# Hybrid-Structure-Oriented Transformer for Arm Musculoskeletal Ultrasound Segmentation

Lingyu Chen[1], Yue Wang[1], Zhe Zhao[2], Hongen Liao[3,4], Daoqiang Zhang[1], Haojie Han[3], Fang Chen[4(✉)]

[1] College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing 211106, China
[2] Orthopedics & Sports Medicine Center, Beijing Tsinghua Changgung Hospital. School of Clinical Medicine, Tsinghua University, Beijing 100084, China
[3] Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing 100084, China
[4] School of Biomedical Engineering and the Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, 200240, China
chenfang_bme@163.com

**Abstract.** Segmenting complex layer structures, including subcutaneous fat, skeletal muscle, and bone in arm musculoskeletal ultrasound (MSKUS), is vital for diagnosing and monitoring the progression of Breast-Cancer-Related Lymphedema (BCRL). Nevertheless, previous researches primarily focus on individual muscle or bone segmentation in MSKUS, overlooking the intricate and hybrid-layer morphology that characterizes these structures. To address this limitation, we propose a novel approach called the hybrid structure-oriented Transformer (HSformer), which effectively captures hierarchical structures with diverse morphology in MSKUS. Specifically, HSformer combines a hierarchical-consistency relative position encoding and a structure-biased constraint for hierarchical structure attention. Our experiments on arm MSKUS datasets demonstrate that HSformer achieves state-of-the-art performance in segmenting subcutaneous fat, skeletal muscle and bone. The code of our implementation is available at:

**Keywords:** Arm Musculoskeletal US Segmentation · Hybrid and Hierarchical Layer Structure · Horizontal and Curvilinear Morphology.

## 1 Introduction

Breast cancer-related lymphedema (BCRL) is a result of treatments for breast cancer, such as surgery or radiation, which cause damage to the lymphatic vessels. BCRL progresses from mild buildup of lymphatic fluid to irreversible edema in the subcutaneous tissue and skeletal muscles, leading to noticeable morphological changes [13]. The thickness of the skin, subcutaneous tissue, and skeletal muscles serve as significant biomarkers for diagnosing BCRL [5]. Traditional approach to assess BCRL using arm musculoskeletal ultrasound (MSKUS) involves the manual delineation of tissue layers by experts, which is inefficient and has

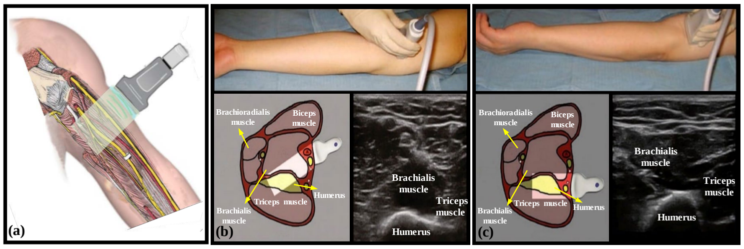

**Fig. 1.** Examples of arm MSKUS with a hybrid and hierarchical structure, including horizontal layers paralleling to the skin surface and curvilinear layers.

limited adoption [18], [6]. Therefore, automatic synchronous segmentation of soft tissues, skeletal muscle and bone structures is crucial for BCRL assessment.

Accurately segmenting arm MSKUS images for BCRL requires a more profound comprehension of the unique structures due to complex and hybrid-layer morphology. Specifically, on the one hand, tissue layers composed of distinct components exhibit varied structural shapes. *As shown in the lateral collection (Figure 1(a)) and arm MSKUS images (Figure 1(b)-(c)), hierarchical layers such as skin, subcutaneous fat, muscles and bones exist.* The upper layers of soft tissue are separated by horizontal or oriented echogenic lines, presenting *parallel layers*; surface of the bone below displays *horizontal boundaries or irregular curves* [5]. The structural bias caused by the above morphology characteristics significantly impact the segmentation performance, particularly ignoring the correlation and differences between horizontal and curvilinear layers. On the other, BCRL also emerges swelling of soft tissues and demonstrates *partially curved boundaries in the parallel layers*, which is caused by the accumulation of liquid volume and even the transformation into fibrous tissue, leading to tissue strain [20]. *Therefore, understanding the hierarchical structures with diverse morphology in MSKUS is key for segmenting subcutaneous fat, skeletal muscles and bones simultaneously.*

