Sparse Coding Inspired LSTM and Self-Attention Integration for Medical Image Segmentation

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

***Abstract*— Accurate and automatic segmentation of medical images plays an essential role in clinical diagnosis and analysis. It has been established that integrating contextual relationships substantially enhances the representational ability of neural networks. Conventionally, Long Short-Term Memory (LSTM) and Self-Attention (SA) mechanisms have been recognized for their proficiency in capturing global dependencies within data. However, these mechanisms have typically been viewed as distinct modules without a direct linkage. This paper presents the integration of LSTM design with SA sparse coding as a key innovation. It uses linear combinations of LSTM states for SA's query, key, and value (QKV) matrices to leverage LSTM's capability for state compression and historical data retention. This approach aims to rectify the shortcomings of conventional sparse coding methods that overlook temporal information, thereby enhancing SA's ability to do sparse coding and capture global dependencies. Building upon this premise, we introduce two innovative modules that weave the SA matrix into the LSTM state design in distinct manners, enabling LSTM to more adeptly model global dependencies and meld seamlessly with SA without accruing extra computational demands. Both modules are separately embedded into the U-shaped convolutional neural network architecture for handling both 2D and 3D medical images. Experimental evaluations on downstream medical image segmentation tasks reveal that our proposed modules not only excel on four extensively utilized datasets across various baselines but also enhance prediction accuracy, even on baselines that have already incorporated contextual modules. Code is available at https://github.com/yeshunlong/SALSTM.**

***Index Terms*— Sparse coding, contextual module, LSTM, self-attention, medical image segmentation.**

I. Introduction

MEDICAL image segmentation is a critical component in the domain of medical imaging, serving as a foundational step for extracting quantitative measurements and facilitating diagnostic and therapeutic procedures [1], [2], [3]. The advent of deep learning has ushered in a new era for this task [4], [5], [6], with numerous researchers leveraging neural networks to achieve state-of-the-art (SOTA) performance on a wide range of medical image segmentation datasets [7], [8], [9], [10].

Sparse coding, a technique aimed at representing signals as sparse linear combinations of basis elements, has been particularly influential in interpreting medical images where the underlying structures are inherently sparse [11], [12]. The primary advantage of sparse coding in medical image segmentation lies in its ability to isolate and emphasize the most relevant features of an image, thereby reducing noise and improving the clarity of important anatomical structures [13], [14], [15], [16]. This sparsity-driven approach aligns well with the characteristics of medical images, which often contain distinct and sparse regions corresponding to different tissues or pathological areas [17], [18], [19], [20]. Moreover, reducing sparsity throughout the layers of a deep network is crucial for effective image segmentation [21]. In each layer, sparse representations help to ensure that only the most salient features are propagated, thus enhancing the network's ability to distinguish between different regions of the image [22]. This layer-wise sparsity can lead to more precise boundary delineation and better segmentation outcomes overall. For instance, by enforcing sparsity constraints, networks can achieve higher robustness against overfitting, which is particularly beneficial given the often limited availability of annotated medical images. Additionally, sparse coding can facilitate more efficient computation, as fewer active neurons lead to reduced computational overhead, which is essential for deploying deep learning models in clinical settings with limited resources [23].

Among the various algorithms developed for sparse coding, the Iterative Shrinkage-Thresholding Algorithm (ISTA) [24] and its learned counterpart, Learned ISTA (LISTA) [25], stand out for their effectiveness. ISTA iteratively refines its representations to achieve sparsity [11], while LISTA transforms ISTA into a Recurrent Neural Network (RNN) architecture, enhancing computation of sparse codes by treating the ISTA process as a sequence of neural layers. However, a significant limitation of both algorithms is the lack of consideration for incorporating historical information into the update rules. This oversight restricts their capacity to leverage temporal or sequential dependencies within data, which can be crucial for understanding the complex structures and anisotropic slices present in medical images [26].

Meanwhile, the Self-Attention (SA) [27] mechanism has an underlying equivalence with sparse coding through its handling of input sequences [21]. It computes output representations as weighted sums of input elements, which can be viewed as a form of sparse coding, where the attention mechanism selectively focuses on relevant parts of the input, effectively encoding it into a sparse representation [21] by using linear projection. Such an approach has shown promise in capturing the intricate dependencies characteristic of medical images, thereby presenting a viable strategy for enhancing

Received 10 May 2024; revised 19 August 2024 and 4 October 2024; accepted 12 October 2024. Date of publication 22 October 2024; date of current version 28 October 2024. This work was supported by the National Science Foundation of China under Grant 62072241. The associate editor coordinating the review of this article and approving it for publication was Dr. Ananda S. Chowdhury. *(Corresponding author: Shunlong Ye.)*

The authors 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).

Digital Object Identifier 10.1109/TIP.2024.3482189

medical image segmentation through the lens of sparse coding principles.

As depicted in Fig. 1 (a), our design intertwines the similar frameworks and migratable interpretability [28] of Long Short-Term Memory (LSTM) [29] with the principles of SA sparse coding. By leveraging the intrinsic state compression and historical data retention capabilities of LSTM units, we amplify the SA mechanism's ability in sparse coding and global dependency modeling. Fig. 1 (b) further demonstrates the improvement of our method compared to network based solely on SA. We measure the sparsity of the output representation after each layer using the L_1 norm, and the results indicate several advantages of our approach. Firstly, our method successfully inherits the original model's ability in sparsification. Secondly, our method exhibits a faster decreasment in sparsity when the network has fewer initial layers. Thirdly, as the network deepens, our method ultimately achieves lower levels of sparsity representation.

In the realm of our design, we introduce two variants, namely SA coupled LSTM (SA-LSTM) and LSTM coupled SA (LSTM-SA), and both modules are tailored to optimally process distinct data dimensionalities with inherent feature densities. We believe a key observation is that SA-LSTM explores the efficacy of employing linear combinations of LSTM states as query, key, and value (QKV) matrices, a technique that shows exceptional promise with 2D data due to its denser feature. This adaptation enables a more granular and rich extraction of contextual relationships within the 2D image, harnessing the depth of LSTM's sequential data processing strengths to complement the broad reach of SA in global dependency mapping. Conversely, in the context of 3D data, where anisotropy between slices presents a challenge to maintaining contextual integrity, a direct application of LSTM states as QKV matrices in SA-LSTM can inadvertently lead to partial context information loss. To counteract this, LSTM-SA employs a strategy where the acceptance degree of the QKV matrices in SA is determined by a combination of LSTM states. The multiplication with the coefficient matrices in this design guarantees the preservation of state integrity and effectiveness, circumventing potential adverse effects without the incorporation of extraneous data.

We perform comprehensive quantitative experiments on four widely used medical image segmentation datasets to validate the effectiveness of our proposed modules. The experiments encompass both 2D and 3D inputs, and the results demonstrate that our approach outperforms SOTA methods on all the datasets. Furthermore, our modules exhibit better performance across a range of baseline models, showcasing their robustness and adaptability in enhancing medical image segmentation accuracy.

This paper's contributions are summarized as follows:

- Building upon widely adopted LSTM and SA designs, we integrate these methodologies to enhance their combined capabilities in sparse coding principles and global dependency modeling, offering a novel integration of LSTM states with SA's QKV matrices for improved context understanding.
- We introduce two tailored modules, optimized for 2D and 3D data, respectively, which not only utilize the unique properties of LSTM and SA but also ensure compatibility with existing models. This design provides a unified framework for medical image segmentation, facilitating seamless integration into various applications.
- Validation across diverse medical image segmentation datasets demonstrates our approach's effectiveness, achieving SOTA performance and showing significant improvements over twenty baseline models across four datasets. These results underscore the practical applicability and generalizability of our method across different dimensionalities of data.

II. Background

In this section, we provide an overview of sparse coding and its classic algorithm, as well as recent research on medical image segmentation. We also discuss the fusion manner of LSTM and SA in our proposed method compared to other approaches in hybrid LSTM and SA design.

A. Sparse Coding and Its Classic Algorithm

In the domain of data representation, the quest for deriving semantic significance from data that is both noisy and high-dimensional has led to the adoption of sparse coding techniques [30]. Sparse coding seeks to discover a dictionary that can sparsely represent data points, thereby revealing the underlying structure of the data.

Given a data matrix $X \in \mathbb{R}^{d_x \times n}$, the goal of sparse coding is to learn a dictionary $B \in \mathbb{R}^{d_x \times d_s}$ and a set of sparse codes $S \in \mathbb{R}^{d_s \times n}$ that represent the data. The optimization problem is expressed as

$$\min_{S,B} \sum_i \|x_i - Bs_i\|_2^2 + \lambda \|s_i\|_1, \tag{1}$$

with the constraint $\|b_j\|_2 \leq 1$ for $j = 1, \ldots, d_s$.

This optimization task is conventionally tackled by alternating between optimizing B and S, dealing separately with dictionary learning and sparse approximation. By fixing S, the dictionary learning reduces to a ridge regression problem with a known solution. Conversely, fixing B shifts the focus to sparse approximation, aiming to represent the input as a sparse linear combination of the dictionary elements.

In the realm of sparse coding, ISTA plays a pivotal role in deriving sparse representations of data. It iteratively refines sparse codes by employing a straightforward update mechanism, formulated as:

$$s^{(t)} = \text{shrink}\left(s^{(t-1)} - \tau \nabla g(s^{(t-1)}), \lambda\tau\right), \tag{2}$$

where $s^{(t)}$ denotes the sparse code at iteration t, τ represents a predetermined learning rate, λ is the regularization parameter, $\nabla g(s)$ corresponds to the gradient of the objective function $g(s) = \frac{1}{2}\|x - Bs\|_2^2$, and the shrink$(\cdot, \lambda\tau)$ function applies a hard thresholding operation to enforce sparsity.

Despite ISTA's success in sparse coding, its application faces significant hurdles. Primarily, ISTA employs a non-adaptive update strategy across dimensions with a fixed

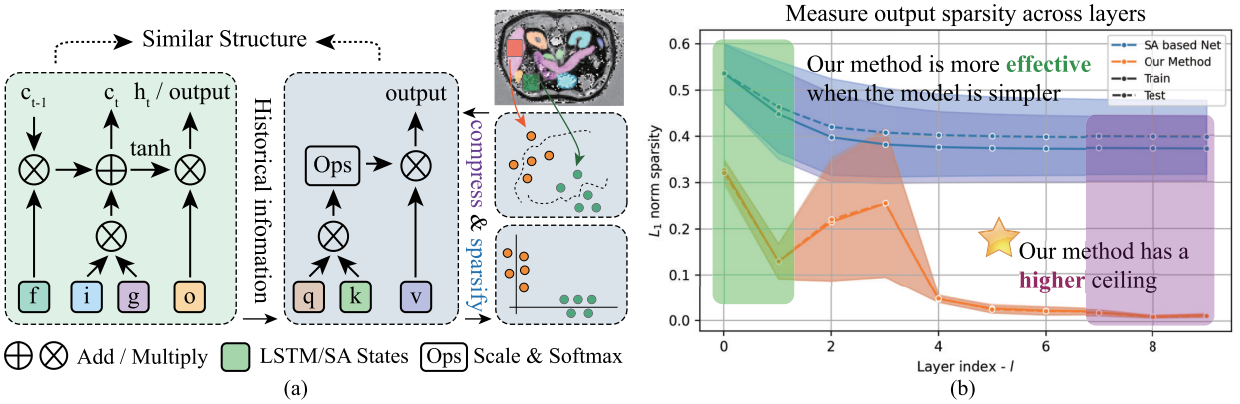


Fig. 1. Our motivation diagram and results. (a) Two motivations for the deep fusion of LSTM and SA: first, there is an intrinsic similarity in the computation process of LSTM states and the QKV matrices in SA; second, using LSTM can effectively provide enhanced representations of historical information for SA in the process of sparse coding. (b) The improvement of our method compared to network based solely on SA in the process of sparse coding.

learning rate, potentially leading to suboptimal performance due to a lack of diversity in parameter updates. To address these limitations, the LISTA is introduced, transforming ISTA into a RNN architecture. LISTA accelerates the computation of sparse codes by reinterpreting the iterative process of ISTA as a sequence of neural network layers. This innovative approach allows LISTA to optimize the dictionary and sparse codes concurrently, offering significant improvements in inference speed by bypassing the need to solve a convex optimization problem directly. Instead, sparse codes can be rapidly acquired through a neural network's forward pass. However, while LISTA marks a significant advance in computational efficiency, it inherits a critical shortcoming from its ISTA lineage. It does not incorporate historical information into its update mechanisms. The fixed learning rate across iterations and the absence of an adaptive updating strategy limit LISTA's ability to fully exploit the wealth of information available from previous iterations [31]. This oversight constrains the potential of LISTA to achieve faster convergence and enhance model performance, highlighting a critical area for future enhancement in the evolution of sparse coding methodologies.

