SSFormer-L [44]	94.72	0.73
FCT [48]	92.84	5.29	**92.02**	90.61	95.89	ESFPNet [34]*	94.90	1.21
LSTMSA (Ours)	**92.86**	**1.07**	91.54	**90.77**	**96.27**	LSTMSA (Ours)	**96.11**	**0.53**

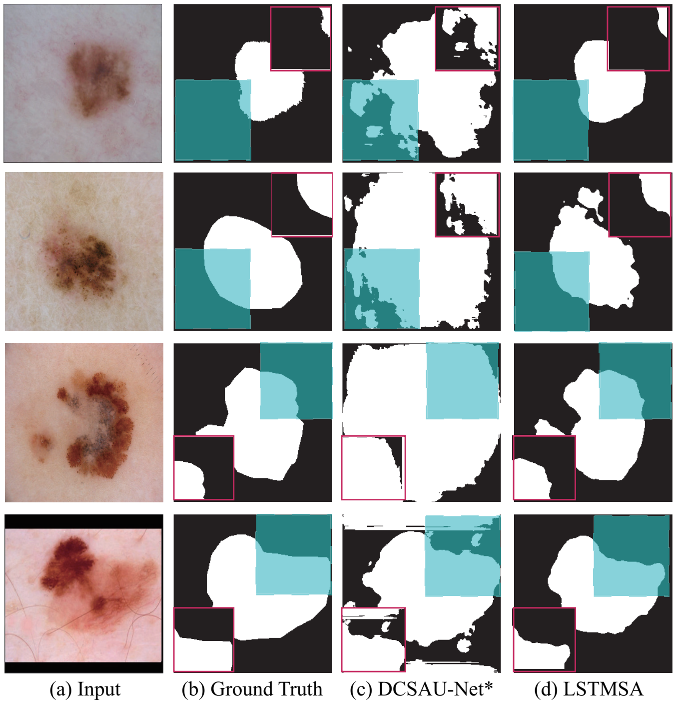

(a) Input (b) Ground Truth (c) DCSAU-Net* (d) LSTMSA

Fig. 5. 2D visualization results on the ISIC2018 dataset. (a) Input, (b) Ground Truth, (c) DCSAU-Net (baseline are starred), (d) LSTMSA (Ours). The red rectangular box indicates the zoomed-in region.

cross attention mechanism in LSTMSA accurately captures key image features, reducing rank collapse and enhancing overall segmentation precision.

We also visualize network performance with the LSTMSA module, as depicted in Fig. 4 and Fig. 5. These figures compare LSTMSA with other baseline methods in organ segmentation tasks, showing that LSTMSA produces notably precise results, particularly in small organ segmentation. The segmentation boundaries generated by LSTMSA are clearer and more accurate, effectively eliminating blurriness and imprecision. Detailed analysis reveals that the LSTMSA module exhibits heightened attention accuracy around small organ edges, capturing even subtle features within the structures.

C. Results on 3D input

The ACDC dataset encompasses cardiac magnetic resonance (MR) images aimed at facilitating precise segmentation

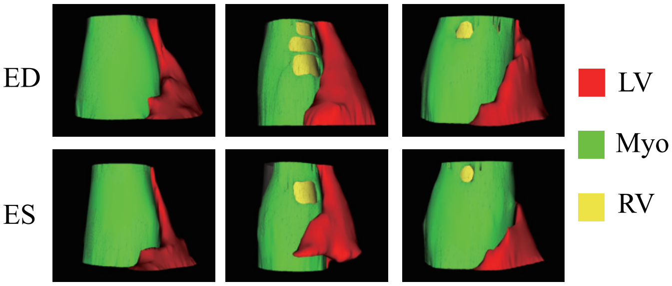


Fig. 6. 3D visualization results on the ACDC dataset. ED is End Diastolic, and ES is End Systolic.

