TSESNet: Temporal-Spatial Enhanced Breast Tumor Segmentation in DCE-MRI Using Feature Perception and Separability

Jiezhou He[1] , **Xue Zhao**[2,3] , **Zhiming Luo**[4] * , **Songzhi Su**[4] , **Shaozi Li**[4] and **Guojun Zhang** [3,5]

[1]Institute of Artificial Intelligence, Xiamen University, China.

[2]National Institute for Data Science in Health and Medicine, Xiamen University, China.

[3]Fujian Key Laboratory of Precision Diagnosis and Treatment in Breast Cancer, Xiang'an Hospital of Xiamen University, School of Medicine, Xiamen University, China.

[4]Department of Artificial Intelligence, Xiamen University, China.

[5]Yunnan Cancer Hospital & The Third Affiliated Hospital of Kunming Medical University & Peking University Cancer Hospital Yunnan, China.

Abstract

Accurate segmentation of breast tumors in dynamic contrast-enhanced magnetic resonance images (DCE-MRI) is critical for early diagnosis of breast cancer. However, this task remains challenging due to the wide range of tumor sizes, shapes, and appearances. Additionally, the complexity is further compounded by the high dimensionality and ill-posed artifacts present in DCE-MRI data. Furthermore, accurately modeling features in DCE-MRI sequences presents a challenge that hinders the effective representation of essential tumor characteristics. Therefore, this paper introduces a novel Temporal-Spatial Enhanced Network (TSESNet) for breast tumor segmentation in DCE-MRI. TSESNet leverages the spatial and temporal dependencies of DCE-MRI to provide a comprehensive representation of tumor features. To address sequence modeling challenges, we propose a Temporal-Spatial Contrastive Loss (TSCLoss) that maximizes the distance between different classes and minimizes the distance within the same class, thereby improving the separation between tumors and the background. Moreover, we design a novel Temporal Series Feature Fusion (TSFF) module that effectively integrates temporal MRI features from multiple time points, enhancing the model's ability to handle temporal sequences and improving overall performance. Finally, we introduce a simple and effective Tumor-Aware (TA) module that enriches feature representation to accommodate tumors of various sizes. We conducted comprehensive experiments to validate the proposed method and demonstrate its superior performance compared to recent state-of-the-art segmentation methods on two breast cancer DCE-MRI datasets.

1 Introduction

Breast cancer is the leading cause of cancer death in women worldwide. Early detection of malignancy is crucial for improving the prognosis of breast cancer patients [Zhao *et al.*, 2023b]. Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is a non-invasive imaging technique that can reveals both temporal and spatial characteristics of the physiological tissue. It plays an important role in the diagnosis and staging of breast tumors [Zhou *et al.*, 2022]. The standard DCE-MRI protocol involves acquiring a precontrast 3D image prior to the injection of an intravenous contrast agent, followed by sequential acquisition of postcontrast 3D images. A representative illustration of DCE-MRI sequences can be seen in the first row in Fig. 1. The response to contrast enhancement varies between tumor and non-tumor tissues. Tumor tissues typically exhibit significantly higher enhancement and increased sensitivity in images compared to non-tumor tissues.

As shown in Fig. 1, the second and third rows illustrate two specific cases consisting precontrast MRI slice V_0 and postcontrast MRI slice $\{V_i, i \in 1 \cdots T\}$, along with the subtraction image between $V_i - V_0$. From the figure, we can find that the subtraction images can provide better tumor visualization (highlighted in the yellow box). However, temporal changes manifest differently at distinct time points. In Case 1 (second row), $V_5 - V_0$ is more appropriate for tumor segmentation, whereas the tumor area in $V_1 - V_0$ is incomplete. On the other hand, in Case 2 (third row), we can observe that the enhanced tumors and glands are adhesion in $V_8 - V_0$, and $V_3 - V_0$ is more suitable choice for tumor segmentation. Consequently, physicians are still unable to accurately differentiate tumors from other enhanced tissues, such as vessels and glands, based solely on intensity changes at a single time point in DCE-MRI.

Despite the extensive focus on general tumor segmentation, the research on breast tumor segmentation in DCE-MR images is relatively limited [Qi *et al.*, 2023; Milletari *et al.*, 2016]. Common methods usually perform the segmentation in a single 3D MRI image based on maximum intensity projection (MIP), which may ignore the temporal characteristics of tumors within the image sequence [Zhang *et al.*, 2019;

*Corresponding author: zhiming.luo@xmu.edu.cn

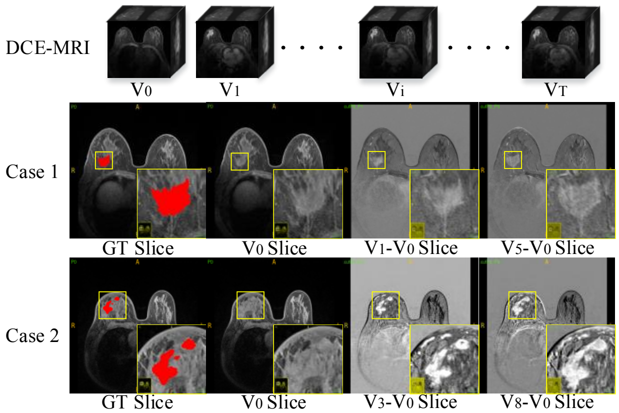

Figure 1: Illustration of the DCE-MRI and two cases. The first row shows a full DCE-MRI sequences, which contains a precontrast V_0 and T postcontrast $V_1 \cdots V_T$ (after injecting). For the second and third rows, V_0 slice is precontrast MRI slice sample, and $V_i - V_0$ slice is the subtraction between the i postcontrast and precontrast slice sample. GT is ground truth. The right corner is the local enlargement result within the yellow box.