Based on the observations above, we have found that the strategy of integrating LISTA through RNN greatly enhances the ability of sparse coding to mine sequential data. Given the same strategy, SA-based sparse coding methods can also further improve their ability to represent historical or neighborhood data [21]. Further considering the structural similarity, this paper focuses on how to couple SA and LSTM to enhance their modeling capabilities.

B. Recent Research on Medical Image Segmentation

In the realm of medical image segmentation, UNet [32] has become a foundational model due to its ability to effectively handle the intricacies of 3D medical imaging data through its encoder-decoder architecture. Given that medical imaging data is often presented in a 3D slice-by-slice format [7], researchers frequently leverage 2D or 3D Convolutional Neural Networks (CNNs) to address segmentation tasks effectively [2], [3]. In response to challenges such as limited training data, the adoption of self-supervised learning approaches has become increasingly prominent. The mean teacher framework [33], which employs consistency regularization to leverage pseudo-labels and enhance learning, exemplifies a robust strategy to mitigate data scarcity issues [34], [35], [36]. These self-supervised models have shown promise in improving segmentation performance by extracting valuable insights from unlabeled data.

Moreover, many models containing contextual information similar to LSTM and SA have been proposed in processing medical image data [37], [38], [39]. Gu et al. [9] indicate that general U-shaped networks may loss sight of high-level information. Thus they utilize a context encoder to capture more high-level information and preserve spatial information. Feng et al. [40] concentrate on the problem of imbalanced class and blurred boundary in medical images, and propose a pyramidal module to fuse multi-scale context information. Girum et al. [8] use a forward system to get the prediction of the segmentation and integrat it into the context feedback system to get the final segmentation.

The application of foundation models in medical image segmentation has expanded with the development of approaches tailored for specific medical challenges. Notably, the adaptation of the Segment Anything Model [41] (SAM) to medical imaging has shown promising results. For instance, the work by Liu et al. [42], building on the foundational SAM architecture, presents a comprehensive framework for general-purpose medical image segmentation. Additionally, the Ma-SAM framework [43], which adapts SAM to handle 3D medical images, presents a significant innovation. By designing a modality-agnostic approach that can efficiently process volumetric data, Ma-SAM addresses one of the key challenges in medical image segmentation—leveraging 3D spatial context while maintaining high segmentation accuracy. This innovative approach stands out for its ability to generalize across different imaging modalities without the need for extensive modality-specific adjustments. Furthermore, the integration of SAM with other advanced techniques [44], [45] leads to a future where foundation models are not only powerful in generalization but also capable of incorporating

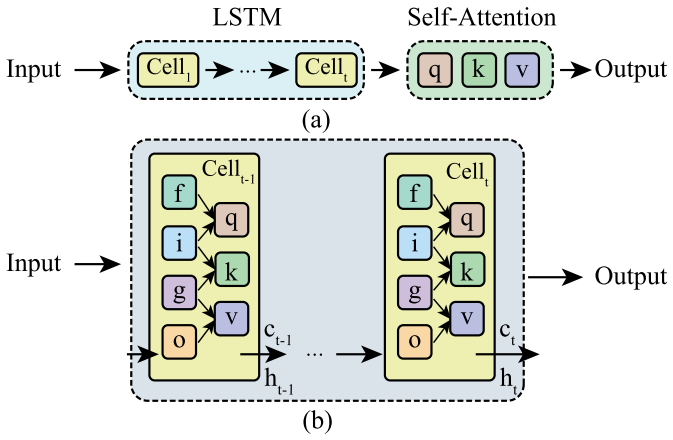


Fig. 2. Comparison and fusion manner of LSTM and SA. (a) Direct fusion of LSTM and SA, using LSTM and SA separately. (b) Deep fusion of LSTM and SA, which fuses the characteristics of LSTM and SA in each cell.

sophisticated context-aware mechanisms to further improve segmentation outcomes in complex medical imaging scenarios.

C. Fusion Manner of LSTM and SA

The most common way to fuse LSTM and SA for medical image segmentation is by stacking them together [38], [46], [47], [48], [49], [50]. Fig. 2 (a) illustrates an architecture where a multi-layer LSTM module is followed by a SA module. The number of layers and hidden state sizes can be adjusted to optimize performance for a specific task. Deeper models with larger hidden states generally perform better, but they also come with increased computational cost. While this combination method can enhance the model's representation ability to some extent, it does not fully consider the inherent differences and connections between LSTM and SA. As a result, it often leads to additional computational overhead and only minor improvements in accuracy [39].

On the other hand, our approach aims to fully fuse LSTM and SA. As shown in Fig. 2 (b), each layer of the module incorporates both LSTM and SA mechanisms. The outputs of these modules are then combined by using a gating mechanism before being fed into the subsequent feed forward neural network. The key distinction between these two approaches is that direct stacking uses separate layers for LSTM and SA, whereas our strategy combines them within each layer to better capture long-term dependencies.

It is worth noting that there are some related works using hybrid LSTM and SA, such as TRANS-BLSTM [37], SAST-LSTM [51], SwinLSTM [38] and RWKV [39], which appear similar to our approach. However, they either ignore the detailed meaning of states in LSTM and SA or primarily focus on introducing a completely different attention mechanism. In contrast, sparse coding is a key factor in our method, which is overlooked by the above methods. We build our module on the basis of this theory, which makes our method more interpretable and explainable in the process of medical image segmentation compared to other methods. Meanwhile, our proposed method is more like a plugin that can enhance most baseline models without constructing an entirely new module. This approach provides more flexibility in improving model performance for medical image segmentation tasks. In addition, the unique feature extraction capability involved in our method is also the main difference between its prompt or adapter fine-tuning approach to the foundation model [41], [42].

III. Method

In this section, we present our proposed modules, SA-LSTM and LSTM-SA, designed to enhance the sparse coding capabilities of LSTM and SA mechanisms. We detail the structural integration of LSTM and SA, elucidating the rationale behind our design choices and the technical underpinnings of our approach.

A. Rethinking the Combination of LSTM and SA for Sparse Coding

Unlike conventional methods, LSTM possesses a stronger memory capability, allowing it to effectively preserve and utilize past input information over extended sequences. This characteristic is particularly beneficial for tasks that rely on the memory of previous inputs, such as medical image segmentation. In such tasks, certain features within the images may become more prominent or complex as the sequence progresses, for example, when analyzing a series of image slices. Conventional methods may focus solely on the features of the current image, failing to fully leverage the information from preceding images, which can result in an incomplete understanding of the image structure. In contrast, LSTM can remember and process this historical information, enabling it to make more accurate segmentation decisions, even when certain features are less apparent in the current image by recalling and utilizing information from previous inputs. This characteristic help LSTM remember and utilize historical data over long sequences, making them more invaluable for tasks requiring memory of past inputs compared with conventional sparse coding algorithm. The incorporation of LSTM states into the SA mechanism's QKV matrices can be construed as a method to harness this memory capability, ensuring that the SA mechanism does not solely rely on the immediate input but also draws upon a rich, historically informed context.

The SA mechanism, the key part of the transformer [27] architecture, excels in modeling interactions across entire sequences, offering a means to directly compute relationships between distant elements on the equivalent sparse coding [21]. By integrating LSTM state information into the SA framework, the model effectively bridges the gap between local coherence and global contextual understanding. This integration allows the SA to dynamically adjust its focus, not just based on the static input sequence, but also influenced by the evolving contextual information encoded in the LSTM states.

Incorporating the intrinsic properties of LSTM units with the SA mechanism, as illustrated in Fig. 3, offers a novel approach to enhancing the model's ability to capture global dependencies within data. This integration aims to leverage the LSTM's inherent capabilities in compressing state information and its facility for retaining historical context, thereby

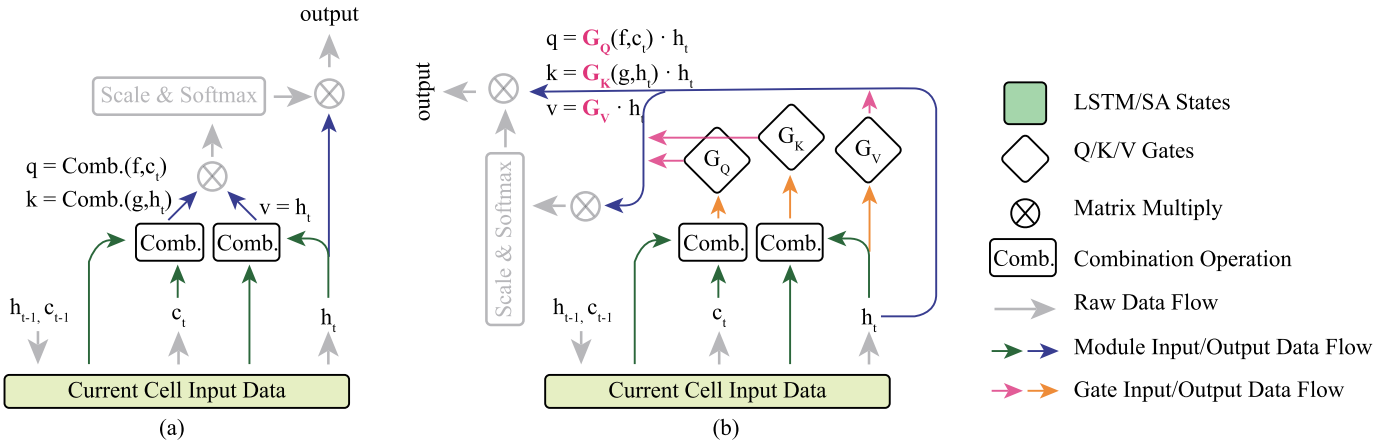


Fig. 3. Two proposed modules of deeply fusing LSTM and SA. (a) Employs the linear combination of LSTM states as the input of the QKV matrices in SA. (b) Adopts the LSTM states as the acceptance degree of SA. The differences between the two modules are highlighted in pink and orange.

augmenting the SA mechanism's capacity for global dependency modeling and sparse coding. We adopt the proposition of employing linear combinations of LSTM states as QKV matrices in SA to amalgamate the LSTM's local information capture with a more globally-aware representation strategy.