Existing MSKUS segmentation studies currently exhibit two main focuses: some emphasize muscle segmentation [14], [8], [7], aiming to quantitatively measure various parameters such as cross-sectional area and thickness, effectively facilitating the diagnosis and follow-up of neuromuscular diseases. Others concentrate on bone segmentation [11], [24], [23], leveraging shape or anatomical structure priors to enhance feature representation, promoting segmentation applications in scenarios like hip joints or spinal curvature. Nevertheless, **the simultaneous segmentation of soft tissue (such as fat and muscle) and bone structures in arm MSKUS [4] remains an area with limited research findings**. But the diagnosis of BCRL requires the segmentation of multiple layers of diverse tissue structures, in order to assess the thickness of the skin, subcutaneous tissue, and skeletal muscles. Effectively leveraging the structural bias, including the parallel structures exhibited by soft tissues, the irregular boundaries resulting from edema, and the curved structures on bone surfaces in the Figure 1, is paramount for achieving accurate arm MSKUS segmentation. Although segmentation frameworks for layer structures have been proposed [15], [16], [19], they are primarily tailored to optical coherence tomography (OCT), which significantly differs from arm MSKUS and limits the application, particularly when confronted with challenges such as soft tissue edema or deformations on the bone surface caused by BCRL.

To effectively address the aforementioned challenges, we propose a novel hybrid structure-oriented Transformer (HSformer) for arm MSKUS segmentation. This model captures effective feature representations for correlation and differences of the interlayer and intralayer structures from both horizontal and vertical perspectives, simultaneously. To address the challenge of curvilinear structures due to the inherent characteristics of the arm MSKUS and the effect of BCRL, hierarchical-consistency relative position encoding (HCPE) imposes structure bias onto the elements within the local window. Meanwhile, to amplify the differences among layers and focus on critical features, a structure-biased constraint (SBC) is designed to calculate attention weights. The contributions of our work can be summarized as follows:

1) We propose a hybrid structure-oriented Transformer framework to simultaneously segment skin, subcutaneous fat, skeletal muscles, and bones, in which the HCPE is designed to perceive and differentiate the structure features of horizontal and curvilinear layers, and the SBC is offered to harness the learned knowledge for boosting the representations.

2) Extensive experiments on arm MSKUS datasets justify the effectiveness of our proposed method. It also exhibits good generalization to an in-house small-scale MSKUS dataset.

## 2 Approach

The proposed HSformer explores an effective approach to represent hybrid structures that facilitates the differentiation of interlayer features and enhances reliance on intralayer characteristics for arm MSKUS segmentation. Figure 2 illustrates the three main components of the HSformer framework: 1) a hybrid structure-oriented Transformer block that extracts contextual features; 2) a hierarchical-consistency position embedding (HCPE) method that incorporates the bias of geometric morphology on the elements within vertical and horizontal local windows; and 3) a structure-biased constraint (SBC), which is designed to calculate attention weights and also compatible with HCPE.

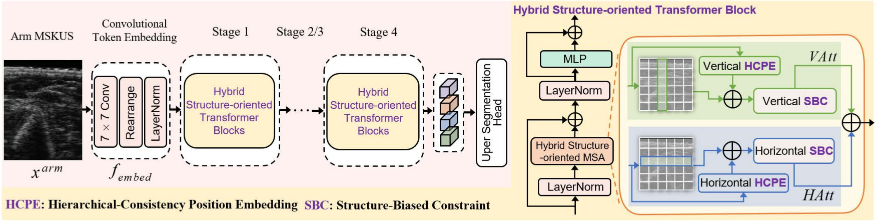
Arm MSKUS
Convolutional Token Embedding
Stage 1
Stage 2/3
Stage 4
7 × 7 Conv
Rearrange
LayerNorm
Hybrid Structure-oriented Transformer Blocks
Hybrid Structure-oriented Transformer Blocks
Uper Segmentation Head
$x^{arm}$
$f_{embed}$
HCPE: Hierarchical-Consistency Position Embedding
SBC: Structure-Biased Constraint
Hybrid Structure-oriented Transformer Block
MLP
LayerNorm
Hybrid Structure -oriented MSA
LayerNorm
Vertical HCPE
Vertical SBC
VAtt
Horizontal SBC
Horizontal HCPE
HAtt

**Fig. 2.** Illustration of the proposed HSformer framework.

### 2.1 Overall Architecture

The $7 \times 7$ convolutional layer with stride 4 takes an arm MSKUS image $x^{arm} \in \mathbb{R}^{H \times W \times 3}$ as input and generates patch token embedding $f_{embed} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$, where $C = 64$. There are four stages for a hierarchical representation in the HSformer, and each stage has $N_i$ sequential hybrid structure-oriented Transformer blocks. In two consecutive stages, the network leverages a $3 \times 3$ convolution layer with stride 2 to multiply the channel dimensions and reduce the number of tokens. Therefore, in the i-th stage, we acquire high-level feature maps with the size of $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times (2^{i-1})C, i = 1, 2, 3, 4$.