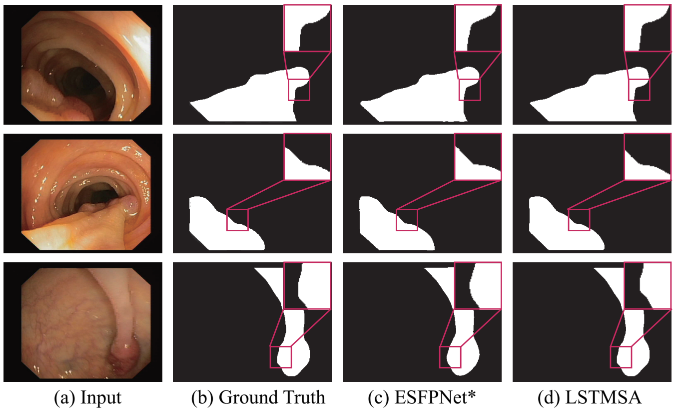

(a) Input (b) Ground Truth (c) ESFPNet* (d) LSTMSA

Fig. 7. 2D visualization results on the CVC-ClinicDB dataset. (a) Input, (b) Ground Truth, (c) ESFPNet (baseline are starred), (d) LSTMSA (Ours). The red rectangular box indicates the zoomed-in region.

of cardiac structures, including the right ventricle (RV), left ventricle (LV), and myocardium (Myo). Comprising a diverse set of 100 cardiac MRI scans, each originating from distinct patients, the dataset serves as the foundation for the segmentation challenge.

The CVC-ClinicDB dataset is a publicly accessible resource curated explicitly for research in the domain of colonoscopy image analysis, with a specific emphasis on tasks related to lesion segmentation. Comprised of a rich collection of high-definition colonoscopy images captured during clinical procedures, the dataset's primary objective revolves around facilitating the development and assessment of algorithms tailored for the automatic detection and segmentation of lesions within colonoscopy images.

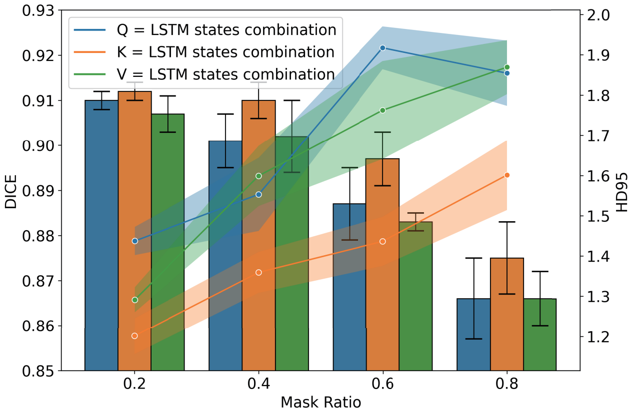


Fig. 8. Time-step dependency analysis on ACDC dataset. The x-axis represents the percentage of slices masked, and the y-axis represents the performance in terms of DICE and HD95.

Our analysis of 3D medical image segmentation, as shown in Table II, demonstrates that the integration of the LSTMSA module significantly improves performance over SOTA methods, particularly those relying solely on attention mechanisms. For example, on the ACDC dataset, the LSTMSA-enhanced model achieves a DICE score of 92.86, surpassing the MT-UNet baseline by nearly 3%, and delivers high precision in segmenting various cardiac structures, with scores of 91.54, 90.77, and 96.27 for RV, Myo, and LV, respectively. These results highlight LSTMSA's effectiveness in capturing complex features and offering nuanced improvements.

On the CVC-ClinicDB dataset, the LSTMSA model achieves a DICE score of 96.11, outperforming advanced models like ESFPNet and TGANet. This performance underscores LSTMSA's capability in handling intricate segmentation tasks, pushing beyond current SOTA levels in 3D segmentation. The model's edge lies in its integration of SA with sequential modeling through LSTM, which enhances the network's ability to capture both spatial and temporal dependencies. This combination refines segmentation accuracy and generalization, particularly visible in the sharper and more accurate boundaries in Fig. 6 and Fig. 7.

