Shape Prior Enhanced Unet Model for Medical Image Segmentation

1st Song Fang
School of Computer & Communication Engineering
University of Science and Technology Beijing
Beijing, China
U202241957@xs.ustb.edu.cn

2nd Jiamin Wang
School of Intelligence Science and Technology
University of Science and Technology Beijing
Beijing, China
D202210387@xs.ustb.edu.cn

3rd Zirun Zhao
School of Computer & Communication Engineering
University of Science and Technology Beijing
Beijing, China
U202241958@xs.ustb.edu.cn

4th Chaopeng Wei
School of Intelligence Science and Technology
University of Science and Technology Beijing
Beijing, China
U202241771@xs.ustb.edu.cn

5th Xudong Zhang
School of Intelligence Science and Technology
University of Science and Technology Beijing
Beijing, China
U202241499@xs.ustb.edu.cn

6th Jiaxin Wang
School of Computer & Communication Engineering
University of Science and Technology Beijing
Beijing, China
U202241911@xs.ustb.edu.cn

7th Mohammad S. Obaidat
King Abdullah II School of Information Technology
University of Jordan,
Amman, Jordan
msobaidat@gmail.com

8th Xiaokun Wang*
School of Intelligence Science and Technology
University of Science and Technology Beijing
Beijing, China
wangxiaokun@ustb.edu.cn

***Abstract*—Medical image segmentation is a critical and challenging task in modern medicine. This task faces issues such as low contrast between the target structures and surrounding tissues, as well as uncertainties in the morphology and location of pathological tissues. These challenges often result in the suboptimal performance of traditional segmentation methods, especially when dealing with complex structures. To address these issues, we propose a UNet model integrated with shape priors. This model utilizes a Shape Prior Module (SPM) to generate and update shape knowledge, which is then integrated as feature maps into the skip connections of the U-shaped network. By combining global and local information, the SPM enhances the network's comprehensive understanding of the image. Additionally, this module can be seamlessly integrated with the convolutional neural network architecture, enhancing the model's practicality and generalization capabilities. Moreover, we design a comprehensive loss function that integrates Dice, CE, and BoundaryDoU metrics, significantly improving the efficiency of model parameter optimization and learning. Experimental results on two medical image datasets demonstrate that the proposed shape prior network model yields better outcomes in enhancing segmentation accuracy and improving classification detection precision.**

***Index Terms*—Medical Image Segmentation, Shape Prior, U-shaped Network, Transformer**

I. Introduction

The application of computer vision technology in healthcare is rapidly expanding and revolutionizing various aspects of medical practice. Among these advancements, medical image segmentation is particularly crucial, playing a pivotal role in early disease diagnosis, continuous condition monitoring, and treatment efficacy evaluation. The ability to accurately delineate anatomical structures and pathological regions within medical images provides clinicians with invaluable insights for informed decision-making and personalized patient care.

Convolutional Neural Networks (CNNs) have long dominated computer vision tasks like image classification and object detection, thanks to their ability to learn hierarchical features from image data through scale expansion, enhanced connectivity, and refined convolution operations [1]. However, CNNs have inherent limitations in capturing global contextual information, crucial for understanding complex inter-regional relationships within an image. This limitation is particularly pronounced in medical image segmentation, where the contrast between target structures and surrounding tissues is often low, and the morphology and location of pathological tissues are highly variable and ambiguous [2]. Consequently, traditional

2024 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI) | 979-8-3503-4983-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/CCCI61916.2024.10736471

CNN-based segmentation methods may struggle, especially with intricate anatomical structures or subtle pathological changes, to achieve satisfactory performance.

Researchers have explored various strategies to enhance CNN performance in medical image segmentation. One approach is increasing the convolution kernel size, effectively expanding the CNN's receptive field to capture more global information [3]. However, this method significantly increases computational complexity and memory consumption, limiting its practicality, especially for high-resolution 3D medical images. Another approach is to deepen the network by adding more convolutional layers [4]. While this can be beneficial to some extent, excessively deep networks are prone to vanishing or exploding gradients, hindering training and potentially affecting model convergence [1].

The emergence of the Transformer architecture, initially groundbreaking in natural language processing, has opened new avenues for global information extraction in computer vision [5]. The Transformer's self-attention mechanism efficiently processes sequential data and captures long-range dependencies. Unlike CNNs, which are inherently local in their receptive field, the self-attention mechanism enables Transformers to weigh information from all image parts, regardless of spatial distance. This capability has proven highly effective in natural language processing tasks and has shown great promise in computer vision applications, such as image segmentation and object detection. The self-attention and multi-head attention mechanisms within the Transformer architecture are particularly advantageous in handling medical images, as they efficiently model complex spatial relationships between different anatomical structures or pathological regions, leading to a more comprehensive understanding of image content.