**The Hybrid Structure-Oriented Transformer** Arm MSKUS reveals a complex hybrid morphological structure, wherein soft tissues demonstrate structures parallel to the skin surface, while tissues with a musculoskeletal surface or edema have a curvilinear pattern. Horizontal or curvilinear structural features warrant particular attention. Therefore, we encode the geometric structure (the layout of each tissue) along the vertical direction and use horizontal strips in the local windows to acquire attention weights $HAtt \in \mathbb{R}^{W \times W}$, **explicitly mining interlayer distinctive features.** At the same time, the feature distribution within each layer is implicitly resolved along the horizontal direction in the local window with vertical strips for scores $VAtt \in \mathbb{R}^{H \times H}$, **aiming to reason about intralayer information through contextual dependencies, especially for non-regular curvilinear structures.** As shown in Figure 2, the Transformer block incorporates attention weights from various orientations to update the vanilla self-attention by using the structural characteristics of arm MSKUS. These are not only suitable for complex hybrid morphology, but the combination of the two window shapes also helps expand the attention region of each token within the network, enabling more effective global self-attention.

**Segmentation Head and Loss Functions** The outputs of the Transformer blocks are fed into a segmentation head to obtain fast inference results. The segmentation head consists of a top-down branch in the form of a feature pyramid, which includes three convolutional layers, followed by a batch normalization layer and an activation layer. We employ the deep supervision strategy by adding meta segmentation losses (the sum of cross-entropy loss and the dice loss).

### 2.2 Hybrid Structure-oriented Multi-head Self-attention with Horizontal or Vertical Windows

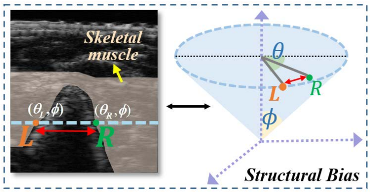

**Fig. 3.** Illustration of the structure bias using local windows with horizontal strips.

In addition to *extracting the global features via local windows of various shapes for achieving a balanced segmentation between interlayer and intralayer structures*, we propose *integrating a hierarchical-consistency position encoding (HCPE) into the $QK^T$ of self-attention, consideration of the structure bias of layers* when extracting the local attention (where $Q$ and $K$ represent the query and key matrices). Furthermore, we *design a structure-bias constraint (SBC) to calculate attention weights that is compatible with HCPE.*

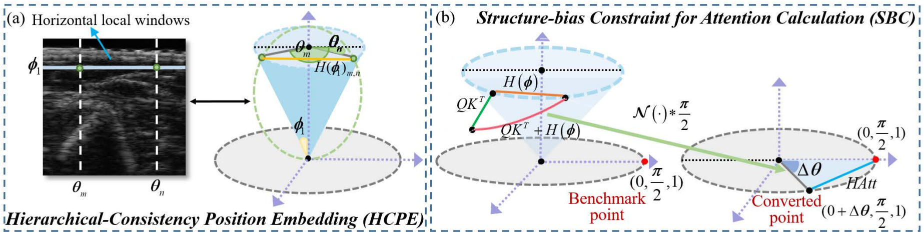
(a) Horizontal local windows
$\phi_1$
$\theta_m$
$\theta_n$
$H(\phi_1)_{m,n}$
$\phi_1$
Hierarchical-Consistency Position Embedding (HCPE)
(b) Structure-bias Constraint for Attention Calculation (SBC)
$H(\phi)$
$QK^T$
$QK^T + H(\phi)$
$\mathcal{N}(\cdot)*\frac{\pi}{2}$
$(0,\frac{\pi}{2},1)$
Benchmark point
$(0,\frac{\pi}{2},1)$
$\Delta\theta$
$HAtt$
Converted point
$(0+\Delta\theta,\frac{\pi}{2},1)$

**Fig. 4.** (a) Acquisition of the horizontal HCPE. (b) SBC for attention weights calculation contains two steps: $Q_i K_i^T + H(\phi_i)$ is converted to the offset angle $\Delta\theta$ by the green arrow, representing the Normalization; the distances of the new converted point from the benchmark point is calculated as $HAtt$, which is indicated by the blue line.

**Structure Bias** The 2D arm MSKUS plane is projected to a sphere image, and each pixel location is represented by $(\theta, \phi, \rho)$, where $\theta \in (0, 2\pi)$ , $\phi \in (0, \pi)$ , $\rho = 1$. In Figure 3, both points L and R are situated within the muscle layer, characterized by the light-yellow area. **Despite they are quite spatially separated in the MSKUS,** their attention weights must remain consistent owing to the underlying structural hierarchy. Fortunately, **in the sphere image, the distance between L and R significantly narrows, thus reinforcing weight consistency.** This provides the inspiration for the HCPE and SBC techniques.