D. Validation of Module Structure Design

To demonstrate that using the LSTM states combination as the key matrix, rather than as the query and value matrices, better accommodates both long-term and short-term memory information, we conduct a time-step dependency analysis. This analysis assesses the model's sensitivity to specific time steps in the input sequence, thereby validating the role of long-term and short-term information within the model. The core approach of this method involves selectively removing or masking certain time steps in the input sequence and observing the resultant changes in model performance. Accordingly, we configure three distinct variants, using the LSTM states combination as the query, key, and value matrices respectively.

Fig. 8 illustrates the performance curves for DICE and HD95 metrics on the ACDC dataset, under different masking proportions (0.2, 0.4, 0.6, 0.8). The experiments are performed three times, and the results are recorded as mean values and standard deviations. The results reveal that the configuration with the LSTM states combination as the key matrix exhibited the least performance degradation upon the removal of time steps and demonstrated superior robustness. This supports the design choice of using the LSTM states combination as the key matrix, indicating that long-term and short-term information is more effectively managed when encoded in the key matrix.

TABLE III
ABLATION STUDY ON ACDC DATASET. DICE SCORES (%) AND HD95 ARE REPORTED. ✓ DONATES USED, AND ✗ DONATES NOT USED. THE BEST RESULTS ARE BOLDED. THE SECOND BEST RESULTS ARE UNDERLINED. SA(X,X,X) DENOTES USING ALL INPUT DATA FOR QKV SELECTION, SA(H,H,H) DENOTES ALL USING HIDDEN STATES FOR QKV SELECTION.

Method	SA	LSTM	DICE ↑	HD95 ↓
MT-UNet [9] (Baseline)	✗	✗	90.43	2.23
Only SA	✓	✗	91.17	1.77
Only LSTM	✗	✓	91.69	1.76
No Cross Attention (SA(x,x,x))	✓	✓	92.43	1.22
No Cross Attention (SA(h,h,h))	✓	✓	92.34	1.17
No Dynamic Positional Encoding	✓	✓	92.38	1.18
LSTMSA (Ours)	✓	✓	**92.86**	**1.07**

E. Abalation Study and Efficiency Analysis

1) Ablation Study: The ablation study conducted on the ACDC dataset, as shown in Table III, provides a detailed analysis of the LSTMSA module's individual components and their impact on segmentation performance. The results indicate that the LSTMSA module's superior performance is contingent on the strategic integration of both SA and LSTM components. Specifically, the ablation study reveals that the exclusion of either SA or LSTM components results in a noticeable decline in segmentation performance, underscoring the critical role played by both components in enhancing the network's segmentation accuracy. Furthermore, the ablation study highlights the importance of cross attention mechanism (including selection of QKV, only using input data and only using hidden states) and dynamic positional encoding in the LSTMSA module, with their exclusion leading to a reduction in segmentation performance.

2) Efficiency Analysis: Table IV presents an efficiency analysis on the Synapse, ISIC2018, ACDC, and CVC-ClinicDB datasets, comparing the computational costs of our LSTMSA module and separated SA and LSTM components within baseline models, using a tensor with shape of $(1, 3, 256, 256)$. The results indicate that the LSTMSA module exhibits a marginal increase compared to the baseline models. However, it does not introduce additional overhead compared to separated SA and LSTM, and the rate of increase in parameters, FLOPs, and inference time is smaller, indicating that the LSTMSA module is a lightweight and efficient solution for medical image segmentation tasks.

3) Input Perturbation Analysis: Table V shows the input perturbation analysis results on the ACDC datasets, assessing the LSTMSA module's robustness to noise. The module demonstrates superior robustness compared to baseline models, with minimal performance decline under varying noise

TABLE IV
Efficiency comparisons between our proposed module and separated SA and LSTM within SOTA method on each dataset. The results are obtained by averaging the outcomes of 10 experiment runs and the rate of increase is also calculated. Params denotes the number of parameters, Flops denotes the number of floating-point operations, and Time denotes the inference time taken for GPU inference.