B. SA-LSTM

As show in Fig. 3 (a), in the pursuit of enhancing the capabilities of SA mechanism, particularly with a focus on achieving a more nuanced understanding and representation of sequential data, we integrate LSTM state information into the computation of query matrix, key matrix, and value matrix deeply. The proposed SA-LSTM, as delineated in Eq. 3, leveraging various combinations of LSTM gate activations and cell states to enrich the SA mechanism's expressiveness and dynamic adaptability.

$$\begin{cases} Q = FQ(i_t, f_t, g_t, o_t, c_t, h_t) \\ K = FK(i_t, f_t, g_t, o_t, c_t, h_t) \\ V = FV(i_t, f_t, g_t, o_t, c_t, h_t) \end{cases} \tag{3}$$

where i_t, f_t, g_t, and o_t represent the input, forget, inner, and output gate activations of the LSTM unit at time step t, respectively. The cell state and output state of the LSTM unit at time step t are denoted by c_t and h_t, respectively.

The rationale behind employing specific LSTM states to constitute the QKV matrices is rooted in the intrinsic properties and roles of these states within the LSTM architecture. The query matrix is conceptualized to aggregate surface-level information, weaving together historical and current inputs, analogous to the LSTM's function of integrating past and present through its cell state (c_t) and forget gate (f_t) [31]. This similarity underscores the choice to link query matrix with elements akin to LSTM's information fusion capabilities and prevent the unbounded cell state to propagate through the network and destabilize learning [52].

Conversely, the key matrix is tasked with measuring the relevance of other elements in relation to the query across both deep and shallow feature layers, mirroring the LSTM's gating mechanisms (g_t) and (h_t) that regulate information flow. This alignment suggests a natural coupling between key matrix and the LSTM's capacity to evaluate inter-element relationships based on accumulated contextual insights.

The value matrix, pivotal in enhancing the SA mechanism's sparsity, is inherently associated with the LSTM's output state (h_t) [39], serving to filter and reinforce the most salient features while suppressing less pertinent information. This association not only preserves but also augments the mechanism's focus on critical attributes, facilitating a more dynamic and expressive computation of SA that is both informed by and adaptable to the spatial and temporal dynamics inherent in sequential data.

The proposition to utilize specific LSTM state combinations for computing QKV matrices within the SA mechanism embodies a strategic approach that leverages LSTM's ability to retain historical information and compress states. This hybrid approach not only aligns structurally but also significantly enhances the SA mechanism by offering a dynamic and expressive method for computing attention weights. The introduction of LSTM states enables the method to incorporate both recent and historical data, allowing for fine-grained control through distinct states. Specifically, the combination of the forget gate and cell state integrates past and present information, focusing on spatiotemporal differences that emerge, thereby guiding the SA mechanism to prioritize sequence segments pertinent to spatial and temporal dynamics. Furthermore, the combination of the input gate and hidden state enhances sparsity, resulting in a more precise attention distribution. The use of the hidden state as the value ensures that core information is preserved, thus improving the robustness of the attention distribution. This approach ultimately enhances the SA's capacity for global dependency modeling, providing a richer, context-aware framework that dynamically integrates both local and global insights for improved sequence representation and processing.

The functional mappings FQ, FK, and FV are initially considered in two distinct formulations to explore the impact of linear versus non-linear combinations of LSTM states,

as shown in Eq. 4.

$$\begin{cases} FQ, FK, FV &= \frac{f_t + c_t}{2}, \frac{g_t + h_t}{2}, h_t \\ FQ, FK, FV &= ReLU(f_t * c_t), ReLU(g_t * h_t), h_t \end{cases} \tag{4}$$

The first set of functions implies a linear combination approach, which is adopted finally, blending the LSTM states through straightforward arithmetic operations. The second set introduces non-linear transformations, facilitated by the ReLU activation function, thus embedding an additional layer of complexity and potential for capturing intricate relationships within the data. More detailed experimental evaluations on different combination strategies can be found in Section IV.

C. LSTM-SA

In the evolving landscape of sequence modeling, particularly when transitioning from 2D to 3D data, the intricacies of capturing contextual information become increasingly complex and challenging for SA-LSTM module proposed in the previous section. As show in Fig. 3 (b), the LSTM-SA module addresses these challenges by modeling the anisotropy present in layers of 3D data. The proposed formulation for the LSTM-SA module is presented in Eq. 5.

$$\begin{cases} W_Q = FQ(i_t, f_t, g_t, o_t, c_t, h_t) \\ W_K = FK(i_t, f_t, g_t, o_t, c_t, h_t) \\ W_V = FV(i_t, f_t, g_t, o_t, c_t, h_t) \\ Q = W_Q \odot h_t \\ K = W_K \odot h_t \\ V = W_V \odot h_t \end{cases} \tag{5}$$

SA-LSTM employs linear combinations of LSTM states as QKV matrices and shows superior performance on 2D data due to the dense nature of its features, while LSTM-SA modifies this integration to adapt with 3D data. This distinction arises from the observation that 3D data exhibits slice-to-slice anisotropy, which can lead to a loss of contextual information when relying solely on linear combinations of LSTM states for QKV in SA-LSTM.

LSTM-SA adapts to this challenge by using the linear combinations of LSTM states to influence the coefficients of W_Q, W_K, and W_V, rather than directly constituting QKV. This method combines the output state of the previous layer into the QKV matrices through element-wise multiplication, thus ensuring the preservation of context and mitigating potential information loss. The multiplication with W_Q, W_K, and W_V guarantees the integrity and effectiveness of the states, without being adversely affected or introducing extraneous information.

This application of LSTM states in LSTM-SA versus SA-LSTM exemplifies a tailored approach to sequence modeling that respects the unique characteristics of the data dimensionality. By integrating LSTM states in a manner that caters specifically to the challenges presented by 3D data, LSTM-SA offers a robust solution for preserving contextual information and ensuring the model's responsiveness to the inherent complexities of the data structure, thereby enhancing the model's overall performance and applicability across diverse data scenarios. More detailed experimental scenario evaluations on SA-LSTM and LSTM-SA can be found in Section IV.

IV. Experiments and Results

In this section, we present the experimental setup, implementation details, and results of our proposed modules on various medical image segmentation datasets. We evaluate the performance of our modules on both 2D and 3D inputs, comparing them against SOTA methods to demonstrate their effectiveness in enhancing segmentation accuracy. Moreover, we conduct several ancillary experiments to discuss the model structure or to demonstrate that the model possesses sufficient robustness and generalisability.

A. Implementation Details

In the case of 2D input, we evaluate the performance of our proposed modules on the following datasets:

- Synapse dataset: This dataset includes 30 abdominal CT scans, each with 85 to 198 slices of 512×512 resolution, targeting segmentation of 8 organs namely the aorta, gallbladder (GB), spleen (SP), left kidney (KL), right kidney (KR), liver, pancreas (PC), and stomach (SM).[1]
- ISIC2018 dataset: This dataset comprises skin lesion images, including both dermoscopic and clinical images. The main task is to develop and evaluate automatic segmentation methods for skin lesions in these images and has been introduced in works by Codella et al. [53] and Tschandl et al. [54].

For the 3D input scenario, we conduct experiments on the following datasets:

- ACDC dataset: This dataset involves the segmentation task of the heart, and consists of 100 cardiac MRI scans for segmenting three cardiac structures: right ventricle (RV), left ventricle (LV), and myocardium (Myo).[2]
- CVC-ClinicDB dataset: This dataset contains a collection of high-definition colonoscopy images captured during clinical procedures and has been used for comparing automatic segmentation methods in the work by Bernal et al. [55].

We evaluate the performance of our proposed modules on these datasets using the DICE score and Hausdorff distance 95% (HD95) together as the evaluation metrics. The DICE score measures the overlap between the predicted and ground truth masks, while the Hausdorff distance quantifies the maximum distance between the predicted and ground truth masks, providing a comprehensive measure of both segmentation quality and edge accuracy. The formula for the DICE score is given by: DICE $= \frac{2\times|X\cap Y|}{|X|+|Y|}$, where X and Y represent the predicted and ground truth masks, respectively. The HD95 is computed as the 95th percentile of the Hausdorff distance between the predicted and ground truth masks.

In all of the aforementioned datasets, we incorporate the U-shaped network with our proposed modules as the baseline

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

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

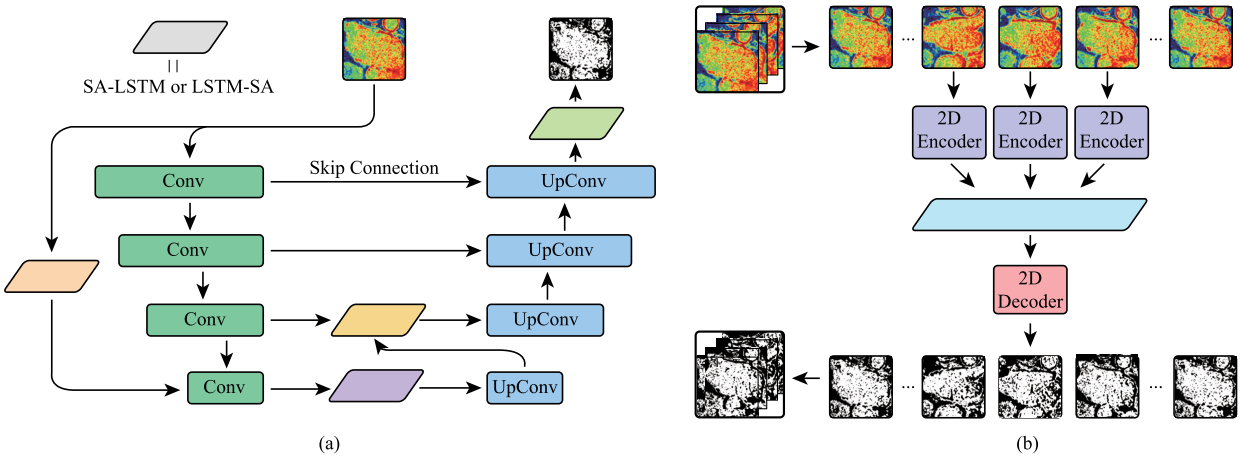


Fig. 4. Illustration of module embbeding strategies. (a) Module embedding strategies for 2D inputs, with light brown, purple, yellow, and green, indicating four different ways to embed the module. (b) Module embedding strategies for 3D inputs, with light blue indicating the position of the module.

model (Fig. 4). It is important to note that we only add the proposed modules to the baseline model's structure without any other modifications in the original network architecture. Additionally, we adopt the same training strategies as the baseline model, meaning that our proposed modules can be seamlessly integrated into existing network structures to enhance model performance, without requiring additional training strategies or hyperparameter tuning.

Since we conduct experiments with twenty baselines across the four datasets, we will present the results of the best baseline model on each dataset in Subsection IV-C and Subsection IV-D. The results of the other baseline models can be found in the subsequent experiments. The specific baseline models used on the Synapse, ISIC2018, ACDC, and CVC-ClinicDB datasets are MERIT [56], DCSAU-Net [57], MT-UNet [58], and ESFPNet [59], respectively. It is crucial to highlight that the compared methods in our experiments are the most advanced ones for each dataset in recent years. Additionally, we incorporate an extra comparison with a classic foundation model MedSAM [42], which is currently one of the most widely used foundation models in medical image segmentation, to provide a more comprehensive evaluation.