Vidal *et al.*, 2022]. Recent methods have begun to explore the segmentation of breast tumors in 3D sequences [Zhang *et al.*, 2023; Lv and Pan, 2021; Lv *et al.*, 2022; Zhao *et al.*, 2023a]. However, the high dimension of the data poses challenges in modeling sequence features, and the tumor segmentation model consumes too much computing resources [Yeung *et al.*, 2022]. Consequently, achieving efficient and effective modeling of tumor features within the DCE-MRI sequence is crucial for accurate segmentation.

To this end, we propose a novel end-to-end Temporal-Spatial Enhanced Network (TSESNet) for DCE-MRI breast tumor segmentation. This approach enhances tumor feature representation by jointly considering the spatial and temporal contextual dependency of inter-sequence, enabling accurate segmentation of tumors of 4D DCE-MRI images. Specifically, the TSESNet utilizes a shared-weight encoder architecture and a Temporal Series Feature Fusion (TSFF) module to address the issue of excessive parameterization. The TSFF module efficiently integrates MRI feature maps from different time points while concurrently reducing the parameters. Additionally, we propose a novel temporal-spatial contrastive loss that constrains the feature representations of different classes in both spatial and temporal domains. The aim is to maximize the discrimination between tumor and background samples. Additionally, we also designed a Tumor-Aware (TA) module based on multi-scale-attention technical to improve the perception of tumors with different scales.

In summary, the main contributions of the proposed method are as follows:

- We proposed a novel TSESNet for breast tumor segmentation in DCE-MRI by exploiting spatial and temporal contextual dependencies. Extensive experiments on two breast DCE-MRI datasets demonstrate the effectiveness of TSESNet, achieving the state-of-the-art performance.
- We propose a Temporal-Spatial Contrastive Loss (TSCLoss), which captures spatial complexity and effectively models the temporal dynamics in DCE-MRI sequences. By addressing the distinctive challenges of capturing temporal-spatial features, TSCLoss elevates feature representation and significantly strengthens the performance in breast tumor segmentation tasks.
- We proposed a novel Temporal Series Feature Fusion (TSFF) module that effectively integrates temporal MRI features from multiple time points. This module enhances the model's ability to handle temporal sequences and improves overall performance.
- We designed a simple and effective Tumor-Aware (TA) module that enriches the feature representation to handle tumors of various sizes. Additionally, the attention mechanism and the residual-like structure in TA assist in minimizing the number of learnable parameters.

2 Related Works

2.1 Breast Tumor Segmentation

Breast tumor segmentation plays a vital role in early diagnosis of breast cancer. With the advent of deep learning, several CNN-based architectures, such as the popular U-Net, are widely used in breast tumor segmentation [Benjelloun *et al.*, 2018; Piantadosi *et al.*, 2018; Qin *et al.*, 2022; Haq *et al.*, 2022]. However, the above methods focus on 2D slices obtained from 3D MRI data, leading to the spatial contextual information missing. To handle this issue, 3D convolutional kernels are performed on 3D MRI volume data.

Vidal et al [Vidal *et al.*, 2022] proposed a 3D U-Net method for breast tumor segmentation in DCE-MRI. Wang et al [Wang *et al.*, 2021] proposed a 3D segmentation method using the comparison of precontrast and postcontrast MRI. Zhou et al [Zhou *et al.*, 2022] proposed a new multi-branch integrated network based on 3D affinity learning for accurate breast tumor segmentation in DCE-MRI. These methods usually rely on Maximum Intensity Projection (MIP) to convert the 4D DCE-MRI sequence into a single MRI. Although this simplification aids in computational efficiency, it neglects the crucial inter-sequence relationships within DCE-MRI.

Recently, some studies have begun to segment breast tumors throughout the DCE-MRI sequence [Zhang *et al.*, 2023; Zhao *et al.*, 2024]. Although these methods enforce global contextual dependencies, they often overlook the spatial and temporal correlations between sequences within DCE-MRI images. Additionally, the high dimensionality of DCE-MRI data aggravates the calculation, and the diversity in tumor shapes and sizes further complicates breast tumor segmentation, making it a challenging task.

2.2 Contrastive Learning

Supervised contrastive learning has emerged as a highly effective technique for representation learning. The key idea is generating positive and negative sample pairs, and then learning discriminative feature representations by minimizing the embedding distance among positive pairs and maximizing that among negative pairs [Zhang *et al.*, 2022;