**Hierarchical-consistency position embedding (HCPE)** To better leverage the abundant hierarchical relationships of arm MSKUS, HCPE is proposed to impose structural morphology bias on the elements within each vertical and horizontal local windows. Horizontal HCPE $H(\phi_i)_{m,n} \in \mathbb{R}^{1\times 1}$ is calculated by measuring the distance between the m-th $\phi_m$ and the n-th elements $\phi_n$ in the i-th horizontal window of Cartesian coordinates (as shown in the orange line in Figure 4(a)). Vertical HCPE $H(\theta_i)_{m,n} \in \mathbb{R}^{1\times 1}$ is calculated by measuring the distance between the m-th $\theta_m$ and the n-th elements $\theta_n$ in the i-th vertical window of Cartesian coordinates. Specifically, $H(\phi_i) \in \mathbb{R}^{W\times W}$ and $H(\theta_i) \in \mathbb{R}^{H\times H}$ for the m-th and n-th elements of i-th window are defined as follows, separately:

$$\boldsymbol{Horizontal\ HCPE} : H(\phi_i)_{m,n} = sign(\theta_m - \theta_n) \cdot \sqrt{2\{1 - \cos(\theta_m - \theta_n)\}} \cdot \sin(\phi_i) \tag{1}$$

$$\boldsymbol{Vertical\ HCPE} : H(\theta_i)_{m,n} = sign(\phi_m - \phi_n) \cdot \sqrt{2\{1 - \cos(\phi_m - \phi_n)\}} \tag{2}$$

where the $sign(\cdot)$ is a sign function used to distinguishes between $H(\phi_i)_{m,n}$ and $H(\phi_i)_{n,m}$, and we denote the sign of them as $+1$ and $-1$, respectively.

Unlike existing position embedding [3], [21], HCPE promotes to assign more similar attention weights at spatial sampling points within the same tissue layer, due to the structure bias in the Figure 3. Then HSformer can better comprehend the curvilinear structures presented in edema or skeletal muscles surfaces.

**Structure-Biased Constraint for Attention calculation (SBC)** We further design a SBC to calculate attention weights, which is compatible with HCPE, as demonstrated in the Figure 4(b). In the polar coordinate, we need to ensure $H(\theta, \phi)_{m,n} = -H(\theta, \phi)_{n,m}$. However, conventional softmax acting on the terms within HCPE might generate inconsistent scores for the (m, n)-th and (n, m)-th elements, potentially leading to misguidance during the training of HSformer [22]. In contrast, the SBC is symmetric, ensuring that HSformer produces consistent scores when combined with HCPE. This characteristic makes SBC more suitable for capturing and modeling curvilinear structures.

Specifically, we first chose a spatial benchmark point $(\theta, \phi, \rho) = (0, \frac{\pi}{2}, 1)$ on the sphere. For the interlayer features with horizontal windows, the elements in $Q_i K_i^T + H(\phi_i)$ are converted to $\Delta\theta$ by performing $\mathcal{N}\{\} \cdot \frac{\pi}{2}$. The new elements can be located as $\left(0 + \Delta\theta, \frac{\pi}{2}, 1\right)$ in the polar coordinate. Ultimately, we calculate the distances between the benchmark point and the converted points as the attention weights $HAtt_i$ as follows:

$$\boldsymbol{Horizontal\ Attention}: HAtt_i = \{1 - \cos(\mathcal{N}\{Q_i K_i^T + H(\phi_i)\} \cdot \frac{\pi}{2})\} \cdot \sin^2(\phi), \phi = \frac{\pi}{2} \tag{3}$$

Similarly, by transforming the $Q_i K_i^T + H(\theta_i)$ to $\Delta\phi$, the new intralayer structure weights $VAtt_i$ is defined as:

$$\boldsymbol{Vertical\ Attention}: VAtt_i = 1 - \cos(\mathcal{N}\{Q_i K_i^T + H(\theta_i)\} \cdot \frac{\pi}{2}) \tag{4}$$

Where $\mathcal{N}$ represents L1 normalization, and we directly utilize the square of the distance as the attention weights, as a way to amplify the weight difference to enhance the characterization of important regions.