B. Module Embbeding Strategies

To test our module's effectiveness on 2D and 3D medical images, we integrated it into various positions within a U-shaped network. For 2D inputs, as illustrated in Fig. 4 (a), the module is embedded in parallel to the encoder, between encoder and decoder, between the skip connection and upsampling layer, and post-decoder. For 3D inputs, the module is positioned between the encoder and decoder, and three adjacent slices are input at a time, which are processed by three 2D encoders separately and then fused using the new module before being fed into the 2D decoder to obtain the final inference result, as shown in Fig. 4 (b). We assess the performance of these different embedding strategies in subsequent experiments.

C. Results on 2D Inputs

Table I demonstrates that both our modules, LSTM-SA and SA-LSTM, achieve SOTA results on the Synapse multi-organ dataset and the ISIC2018 segmentation dataset. In particular, SA-LSTM outperforms LSTM-SA across most metrics, indicating its superior performance on 2D data. For example, on the Synapse dataset, SA-LSTM achieves a DICE score of 85.13% and an HD95 of 13.25, compared to the baseline model MERIT, which achieves a DICE score of 84.22% and an HD95 of 16.51. Similarly, on the ISIC2018 dataset, LSTM-SA achieves a DICE score of 91.08% and an HD95 of 2.53, outperforming the baseline model DCSAU-Net, which achieves a DICE score of 90.41% and an HD95 of 2.21. In addition, both our modules exhibit better performance compared to the foundation model MedSAM, highlighting the effectiveness of our modules in dealing with specific medical imaging challenges.

Comparing our modules with other SOTA methods in the table, we can see that both LSTM-SA and SA-LSTM consistently outperform the baselines and many other methods across various organs and structures. Such as the aorta (88.93% vs. 88.38%), gallbladder (75.87% vs. 73.48%), liver (95.57% vs. 95.06%), and spleen (92.02% vs. 91.21%) on the Synapse dataset. This demonstrates the effectiveness of our modules in enhancing segmentation accuracy and improving the model's ability to capture detailed information from input images.

The visual analysis shown in Fig. 5 further illustrates the effectiveness of our modules in dealing with various small targets, such as the gallbladder, spleen, and pancreas. These targets are accurately segmented, highlighting the modules' proficiency in capturing and utilizing contextual information from input images. This enhancement in segmentation performance is crucial in medical imaging applications where the precise delineation of such structures is essential for diagnosis and treatment planning.

D. Results on 3D Inputs

Table II presents the segmentation results of our modules on the ACDC and CVC-ClinicDB datasets, where they achieve

TABLE I

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

Synapse											ISIC2018		
Method	DICE ↑	HD95 ↓	Aorta	GB	KL	KR	Liver	PC	SP	SM	Method	DICE ↑	HD95 ↓
UNet [32]	70.11	44.69	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96	UNet [32]	87.41	4.03
TransUNet [60]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62	DWUNet [61]	87.47	4.55
MT-UNet [58]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81	ResUNet [62]	87.91	3.49
SwinUNet [63]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60	UNet++ [64]	88.32	3.83
MISSFormer [65]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81	R2UNet [66]	90.13	3.62
TransCASCADE [67]	82.68	17.34	86.63	68.48	87.66	84.56	94.43	65.33	90.79	83.52	DCSAU-Net [57]*	90.41	2.21
MERIT [56]*	84.22	16.51	88.38	73.48	87.21	84.31	95.06	69.97	91.21	84.15	MSCA-Net [68]	90.52	2.79
MedSAM [42]	81.88	19.40	87.77	69.11	80.45	79.95	94.80	**72.17**	88.72	82.06	MedSAM [42]	87.30	4.32
SA-LSTM (Ours)	**85.13**	**13.25**	**88.93**	**75.87**	87.47	84.06	**95.57**	71.52	91.61	**85.98**	SA-LSTM (Ours)	**91.17**	**2.02**
LSTM-SA (Ours)	84.50	13.62	88.24	74.97	**88.45**	**85.47**	95.41	67.23	**92.02**	84.21	LSTM-SA (Ours)	91.08	2.53

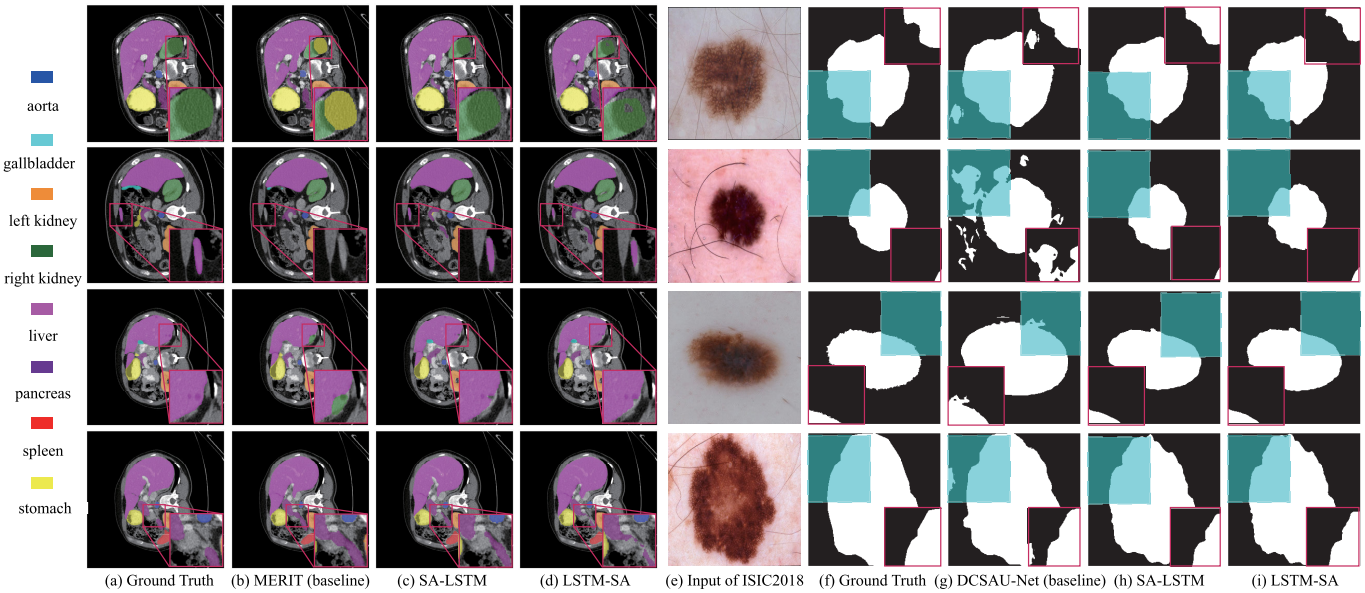


Fig. 5. Visualization results on the Synapse and ISIC2018 dataset. (a) Ground truths of Synapse. (b)-(d) Segmentation results obtained by MERIT (baseline), SA-LSTM and LSTM-SA on Synapse, respectively. (e) and (f) Original images and ground truths of ISIC2018. (g)-(i) Segmentation results obtained by DCSAU-Net (baseline), SA-LSTM and LSTM-SA on ISIC2018, respectively. The red rectangular box indicates the region that have visible improvements between the baseline and our proposed modules.

TABLE II

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

ACDC						CVC-ClinicDB		
Method	DICE ↑	HD95 ↓	RV	Myo	LV	Method	DICE ↑	HD95 ↓
TransUNet [60]	89.71	2.54	88.86	84.53	95.73	ColonSegNet [69]	88.62	4.56
SwinUNet [63]	90.00	4.52	88.55	85.62	95.83	FCBFormer [70]	92.53	3.21
MT-UNet [58]*	90.43	2.23	86.64	89.04	95.62	SSFormer-S [71]	92.68	1.45
MISSFormer [65]	90.86	2.13	89.55	88.04	94.99	HarDNet-DFUS [72]	93.32	1.29
PVT-CASCADE [67]	91.46	1.09	88.9	89.97	95.50	FANet [73]	93.55	1.15
TransCASCADE [67]	91.63	1.08	89.14	90.25	95.50	TGANet [74]	94.57	1.47
MERIT [56]	92.32	1.08	90.87	90.00	96.08	SSFormer-L [71]	94.72	0.73
FCT [75]	92.84	5.29	92.02	90.61	95.89	ESFPNet [59]*	94.90	1.21
MedSAM [42]	92.30	1.22	91.37	90.14	95.39	MedSAM [42]	94.50	1.53
SA-LSTM (Ours)	93.06	1.05	91.67	**90.90**	**96.60**	SA-LSTM (Ours)	96.03	0.48
LSTM-SA (Ours)	**93.08**	**1.03**	**92.05**	90.69	96.52	LSTM-SA (Ours)	**96.35**	**0.41**

SOTA performance, notably improving the segmentation of target structures. Compared to the baseline models, MT-UNet and ESFPNet and foundational model MedSAM, our modules show significant improvements across all metrics on the ACDC dataset, with an improvement of approximately 3%. This confirms the efficacy of our modules in boosting segmentation accuracy and enhancing global dependency modeling.

Furthermore, the comparison between LSTM-SA and SA-LSTM reveals that LSTM-SA is more suitable for processing 3D data, as it outperforms SA-LSTM on most metrics. This

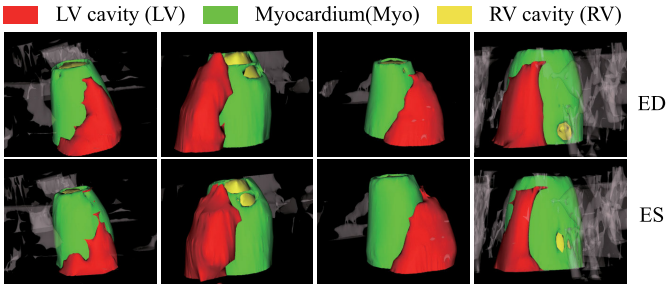


Fig. 6. 3D Visualization results on the ACDC dataset. ED is End Diastolic, and ES is End Systolic.

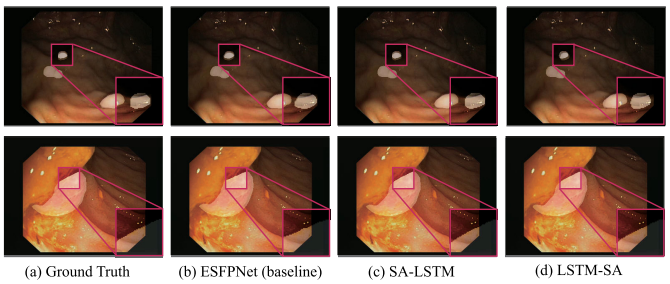


Fig. 7. 2D Visualization results on the CVC-ClinicDB dataset. (a) Ground Truth. (b) ESFPNet (baseline). (c) SA-LSTM. (d) LSTM-SA. The red rectangular box indicates the zoomed-in region.

indicates that LSTM-SA has superior performance in capturing spatial dependencies and context information in volumetric data, making it more adept at handling 3D medical imaging tasks.

Visual analysis, as shown in Fig. 6 and Fig. 7, the model excels in segmenting detailed areas like the valve regions, ventricle tops, and polyp outlines. This precision is attributed to our modules' enhanced ability to capture and utilize input information effectively, leading to improved segmentation accuracy, especially in areas with sparse or intricate structures.

E. Efficiency Analysis and Ablation Studies

1) Computational Cost Analysis: Table III provides an assessment of the efficiency of our proposed modules in representing global features on the Synapse dataset. The comparison involves evaluating the parameters, FLOPs, and inference time of the baseline models and the models with our proposed modules by using a tensor of size $\mathcal{Z} \in \mathbb{R}^{1\times3\times256\times256}$ as input. Notably, our proposed modules maintain a lightweight profile in terms of parameters, FLOPs, and inference time, with a minimal increase in computational cost. This demonstrates the efficiency of our proposed modules in capturing global dependencies without introducing significant computational overhead. In comparison to the approach of separate SA and LSTM structures, our proposed module does not introduce any additional spatio-temporal cost during combination.