Inspired by the success of Transformers in capturing global contextual information, we propose a novel Transformer-driven Shape Prior Module and seamlessly integrate it into the UNet architecture, a widely adopted and highly effective model for medical image segmentation. Our proposed SPM synergistically combines the strengths of both UNet and the Transformer, leveraging the local feature extraction capabilities of UNet and the global information processing power of the Transformer. By embedding shape priors in the UNet decoder pathway, our method guides segmentation towards more anatomically plausible and accurate results, especially for complex structures. This approach not only enhances segmentation accuracy but also contributes to greater consistency and robustness in the segmentation output.

Our main contributions are summarized as follows: (1) We propose an innovative Shape Prior Module, comprising a Self-updating Module and a Cross-updating Module, which are capable of extracting global features and integrating local features from the encoder, respectively. (2) The SPM seamlessly integrates with existing UNet architectures, enhancing both flexibility and practicality. (3) We introduce a comprehensive loss function that combines Dice Loss, Cross-Entropy (CE) Loss, and Boundary Dice Loss to achieve balanced learning of detailed features and overall structure. Experimental results demonstrate that our proposed method effectively addresses the challenges faced by CNNs in handling complex medical image segmentation tasks.

II. Related work

A. Medical Image Segmentation

The U-Net network architecture has become a widely adopted method in medical image segmentation owing to its exceptional performance and broad applicability. Introduced by Ronneberger et al. in 2015 [6], the U-Net model is an end-to-end convolutional neural network meticulously designed for image segmentation tasks. This model comprises an encoder and a decoder. The encoder performs down-sampling operations to extract abstract features, whereas the decoder reconstructs image details through up-sampling processes. Collectively, these components form the distinctive "U"-shaped structure.

In recent years, researchers have made numerous innovative improvements to the original U-Net architecture in order to further enhance model performance. For instance, U-Net++ introduces deep supervision mechanisms and dense skip connections based on the original U-Net. These enhancements are designed to optimize feature transfer paths and enhance feature fusion, thereby improving the model's ability to capture detailed information [7]. Other models that have incorporated enhanced skip connections include H-DenseUNet [8] and UNet 3+ [9], among others.

The ResNet architecture integrates residual blocks within both the encoder and decoder, thereby facilitating stable training of deep neural networks [1]. The authors further improved the architecture by introducing pre-activation adjustments, resulting in the development of ResNetv2 [10]. Recent advancements include modifying the topology of residual blocks through the implementation of RobustResBlock, which boosts the model's robustness against adversarial attacks [11]. Additionally, channel-wise attention has been implemented to improve performance in the ResNeSt architecture [12].

These advancements have significantly enriched the theory and practice in the field of image segmentation. For instance, EGE-UNet effectively achieves lightweight feature processing by introducing the Grouped Multi-Axis Hadamard Product Attention module (GHPA) and the Group Aggregation Bridge module (GAB), thereby improving segmentation accuracy while maintaining low parameter and computational costs [13]. Furthermore, the Multi-ResUNet model employs multi-scale attention and hybrid dilated strategies, enabling the model better handle features of different scales and enhance the accuracy of medical image segmentation [14].

Although these improved methods have enhanced U-Net's performance to some extent, they also bring about new limitations or challenges. For instance, increasing the convolution kernel size may increase computational burden, and excessively deep networks can lead to issues such as gradient explosion [1].

B. Transformer-based Segmentation

Since the introduction of the Transformer architecture by Vaswani et al. in 2017, it has had a significant impact on various fields, including computer vision and image segmentation. The self-attention mechanism, which is at the core of the Transformer architecture, has been proposed as an alternative to traditional recurrent neural networks (RNNs) and convolutional neural networks (CNNs). It enables parallel processing of sequential data and enhances training efficiency in many applications [5].

Several innovative Transformer-based models have been proposed for image segmentation tasks, leveraging the unique capabilities of the Transformer architecture. The CGFormer model introduces a contrastive grouping Transformer framework specifically tailored for directional image segmentation tasks. This model employs learnable query tokens to represent objects and achieves object-oriented cross-modal reasoning through the iterative use of query tokens and visual features [15]. On the other hand, the OneFormer model, introduced by Hassani et al., demonstrates exceptional performance by processing and optimizing panoramic, semantic, and instance segmentation tasks through the collaborative intervention of task-conditioned input, multi-scale feature extraction, task-conditioned query generation, and a Transformer decoder [16; 17; 18].

In medical image segmentation, the Swin UNETR model amalgamates the hierarchical window attention mechanism of the Swin Transformer with the traditional UNet architecture, specifically targeting the semantic segmentation of brain tumor MRI images. This approach effectively integrates local and global information, achieving efficiency and accuracy in 3D image segmentation [19; 20]. Chen et al. [21] proposed a TransUNet model, which employs a hybrid CNN-Transformer architecture. It leverages the strengths of both CNNs for low-level feature extraction and Transformers for capturing long-range dependencies and global context, thus enhancing the performance of medical image segmentation.