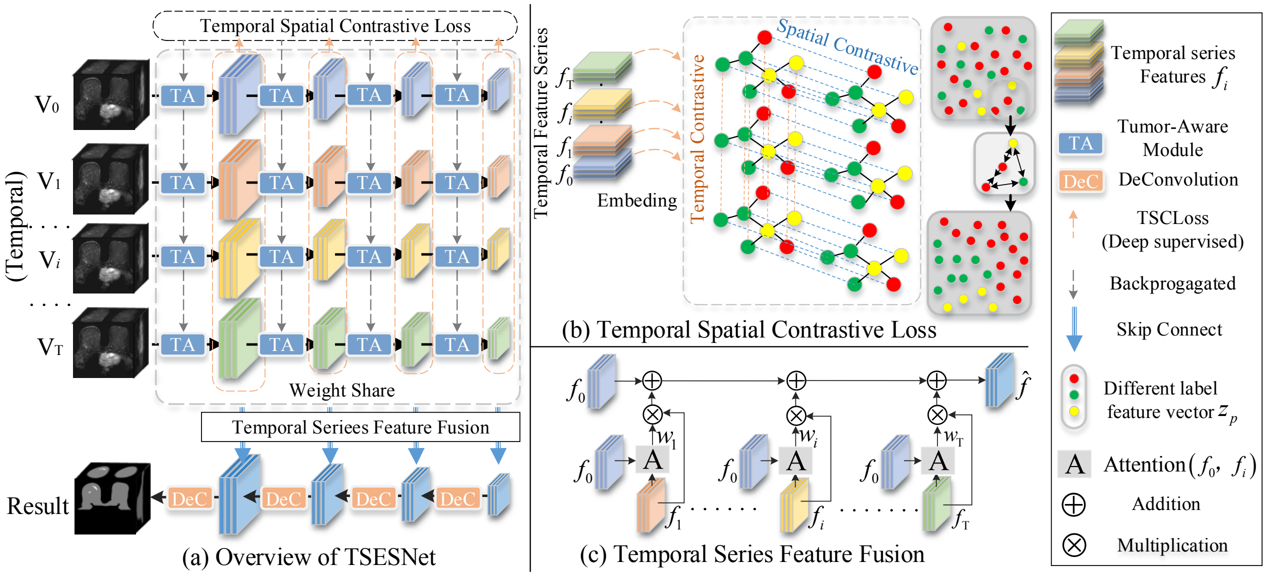

Figure 2: Architecture of the proposed TSESNet. (a) is the overall structure, (b) is a schematic diagram of temporal-spatial contrastive loss, and (c) is the temporal series feature fusion module

Lee *et al.*, 2023; You *et al.*, 2022]. By incorporating pixel-level supervision, these models can effectively capture discriminative features and thus improve segmentation accuracy.

Instance discrimination serves as an approach to supervised contrastive learning for medical image segmentation. In this approach, each pixel is treated as an instance, and the task is to learn the discrimination between pixels belonging to the same class and pixels from different classes. [Wang *et al.*, 2023] proposed a pixel-level contrastive branch to pretrain both the encoder and decoder. This offers the advantage of improving the reconstruction of high-resolution features at different scales. [Zeng *et al.*, 2021] propose a novel positional contrastive learning (PCL) framework, which generates contrastive data pairs by leveraging the position information within volumetric medical images. However, previous methods have primarily focused on either single-image segmentation or 3D volume segmentation, overlooking the spatial correlations between slices and the temporal correlations between sequences in DCE-MRI images.

3 Method

3.1 Overview

In Fig. 2 (a), the proposed TSESNet consists of four main components: Weight-share encoder, Temporal-Spatial Contrastive Loss, Temporal Series Feature Fusion, and Decoder. Specifically, for a DCE-MRI represented as $V = \{V_0, V_1, \ldots, V_T\}$, where T represents the number of contrast-enhanced samples, we apply a weight-shared encoder to each volume V_i. This encoder extracts diverse hierarchical feature maps $\mathbb{F} = F^l$ at different downsampling levels l. Then, these temporally sampled feature maps at the same level are merged into temporal-spatial features $\{f_i \in F^l, i \in 0 \ldots T\}$. The entire encoder leverages the Tumor-Aware (TA) module introduced in Sec. 3.2 to enhance the capture capability of tumor features across various shapes and sizes while reducing computational complexity. Simultaneously, the encoder is guided by the Temporal-Spatial Contrastive Loss, ensuring that the represented spatiotemporal feature sequence exhibits higher separability.

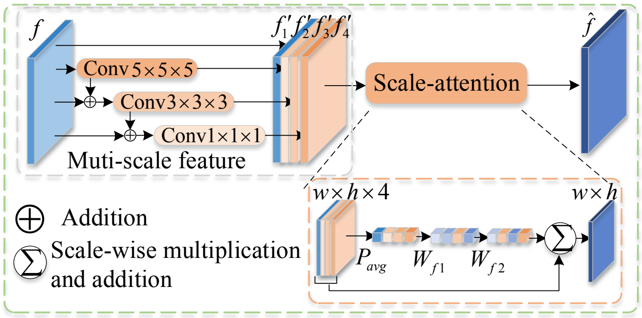

Figure 3: Tumor-Aware Module (TA).

Subsequently, by employing temporal attention mechanisms, the model effectively leverages inter-sequence correspondences while preserving the incremental nature of temporal context. As a result, the temporal feature sequence is fused into a spatiotemporal feature map. Finally, this fused feature map is propagated to the decoder to generate predictions for DCE-MRI tumor segmentation.

3.2 Tumor-Aware Module

Multi-scale information can provide rich semantic features for the segmentation of breast tumor with various sizes. Thus, we proposed a TA module that use attention strategy to fuse the channel dependence between different feature maps of varying resolutions and scales, as depicted in Fig. 3.

Specifically, we used three atrous convolutions with different kernel sizes (*i.e.*, $5 \times 5 \times 5, 3 \times 3 \times 3$, and $1 \times 1 \times 1$)

to extract multi-scale feature maps to increase the attention for tumors. Meanwhile, residual mechanisms are used to improve the communication between features, thus promoting the fusion. The detailed calculation is as follows:

$$f_i' = \begin{cases} f & i = 1 \\ \text{Conv}_1(f) & i = 2 \\ \text{Conv}_i(f + f_{i-1}') & 2 < i \leq 4. \end{cases} \tag{1}$$

$$f' = C[f_1', f_2', f_3', f_4'], \tag{2}$$

where $f \in f_i$ is a temporal-spatial feature map, $C[\cdot]$ is the concatenation operation.

The scale attention is used to weigh the feature channels at different scales, thus making the network focused on the effective features [Chen *et al.*, 2017]. The scale-attention mechanism is calculated as follows. Firstly, the global average pooling is used to embed the global spatial information into the scale vector:

$$g_s = \frac{1}{W \times H \times D} \sum_{d=1}^{D} \sum_{h=1}^{H} \sum_{w=1}^{W} f'(h, w), \tag{3}$$

where D, H, W represent the depth, height and width, respectively. f' is the multi-scale feature map.

The scale-attention weight vector is calculated based on the average pooling vector as follows:

$$w = Sigmoid(W_{f2}(ReLU(W_{f1}(g_s)))), \tag{4}$$

where W_{f1} and W_{f2} are the weights of two linear layers.