2) Effectiveness of Combination Strategies: Table IV shows the results of the ablation experiment performed on the ACDC dataset. The outcomes demonstrate that when SA or LSTM is used independently or separately, satisfactory results are not achieved. This finding can be attributed to the fact that neither method alone can effectively represent global dependencies, resulting in suboptimal segmentation performance. However, when the two methods are combined and deeply integrated as the proposed modules (SA-LSTM and LSTM-SA), the results are significantly improved. This highlights the complementary nature of SA and LSTM, indicating that both methods contribute unique strengths to the model. Meanwhile, the ablation study on the combination strategies reveals that the linear combination of LSTM states for QKV computation in SA-LSTM and LSTM-SA both outperform the non-linear combination. This result underscores the importance of leveraging the inherent properties of LSTM states to enhance the model's ability to capture global dependencies effectively. The linear combination strategy in LSTM-SA enables the model to leverage the LSTM states' historical information retention and state compression capabilities, resulting in improved segmentation performance. Analyzing the impact of these combinations on the sparse coding process reveals that non-linear combinations with activation functions elevate the optimization challenge due to the introduction of non-linearity and the potential for creating more complex decision boundaries, while the linear combination alone already affords substantial expressiveness.

3) Comparison of Module Embbeding Strategies: We compare different embedding strategies for 2D data inputs on the ISIC2018 dataset, as detailed in Table V. The optimal performance occurs when the module is placed between the encoder and decoder. Other strategies also produce satisfactory outcomes. We analyze the reasons for these results from two theoretical perspectives. First, from the perspective of contextual information integration, placing the module between the encoder and decoder allows for the effective integration of contextual information from deeper feature representation. This positioning enables the model to refine its segmentation decisions based on a more comprehensive understanding of the image context, leading to improved segmentation accuracy. This observation aligns with the theoretical underpinning of U-shape networks, which aim to capture both local and global features for accurate segmentation. Second, from the perspective of feature reuse and refinement, by inserting the module at different stages of the U-shape network, we are essentially altering the model's ability to reuse and refine features. Placing the module between the skip connection and the upsampling layer, for instance, allow the model to refine features at a more abstract level, potentially capturing higher-level semantic information compared to other strategies.

F. Effect on Different Baselines and Statistical Significance Analysis

To demonstrate the generalizability of our proposed module, we perform supplementary experiments on 2D and 3D inputs. We adopt five baselines for each dataset to evaluate if integrating our module improves prediction accuracy. Notably, some methods are not open source, so we reproduce them from descriptions in their respective papers. The data shown in this paper come from experiments using these independently reproduced code.

The results from the experiments on both 2D and 3D inputs, as shown in Table VI, Table VII, demonstrate the generalizability of our proposed module. Across all baselines,

TABLE III

EFFICIENCY COMPARISONS BETWEEN OUR PROPOSED MODULES 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 INFERENCE TIME DENOTES THE TIME TAKEN FOR GPU INFERENCE

Matrices	Dataset	Baseline	Separated SA and LSTM	SA-LSTM (Ours)	LSTM-SA (Ours)
Params (M)	Synapse	146.51	147.62 (+0.75%)	147.62 (+0.75%)	147.67 (+0.79%)
	ISIC2018	2.59	2.71 (+4.63%)	2.70 (+4.24%)	2.74 (+5.79%)
	ACDC	49.31	49.52 (+0.41%)	49.51 (+0.40%)	49.57 (+0.52%)
	CVC-ClinicDB	3.53	3.64 (+3.11%)	3.63 (+2.83%)	3.67 (+3.97%)
Flops (G)	Synapse	28.32	29.44 (+3.95%)	29.32 (+3.54%)	29.42 (+3.84%)
	ISIC2018	15.88	16.98 (+6.92%)	16.79 (+5.73%)	17.01 (+7.11%)
	ACDC	14.35	15.15 (+5.57%)	15.16 (+5.64%)	15.42 (+7.46%)
	CVC-ClinicDB	0.48	0.68 (+41.67%)	0.64 (+33.33%)	0.69 (+43.75%)
Inference Time (ms)	Synapse	45.41	46.51 (+2.42%)	46.21 (+1.76%)	46.53 (+2.46%)
	ISIC2018	10.95	11.97 (+9.31%)	11.75 (+9.31%)	12.02 (+9.77%)
	ACDC	33.91	35.21 (+3.83%)	35.01 (+3.24%)	35.32 (+4.16%)
	CVC-ClinicDB	5.19	5.88 (+13.29%)	5.59 (+7.07%)	6.01 (+15.79%)

TABLE IV

ABLATION STUDY ON ACDC DATASET. DICE SCORES (%) AND HD95 ARE REPORTED. ✓ DONATES USED, AND ✗ DONATES NOT USED. THE BEST RESULTS ARE IN BOLD. THE SECOND BEST RESULTS ARE UNDERLINED. LINEAR COMB. AND NON-LINEAR COMB. DENOTE THE LINEAR AND NON-LINEAR COMBINATION STRATEGIES IN EQ. 4

Method	SA	LSTM	DICE ↑	HD95 ↓
MT-UNet [58] (Baseline)	✗	✗	90.43	2.23
Only SA	✓	✗	91.17	1.77
Only LSTM	✗	✓	92.69	1.76
Separate SA and LSTM	✓	✓	92.73	1.45
SA-LSTM (Linear Comb.)	✓	✓	<u>93.06</u>	<u>1.05</u>
SA-LSTM (Non-Linear Comb.)	✓	✓	92.59	1.11
LSTM-SA (Linear Comb.)	✓	✓	**93.08**	**1.03**
LSTM-SA (Non-Linear Comb.)	✓	✓	92.64	1.10

TABLE V

COMPARISON STUDY ON ISIC2018 DATASET. DICE SCORES (%) AND HD95 ARE REPORTED. THE BEST RESULTS ARE IN BOLD. STRATEGIES 1, 2, 3, AND 4 DENOTE EMBEDDING THE MODULE IN PARALLEL WITH THE ENCODER, BETWEEN THE ENCODER AND DECODER, BETWEEN THE SKIP CONNECTION AND THE UPSAMPLING LAYER, AND AFTER THE DECODER, RESPECTIVELY

Method	DICE ↑	HD95 ↓
DCSAU-Net [57] (Baseline)	90.41	2.21
SA-LSTM Embbeding Using Strategy 1	90.52	2.54
SA-LSTM Embbeding Using Strategy 2	**91.17**	**2.02**
SA-LSTM Embbeding Using Strategy 3	90.42	2.68
SA-LSTM Embbeding Using Strategy 4	90.98	2.31
LSTM-SA Embbeding Using Strategy 1	90.60	2.53
LSTM-SA Embbeding Using Strategy 2	**91.08**	2.53
LSTM-SA Embbeding Using Strategy 3	90.89	**2.01**
LSTM-SA Embbeding Using Strategy 4	90.09	2.32

our model consistently exhibits significant improvements in all evaluation metrics, highlighting the effectiveness and versatility of the proposed module. The variation in improvement observed across different baselines can be attributed to the varying degrees to which each baseline already considers global dependencies. Baselines that have already addressed global dependencies effectively, such as MERIT with a cascaded attention decoder, DAEFormer with dual attention guided transformer, and MISSFormer with efficient SA applied to high-resolution feature maps, show comparatively less improvement with the introduction of our proposed module. On the other hand, for baselines that do not inherently consider global dependencies, such as TransUNet with a raw ViT model and MT-UNet with only axial attention to reduce time complexity, the improvement achieved by our model is substantial. This observation reinforces the significance of our proposed module in enhancing models that lack global context awareness.

The integration of statistical significance testing in this study plays an important role in providing a thorough and objective validation of the proposed modules' performance across various datasets. In segmentation tasks, even minor improvements in metrics like DICE or HD95 can seem noteworthy; however, without proper statistical analysis, it becomes challenging to ascertain whether these improvements are genuinely meaningful or merely a result of random variation. We selected the paired t-test for our analysis, given the nature of our experimental setup, which involves evaluating each dataset under two different conditions: before and after the application of the proposed modules. This test is particularly appropriate for our context, as it considers the correlation between paired observations, thereby offering a more nuanced assessment of whether the performance differences are statistically significant. In our experimental design, we calculated the t-statistic and the corresponding p-value for each dataset, using a conventional significance level of 0.05. As shown in Table VIII, the resulting p-values are consistently below this threshold, suggesting that the observed performance improvements are statistically significant. This systematic approach to validation is particularly relevant in the realm of medical image segmentation, where even modest enhancements may contribute to improved clinical outcomes. By establishing that the observed improvements are statistically significant, this analysis not only highlights the reliability of the proposed modules but also supports their applicability across different datasets and baseline models. The low p-values across all datasets indicate that the improvements are not limited to specific datasets, suggesting broader potential for effectiveness in various applications. This careful validation process ultimately contributes to the credibility and relevance of the study's findings.

G. Input Perturbation Analysis

We conduct an input perturbation analysis to evaluate the robustness of our proposed modules against input

TABLE VI

RESULTS ON 2D INPUT DATASET. DICE AND PRECISION SCORES (%) ARE REPORTED. ↑ DONATES DICE INCREASE LESS THAN 1%, ⇑ DONATES DICE INCREASE MORE THAN 1%

Synapse				ISIC2018			
Method	DICE ↑	SA-LSTM added	LSTM-SA added	Method	DICE	SA-LSTM added	LSTM-SA added
TransUNet [60]	77.48	78.72 ⇑	78.77 ⇑	UNet [32]	87.41	89.57 ⇑	88.17 ↑
MT-UNet [58]	77.87	79.35 ⇑	78.40 ↑	DWUNet [61]	87.47	88.75 ⇑	88.37 ↑
MISSFormer [65]	80.32	80.92 ↑	80.91 ↑	ResUNet [62]	87.91	88.61 ↑	89.22 ⇑
DAEFormer [76]	81.87	81.98 ↑	82.47 ↑	UNet++ [64]	88.32	89.85 ⇑	89.98 ⇑
MERIT [56]	84.22	85.13 ↑	84.50 ↑	DCSAU-Net [57]	90.41	91.17 ↑	91.08 ↑

TABLE VII

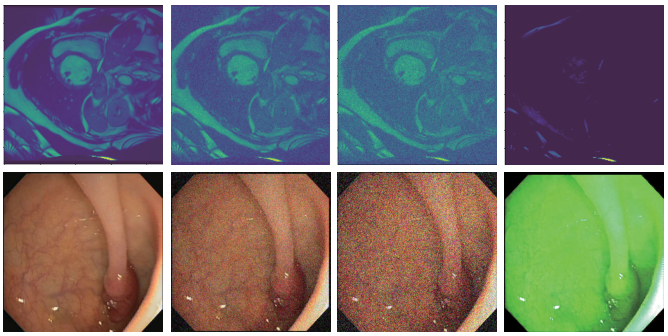
RESULTS ON 3D INPUT DATASET. DICE SCORES (%) ARE REPORTED. ↑ DONATES DICE INCREASE LESS THAN 1%, ⇑ DONATES DICE INCREASE MORE THAN 1%