## 3 Experiments

**Materials.** We used the public arm MSKUS dataset [4] to verify the proposed HSformer. B-mode arm MSKUS contains different tissue layers, from top to bottom: gel, skin, subcutaneous fat, skeletal muscle, and bone. It comprises 468 arm MSKUS images from 39 subjects, and all the images are resized into a resolution of 224 × 224 for experiments. During training, we perform routine data augmentation operations, such as random horizontal and vertical flips.

**Implementation Details.** In the proposed HSformer, the block numbers $N_i$ of each hybrid structure-oriented Transformer stage are separately 1, 2, 21, 1, the head number of four stages is assigned to 2, 4, 8, 16, and the settings of other parameters refer to [3]. HSformer is trained for 300 epochs and uses the Adam optimizer with the initial learning rate of 0.0002 and a batch size of 6, and the cosine learning rate scheduler with 30 epochs linear warm-up is leveraged. We carry out a 5-fold cross-validation to evaluate the HSformer, and only record the mean value. HSformer is implemented by PyTorch on one NVIDIA GeForce RTX 3060. Dice coefficient (DSC), Jaccard Index (JI), Hausdorff Distance (HD) and Average Surface Distance (ASD) are utilized as evaluation metrics.

**Compared with Advanced Segmentation Approaches.** Considering the significant characteristics of arm MSKUS, we design comparative experiments with three types: (1) representative image segmentation models, including UNet [17], DeepLabV3 [2], TransUNet [1]; (2) recent MSKUS image segmentation works, such as GSCNN [4] and AMCNet [23]; (3) outstanding layer-structure segmentation frameworks, involving TCCT [19] and CoherentSeg [10]. Table 1 and Figure 6 give the performance of the HSformer and seven advanced competing methods. The results demonstrate that our model outperforms all other methods consistently, as proved by higher DSC and JI on the five layers. Notably, despite with medical segmentation networks [19], [10] being more focused on layer structures or MSKUS images, the precision of the corresponding five-layer target is still lower than HSformer. This suggests that understanding the hierarchical layer structures with horizontal and curvilinear morphology can improve the arm MSKUS segmentation performance with hybrid structural bias. Simultaneously, the lower HD and ASD indicate that HSformer provides highly accurate boundaries and acquires effective feature representations by distinguishing and exploring interlayer and intralayer structure relationships. Especially, the HD and ASD of the bone structure are 7.39 and 1.122 lower than smallest values of other models, indicating that HSformer accurately predicts curvilinear shapes. Furthermore, we test on the recently popular SAM (Segment Anything) [9], and the mean DSC and mean HD values of our HSformer are 0.13 higher and 14.00 lower than the SAM-based models, respectively. The qualitative and quantitative results are demonstrated in the Figure A2 and Table A1 of the appendix.