The Uni3DETR model, proposed by Wang et al., is a unified 3D detection model that adopts a point-voxel interaction detection transformer for object prediction. It enhances the model's generalization ability in diverse scenarios through the use of hybrid query points and decoupled IoU techniques [22; 23]. The nnFormer model, introduced by Zhou et al., is a volumetric segmentation framework that combines the benefits of both CNNs and Transformers. It employs a CNN-based encoder for feature extraction and a Transformer-based decoder for capturing long-range dependencies, achieving state-of-the-art performance in various 3D medical image segmentation tasks [24].

While Transformer models have demonstrated considerable promise in image segmentation tasks, they are not without limitations. These architectures can be computationally intensive when processing large-scale datasets or high-resolution images. Additionally, the generalization ability of Transformer models often depends on pre-training with large-scale datasets, which may limit their flexibility in applications with limited data [25]. To address these challenges, researchers have proposed hybrid architectures that combine the strengths of Transformers with other neural network frameworks, such as CNNs. These hybrid approaches aim to balance the trade-off between computational efficiency and segmentation accuracy [21; 24].

In summary, Transformer models have significantly influenced the field of image segmentation, offering unique architectural advantages and providing effective solutions to various challenges. Researchers are continuously exploring new ways to leverage the power of Transformers in image segmentation tasks, advancing the capabilities within this evolving field. Though the future advancements in hardware acceleration and algorithm optimization hold potential, a cautious approach is warranted to manage current computational constraints and the need for large-scale datasets.

III. Methods

As illustrated in Fig. 1, the model consists of the encoder and decoder of a conventional U-shaped network, along with our designed hierarchical shape prior module. Unlike a typical U-shaped network, where feature maps $\boldsymbol{F}$ extracted by the encoder are directly passed to the decoder via skip connections, our model integrates $\boldsymbol{F}$ with an iteratively updated shape prior $\boldsymbol{S}_g$ within the shape prior module, resulting in an updated feature map $\boldsymbol{F}_{\text{updated}}$. This updated feature map, which retains the same shape as $\boldsymbol{F}$ but incorporates stronger global information about the target, achieves multi-scale feature fusion and replaces $\boldsymbol{F}$ in guiding the decoder for more accurate segmentation predictions.

A. Self-Updating Module

The self-updating module employs a self-attention mechanism to update the shape prior, enhancing its ability to capture global information and internal structural relationships, then outputs it to the cross-update module.

The input to the self-updating module is $\boldsymbol{S}$, which is randomly initialized as $\boldsymbol{S}_0$ for the first layer, and as $\boldsymbol{S}_e$, the output from the cross-update module of the previous layer, for subsequent layers. The shape of $\boldsymbol{S}$ is (B, C, H, W, D), where B is the batch size, C is the number of channels, and H, W, D denote the height, width, and depth, respectively.

As a preprocessing step, the spatial dimensions of $\boldsymbol{S}$ are flattened to one dimension, transposed to yield a shape prior $\boldsymbol{S}_r$ with dimensions (B, HWD, C), and layer normalized to reduce internal covariate shift and enhance training stability:

$$\boldsymbol{S}_r = \mathit{LayerNorm}(\boldsymbol{S}_r) \tag{1}$$

Subsequently, $\boldsymbol{S}_r$ undergoes preprocessing to compute the query matrix $\boldsymbol{Q}$, key matrix $\boldsymbol{K}$, and value matrix $\boldsymbol{V}$ with weights $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$, respectively:

$$\boldsymbol{Q} = \boldsymbol{S}_r\boldsymbol{W}_Q \quad \boldsymbol{K} = \boldsymbol{S}_r\boldsymbol{W}_K \quad \boldsymbol{V} = \boldsymbol{S}_r\boldsymbol{W}_V \tag{2}$$