ACDC				CVC-ClinicDB			
Method	DICE ↑	SA-LSTM added	LSTM-SA added	Method	DICE	SA-LSTM added	LSTM-SA added
MISSFormer [65]	87.85	90.25 ⇑	88.94 ⇑	FCBFormer [70]	92.53	92.98 ↑	92.81 ↑
TransUNet [60]	89.71	91.93 ⇑	91.86 ⇑	SSFormer-L [71]	92.68	93.41 ↑	92.95 ↑
DAEFormer [76]	90.34	91.33 ↑	90.96 ↑	HarDNet-DFUS [72]	93.32	93.73 ↑	94.58 ⇑
MT-UNet [58]	90.43	93.06 ⇑	93.08 ⇑	FANet [73]	93.55	95.73 ⇑	95.54 ↑
MERIT [56]	92.32	92.82 ↑	92.74 ↑	ESFPNet [59]	94.90	96.03 ⇑	96.35 ⇑

TABLE VIII

STATISTICAL SIGNIFICANCE TESTING RESULTS FOR DIFFERENT DATASETS WITH T-STATISTICS, P-VALUES, AND SIGNIFICANCE. A P-VALUE LESS THAN 0.05 INDICATES A STATISTICALLY SIGNIFICANT IMPROVEMENT

Dataset	Method	t-statistic	p-value	Significant
Synapse	SA-LSTM	3.6025	0.0227	Yes
ISIC2018	SA-LSTM	4.7868	0.0087	Yes
ACDC	SA-LSTM	4.1444	0.0143	Yes
CVC-ClinicDB	SA-LSTM	3.0024	0.0398	Yes
Synapse	LSTM-SA	3.9091	0.0174	Yes
ISIC2018	LSTM-SA	5.7060	0.0047	Yes
ACDC	LSTM-SA	3.1843	0.0334	Yes
CVC-ClinicDB	LSTM-SA	3.1037	0.0361	Yes



(a) image (b) weakly perturbed (c) moderately perturbed (d) strongly perturbed

Fig. 8. The sample images of input perturbation analysis. (a) Original image. (b) Image with noise level 0.1 (weakly perturbed). (c) Image with noise level 0.2 (moderately perturbed). (d) Image with strong noise level (strongly perturbed).

variations using ACDC and CVC-ClinicDB dataset. The analysis involves introducing different level noise to the input images and assessing the model's performance under these perturbed conditions. As shown in Fig. 8, we introduce three levels of noise to the input images: 0.1, 0.2, and strong. The first two noise are added to the input images by applying gaussian noise with a standard deviation of 0.1, 0.2, respectively. The strong noise is added by applying a random color jitter transformation to the input images.

Table IX presents the input perturbation analysis results, showing that our modules are more robust against input perturbations. For example, when the input images are perturbed with a noise level of 0.2, our model achieves a DICE score of 84.16% and 86.41%, while the baseline models only achieve accuracies of 81.02%. Such results demonstrates that our model retains higher accuracy under noisy conditions with less performance degradation than baseline models. We attribute this to the benefits of using LSTM to introduce historical information into SA, which updates the rules for handling context dependencies. This combination allows the model to better utilize historical information to handle input variations, thereby enhancing the model's robustness.

H. Convergence Results and Attention Visualization

Fig. 9 provides a comparison of the convergence behavior of two baselines that utilize the same method to calculate the loss. The loss function is a combination of the cross-entropy (CE) loss and the DICE loss, calculated as $Loss = 0.5 \cdot Loss_{DICE} + 0.5 \cdot Loss_{CE}$.

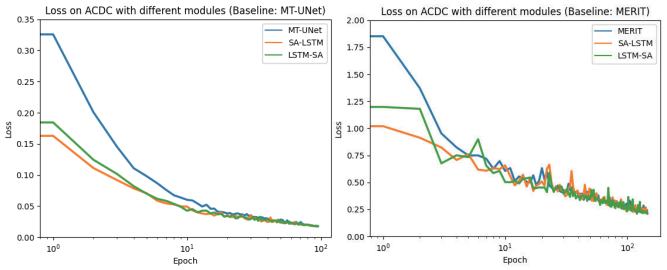
The results from Fig. 9 suggest that the addition of our proposed modules appears to support the model's ability to converge faster during training. The observed improvement in convergence may be related to the ability of our proposed modules to more effectively represent global dependencies. By capturing global contextual information, the model can potentially learn more meaningful features in the early stages of training, which might facilitate faster convergence.

Fig. 10 illustrates the attention maps generated by our proposed modules using different linear combination strategies. We get the attention map by indexing the layers using natural numbers. The activation maps at a given index is a function of the form $f : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{H' \times W' \times C}$. After getting the activations in the shape (H', W', C), we represent what parts of the image is the activation paying attention to the

TABLE IX

INPUT PERTURBATION ANALYSIS ON ACDC AND CVC-CLINICDB DATASET. DICE SCORES (%) AND HD95 ARE REPORTED. MINIMAL DEGRADATION IS BOLDED. SUB-MINOR DEGRADATION IS UNDERLINED

Method	Noise level	DICE ↑ delta	HD95 ↓ delta
MT-UNet [58]	0.1	86.21 (-4.22)	3.04 (+0.81)
SA-LSTM	0.1	**89.27 (-3.79)**	**1.27 (+0.22)**
LSTM-SA	0.1	89.13 (-3.95)	1.79 (+0.76)
MT-UNet [58]	0.2	81.02 (-9.41)	4.55 (+2.22)
SA-LSTM	0.2	84.16 (-8.90)	**1.85 (+0.80)**
LSTM-SA	0.2	**86.41 (-6.67)**	1.99 (+0.96)
MT-UNet [58]	strong	12.54 (-77.89)	29.45 (+27.22)
SA-LSTM	strong	17.59 (-75.47)	25.46 (+24.41)
LSTM-SA	strong	**19.22 (-73.86)**	**25.11 (+24.08)**
ESFPNet [59]	0.1	94.35 (-0.55)	1.42 (+0.21)
SA-LSTM	0.1	**95.95 (-0.08)**	**0.49 (+0.01)**
LSTM-SA	0.1	96.15 (-0.20)	0.45 (+0.04)
ESFPNet [59]	0.2	93.04 (-1.86)	1.53 (+0.32)
SA-LSTM	0.2	95.67 (-0.36)	**0.55 (+0.07)**
LSTM-SA	0.2	**96.13 (-0.22)**	0.53 (+0.12)
ESFPNet [59]	strong	92.65 (-2.25)	2.38 (+1.17)
SA-LSTM	strong	**95.02 (-1.01)**	**1.37 (+0.89)**
LSTM-SA	strong	95.07 (-1.28)	1.39 (+0.98)



(a) MT-UNet (b) MERIT

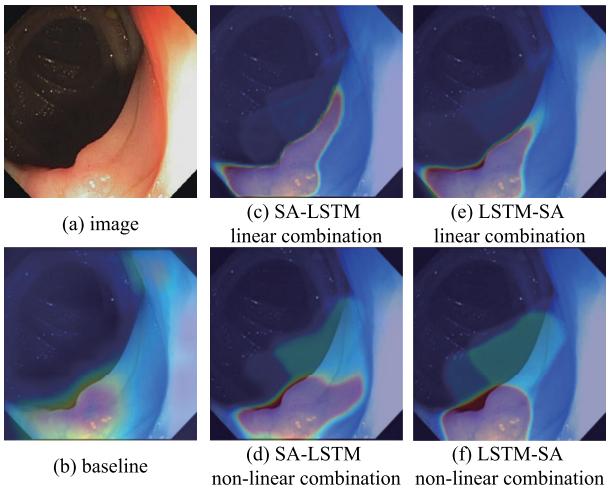
Fig. 9. The convergence of different baselines on ACDC dataset. (a) Using MT-UNet as baseline. (b) Using MERIT as baseline.

most by converting the shape of the activations to the form $g : \mathbb{R}^{H' \times W' \times C} \rightarrow \mathbb{R}^{H' \times W'}$. Formally, given an activation A we define $A_i = A[:, :, i]$ which represents an index across the channel dimension, we evaluate the attention map as $g(A) = \sum_{i=1}^{C} |A_i|$ [77].

The attention maps provide insights into the model's decision-making process and illustrate the regions of the input image that are most relevant for segmentation. The attention maps generated by SA-LSTM and LSTM-SA demonstrate a more focused and precise attention distribution compared to the baseline attention map. This improvement in attention distribution could be contributed to the model's enhanced ability by capturing the global dependencies and leveraging the contextual information. Moreover, the attention maps generated using the linear combination strategy in SA-LSTM and LSTM-SA reveal a slightly more refined and detailed attention distribution, which suggests the potential effectiveness of this strategy in enhancing the model's attention mechanism.

I. Validation of LSTM State Combinations for Self-Attention Mechanism

In Section III, we subjectively analyse the logic of combining LSTM states and QKV matrices through their meanings. To substantiate the effectiveness of our proposed LSTM state



(a) image (c) SA-LSTM linear combination (e) LSTM-SA linear combination

(b) baseline (d) SA-LSTM non-linear combination (f) LSTM-SA non-linear combination

Fig. 10. Comparison of the attention map using different linear combination strategies on CVC-ClinicDB dataset. (a) Original image. (b) Baseline attention map. (c) Attention map using linear combination strategy in SA-LSTM. (d) Attention map using non-linear combination strategy in SA-LSTM. (e) Attention map using linear combination strategy in LSTM-SA. (f) Attention map using non-linear combination strategy in LSTM-SA. Linear combination and non-linear combination refer to the combination strategies in Eq. 4.

combinations for the Self-Attention mechanism, we conduct a series of experiments to validate the following conclusions:

- The combination of f state and c state effectively integrates past and present information.
- The combination of f state and c state as Q matrix enhances sparse coding; similarly, the combination of g and h as K matrix enhances sparse coding.
- The combination of f state and c state predominantly contains shallow features, whereas the combination of g and h predominantly contains deep features.
- The combination of (f, c), (g, h), and h performs better than the combination of h, h, and h.

To validate these conclusions, we use the ACDC dataset and adopt various methods such as t-SNE [78] visualization, sparsity analysis using L_1 norm, and performance comparison on the DICE metric.

1) Validation of f and c Combination for Integrating Past and Present Information: To validate the effectiveness of the f state and c state combination in integrating past and present information, we perform a detailed analysis using the ACDC dataset. Specifically, we visualize the feature distributions of the LSTM states f and c, along with the feature extracted from original data, using t-SNE. This visualization is conducted on adjacent slices to observe how f and c capture temporal information from the previous and current slices.

As shown in Fig. 11, by comparing the feature distributions of previous and current slices, we observe two key findings. First, the feature distribution of c state is more aligned with the current slice's original data distribution, indicating that c state primarily focuses on present information. Second, the feature distribution of f contains significant information from the previous slice, demonstrating that f effectively inherits past information.