Matrices	Dataset	Baseline	Separated δ	LSTMSA δ
Params (M)	Synapse	146.51	+0.75%	+0.56%
	ISIC2018	2.59	+4.63%	+4.15%
	ACDC	49.31	+0.41%	+1.65%
	CVC-ClinicDB	3.53	+3.11%	+8.97%
Flops (G)	Synapse	28.32	+3.95%	+0.28%
	ISIC2018	15.88	+6.92%	+1.50%
	ACDC	14.35	+5.57%	+0.56%
	CVC-ClinicDB	0.48	+41.67%	+34.29%
Time (ms)	Synapse	45.41	+2.42%	+1.44%
	ISIC2018	10.95	+9.31%	+5.51%
	ACDC	33.91	+3.83%	+0.56%
	CVC-ClinicDB	5.19	+13.29%	+10.05%

TABLE V
Input perturbation analysis on ACDC dataset. DICE scores (%) and HD95 are reported. Low and moderate noise is added using guassian noise with $\sigma = 0.1$ and $\sigma = 0.2$, respectively. Strong noise is added using random contrast, saturation, hue and brightness transformations.

Method	Noise level	DICE δ	HD95 δ
MT-UNet [9]	low	-4.22	+0.81
LSTMSA (Ours)	low	-2.63	+0.29
MT-UNet [9]	moderate	-9.41	+2.22
LSTMSA (Ours)	moderate	-6.98	+0.90
MT-UNet [9]	strong	-77.89	+27.22
LSTMSA (Ours)	strong	-72.45	+20.11

TABLE VI
Baseline comparison and statistical significance testing results across datasets and baselines. DICE scores (%) are reported. ↑ indicates an increase in DICE of less than 1% and ⇑ indicates an increase in DICE of greater than 1%. A p-value less than 0.05 indicates statistical significance.

Dataset	Method	DICE δ	Significance
Synapse	TransUNet [4]	78.89 ⇑	
	MT-UNet [9]	78.41 ↑	t-statistic: 3.3665
	MISSFormer [39]	80.49 ↑	p-value: 0.0281
	DAEFormer [49]	82.52 ↑	Significant: Yes
	MERIT [10]	84.85 ↑	
ISIC2018	UNet [19]	88.37 ↑	
	DWUNet [35]	88.56 ⇑	t-statistic: 19.5116
	ResUNet [36]	89.01 ⇑	p-value: 0.0004
	UNet++ [38]	89.14 ↑	Significant: Yes
	DCSAU-Net [33]	91.42 ⇑	
ACDC	MISSFormer [39]	90.75 ⇑	
	TransUNet [4]	91.64 ⇑	t-statistic: 3.7350
	DAEFormer [49]	91.82 ⇑	p-value: 0.0202
	MT-UNet [9]	92.86 ⇑	Significant: Yes
	MERIT [10]	92.45 ↑	
CVC-ClinicDB	FCBFormer [14]	95.63 ⇑	
	SSFormer-S [44]	94.10 ⇑	t-statistic: 6.0501
	HarDNet-DFUS [45]	95.65 ⇑	p-value: 0.0037
	FANet [46]	95.88 ⇑	Significant: Yes
	ESFPNet [34]	96.11 ⇑	

levels, indicating its ability to maintain segmentation accuracy in the presence of noise.

4) Effect on Different Baselines: To assess the LSTMSA module's universality, Table VI compares performance across different baseline networks with and without the LSTMSA module. It is worth noting that for some models without publicly available code, we reimplement the corresponding models according to the respective papers. The results demonstrate that our module consistently improves segmentation accuracy across datasets, confirming its versatility. Moreover, we adopt significance testing to evaluate the improvements in segmentation performance achieved by the LSTMSA module across different datasets and baseline models. The paired t-test is chosen due to the nature of the experimental setup, where each dataset is evaluated under two different conditions before and after the application of the proposed module. The results show that the improvements in segmentation performance are statistically significant, indicating the robustness and reliability of the LSTMSA module.