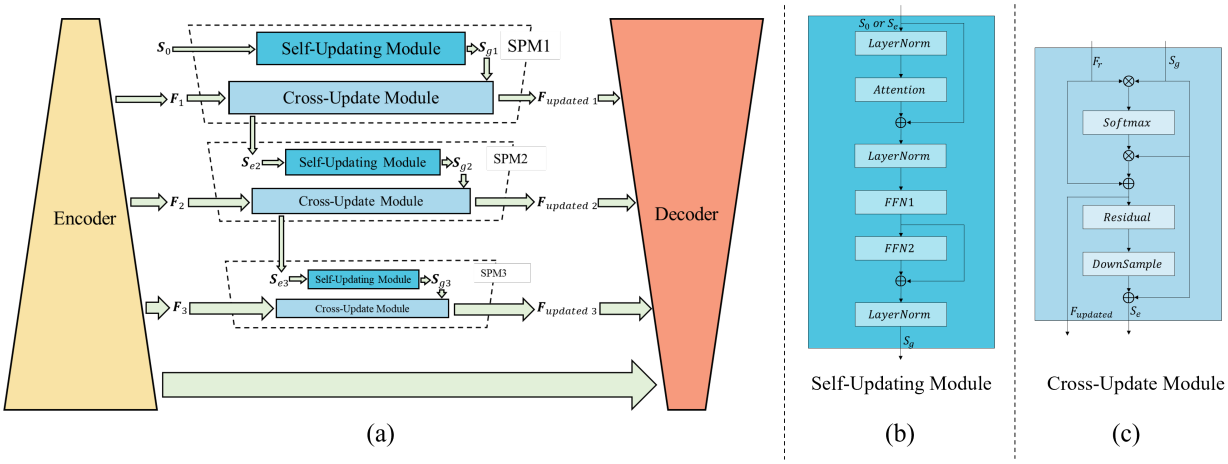

Fig. 1. Structure of the proposed network. (a) The network is designed based on U-shaped network structure. The skip connections in it are replaced by a shape prior module composed of a self-update module and a cross-update module. (b) Structure of the Self-Updating Module. (c) Structure of the Cross-Update Module.

The attention probability matrix $\boldsymbol{A}_P$ is computed, incorporating Dropout regularization to prevent overfitting:

$$\boldsymbol{A}_P = Dropout\left(Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{N}} + \boldsymbol{P}_E\right)\right) \tag{3}$$

where N denotes the number of classes, and $\boldsymbol{P}_E$ is the positional embedding matrix.

Using $\boldsymbol{A}_P$ and the value matrix $\boldsymbol{V}$, the context vector $\boldsymbol{C}_V$ is derived as follows:

$$\boldsymbol{C}_V = \boldsymbol{A}_P \cdot \boldsymbol{V} \tag{4}$$

This process captures global dependencies and long-range feature associations in the input data via the self-attention mechanism[26], dynamically adjusting feature representations for each position by computing the similarity between input features.

The context vector $\boldsymbol{C}_V$ updates the module's input $\boldsymbol{S}_r$ by adding $\boldsymbol{C}_V$ to $\boldsymbol{S}_r$, implementing a residual connection that facilitates better information and gradient flow. The result is further layer normalized, producing the updated shape prior $\boldsymbol{S}'_r$:

$$\boldsymbol{S}'_r = LN(\boldsymbol{S}_r + \boldsymbol{C}_V) \tag{5}$$

To introduce nonlinearity and enhance feature expression capabilities, $\boldsymbol{S}'_r$ is fed into a two-layer fully connected network and processed using the GELU activation function. The output $\boldsymbol{S}_{ffn2}$ is added back to the original input $\boldsymbol{S}'_r$ to form another residual connection, aiding efficient information and gradient transmission. The final result is layer normalized to yield the output $\boldsymbol{S}_g$ of the self-updating module:

$$\boldsymbol{S}_{ffn1} = GELU(\boldsymbol{S}'_r\boldsymbol{W}_1 + b_1) \tag{6}$$

$$\boldsymbol{S}_{ffn2} = \boldsymbol{S}_{ffn1}\boldsymbol{W}_2 + b_2 \tag{7}$$

$$\boldsymbol{S}_g = LN(\boldsymbol{S}'_r + \boldsymbol{S}_{ffn2}) \tag{8}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are weights for the first and second linear transformations and$b_1$and$b_2$are the respective biases.

B. Cross-Update Module

The cross-update module integrates the shape prior $\boldsymbol{S}_g$ output from the self-updating module with feature maps $\boldsymbol{F}$ extracted by the encoder, introducing shape prior information into the decoding process while updating and optimizing the shape prior.

Initially, the module preprocesses $\boldsymbol{F}$ and $\boldsymbol{S}_g$, resulting in $\boldsymbol{F}_r$ and $\boldsymbol{S}_g$ with shapes $(B,\ C,\ DHW)$ and $(B,\ N,\ DHW)$, respectively. These are fused based on the computation of class features [22], with the result normalized:

$$\boldsymbol{C}_F = Softmax\left(\frac{\boldsymbol{F}_r \cdot \boldsymbol{S}_g^T}{\sqrt{N}}\right) \tag{9}$$

$\boldsymbol{C}_F$ is then combined with $\boldsymbol{S}_g$ using the Einstein summation convention, producing a new feature map which is added to the original feature map $\boldsymbol{F}$ in a residual connection manner. This retains the original convolutional features while incorporating shape prior information, enhancing the feature map's expressiveness and sensitivity to target shapes and structures:

$$\boldsymbol{F}_{\text{updated}} = \boldsymbol{F}_r + \boldsymbol{C}_F \cdot \boldsymbol{S}_g \tag{10}$$

The updated feature map $\boldsymbol{F}_{\text{updated}}$ is reshaped back to its original form and passed to the encoder via skip connections.