**Table 1.** The comparison with other methods.

| Layer Structure | Methods | Representative Segmentation Models | | | MKUS Image Segmentation | | Layer-Structure Segmentation | | Ours |
|---|---|---|---|---|---|---|---|---|---|
| | | UNet | DeepLabV3 | TransUNet | GSCNN (TBME 2023) | AMCNet (MICCAI 2023) | TCCT (TMI 2023) | CoherentSeg (MIA 2024) | HSformer |
| gel | DSC | 0.9034 | 0.9337 | 0.9694 | 0.9494 | 0.9170 | 0.9368 | 0.9051 | **0.9752** |
| | JI | 0.8258 | 0.8788 | 0.9447 | 0.9096 | 0.8531 | 0.8861 | 0.8319 | **0.9529** |
| | HD | 2.3644 | 2.3639 | 2.3939 | 2.3418 | 2.5733 | 2.5272 | 2.4552 | **2.2727** |
| | ASD | 0.5402 | 0.5180 | 0.5429 | 0.5210 | 0.5251 | 0.5517 | 0.4919 | **0.4795** |
| skin | DSC | 0.7846 | 0.7925 | 0.8577 | 0.7903 | 0.7987 | 0.7976 | 0.7823 | **0.8722** |
| | JI | 0.6602 | 0.6702 | 0.7629 | 0.6632 | 0.6749 | 0.6736 | 0.6549 | **0.7847** |
| | HD | 3.8346 | 3.4749 | 2.8631 | 3.5666 | 3.4505 | 3.9399 | 3.9573 | **2.7669** |
| | ASD | 1.2272 | 1.2235 | 1.0975 | 1.1818 | 1.1367 | 1.2596 | 1.225 | **1.0576** |
| subcutaneous fat | DSC | 0.8349 | 0.8444 | 0.8318 | 0.8604 | 0.8680 | 0.8642 | 0.8356 | **0.9004** |
| | JI | 0.7345 | 0.7506 | 0.7554 | 0.7711 | 0.7846 | 0.7742 | 0.7292 | **0.8299** |
| | HD | 18.2158 | 16.0406 | 18.3368 | 15.6484 | 14.2870 | 16.5439 | 20.9985 | **13.2110** |
| | ASD | 4.1608 | 3.9346 | 4.4826 | 3.8932 | 3.486 | 4.2678 | 4.1733 | **3.1869** |
| muscles | DSC | 0.8081 | 0.8261 | 0.8441 | 0.8430 | 0.8524 | 0.8617 | 0.8090 | **0.8922** |
| | JI | 0.7039 | 0.7295 | 0.7795 | 0.7572 | 0.7630 | 0.7651 | 0.6869 | **0.8134** |
| | HD | 32.1766 | 30.5216 | 26.3134 | 26.2865 | 28.0218 | 25.8736 | 48.0643 | **25.0155** |
| | ASD | 6.8631 | 6.5158 | 6.8182 | 6.0242 | 5.3709 | 5.2882 | 8.9133 | **5.1583** |
| bones | DSC | 0.7891 | 0.8011 | 0.8192 | 0.8149 | 0.8255 | 0.8458 | 0.7921 | **0.8873** |
| | JI | 0.6934 | 0.7014 | 0.7522 | 0.7287 | 0.7339 | 0.7557 | 0.6751 | **0.8157** |
| | HD | 34.149 | 30.0704 | 29.6025 | 28.4942 | 27.171 | 27.1545 | 47.1655 | **19.7612** |
| | ASD | 6.7805 | 5.8932 | 6.2708 | 5.5064 | 5.0220 | 5.6406 | 8.2199 | **3.9000** |
| mean | DSC | 0.8245 | 0.8455 | 0.8644 | 0.8516 | 0.8523 | 0.8612 | 0.8248 | **0.9055** |
| | JI | 0.7239 | 0.7461 | 0.7957 | 0.7660 | 0.7619 | 0.7709 | 0.7156 | **0.8393** |
| | HD | 17.9749 | 16.4943 | 15.9019 | 15.2675 | 15.1007 | 15.2078 | 24.5282 | **12.6054** |
| | ASD | 3.8747 | 3.6170 | 3.6170 | 3.4253 | 3.1081 | 3.4016 | 4.6047 | **2.7564** |

**Ablations.** To evaluate the contributions of HCPE and SBC, we implement four variants: (1) Ours w/o HCPE, where we discard the HCPE; (2) Ours w/o SBC, where we utilize the softmax function to replace SBC and calculate the attention weights; (3) Ours w LPE or RPE, where we replace HCPE with learnable PE (LPE) [21] or relative PE (RPE) [12], as listed in Table 2. It reveals that, when the HCPE was removed, the segmentation performance drops obviously, especially on the skin, muscle and bone structure. Meanwhile, compared with LPE and RPE, the HCPE both increases JI scores from 0.74, 0.73 to 0.82 for

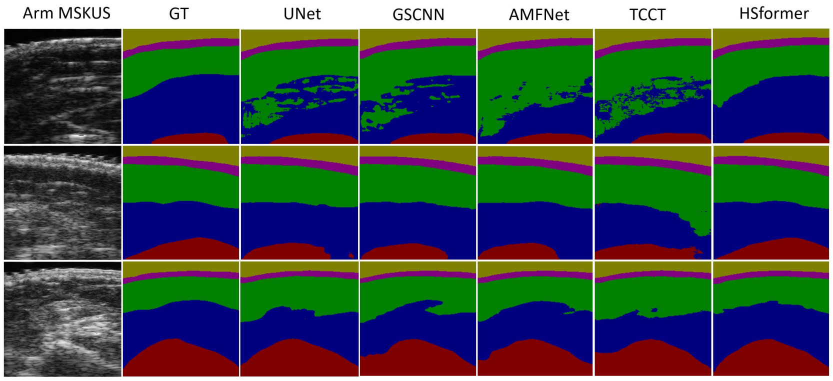

**Fig. 5.** The segmentation examples of four more remarkable models and Hsformer. To provide a more comprehensive comparison with all models, we show the segmentation results of all models corresponding to two cases in the Figure A1 of the appendix.

the representative curvilinear bone segmentation. It indicates that the HCPE contributes substantially to explore the hybrid structure bias, and the SBC can effectively produce more consistent attention weights than softmax function.