V. Conclusion

In this paper, we present an innovative deep learning model, SA-embedded LSTM (LSTMSA), designed to address critical challenges in conventional contextual models in medical image segmentation. By seamlessly integrating the LSTM and SA mechanism, LSTMSA achieves a harmonious balance between global and local context modeling, resulting in enhanced segmentation performance. Our research has demonstrated the effectiveness of LSTMSA in mitigating the issues of premature loss convergence and rank collapse, which have posed substantial impediments to pure SA networks. Through extensive evaluations on various medical image datasets, LSTMSA consistently outperforms existing methodologies, and its universal applicability across diverse baseline architectures has been validated.

Despite these significant achievements, it is essential to acknowledge a limitation in the current approach. The fusion of LSTM and SA in LSTMSA does not involve parallel training, leading to a sacrifice in training speed. To further enhance the practicality and efficiency of LSTMSA, we will explore strategies for incorporating parallel training structures, thereby optimizing both training time and segmentation performance.

In conclusion, the introduction of LSTMSA represents a noteworthy advancement in the field of medical image segmentation. Our proposed model not only resolves long-standing challenges but also sets the stage for future improvements that promise to revolutionize the domain.

References

[1] S. Shirmohammadi and A. Ferrero, "Camera as the instrument: The rising trend of vision based measurement," *IEEE Instrumentation & Measurement Magazine*, vol. 17, no. 3, pp. 41–47, 2014.

[2] Q. Hao, Y. Pei, R. Zhou, B. Sun, J. Sun, S. Li, and X. Kang, "Fusing multiple deep models for in vivo human brain hyperspectral image classification to identify glioblastoma tumor," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2021.

[3] X. Wei, W. Li, M. Zhang, and Q. Li, "Medical hyperspectral image classification based on end-to-end fusion deep neural network," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 11, pp. 4481–4492, 2019.

[4] 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," *arXiv preprint arXiv:2102.04306*, 2021.

[5] M. D. Alahmadi, "Medical image segmentation with learning semantic and global contextual representation," *Diagnostics*, vol. 12, no. 7, p. 1548, 2022.

[6] J.-H. Chang, K.-H. Lin, T.-H. Wang, Y.-K. Zhou, and P.-C. Chung, "Image segmentation in 3d brachytherapy using convolutional lstm," *Journal of Medical and Biological Engineering*, vol. 41, pp. 636–651, 2021.

[7] E. Shibuya and K. Hotta, "Cell image segmentation by using feedback and convolutional lstm," *The Visual Computer*, vol. 38, no. 11, pp. 3791–3801, 2022.

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

[9] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X.-H. Han, Y.-W. Chen, and R. Tong, "Mixed transformer u-net for medical image segmentation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2390–2394.

[10] M. M. Rahman and R. Marculescu, "Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation," *arXiv preprint arXiv:2303.16892*, 2023.

[11] Y. Chen, H. Zhang, Y. Wang, L. Liu, Q. J. Wu, and Y. Yang, "Tae-seg: Generalized lung segmentation via tilewise autoencoder enhanced network," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–13, 2022.

[12] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[13] Q. Zhou, Y. Feng, Z. Huang, M. Ding, and X. Zhang, "Specific instance and cross prompt based robust 3d semi-supervised medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[14] E. Sanderson and B. J. Matuszewski, "Fcn-transformer feature fusion for polyp segmentation," in *Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27–29, 2022, Proceedings*. Springer, 2022, pp. 892–907.

[15] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE journal of biomedical and health informatics*, vol. 25, no. 1, pp. 121–130, 2020.

[16] J. Zhao, F. Huang, J. Lv, Y. Duan, Z. Qin, G. Li, and G. Tian, "Do rnn and lstm have long memory?" in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 365–11 375.

[17] Y. Dong, J.-B. Cordonnier, and A. Loukas, "Attention is not all you need: Pure attention loses rank doubly exponentially with depth," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2793–2803.

[18] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

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

[20] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer, 2021, pp. 36–46.

[21] R. Azad, R. Arimond, E. K. Aghdam, A. Kazerouni, and D. Merhof, "Dae-former: Dual attention-guided efficient transformer for medical image segmentation," in *International Workshop on PRedictive Intelligence In MEdicine*. Springer, 2023, pp. 83–95.