The final output feature map is a weighted sum of feature maps of different scales and their attention weights:

$$\hat{f} = \sum_{i=1}^{4} w_i * f'. \tag{5}$$

3.3 Temporal-Spatial Contrastive Loss

Breast tumor segmentation in DCE-MRI, it is important to consider not only the spatial relationship between individual MRIs, but also the temporal features across sequences. Here, we propose a novel Temporal-Spatial Contrastive Loss (TSCLoss) to enhance feature separability within the encoder. The goal of TSCLoss is to enforce closer feature distances for pixels of the same class in both spatial and temporal domains, while enlarging the feature distances of different classes. This helps enhance the discriminative capabilities of the encoder's feature maps, and thus increase the performance of breast tumor segmentation. The TSCLoss is shown in Fig. 2 (b).

Spatial Domain: In the spatial domain, we utilize the feature maps f_i at a given time point to encourage aggregation of features from the same class and dispersion of features from different classes. Consider the feature vectors $z_p \in R^d$ for each pixel p in f_i, where d is the dimensionality of the feature vectors. Pixel pairs of z_p and $z_{p'}$ belonging to the same class are regarded as positive pairs $(z_p, z_{p'}) \in \Omega^+$. Pixel pairs of z_p and z_q belong to different classes and are treated as negative pairs $(z_p, z_q) \in \Omega^-$. We use the cosine similarity $sim(z_i, z_j) = \frac{z_i \cdot z_j}{||z_i||||z_j||}$ to measure the similarity between feature vectors z_i and z_j within a single feature map f_i [Caliskan, 2023].

Temporal Domain: Across the temporal sequence f_0 to f_T, we treat pixels at the same spatial location as positive pairs, denoted as $(z_p, z_p^t) \in \Omega^+$. Here, z_p and z_p^t represent feature vectors for pixels at the same spatial location but different time points. Likewise, the cosine similarity $sim(z_p, z_p^t)$ measures the similarity between feature vectors z_p, z_p^t across different time points f_i and f_j.

Temporal-Spatial Contrastive Loss: The Temporal-Spatial Contrastive Loss (TSCLoss) aims to minimize the feature distance between positive pairs $(z_p, z_p') \in \Omega^+$ and maximize the feature distance between negative pairs $(z_p, z_q) \in \Omega^-$ in the spatial domain, as well as enforcing proximity between positive pairs $(z_p, z_p^t) \in \Omega^+$ across in the temporal domain. This is achieved by maximizing similarity between similar features and minimizing similarity between dissimilar features:

$$L(z, \widetilde{z}) = -\log \frac{e^{sim(z,\widetilde{z})/\tau}}{e^{sim(z,\widetilde{z})/\tau} + \sum_{\hat{z}\in\Omega^-} e^{sim(z,\hat{z})/\tau}}, \tag{6}$$

where Ω^- is the set of indices of negative samples. τ is the temperature used to smooth or sharpen the distribution.

To obtain the global loss of the entire training set, we averaged the single loss value of all pixel areas and applied the loss to different sampling layers (deep supervised):

$$L_{TSC} = \sum_{i=1}^{l} \frac{1}{|Z|} \sum_{z\in Z} \frac{1}{|\Omega^+|} \sum_{\widetilde{z}\in\Omega^+} L(z, \widetilde{z}), \tag{7}$$

where Ω^+ is the set of indices of positive samples, $|\Omega^+|$ stands for the size, $|Z|$ represents the sample count of pixel, l is the encoder layer number.

By minimizing the temporal-spatial contrastive loss function, we effectively constrain the spatial and temporal features to enhance separability, improving the representation of the encoded feature maps.

3.4 Temporal Series Feature Fusion

In order to effectively capture the temporal contextual dependency of MRI sequences and reduce model complexity. We propose the introduction of a Temporal-Spatial Feature Fusion (TSFF) module, which facilitates interactions among the extracted features and calculates their interdependencies, as illustrated in Fig. 2 (c). In clinical practice, radiologists typically confirm the presence of tumors by observing the dynamic changes between post-contrast MRIs and pre-contrast MRI scans. Leveraging this prior knowledge, we model the relationship between the features of post-contrast MRIs (f_i) and pre-contrast MRI (f_0) for feature fusion.

This module is designed to learn the temporal context and weights to fuse the features effectively. Specially, calculate attention weights based on the temporal dependencies between the sequential feature maps $f_1, \cdots, f_T$ with f_0 to emphasize the importance of each time step in the sequence. The relationship weight matrix W_i is calculated as follows:

$$W_i(f_0, f_i) = \text{Softmax}(\frac{f_0 W_q (f_0 W_k)^T}{d}) f_i W_v, \tag{8}$$

where W_k, W_q and W_v are corresponding learned weight matrices, d is feature dimension. $f_i \in \{f_1 \cdots f_T\}$

The final fused feature map $\hat{f}$ is obtained by weighting each f_i with its corresponding attention weight matrix W_i and summing them up:

$$\hat{f} = f_0 + \sum_{i=1}^{T}(W_i \cdot f_i), \tag{9}$$

where $\cdot$ denotes element-wise multiplication.

3.5 Loss Function

The model is trained end-to-end with dice loss and TSCLoss. The loss function is formulated as:

$$L = \alpha L_{Dice} + \beta \sum_{i=1}^{l} L_{TSC}, \tag{10}$$

where L_{Dice} is dice loss [Ma *et al.*, 2021], L_{TSC} is TSCLoss, and l is the number of encoder layers. In our experiment, we set $\alpha = 0.7, \beta = 0.5$, and $\tau = 10^{-5}$.