These observations support our hypothesis that the combination of f and c can effectively integrate past and present

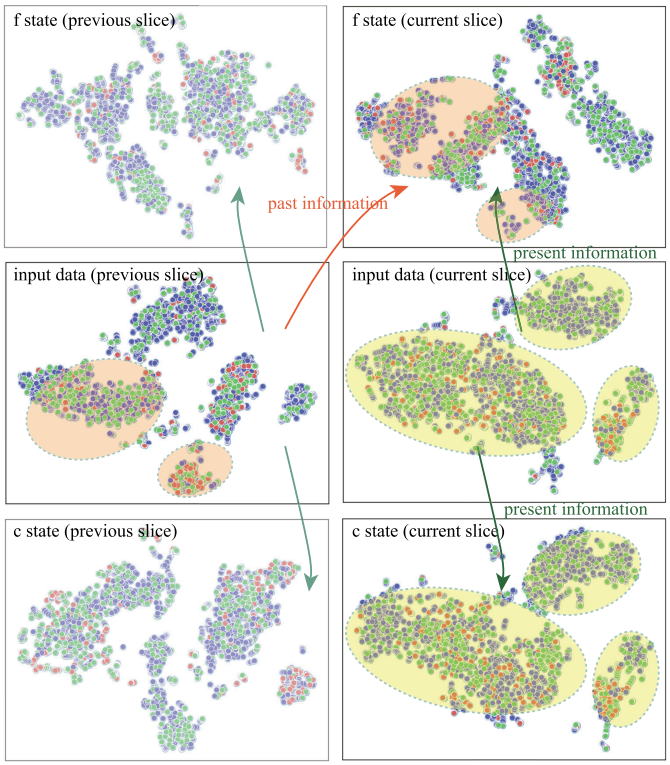


Fig. 11. t-SNE visualization of LSTM states on ACDC dataset. (a) f state, input data, and c state for previous slice from top to bottom. (b) f state, input data, and c state for current slice from top to bottom. Orange arrow indicates the past information flow, while green arrows indicate the present information flow.

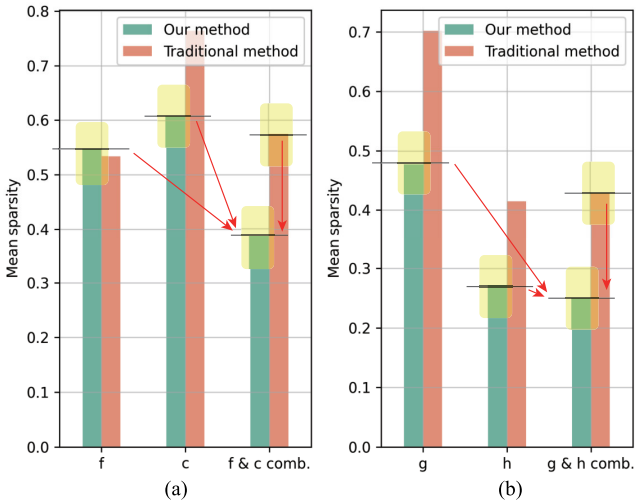


Fig. 12. Comparison of our method and conventional method based on sparsity analysis on ACDC dataset. (a) Sparsity analysis of h, f, c, and $f\&c$ states combination for Q matrix. (b) Sparsity analysis of h, g, and $g\&h$ states combination for K matrix. Red arrows indicate the enhancements.

information, making it suitable for generating the Q matrix in the SA mechanism. By averaging f and c and combining them, we obtain a feature representation that captures both the temporal context (past) and the current state (present), thus enhancing the model's ability to make informed predictions based on a comprehensive understanding of the data.

2) Validation of Enhanced Sparse Coding Ability in Q and K Matrices Through (f, c) and (g, h) Combinations: We conducte a comparative sparsity analysis using the L_1 norm. This analysis is performed on the ACDC dataset, where we compare the sparsity levels of individual states (f, c, g, h) and their combinations against a standard attention mechanism using h state for Q matrix, K matrix, and V matrix.

As illustrated in Fig. 12, we generate bar charts to compare the average sparsity of h, f, c, and $f\&c$ combination for Q matrix, and h, g, and $g\&h$ combination for K matrix, under the same input conditions. The results show that the L_1 norm sparsity of $f\&c$ combination is significantly lower than that of f state, and c state individually, indicating that the combination of f and c for Q enhances the ability of the model to focus on more relevant features while reducing redundancy. Similarly, the L_1 norm sparsity of $g\&h$ combination is lower than that of g state and h state, confirming that the combination of g and h for K matrix enhance the ability of the model to do sparse coding.

These results validate our hypothesis that combining f state and c state for Q matrix and g state and h state for K matrix leads to better sparse coding and more efficient feature representation in the SA mechanism, contributing to improved performance in medical image segmentation tasks.

Additionally, the bar charts in Fig. 12 include a comparison between our method and conventional methods that combine LSTM and SA. The results show that our approach achieves approximately a 20% improvement in sparsity over conventional methods for Q matrix and a 15% improvement for K matrix, which further demonstrates the effectiveness of our proposed method.

3) Validation of Shallow and Deep Features in (f, c) and (g, h) Combinations: To validate that the combination of f and c predominantly contains shallow features, while the combination of g and h predominantly contains deep features, we conduct a comprehensive analysis involving t-SNE visualization and gradient distribution comparison.

We utilize t-SNE to visualize the feature distributions of f, c, $f\&c$ states combination, g, h, and $g\&h$ states combination in a lower-dimensional space. This allows us to observe how these features are distributed and to determine whether they capture shallow or deep characteristics. The results, as shown in Fig. 13, indicate that the feature distributions of f state and c state are more dispersed and can clearly differentiate between different categories of data. This dispersion suggests that f state and c state are capturing more localized, detail-oriented shallow features. In contrast, the feature distributions of g state and h state are more concentrated, indicating that these states capture more abstract, global deep features. Moreover, the combinations $f\&c$ states combination and $g\&h$ states combination show intermediate characteristics. The $f\&c$ states combination retains the ability to differentiate between categories while becoming more structured, indicating a mix of shallow features with some higher-level abstraction. Similarly, $g\&h$ states combination shows a more refined distribution, integrating deeper features.

To further analyze the depth of features, we visualize the gradient distributions for $f\&c$ states combination and $g\&h$ states combination. Gradient distribution helps in understanding how sensitive these features are to input

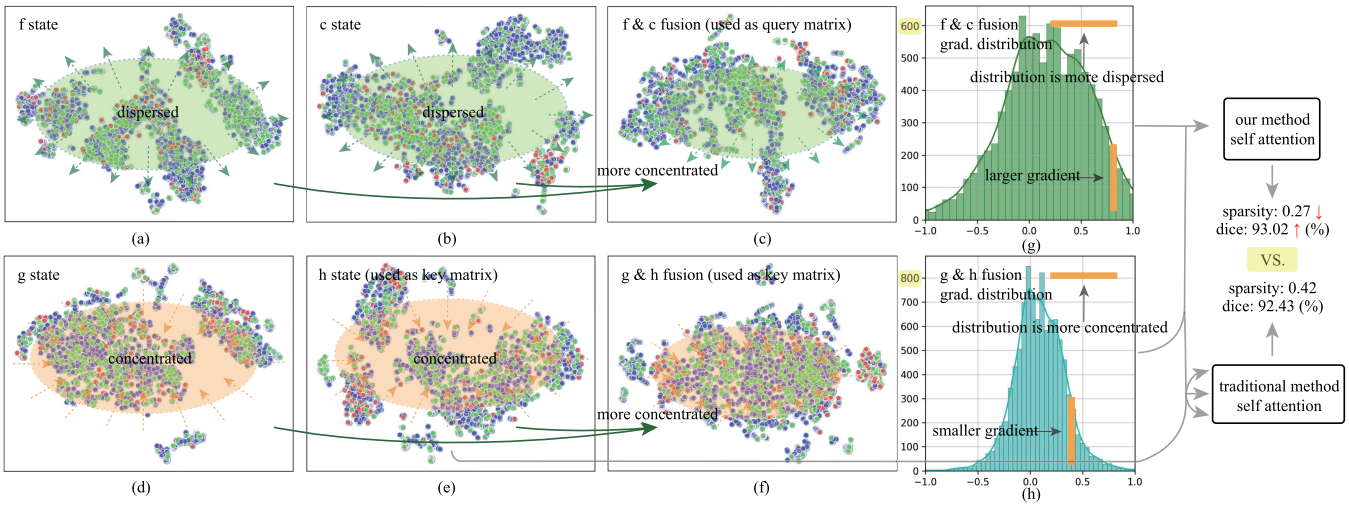


Fig. 13. t-SNE visualization and gradient distribution comparison of LSTM states on ACDC dataset. (a)-(f) t-SNE visualization of f, c, $f\&c$ states combination, g, h, and $g\&h$ states combination. (g)-(h) Gradient distribution comparison of $f\&c$ combination and $g\&h$ combination. Grey arrows indicate the composition of the states for Q, K, and V matrices.

variations, thereby indicating whether they capture shallow or deep information [79]. The results, also depicted in Fig. 13, suggest that gradients of $f\&c$ states combination are larger and more dispersed, reflecting higher sensitivity to input changes, which is characteristic of shallow features. While gradients of $g\&h$ states combination are smaller and more concentrated, indicating less sensitivity to input changes, a hallmark of deep features.

Through this detailed feature analysis, we can summarize the following points. The t-SNE visualization shows that the features from f state and c state are closer to the input data, capturing fine, local details. These features are dispersed, allowing clear differentiation between categories. The gradient distribution further supports this by showing higher sensitivity to input changes. The features from g state and h state capture more abstract, high-level information. Their t-SNE visualization shows a more compact distribution, indicating a comprehensive representation of the data's overall structure. The gradient distribution confirms this by showing lower sensitivity to input changes. The $f\&c$ states combination integrates both past and present information, creating a more structured shallow feature set, and the $g\&h$ states combination refines deep features, making them more representative of the global structure. Both combinations demonstrate intermediate characteristics that balance shallow and deep features effectively. These findings validate our hypothesis that f state and c state primarily contain shallow features, while g state and h state contain deep features. By combining these states, we achieve a balanced representation that captures both local details and global abstractions, enhancing the model's performance in medical image segmentation tasks. This comprehensive analysis highlights the effectiveness of our approach in leveraging the unique strengths of different LSTM states to improve feature representation and model accuracy.

4) Comparison of (f, c), (g, h), and h Combinations for Q, K, and V Matrices: To validate that the combination of f, c, g, and h for Q, K, and V matrices performs better than using h alone, we compare the sparsity and DICE metric results on the ACDC dataset.

Using the L_1 norm, we compare the sparsity of the (f, c), (g, h), and h combinations. The results, depicted in Fig. 12, show that the $f\&c$ combination for Q matrix significantly improve model's sparse coding ability compared to h alone. On the other hand, the $g\&h$ combination for K matrix also demonstrates lower sparsity than h alone. Meanwhile, we evaluate the DICE metric for both methods. The results in Fig. 13 (rightmost part) indicate that our method, using (f, c), (g, h), and h combination, outperforms the conventional method using h alone in both sparsity and DICE metric. These results validate that the combination of (f, c), (g, h), and h for Q, K, and V matrices enhances feature sparsity and model performance, resulting in better segmentation accuracy compared to using h alone.

V. Conclusion

We introduce a novel approach to medical image segmentation by rethinking the interplay between LSTM and SA mechanism through the view of sparse coding, and propose a novel methodology that harnesses the intrinsic capabilities of LSTM for state compression and historical data retention, integrating these aspects with SA's ability in capturing global dependencies. Our innovative modules, SA-LSTM and LSTM-SA, are meticulously designed to leverage the unique strengths of both LSTM and SA, providing a synergistic framework that enhances the segmentation process for 2D and 3D medical images, respectively. Experimental validations across multiple datasets demonstrate the superior performance of our approach, setting new benchmarks in medical image segmentation.

We emphasize that the combination of contextual modules is not restricted to the two proposed modules. Indeed, there are various ways to combine them, including leveraging other well-known modules such as multi-head SA and GRU. Moreover, while we focus on medical image segmentation tasks in this paper, we believe that our proposed modules can also be applied to other tasks, such as image classification and object detection. Looking forward, our future work will not only explore more sophisticated combinations of these

contextual modules but also investigate how these methods can be effectively compared and fused with foundation models at a finer granularity.