Simultaneously, $\boldsymbol{F}_{\text{updated}}$ is used to generate a shape prior $\boldsymbol{S}_e$ containing both global and local information, which serves as input to the self-updating module in the next layer:

$$\boldsymbol{S}_e = DownSample(Residual(\boldsymbol{F}_{\text{updated}}), s) + \boldsymbol{S}_g \tag{11}$$

where $Resblock(\cdot)$ denotes a residual convolution block, and $DownSample(\cdot, s)$ uses the interpolate function with a scaling factor s.

C. Loss Functions

Our model incorporates three loss functions: Dice Loss, Cross-Entropy Loss, and BoundaryDoU Loss [27].

The Dice Loss measures the similarity between two sets of samples, guiding the segmentation task in our model to increase the overlap between predicted and ground truth regions. The formula is:

$$L_{Dice} = 1 - \frac{2|P_m \cap G_m|}{|P_m| + |G_m|} \tag{12}$$

where P_m represents the predicted mask and G_m represents the ground truth mask.

The Cross-Entropy Loss quantifies the difference between predicted and actual distributions, guiding the multi-class classification task in our model. Its formula is:

$$L_{Cross-Entropy} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{n=1}^{N} y_{i,n}\log(p_{i,n}) \tag{13}$$

where M is the number of samples, and $y_{i,n}$ and $p_{i,n}$ are the true label and predicted probability for sample i in class n, respectively.

The BoundaryDoU Loss measures the discrepancy between predicted and ground truth boundaries. This loss function is included to prevent the neglect of small but clinically significant structures, such as fine vessels and tissues, which might be overlooked. The formula is:

$$L_{BoundaryDoU} = \frac{|P_b \Delta G_b|}{|P_b \Delta G_b| + |P_b \cap G_b|} \tag{14}$$

where P_b and G_b represent the predicted and ground truth boundaries, and $|P_b \Delta G_b|$ represents the symmetric difference between the predicted and ground truth boundaries. The symmetric difference, denoted by Δ, includes elements that are in either of the sets but not in their intersection.

IV. Experiment

A. Dataset and Preprocessing

We utilized datasets from the MICCAI 2022 Kidney Multi-Organ Segmentation Challenge (KiPA 22) [28; 29; 30; 31] and the MICCAI 2017 Automated Cardiac Detection Challenge (ACDC) [32]. KiPA 22 is an abdominal kidney segmentation CTA dataset containing 130 cases, aiming to segment four categories: 3D kidneys, kidney tumors, renal arteries, and renal veins. ACDC is a cardiac cine MRI dataset containing 100 cases, targeting the segmentation of three categories: 3D left ventricle, right ventricle, and myocardium.

To streamline and standardize the training process, and to minimize the interference from irrelevant edge regions, we performed resampling and normalization on the data beforehand. We applied cubic spline interpolation [33] to the raw data and nearest neighbor interpolation [34] to the label data. As a result, the KiPA 22 dataset was cropped to 160×128×128 centered on the kidney region, and the ACDC dataset was cropped to 32×128×128 centered on the heart region. To enhance feature representation, accelerate convergence, and improve model performance, we applied Z-score normalization to the resampled data. During training, we employed three data augmentation methods: random perturbation of pixel values, random rotation, and flipping of images and labels, which increased data diversity and improved model robustness.

B. Implementation Details

The segmentation model adopts a U-shaped network as the backbone, augmented with ResNet residual blocks. The encoder structure resembles that of ResUNet. For the decoder, we define four stages with channel numbers set to 256, 128, 64, and 16 for each stage. Three shape prior modules are integrated into the model, which update features using self-attention mechanisms with 8 attention heads. To accommodate varying channel numbers of skip connection feature maps across layers in the model, we design different scaling factors for the shape prior modules, specifically 2, 4, and 8. All models in the experiments were optimized using the AdamW [35] optimizer, initialized with a learning rate of 5e-4. A cosine annealing strategy dynamically adjusts the learning rate, and a warm restart is performed every 25 epochs to help the model escape local optima and improve generalization.

All experiments are conducted in a PyTorch 1.13.1 and CUDA 11.7 environment, with training performed on an NVIDIA GeForce RTX 4090 with 24GB of VRAM.