4 Experiment

4.1 Experimental Setup

Datasets: To assess the effectiveness of our proposed method for breast tumor segmentation, we conduct experiments on two datasets. The first dataset, known as AI-assistant-for-breast-tumor-segmentation (BTS) dataset [Zhang *et al.*, 2023], is a publicly available collection comprising 100 DCE-MRI cases from seven different institutions. The second one, named ST177, is a private collection consisting of 177 labeled DCE-MRI cases. In ST177, all patients underwent preoperative DCE-MRI using a 3.0 Tesla scanner (Signa, GE Healthcare, Milwaukee, WI) with a dedicated 8-channel breast coil. All included studies contained a fat-saturated gradient echo T1-weighted pre-contrast sequence and typically five post-contrast T1-weighted sequences acquired after the administration of the contrast agent. Each sequence contained 88 to 108 2D slices acquired with the following parameters: repetition time (TR) ranged [3.9, 4.8] ms, echo time (TE) ranged [1.7, 1.8] ms, flip angle = $5°$, field of view (FOV) = 340×340 mm, matrix size = 320×320, and slice thickness = 1.4 or 1.6 mm. Both datasets comprise 3D DCE-MRI volumes obtained from multiple contrasts ($V_0 - V_5$). These contrasts were acquired before the intravenous injection (V_0) and during the post-injection phase ($V_1 - V_5$) using a positive paramagnetic contrast agent at a dosage of 0.1 mmol/kg.

Preprocessing: Before training, we conducted denoising by removing pixels with values falling within the first and last 0.1% of the range. Following denoising, we performed grayscale normalization.

Implementation details: To ensure a fair comparison, all models were implemented using PyTorch and trained on four NVIDIA RTX 3090Ti GPUs. The ADAM optimizer was employed for optimization, with an initial learning rate of 1e-4. Each GPU had a batch size of 1. The training process lasted for 300 epochs, with an early stop criterion set at 30 steps. Following [Zhang *et al.*, 2023], sub-volumes of size 96× 96× 48 were randomly extracted from DCE-MRI scans as the input for the training stage. In the testing stage, we utilized a sliding window-based strategy, with the same window size as in the training stage.

Evaluation metrics: To evaluate the performance of our method, three commonly used metrics are utilized, *i.e.*, Dice Similarity Coefficient **(DSC: %)**, 95% Hausdorff Distance **(95% HD: mm)**, and Sensitivity **(Sen: %)**.

4.2 Comparisons With the State-of-the-Art Methods

We compare our TSESNet with seven existing DCE-MRI breast tumor segmentation methods. These methods include: **(1) 2D Slice-based:** MA-Net [Peng *et al.*, 2022] and Att-U-Node [Ru *et al.*, 2023], **(2) 3D U-Net based:** [Vidal *et al.*, 2022], **(3) Pre- and post-contrast enhancement MRIs as input-based:** ALMN [Zhou *et al.*, 2022] and Tumor-sen [Wang *et al.*, 2021], **(4) Whole DCE-MRI sequence as input-based:** ST-Tumor-Seg [Zhang *et al.*, 2023] and SwinHR [Zhao *et al.*, 2024].

Quantitative Comparison

ST177 Dataset: As shown in Tab. 1, the TSESNet achieves the highest DSC of 88.37% on the ST177 dataset, outperforming all comparison methods by a large margin. Additionally, TSESNet demonstrates superior performance in HD (7.154) and Sen. (83.19%), surpassing the existing state-of-the-art methods. Notably, methods designed for temporal analysis, such as SwinHR and ST-Tumor-Seg, leverage temporal information and achieve competitive results. However, the TSESNet utilizing temporal-spatial contrastive learning, exhibits a remarkable advancement, notably reducing the HD and improving Sensitivity significantly.

BTS Dataset: Similar trends can be observed on the BTS dataset, wherein TSESNet achieves the highest DSC of 78.95%. Further more, our proposed method demonstrates outstanding performance in terms of Hausdorff Distance (HD) with a value of 6.14 and Sensitivity (Sen) at 79.13%, surpassing the comparative approaches.

In summary, the experimental results on both ST177 and BTS datasets validate the efficacy of our proposed TSESNet for breast tumor segmentation in DCE-MRI. The incorporation of temporal-spatial contrastive learning contributes significantly to improved segmentation accuracy and delineation of tumor boundaries when compared to state-of-the-art methods. The consistent outperformance evaluation metrics underline the robustness and effectiveness of TSESNet in handling the challenges posed by breast tumor heterogeneity and complex DCE-MRI data.

Qualitative Comparison

In Fig. 4, our proposed TSESNet demonstrates highly consistent segmentation results that closely align with the ground truth. In contrast, 2D-based models, such as Att-U-Node and LMA-Net, exhibit intermittent slice losses, resulting partial segmentation discrepancies between slices. On the other hand, methods relying solely on single 3D MRI inputs, such as Tumor-sen, ALMN, and 3D U-Net, encounter difficulties in capturing inter-sequence relationships, leading to segmentation errors. In contrast, our TSESNet, leveraging the temporal series feature fusion module and temporal-spatial con-