REFERENCES

[1] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "DRINet for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2453–2462, Nov. 2018.

[2] G. Wang et al., "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1562–1573, Jul. 2018.

[3] T. Hassanzadeh, D. Essam, and R. Sarker, "2D to 3D evolutionary deep convolutional neural networks for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 712–721, Feb. 2021.

[4] X. Wei, F. Ye, H. Wan, J. Xu, and W. Min, "TANet: Triple attention network for medical image segmentation," *Biomed. Signal Process. Control*, vol. 82, Apr. 2023, Art. no. 104608.

[5] K. Li, Y. Zhu, L. Yu, and P.-A. Heng, "A dual enrichment synergistic strategy to handle data heterogeneity for domain incremental cardiac segmentation," *IEEE Trans. Med. Imag.*, vol. 43, no. 6, pp. 2279–2290, Jun. 2024.

[6] D. N. Abdul Kareem, M. Fiaz, N. Novershtern, J. Hanna, and H. Cholakkal, "Improving 3D medical image segmentation at boundary regions using local self-attention and global volume mixing," *IEEE Trans. Artif. Intell.*, vol. 5, no. 6, pp. 3233–3244, Jun. 2024.

[7] S. Pang et al., "SpineParseNet: Spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 262–273, Jan. 2021.

[8] K. B. Girum, G. Créhange, and A. Lalande, "Learning with context feedback loop for robust medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 6, pp. 1542–1554, Jun. 2021.

[9] Z. Gu et al., "CE-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[10] J. Zhang, Y. Xie, Y. Wang, and Y. Xia, "Inter-slice context residual learning for 3D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 661–672, Feb. 2021.

[11] M. Afzali, A. Ghaffari, E. Fatemizadeh, and H. Soltanian-Zadeh, "Medical image registration using sparse coding of image patches," *Comput. Biol. Med.*, vol. 73, pp. 56–70, Jun. 2016.

[12] B. Bozorgtabar, M. Abedini, and R. Garnavi, "Sparse coding based skin lesion segmentation using dynamic rule-based refinement," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, Athens, Greece. Cham, Switzerland: Springer, Oct. 17, 2016, pp. 254–261.

[13] L. Zeng and K. Wu, "Medical image segmentation via sparse coding decoder," 2023, *arXiv:2310.10957*.

[14] E. Ahn, "Sparse coding for medical image analysis: Applications to image segmentation and classification," Master's thesis, Faculty Eng. Inf. Technol., School Inf. Technol., Univ. Sydney, Sydney, NSW, Australia, 2016.

[15] S. Zhang, Y. Zhan, and D. N. Metaxas, "Deformable segmentation via sparse representation and dictionary learning," *Med. Image Anal.*, vol. 16, no. 7, pp. 1385–1396, Oct. 2012.

[16] S. D. Salman Al-Shaikhli, M. Y. Yang, and B. Rosenhahn, "Brain tumor classification and segmentation using sparse coding and dictionary learning," *Biomed. Eng./Biomedizinische Technik*, vol. 61, no. 4, pp. 413–429, Aug. 2016.

[17] J. Tong, Y. Zhao, P. Zhang, L. Chen, and L. Jiang, "MRI brain tumor segmentation based on texture features and kernel sparse coding," *Biomed. Signal Process. Control*, vol. 47, pp. 387–392, Jan. 2019.

[18] Y. Li, Q. Dou, J. Yu, F. Jia, J. Qin, and P.-A. Heng, "Automatic brain tumor segmentation from MR images via a multimodal sparse coding based probabilistic model," in *Proc. Int. Workshop Pattern Recognit. Neuroimag.*, Jun. 2015, pp. 41–44.

[19] G. Sandhya, A. Srinag, G. B. Pantangi, and J. A. Kanaparthi, "Sparse coding for brain tumor segmentation based on the non-linear features," *J. Biomimetics, Biomater. Biomed. Eng.*, vol. 49, pp. 63–73, Feb. 2021.

[20] N. Moradi and N. Mahdavi-Amiri, "Kernel sparse representation based model for skin lesions segmentation and classification," *Comput. Methods Programs Biomed.*, vol. 182, Dec. 2019, Art. no. 105038.

[21] Y. Yu et al., "White-box transformers via sparse rate reduction," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 9422–9457.

[22] E. Ahn, A. Kumar, M. Fulham, D. Feng, and J. Kim, "Convolutional sparse kernel network for unsupervised medical image analysis," *Med. Image Anal.*, vol. 56, pp. 140–151, Aug. 2019.

[23] H. Tang et al., "CSC-Unet: A novel convolutional sparse coding strategy based neural network for semantic segmentation," *IEEE Access*, vol. 12, pp. 35844–35854, 2024.

[24] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *J. Fourier Anal. Appl.*, vol. 14, nos. 5–6, pp. 629–654, Dec. 2008.

[25] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, Jun. 2010, pp. 399–406.

[26] R. Zhang, J. Shen, F. Wei, X. Li, and A. K. Sangaiah, "Medical image classification based on multi-scale non-negative sparse coding," *Artif. Intell. Med.*, vol. 83, pp. 44–51, Nov. 2017.

[27] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[28] G. Paulo, T. Marshall, and N. Belrose, "Does transformer interpretability transfer to RNNs?" 2024, *arXiv:2404.05971*.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] X. Peng, C. Lu, Z. Yi, and H. Tang, "Connections between nuclear-norm and frobenius-norm-based representations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 218–224, Jan. 2018.

[31] J. T. Zhou et al., "SC2Net: Sparse LSTMs for sparse coding," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 4588–4595.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[33] J. Deng, W. Li, Y. Chen, and L. Duan, "Unbiased mean teacher for cross-domain object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sep. 2021, pp. 4091–4101.

[34] H. Huang et al., "Medical image segmentation with deep atlas prior," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3519–3530, Dec. 2021.

[35] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert, "Self-supervised learning for few-shot medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 7, pp. 1837–1848, Jul. 2022.

[36] R. Zheng, Y. Zhong, S. Yan, H. Sun, H. Shen, and K. Huang, "MsVRL: Self-supervised multiscale visual representation learning via cross-level consistency for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 1, pp. 91–102, Jan. 2023.

[37] H. Xu, Y. Song, Q. Liu, J. van Genabith, and D. Xiong, "Rewiring the transformer with depth-wise LSTMs," 2020, *arXiv:2007.06257*.

[38] S. Tang, C. Li, P. Zhang, and R. Tang, "SwinLSTM: Improving spatiotemporal prediction accuracy using swin transformer and LSTM," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 13470–13479.

[39] B. Peng et al., "RWKV: Reinventing RNNs for the transformer era," 2023, *arXiv:2305.13048*.

[40] S. Feng et al., "CPFNet: Context pyramid fusion network for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3008–3018, Oct. 2020.

[41] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[42] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Commun.*, vol. 15, no. 1, p. 654, Jan. 2024.

[43] C. Chen et al., "MA-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation," *Med. Image Anal.*, vol. 98, Dec. 2024, Art. no. 103310.

[44] B. Towle, X. Chen, and K. Zhou, "SimSAM: Zero-shot medical image segmentation via simulated interaction," 2024, *arXiv:2406.00663*.

[45] B. Xie, H. Tang, B. Duan, D. Cai, and Y. Yan, "MaskSAM: Towards auto-prompt SAM with mask classification for medical image segmentation," 2024, *arXiv:2403.14103*.

[46] K. Cao, T. Zhang, and J. Huang, "Advanced hybrid LSTM-transformer architecture for real-time multi-task prediction in engineering systems," *Sci. Rep.*, vol. 14, no. 1, p. 4890, Feb. 2024.

[47] 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, Sep. 2023.

[48] 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," *Knowl.-Based Syst.*, vol. 293, Jun. 2024, Art. no. 111637.

[49] H. Zhang, Z. Wang, and H. Vallery, "Learning-based NLOS detection and uncertainty prediction of GNSS observations with transformer-enhanced LSTM network," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, vol. 8, Sep. 2023, pp. 910–917.

[50] X. Kang, F. Han, A. Fayjie, and D. Gong, "FocDepthFormer: Transformer with LSTM for depth estimation from focus," 2023, *arXiv:2310.11178*.

[51] Z. Yang, H. Wu, Q. Liu, X. Liu, Y. Zhang, and X. Cao, "A self-attention integrated spatiotemporal LSTM approach to edge-radar echo extrapolation in the Internet of Radars," *ISA Trans.*, vol. 132, pp. 155–166, Jan. 2023.

[52] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space Odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2016.

[53] N. C. F. Codella 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 *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 168–172.

[54] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, Aug. 2018.

[55] 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 Med. Imag. Graph.*, vol. 43, pp. 99–111, Jul. 2015.

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

[57] Q. Xu, Z. Ma, N. He, and W. Duan, "DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation," *Comput. Biol. Med.*, vol. 154, Mar. 2023, Art. no. 106626.

[58] H. Wang et al., "Mixed transformer U-Net for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2390–2394.

[59] 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," 2022, *arXiv:2207.07759*.

[60] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[61] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[63] H. Cao et al., "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 205–218.

[64] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, vol. 11045. Cham, Switzerland: Springer, 2018, pp. 3–11.

[65] X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, "MISSFormer: An effective transformer for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 5, pp. 1484–1494, May 2023.

[66] M. Zahangir Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-Net (R^2U-Net) for medical image segmentation," 2018, *arXiv:1802.06955*.

[67] M. M. Rahman and R. Marculescu, "Medical image segmentation via cascaded attention decoding," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 6222–6231.

[68] 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 Recognit.*, vol. 139, Jul. 2023, Art. no. 109524.

[69] D. Jha et al., "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *IEEE Access*, vol. 9, pp. 40496–40510, 2021.

[70] E. Sanderson and B. J. Matuszewski, "FCN-transformer feature fusion for polyp segmentation," in *Proc. Annu. Conf. Med. Image Understand. Anal.*, vol. 13413. Cham, Switzerland: Springer, 2022, pp. 892–907.

[71] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, and S. Song, "Stepwise feature fusion: Local guides global," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 110–120.

[72] 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," 2022, *arXiv:2209.07313*.

[73] N. K. Tomar et al., "FANet: A feedback attention network for improved biomedical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 9375–9388, Mar. 2022.

[74] N. K. Tomar, D. Jha, U. Bagci, and S. Ali, "TGANet: Text-guided attention for improved polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Sep. 2022, pp. 151–160.

[75] A. Tragakis, C. Kaul, R. Murray-Smith, and D. Husmeier, "The fully convolutional transformer for medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3649–3658.

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

[77] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2016, *arXiv:1612.03928*.

[78] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[79] T. S. Cho, C. L. Zitnick, N. Joshi, S. B. Kang, R. Szeliski, and W. T. Freeman, "Image restoration by matching gradient distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 683–694, Apr. 2012.

Zexuan Ji (Member, IEEE) received the B.E. degree in computer science and technology and the Ph.D. degree in pattern recognition and intelligence system from Nanjing University of Science and Technology (NUST), Nanjing, China, in 2007 and 2012, respectively. He is currently an Associate Professor with the School of Computer Science and Engineering, NUST. His current research interests include medical imaging, image processing, and pattern recognition.

Shunlong Ye received the bachelor's degree in software engineering from Nanjing Normal University. He is currently pursuing the master's degree with Nanjing University of Science and Technology. His research interests include the field of medical image processing.

Xiao Ma (Student Member, IEEE) received the M.Sc. degree from the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China, in 2021, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering. His research interests include weakly supervised learning and medical image processing.