**In-house Small-scale Arm MSKUS Test.** To validate the generalization of HSformer, we directly test 100 arm MSKUS images without model training, which were collected from Beijing Tsinghua Changgung Hospital using a SonoScape E2 machine. While the public dataset uses Alpinion E-Cube 12 system (Bothell, WA, USA) with L3-12H high-density linear probe for MSKUS imaging. And anatomical structures of the MSKUS images from the two datasets are consistent, with a large amount of speckle noise and shadow artifacts that pose greater challenges to the segmentation task. For five-layer tissue structures, we obtain the high mean DSC and low measn ASD of 0.82 and 5.31, as shown in the Table 3. We also visualize the segmentation maps in Figure A3 of the appendix. Although there are significant differences in image quality and style between the two datasets, the proposed HSformer both achieves excellent performance and demonstrates the strong generalization.

**Table 2.** Ablation results of the HSformer with four variants.

| Layer Structures | gel | | | | skin | | | | subcutaneous fat | | | | muscles | | | | bones | | | | mean | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| variants | DSC | JI | HD | ASD | DSC | JI | HD | ASD | DSC | JI | HD | ASD | DSC | JI | HD | ASD | DSC | JI | HD | ASD | DSC | JI | HD | ASD |
| w/o HCPE | 0.96 | 0.93 | 3.12 | 0.72 | 0.8 | 0.7315 | 4.00 | 1.30 | 0.87 | 0.79 | 15.51 | 3.77 | 0.84 | 0.75 | 27.36 | 6.43 | 0.81 | 0.72 | 32.45 | 6.41 | 0.86 | 0.79 | 16.49 | 3.73 |
| w/o SBC | 0.97 | 0.95 | 2.29 | 0.50 | 0.86 | 0.77 | 3.72 | 1.17 | 0.90 | 0.82 | 14.74 | 3.79 | 0.86 | 0.78 | 30.98 | 6.83 | 0.87 | 0.79 | 27.91 | 4.99 | 0.89 | 0.82 | 15.93 | 3.46 |
| w LPE | 0.97 | 0.94 | 3.79 | 0.77 | 0.84 | 0.75 | 3.59 | 1.21 | 0.87 | 0.80 | 13.34 | **3.15** | 0.86 | 0.78 | 26.59 | 6.19 | 0.83 | 0.74 | 30.14 | 6.04 | 0.87 | 0.80 | 15.49 | 3.47 |
| w RPE | 0.97 | 0.94 | 2.69 | 0.57 | 0.84 | 0.75 | 3.74 | 1.29 | 0.86 | 0.78 | 18.01 | 4.14 | 0.85 | 0.78 | 26.80 | 6.36 | 0.82 | 0.73 | 30.74 | 6.18 | 0.87 | 0.80 | 16.40 | 3.71 |
| HSformer | **0.98** | **0.95** | **2.27** | **0.48** | **0.87** | **0.78** | **2.77** | **1.06** | **0.90** | **0.83** | **13.21** | 3.19 | **0.89** | **0.81** | **25.02** | **5.16** | **0.89** | **0.82** | **19.76** | **3.90** | **0.91** | **0.84** | **12.61** | **2.76** |

**Table 3.** Performance for the In-house Small-scale Arm MSKUS Dateset.

| | gel | skin | subcutaneous fat | muscles | bones | mean |
|---|---|---|---|---|---|---|
| DSC | 0.95 | 0.74 | 0.72 | 0.77 | 0.93 | 0.82 |
| JI | 0.91 | 0.58 | 0.56 | 0.62 | 0.86 | 0.71 |
| HD | 4.00 | 4.52 | 39.53 | 39.36 | 27.41 | 22.95 |
| ASD | 1.21 | 1.81 | 7.46 | 11.36 | 4.71 | 5.31 |

## 4 Conclusion

This paper introduces a novel hybrid structure-oriented Transformer for arm MSKUS segmentation to aid for diagnosis and screening BCRL. To address the challenge of the hierarchical structures with diverse morphology in MSKUS, HSformer employs horizontal and vertical local windows to capture optimal feature representations. Particularly, HSformer designs a HCPE to impose structural bias onto the elements within local window, and a SBC to calculate attention weights and is more suitable for horizontal or curvilinear hybrid structures. The empirical experiments demonstrate that the HSformer outperforms existing segmentation models and has good generalization for arm MSKUS segmentation.