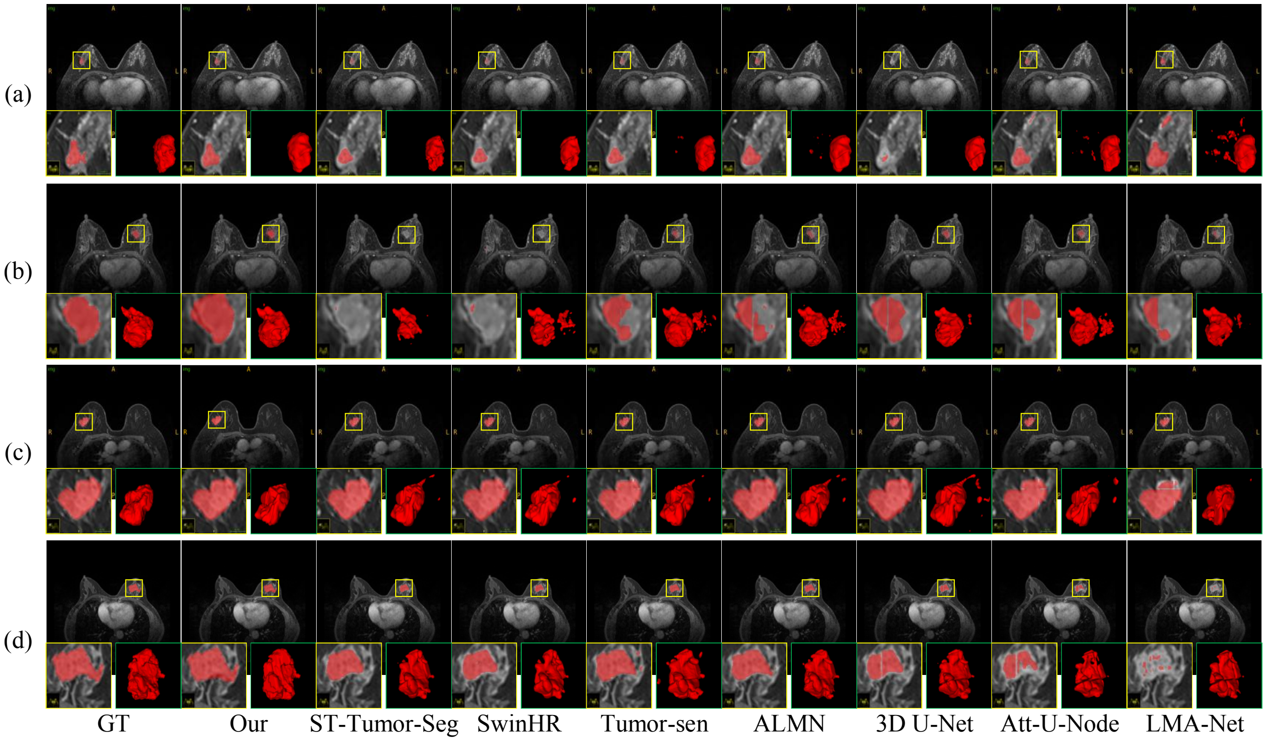

Figure 4: The qualitative segmentation results of two datasets, (a) and (b) from ST177, (c) and (d) from BTS. GT is Ground truth. The red shows the predictions of tumors. The lower left corner is the local enlargement result within the yellow box. The lower right corner is the 3D effect of tumor segmentation.

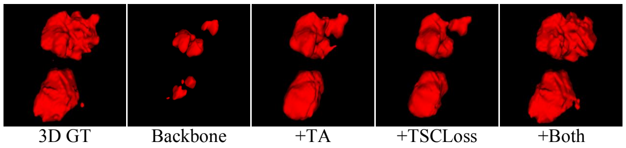

Figure 5: Ablation studies of core components of TA module and TSCLoss. GT: Ground truth.

trastive loss, achieves more accurate segmentation than other 4D segmentation models.

Parameters and Complexity

The analysis of model size and complexity (Tab. **??**) highlights the efficiency trade-offs among different 4D DCE-MRI segmentation methods. SwinHR exhibits a larger model size and high computational demand (146.20M parameters, 314.72G FLOPs). ST-Tumor-Seg, with a larger model size than SwinHR, showcases reduced computational complexity (175.54M parameters, 134.37G FLOPs). Our TSESNet achieves a smaller model size (103.58M parameters) and reduced computational complexity (98.66G FLOPs). This efficiency is attributed to the encoder layer with shared weights and the temporal feature fusion module.

4.3 Ablation Study

To investigate the effectiveness of different components in the proposed framework, we conduct a series of ablation studies on BTS dataset. Quantitative and qualitative comparisons are shown in Tab. 3 and Fig. 5.

Effectiveness of the TA Module: In Tab. 3, in the second line, the inclusion of the Tumor-Aware module leads to an increase in segmentation accuracy (DSC) from 63.41% to 70.14% by fusing features from different receptive fields. This module also refines boundary delineation (95% HD) from 21.179 to 13.192 and increase sensitivity from 66.36% to 71.47% by enhancing the perception of small tumors and complex boundaries. These improvements can be attributed to the strong learning ability of the Tumor-Aware module in capturing tumor features. Moreover, by effectively fusing information from multi-receptive fields, this module enhances its capability to discern subtle features of small tumors and intricate boundaries, resulting in a significant improvement in overall segmentation performance.

Effectiveness of the TSCLoss: In Tab. 3, in the third row, the utilization of Temporal-Spatial Contrastive Loss leads to an increase in DSC to 71.51% by enforcing constraints at the feature representation level, thereby enhancing the discriminative features between the tumor and background. Furthermore, while maintaining a low 95% HD (13.914), TSCLoss slightly improves sensitivity to 72.18%, thus enhancing the separability between tumor and non-tumor features.

Combined Impact: In Tab. 3, in the fourth line, the combination of both modules results in a significant performance boost, achieving a DSC of 78.95%, 95% HD of 6.14, and Sensitivity of 79.13%. The Tumor-Aware module enhances fea-

DataSets	Methods	Input Types	DSC (%) $\uparrow$	95% HD (mm) $\downarrow$	Sen. (%) $\uparrow$
ST177	LMA-Net [Peng *et al.*, 2022]	2D	77.38	26.193	74.19
	Att-U-Node [Ru *et al.*, 2023]	2D	81.25	12.917	76.54
	3D U-Net [Vidal *et al.*, 2022]	3D	70.88	43.305	70.58
	ALMN [Zhou *et al.*, 2022]	Two 3D	81.27	17.618	77.18
	Tumor-sen [Wang *et al.*, 2021]	Two 3D	84.16	15.009	77.36
	SwinHR [Zhao *et al.*, 2024]	4D	83.92	14.517	76.31
	ST-Tumor-Seg [Zhang *et al.*, 2023]	4D	84.39	14.037	80.51
	TSESNet	4D	**88.37**	**7.154**	**83.19**
BTS	LMA-Net [Peng *et al.*, 2022]	2D	63.16	35.708	64.49
	Att-U-Node [Ru *et al.*, 2023]	2D	65.17	24.317	68.47
	3D U-Net [Vidal *et al.*, 2022]	3D	59.35	59.537	61.91
	ALMN [Zhou *et al.*, 2022]	Two 3D	65.18	21.175	64.93
	Tumor-sen [Wang *et al.*, 2021]	Two 3D	67.13	17.413	65.81
	SwinHR [Zhao *et al.*, 2024]	4D	67.32	16.509	67.19
	ST-Tumor-Seg [Zhang *et al.*, 2023]	4D	69.81	14.170	71.03
	TSESNet	4D	**78.95**	**6.140**	**79.13**

Table 1: Quantitative comparison results on ST177 and BTS. For each column, the best result is highlighted in bold.