[22] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, "U-net transformer: Self and cross attention for medical image segmentation," in *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*. Springer, 2021, pp. 267–276.

[23] K. Cao, T. Zhang, and J. Huang, "Advanced hybrid lstm-transformer architecture for real-time multi-task prediction in engineering systems," *Scientific Reports*, vol. 14, no. 1, p. 4890, 2024.

[24] S. Kim and S.-P. Lee, "A bilstm–transformer and 2d cnn architecture for emotion recognition from speech," *Electronics*, vol. 12, no. 19, p. 4034, 2023.

[25] Q. Luo, S. He, X. Han, Y. Wang, and H. Li, "Lsttn: A long-short term transformer-based spatiotemporal neural network for traffic flow forecasting," *Knowledge-Based Systems*, vol. 293, p. 111637, 2024.

[26] H. Zhang, Z. Wang, and H. Vallery, "Learning-based nlos detection and uncertainty prediction of gnss observations with transformer-enhanced lstm network," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 910–917.

[27] X. Kang, F. Han, A. Fayjie, and D. Gong, "Focdepthformer: Transformer with lstm for depth estimation from focus," *arXiv preprint arXiv:2310.11178*, 2023.

[28] S. Tang, C. Li, P. Zhang, and R. Tang, "Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 470–13 479.

[29] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, H. Cao, X. Cheng, M. Chung, M. Grella, K. K. GV *et al.*, "Rwkv: Reinventing rnns for the transformer era," *arXiv preprint arXiv:2305.13048*, 2023.

[30] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.

[31] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[32] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.

[33] Q. Xu, Z. Ma, H. Na, and W. Duan, "Dcsau-net: A deeper and more compact split-attention u-net for medical image segmentation," *Computers in Biology and Medicine*, vol. 154, p. 106626, 2023.

[34] Q. Chang, D. Ahmad, J. Toth, R. Bascom, and W. E. Higgins, "Esfpnet: efficient deep learning architecture for real-time lesion segmentation in autofluorescence bronchoscopic video," *arXiv preprint arXiv:2207.07759*, 2022.

[35] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] 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," in *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*. Springer, 2023, pp. 205–218.

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

[39] X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, "Missformer: An effective transformer for 2d medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1484–1494, 2023.

[40] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," *arXiv preprint arXiv:1802.06955*, 2018.

[41] M. M. Rahman and R. Marculescu, "Medical image segmentation via cascaded attention decoding," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6222–6231.

[42] Y. Sun, D. Dai, Q. Zhang, Y. Wang, S. Xu, and C. Lian, "Msca-net: Multi-scale contextual attention network for skin lesion segmentation," *Pattern Recognition*, vol. 139, p. 109524, 2023.

[43] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *Ieee Access*, vol. 9, pp. 40 496–40 510, 2021.

[44] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, and S. Song, "Stepwise feature fusion: Local guides global," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*. Springer, 2022, pp. 110–120.

[45] T.-Y. Liao, C.-H. Yang, Y.-W. Lo, K.-Y. Lai, P.-H. Shen, and Y.-L. Lin, "Hardnet-dfus: An enhanced harmonically-connected network for diabetic foot ulcer image segmentation and colonoscopy polyp segmentation," *arXiv preprint arXiv:2209.07313*, 2022.

[46] N. K. Tomar, D. Jha, M. A. Riegler, H. D. Johansen, D. Johansen, J. Rittscher, P. Halvorsen, and S. Ali, "Fanet: A feedback attention network for improved biomedical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[47] N. K. Tomar, D. Jha, U. Bagci, and S. Ali, "Tganet: text-guided attention for improved polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*. Springer, 2022, pp. 151–160.

[48] A. Tragakis, C. Kaul, R. Murray-Smith, and D. Husmeier, "The fully convolutional transformer for medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3660–3669.

[49] R. Azad, R. Arimond, E. K. Aghdam, A. Kazerouni, and D. Merhof, "Dae-former: Dual attention-guided efficient transformer for medical image segmentation," *arXiv preprint arXiv:2212.13504*, 2022.