C. Baseline methods and metric

We compared our method with the basic 3D UNet [36] and several state-of-the-art UNet methods, which can be categorized into three types: 1) three-dimensional extension models of U-Net, namely 3D UNet; 2) improved methods without incorporating transformers, including nnUNet [37] and UXNet [38]; 3) improved methods incorporating transformers, namely SwinUNETR [19]. For evaluation metrics, we used the Dice Similarity Coefficient (DSC), the 95th percentile Hausdorff Distance (HD_{95}), and the Average Symmetric Surface Distance (ASSD). These metrics comprehensively evaluate the model's performance in terms of overall overlap between segmentation results and labels, maximum boundary error, and overall surface distance.

$$DSC(A,B) = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{15}$$

$$HD_{95}(A,B) = \max\left\{\sup_{a\in A}\inf_{b\in B}\|a-b\|_2, \sup_{b\in B}\inf_{a\in A}\|b-a\|_2\right\}_{95\%} \tag{16}$$

$$ASSD(A,B) = \frac{1}{|A|+|B|}\left(\sum_{a\in A}\inf_{b\in B}\|a-b\|_2 + \sum_{b\in B}\inf_{a\in A}\|b-a\|_2\right) \tag{17}$$

where A denotes the set of segmentation results, B denotes the set of ground truth labels, a and b represent elements from A and B respectively, and $\max\{\cdot\}_{95\%}$ denotes the 95th percentile maximum value of the calculated distance set.

D. Results and discussion

KiPA 22: This dataset presents segmentation challenges due to morphological variations, distribution changes, class imbalance, and low salience across four categories. The model needs to extract and utilize global information and capture long-range spatial dependencies within the image.

Our model was evaluated on the KiPA 22 dataset, and the trends of the loss function and Dice Coefficient are shown in Fig. 2. Both the training and validation losses drop rapidly and then stabilize, while the Dice Coefficient steadily approaches 0.9. This shows that our model achieves accurate segmentation of medical images and exhibits good generalizability.

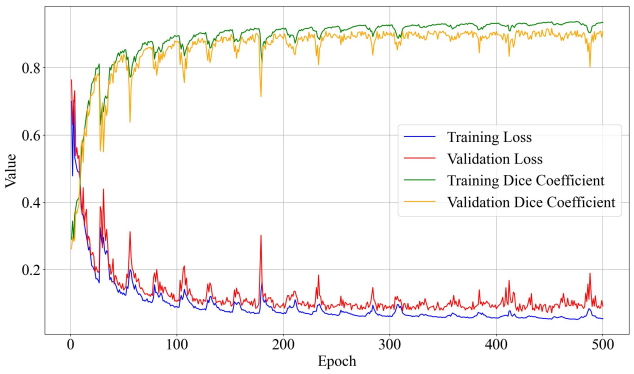

Fig. 2. Trends of loss and DSC during training and validation of KiPA 22.

As shown in Fig. 3, previous models exhibited segmentation errors to varying degrees. For nnUNet and UXNet, segmentation errors primarily occurred in the renal arteries and veins. We attribute this to the lack of global information guidance, which prevents capturing long-range spatial dependencies in the images. SwinUNETR, on the other hand, particularly misidentified kidney tissue as tumors, performing only slightly better than 3D UNet. We believe this is due to insufficient integration of U-shaped network and transformer elements in SwinUNETR, resulting in a loss of the U-shaped network's capability for local feature extraction and segmentation. In contrast, our model did not exhibit significant segmentation errors.

Further quantitative analysis was conducted, and the experimental results are shown in Table I. Our model achieved the best results, with a DSC improvement of 5.2636%, a reduction in HD95 by 9.9105mm, and a reduction in ASSD by 1.9110mm compared to the baseline 3D UNet. Compared to the second-best performing nnUNet, our model's DSC improved by 1.4587%, HD95 decreased by 0.4041mm, and ASSD decreased by 0.1526mm. The results indicate that our model's improvements, incorporating shape priors, outperform other methods, demonstrating stronger global feature extraction and utilization capabilities while retaining the local detail segmentation advantages of the U-shaped network.

ACDC: As shown in Fig. 4, all models except ours exhibited significant errors such as missing right ventricles and excessively thin or even missing myocardium in their segmentation results. From the segmentation evaluation metrics presented in Table II, it is evident that, apart from the baseline model, the results of different models were highly similar, especially for DSC, which approached 90.00%. Our model, while avoiding significant segmentation errors, achieved a DSC exceeding 92.00%, reduced HD95 to below 2mm, and decreased ASSD to below 0.5mm.

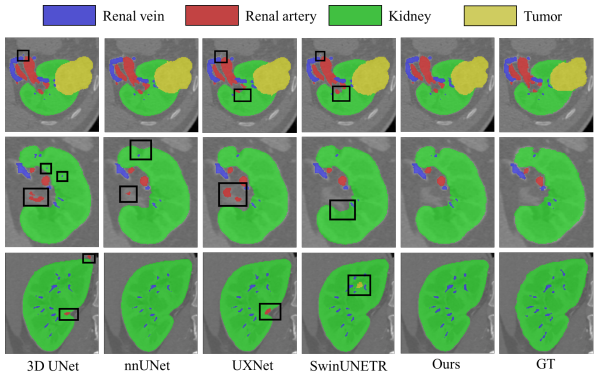

Fig. 3. Examples of segmentation results of different models on the KiPA 22 dataset.