Methods	Params(M)	FLOPs(G)
SwinHR	146.20	314.72
ST-Tumor-Seg	175.54	134.37
TSESNet	103.58	98.66

Table 2: Model size and complexity.

Methods	DSC (%) $\uparrow$	95% HD $\downarrow$	Sen. (%) $\uparrow$
backbone	63.41	21.179	66.36
+TA module	70.14	13.192	71.47
+TSCLoss	71.51	13.914	72.18
+Both	78.95	6.140	79.13

Table 3: Ablation studies of core components of TA module and TSCLoss. The performance is evaluated on the BTS dataset.

ture perception by utilizing varied receptive fields. Simultaneously, the Temporal-Spatial Contrastive Loss improves discriminative feature representation. The combination of these two modules contributes collectively to the improved accuracy and boundary delineation in breast tumor segmentation from DCE-MRI images.

4.4 Performance Analysis for TSFF Module

To assess the impact of Temporal Series Feature Fusion (TSFF) on the performance of the TSESNet, we experimentally compared TSFF with four different feature fusion methods. 1) **MeanPool**: Temporal feature averaging in every skip connection to reduce dimensionality. 2) **CNN**: Adaptive temporal feature fusion. 3) **TSFF(Bottleneck)**: Incorporating TSFF in bottleneck layers, calculating and applying weights to skip connections. 4) **STT (Spatial-Temporal Transformer)**: Temporal feature fusion based on the ST-Tumor-Seg [Zhang *et al.*, 2023]. The detailed experimental results are reported in Tab. 4.

The results demonstrate the effectiveness of our proposed TSESNet with the TSFF module, surpassing alternative fusion strategies in terms of all evaluation metrics. Specifically, in terms of DSC, utilizing the TSFF module only at the bottleneck layer, the TSFF + Bottleneck approach achieved a significant increase in the DSC score (72.43) compared to STT (69.54). Despite obtaining a lower score than TSESNet, this improvement validates the contribution of the TSFF module in enhancing segmentation precision. In computational efficiency, the TSFF consumes only 98.93M parameters and performs 75.42G FLOPs, which is substantially less than STT at 169.17M parameters and 124.58G FLOPs. Notably, the Bottleneck+TSFF configuration yields competitive results, highlighting the pivotal role of TSFF in not only reducing model complexity, but also improving the model's ability to effectively capture sequential features in DCE-MRI. The comparison revealed a distinct advantage over existing approaches, such as STT, thereby demonstrating the effectiveness of our proposed TSFF module in enhancing the performance of breast tumor segmentation.

	DSC (%)	95% HD	Sen. (%)	Params(M)	FLOPs(G)
TSFF	78.95	6.140	79.13	103.58	98.66
MeanPool	66.31	18.157	64.19	98.31	67.42
CNN	63.17	24.160	61.83	101.16	72.32
STT	69.54	12.415	69.81	169.17	124.58
TSFF (Bottleneck)	72.43	11.325	72.54	98.93	75.42

Table 4: Comparison between TSFF module and other fusion strategies on the BTS dataset.

5 Conclusion

In this study, we propose a novel Temporal-Spatial Enhanced Network (TSESNet) for Breast Tumor Segmentation in DCE-MRI. Our approach leverages temporal-spatial contrastive learning to enhance the model's feature representation. It integrates contrast-enhanced features of breast tumors across different time points and incorporates a tumor-aware module that adapts to diverse tumor shapes and sizes. Additionally, the weight-sharing encoder structure and temporal series feature fusion module effectively capture sequence features in DCE-MRI while reducing the model's parameter count and complexity. Extensive evaluations on two DCE-MRI datasets validate the effectiveness, adaptability, and robustness of our proposed method in achieving accurate breast tumor segmentation.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62276221, No. 62376232), the Science and Technology Plan Project of Xiamen (No. 3502Z20221025), and the key technologies for intelligent medical equipment (NO.20223160A0621)

References

[Benjelloun *et al.*, 2018] Mohammed Benjelloun, Mohammed El Adoui, Mohamed Amine Larhmam, and Sidi Ahmed Mahmoudi. Automated breast tumor segmentation in dce-mri using deep learning. In *2018 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech)*, pages 1–6. IEEE, 2018.

[Caliskan, 2023] Aylin Caliskan. Artificial intelligence, bias, and ethics. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.

[Chen *et al.*, 2017] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017.

[Haq *et al.*, 2022] Imran Ul Haq, Haider Ali, Hong Yu Wang, Lei Cui, and Jun Feng. Bts-gan: computer-aided segmentation system for breast tumor using mri and conditional adversarial networks. *Engineering Science and Technology, an International Journal*, 36:101154, 2022.

[Lee *et al.*, 2023] Ho Hin Lee, Yucheng Tang, Qi Yang, Xin Yu, Leon Y Cai, Lucas W Remedios, Shunxing Bao, Bennett A Landman, and Yuankai Huo. Semantic-aware contrastive learning for multi-object medical image segmentation. *IEEE journal of biomedical and health informatics*, 2023.

[Lv and Pan, 2021] Tianxu Lv and Xiang Pan. Temporal-spatial graph attention networks for dce-mri breast tumor segmentation. 2021.