**Acknowledgments.** The authors acknowledge supports from National Nature Science Foundation of China Grants (82027807, U20A20389, 62271246), National Key Research and Development Program of China (2022YFC2405200), and Natural Science Foundation of Jiangsu Province (BK20221477).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
2. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
3. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12124–12134 (2022)
4. Goudarzi, S., Whyte, J., Boily, M., Towers, A., Kilgour, R.D., Rivaz, H.: Segmentation of arm ultrasound images in breast cancer-related lymphedema: A database and deep learning algorithm. IEEE Transactions on Biomedical Engineering (2023)
5. Hashemi, H.S., Fallone, S., Boily, M., Towers, A., Kilgour, R.D., Rivaz, H.: Assessment of mechanical properties of tissue in breast cancer-related lymphedema using ultrasound elastography. IEEE transactions on ultrasonics, ferroelectrics, and frequency control **66**(3), 541–550 (2018)
6. Hidding, J.T., Viehoff, P.B., Beurskens, C.H., van Laarhoven, H.W., Nijhuis-van der Sanden, M.W., van der Wees, P.J.: Measurement properties of instruments for measuring of lymphedema: systematic review. Physical therapy **96**(12), 1965–1981 (2016)
7. Katakis, S., Barotsis, N., Kakotaritis, A., Economou, G., Panagiotopoulos, E., Panayiotakis, G.: Automatic extraction of muscle parameters with attention unet in ultrasonography. Sensors **22**(14), 5230 (2022)
8. Katakis, S., Barotsis, N., Kakotaritis, A., Tsiganos, P., Economou, G.: Muscle cross-sectional area segmentation in transverse ultrasound images using vision transformers. Diagnostics **13**(2), 217 (2023)

9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
10. Liu, H., Wei, D., Lu, D., Tang, X., Wang, L., Zheng, Y.: Simultaneous alignment and surface regression using hybrid 2d–3d networks for 3d coherent layer segmentation of retinal oct images with full and sparse annotations. Medical Image Analysis **91**, 103019 (2024)
11. Liu, R., Liu, M., Sheng, B., Li, H., Li, P., Song, H., Zhang, P., Jiang, L., Shen, D.: Nhbs-net: A feature fusion attention network for ultrasound neonatal hip bone segmentation. IEEE Transactions on Medical Imaging **40**(12), 3446–3458 (2021)
12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
13. of Lymphology, I.S.: The diagnosis and treatment of peripheral lymphedema. consensus document of the international society of lymphology. Lymphology **36**(2), 84–91 (2003)
14. Marzola, F., van Alfen, N., Doorduin, J., Meiburger, K.M.: Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment. Computers in Biology and Medicine **135**, 104623 (2021)
15. Mishra, Z., Ganegoda, A., Selicha, J., Wang, Z., Sadda, S.R., Hu, Z.: Automated retinal layer segmentation using graph-based algorithm incorporating deep-learning-derived information. Scientific Reports **10**(1), 9541 (2020)
16. Rasti, R., Biglari, A., Rezapourian, M., Yang, Z., Farsiu, S.: Retifluidnet: A self-adaptive and multi-attention deep convolutional network for retinal oct fluid segmentation. IEEE Transactions on Medical Imaging (2022)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, October, Part III 18. pp. 234–241. Springer (2015)
18. Seward, C., Skolny, M., Brunelle, C., Asdourian, M., Salama, L., Taghian, A.G.: A comprehensive review of bioimpedance spectroscopy as a diagnostic tool for the detection and measurement of breast cancer-related lymphedema. Journal of surgical oncology **114**(5), 537–542 (2016)
19. Tan, Y., Shen, W.D., Wu, M.Y., Liu, G.N., Zhao, S.X., Chen, Y., Yang, K.F., Li, Y.J.: Retinal layer segmentation in oct images with boundary regression and feature polarization. IEEE Transactions on Medical Imaging (2023)
20. Van der Veen, P., Vermeiren, K., Von Kemp, K., Lamote, J., Sacre, R., Lievens, P.: A key to understanding postoperative lymphoedema: a study on the evolution and consistency of oedema of the arm using ultrasound imaging. The Breast **10**(3), 225–230 (2001)
21. Wan, Q., Huang, Z., Lu, J., Yu, G., Zhang, L.: Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation. arXiv preprint arXiv:2301.13156 (2023)
22. Yun, I., Shin, C., Lee, H., Lee, H.J., Rhee, C.E.: Egformer: Equirectangular geometry-biased transformer for 360 depth estimation. arXiv preprint arXiv:2304.07803 (2023)
23. Zeng, B., Chen, L., Zheng, Y., Kikinis, R., Chen, X.: Fine-grained hand bone segmentation via adaptive multi-dimensional convolutional network and anatomy-constraint loss. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 395–404. Springer (2023)

24. Zhao, R., Huang, Z., Liu, T., Leung, F.H., Ling, S.H., Yang, D., Lee, T.T.Y., Lun, D.P., Zheng, Y.P., Lam, K.M.: Structure-enhanced attentive learning for spine segmentation from ultrasound volume projection images. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1195–1199. IEEE (2021)