TABLE I
SEGMENTATION ACCURACY OF DIFFERENT MODELS ON THE KIPA 22 DATASET.

Segmentation Method	MIICA KiPA 22 Dataset		
	DSC↑ *(%)*	*HD_{95}*↓ *(mm)*	*ASSD*↓ *(mm)*
3D U-Net	0.8192	12.2498	2.5767
nnU-Net	0.8609	5.9381	1.2620
UX-NET	0.8600	5.9266	1.1605
SwinUNETR	0.8428	6.2783	1.5369
Ours	**0.8718**	**2.3394**	**0.6656**

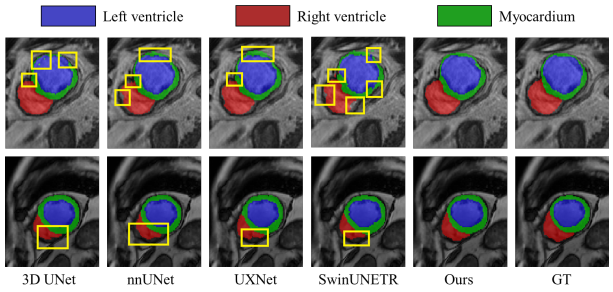

Fig. 4. Examples of segmentation results of different models on the ACDC dataset.

Our model demonstrates superior segmentation performance on the ACDC dataset after 500 training iterations, effectively tackling challenges such as brightness non-uniformity within the left and right ventricular cavities caused by blood flow, similarity in intensity among trabeculae, papillary muscles, and the myocardium, as well as notable variations in the

TABLE II
SEGMENTATION ACCURACY OF DIFFERENT MODELS ON THE ACDC DATASET.

Segmentation Method	MIICA ACDC Dataset		
	***DSC*↑ (%)**	***HD*$_{95}$↓ (mm)**	***ASSD*↓ (mm)**
3D U-Net	0.9048	2.1060	0.5848
nnU-Net	0.9054	2.0633	0.5248
UX-NET	0.8937	2.4225	0.7744
SwinUNETR	0.8914	2.6329	0.7704
Ours	**0.9202**	**1.6592**	**0.4322**

TABLE III
ABLATION EXPERIMENT ON THE KIPA 22 DATASET.

Segmentation Method	MIICA KiPA 22 Dataset		
	***DSC*↑ (%)**	***HD*$_{95}$↓ (mm)**	***ASSD*↓ (mm)**
Baseline U-Net	0.8614	3.4544	0.8802
Baseline U-Net + Loss	0.8618	2.7183	0.8096
Baseline U-Net + Shape prior	0.8641	3.3163	0.8481
Complete model	**0.8718**	**2.3393**	**0.6656**

TABLE IV
ABLATION EXPERIMENT ON THE ACDC DATASET.

Segmentation Method	MIICA ACDC Dataset		
	***DSC*↑ (%)**	***HD*$_{95}$↓ (mm)**	***ASSD*↓ (mm)**
Baseline U-Net	0.9078	1.9553	0.4987
Baseline U-Net + Loss	0.9093	1.9329	0.4429
Baseline U-Net + Shape prior	0.9103	1.9386	0.4885
Complete model	**0.9202**	**1.6592**	**0.4322**

shape and intensity of cardiac structures across diverse patients and pathologies. This is achieved by integrating the global information extracted by the shape prior module with the local information extracted by the U-shaped network, thus demonstrating the effectiveness of our approach.

E. Ablation studies

To validate the contributions of our designed shape prior module and loss function to the model's performance, we conducted systematic ablation experiments on the KiPA 22 dataset and the ACDC dataset. The experiments were divided into four groups. The first group served as the baseline model, utilizing a 3D U-Net with residual blocks but without the shape prior module or the comprehensive loss function, acting as a comparison benchmark. The second group combined the newly designed comprehensive loss function with the baseline model to assess the contribution of the loss function to segmentation performance. The third group integrated the shape prior module into the baseline model without the new loss function to evaluate the independent impact of the shape prior module. The fourth group incorporated both the shape prior module and the comprehensive loss function, representing our proposed complete model.

The detailed experimental results are presented in Table III and Table IV. The addition of the comprehensive loss function improved segmentation accuracy to a certain extent. In particular, the significant reductions in HD_{95} and ASSD metrics indicate the loss function's ability to handle local details, enhancing the segmentation performance in boundary regions. Comparing the shape prior-enhanced model with the original baseline model, it can be seen that the shape prior module improved global geometric consistency. Furthermore, comparing the results of the complete model with other models demonstrates the synergistic effect of the shape prior module and the loss function in enhancing model performance, effectively balancing global shape consistency with local boundary accuracy.