[Lv *et al.*, 2022] Tianxu Lv, Youqing Wu, Yihang Wang, Yuan Liu, Lihua Li, Chuxia Deng, and Xiang Pan. A hybrid hemodynamic knowledge-powered and feature reconstruction-guided scheme for breast cancer segmentation based on dce-mri. *Medical Image Analysis*, 82:102572, 2022.

[Ma *et al.*, 2021] Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L Martel. Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71:102035, 2021.

[Milletari *et al.*, 2016] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[Peng *et al.*, 2022] Chengtao Peng, Yue Zhang, You Meng, Yang Yang, Bensheng Qiu, Yuzhu Cao, and Jian Zheng. Lma-net: A lesion morphology aware network for medical image segmentation towards breast tumors. *Computers in Biology and Medicine*, 147:105685, 2022.

[Piantadosi *et al.*, 2018] Gabriele Piantadosi, Mario Sansone, and Carlo Sansone. Breast segmentation in mri via u-net deep convolutional neural networks. In *2018 24th international conference on pattern recognition (ICPR)*, pages 3917–3922. IEEE, 2018.

[Qi *et al.*, 2023] Wenbo Qi, HC Wu, and SC Chan. Mdf-net: A multi-scale dynamic fusion network for breast tumor segmentation of ultrasound images. *IEEE Transactions on Image Processing*, 2023.

[Qin *et al.*, 2022] ChuanBo Qin, JingYin Lin, JunYing Zeng, YiKui Zhai, LianFang Tian, ShuTing Peng, and Fang Li. Joint dense residual and recurrent attention network for dce-mri breast tumor segmentation. *Computational Intelligence and Neuroscience*, 2022, 2022.

[Ru *et al.*, 2023] Jintao Ru, Beichen Lu, Buran Chen, Jialin Shi, Gaoxiang Chen, Meihao Wang, Zhifang Pan, Yezhi Lin, Zhihong Gao, Jiejie Zhou, et al. Attention guided neural ode network for breast tumor segmentation in medical images. *Computers in Biology and Medicine*, 159:106884, 2023.

[Vidal *et al.*, 2022] Joel Vidal, Joan C Vilanova, Robert Martí, et al. A u-net ensemble for breast lesion segmentation in dce mri. *Computers in Biology and Medicine*, 140:105093, 2022.

[Wang *et al.*, 2021] Shuai Wang, Kun Sun, Li Wang, Liangqiong Qu, Fuhua Yan, Qian Wang, and Dinggang Shen. Breast tumor segmentation in dce-mri with tumor sensitive synthesis. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[Wang *et al.*, 2023] Yu Wang, Bo Liu, and Fugen Zhou. Pcmask: A dual-branch self-supervised medical image segmentation method using pixel-level contrastive learning and masked image modeling. In *Image and Vision Computing: 37th International Conference, IVCNZ 2022, Auckland, New Zealand, November 24–25, 2022, Revised Selected Papers*, pages 501–510. Springer, 2023.

[Yeung *et al.*, 2022] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95:102026, 2022.

[You *et al.*, 2022] Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(9):2228–2237, 2022.

[Zeng *et al.*, 2021] Dewen Zeng, Yawen Wu, Xinrong Hu, Xiaowei Xu, Haiyun Yuan, Meiping Huang, Jian Zhuang, Jingtong Hu, and Yiyu Shi. Positional contrastive learning for volumetric 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 II 24*, pages 221–230. Springer, 2021.

[Zhang *et al.*, 2019] Lei Zhang, Zhimeng Luo, Ruimei Chai, Dooman Arefan, Jules Sumkin, and Shandong Wu. Deep-learning method for tumor segmentation in breast dce-mri. In *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications*, volume 10954, pages 97–102. SPIE, 2019.

[Zhang *et al.*, 2022] Jiyang Zhang, Jianxiao Zou, Zhiheng Su, Jianxiong Tang, Yuhao Kang, Hongbing Xu, Zhiliang Liu, and Shicai Fan. A class-aware supervised contrastive learning framework for imbalanced fault diagnosis. *Knowledge-Based Systems*, 252:109437, 2022.

[Zhang *et al.*, 2023] Jiadong Zhang, Zhiming Cui, Zhenwei Shi, Yingjia Jiang, Zhiliang Zhang, Xiaoting Dai, Zhenlu Yang, Yuning Gu, Lei Zhou, Chu Han, et al. A robust and efficient ai assistant for breast tumor segmentation from dce-mri via a spatial-temporal framework. *Patterns*, 4(9), 2023.

[Zhao *et al.*, 2023a] Xiaoming Zhao, Yuehui Liao, Jiahao Xie, Xiaxia He, Shiqing Zhang, Guoyu Wang, Jiangxiong Fang, Hongsheng Lu, and Jun Yu. Breastdm: A dce-mri dataset for breast tumor image segmentation and classification. *Computers in Biology and Medicine*, 164:107255, 2023.

[Zhao *et al.*, 2023b] Xue Zhao, Jing-Wen Bai, Qiu Guo, Ke Ren, and Guo-Jun Zhang. Clinical applications of deep learning in breast mri. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, page 188864, 2023.

[Zhao *et al.*, 2024] Zhihe Zhao, Siyao Du, Zeyan Xu, Zhi Yin, Xiaomei Huang, Xin Huang, Chinting Wong, Yanting Liang, Jing Shen, Jianlin Wu, et al. Swinhr: Hemodynamic-powered hierarchical vision transformer for breast tumor segmentation. *Computers in Biology and Medicine*, page 107939, 2024.

[Zhou *et al.*, 2022] Lei Zhou, Shuai Wang, Kun Sun, Tao Zhou, Fuhua Yan, and Dinggang Shen. Three-dimensional affinity learning based multi-branch ensemble network for breast tumor segmentation in mri. *Pattern Recognition*, 129:108723, 2022.