V. CONCLUSION

We propose a 3D medical image segmentation framework that incorporates layer-by-layer updated shape priors into a traditional U-shaped network architecture. Our method has been validated on the KiPA 22 and ACDC datasets. These datasets present significant challenges for medical image segmentation due to structural irregularities, variations in appearances, severe class imbalances, and difficulties in distinguishing specific anatomical features from the background. Additionally, manual expert segmentation and patient variability contribute to the complexity. According to three evaluation metrics, our method outperforms all comparative approaches. With its robustness and superior performance, our model shows great promise in enhancing clinical examinations and treatments by providing more accurate diagnostic tools.

Despite these advancements, our method faces several computational challenges: 1) the integration of shape priors increases computational complexity, and 2) the training process exhibits slower convergence. In future work, we will explore methods to decrease the computational load while maintaining segmentation accuracy.

REFERENCES

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

[2] R. A. Zeineldin, M. E. Karar, O. Burgert, and F. Mathis-Ullrich, "Multimodal cnn networks for brain tumor segmentation in mri: a brats 2022 challenge solution," in *International MICCAI Brainlesion Workshop*. Springer, 2022, pp. 127–137.

[3] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt, "Automatic multi-organ segmentation on abdominal ct with dense v-networks," *IEEE transactions on medical imaging*, vol. 37, no. 8, pp. 1822–1834, 2018.

[4] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

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

[8] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[9] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 1055–1059.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016, pp. 630–645.

[11] S. Huang, Z. Lu, K. Deb, and V. N. Boddeti, "Revisiting residual networks for adversarial robustness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8202–8211.

[12] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2736–2746.

[13] J. Ruan, M. Xie, J. Gao, T. Liu, and Y. Fu, "Ege-unet: an efficient group enhanced unet for skin lesion segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2023, pp. 481–490.

[14] C. Williams, F. Falck, G. Deligiannidis, C. C. Holmes, A. Doucet, and S. Syed, "A unified framework for u-net design and analysis," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[15] J. Tang, G. Zheng, C. Shi, and S. Yang, "Contrastive grouping with transformer for referring image segmentation," pp. 23 570–23 580, 2023.

[16] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[17] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "Max-deeplab: End-to-end panoptic segmentation with mask transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5463–5474.

[18] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. Goh, "Medical image segmentation using squeeze-and-expansion transformers," *arXiv preprint arXiv:2105.09511*, 2021.

[19] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI brainlesion workshop*. Springer, 2021, pp. 272–284.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[21] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[23] Z. Wang, Y.-L. Li, X. Chen, H. Zhao, and S. Wang, "Uni3detr: Unified 3d detection transformer," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[24] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and

Y. Yu, "nnformer: Interleaved transformer for volumetric segmentation," *arXiv preprint arXiv:2109.03201*, 2021.

[25] Y. Fang, B. Liao, X. Wang, J. Fang, J. Qi, R. Wu, J. Niu, and W. Liu, "You only look at one sequence: Rethinking transformer in vision through object detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26 183–26 197, 2021.

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

[27] F. Sun, Z. Luo, and S. Li, "Boundary difference over union loss for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 292–301.

[28] Y. He, G. Yang, J. Yang, R. Ge, Y. Kong, X. Zhu, S. Zhang, P. Shao, H. Shu, J.-L. Dillenseger *et al.*, "Meta grayscale adaptive network for 3d integrated renal structures segmentation," *Medical image analysis*, vol. 71, p. 102055, 2021.

[29] Y. He, G. Yang, J. Yang, Y. Chen, Y. Kong, J. Wu, L. Tang, X. Zhu, J.-L. Dillenseger, P. Shao *et al.*, "Dense biased networks with deep priori anatomy and hard region adaptation: Semi-supervised learning for fine renal artery segmentation," *Medical image analysis*, vol. 63, p. 101722, 2020.

[30] P. Shao, C. Qin, C. Yin, X. Meng, X. Ju, J. Li, Q. Lv, W. Zhang, and Z. Xu, "Laparoscopic partial nephrectomy with segmental renal artery clamping: technique and clinical outcomes," *European urology*, vol. 59, no. 5, pp. 849–855, 2011.

[31] P. Shao, L. Tang, P. Li, Y. Xu, C. Qin, Q. Cao, X. Ju, X. Meng, Q. Lv, J. Li *et al.*, "Precise segmental renal artery clamping under the guidance of dual-source computed tomography angiography during laparoscopic partial nephrectomy," *European urology*, vol. 62, no. 6, pp. 1001–1008, 2012.

[32] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[33] S. McKinley and M. Levine, "Cubic spline interpolation," *College of the Redwoods*, vol. 45, no. 1, pp. 1049–1060, 1998.

[34] N. Bhatia *et al.*, "Survey of nearest neighbor techniques," *arXiv preprint arXiv:1007.0085*, 2010.

[35] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[36] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 424–432.

[37] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[38] H. H. Lee, S. Bao, Y. Huo, and B. A. Landman, "3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation," *arXiv preprint arXiv:2209.15076*, 2022.