Learning with Explicit Shape Priors for Medical Image Segmentation

Xin You, Junjun He, Jie Yang, *Senior Member, IEEE*, Yun Gu, *Member, IEEE*

***Abstract*— Medical image segmentation is a fundamental task for medical image analysis and surgical planning. In recent years, UNet-based networks have prevailed in the field of medical image segmentation. However, convolutional neural networks (CNNs) suffer from limited receptive fields, which fail to model the long-range dependency of organs or tumors. Besides, these models are heavily dependent on the training of the final segmentation head. And existing methods can not well address aforementioned limitations simultaneously. Hence, in our work, we proposed a novel shape prior module (SPM), which can explicitly introduce shape priors to promote the segmentation performance of UNet-based models. The explicit shape priors consist of global and local shape priors. The former with coarse shape representations provides networks with capabilities to model global contexts. The latter with finer shape information serves as additional guidance to relieve the heavy dependence on the learnable prototype in the segmentation head. To evaluate the effectiveness of SPM, we conduct experiments on three challenging public datasets. And our proposed model achieves state-of-the-art performance. Furthermore, SPM can serve as a plug-and-play structure into classic CNNs and Transformer-based backbones, facilitating the segmentation task on different datasets. Source codes are available at https://github.com/AlexYouXin/Explicit-Shape-Priors.**

***Index Terms*— Medical image segmentation, explicit shape prior, UNet, BraTS 2020.**

I. Introduction

Medical image segmentation is regarded as one of the most essential and challenging tasks for medical image analysis, which is a prerequisite for image-guided diagnosis and computer-assisted intervention [23]. It provides anatomical shape information by making per-pixel predictions for organs or lesions in images [31]. Recently, deep learning techniques [45] have dominated medical image segmentation. The most successful architectures are U-shape networks [43]. The multi-scale features are extracted for specific regions in the encoder, which contain semantic and detail information. Then deep features from the bottleneck are fused with encoded features via skip connections in the decoder structure. Lastly, per-pixel classifications are carried out on decoded features via the segmentation head [59], which is a learnable prototype. Deep networks are free from hand-crafted features to achieve outstanding performance for segmentation.

This work is supported in part by the Open Funding of Zhejiang Laboratory under Grant 2021KH0AB03, in part by the Shanghai Sailing Program under Grant20YF1420800, and in part by NSFC under Grant 62003208, and in part by Shanghai Municipal of Science and Technology Project, under Grant 20JC1419500 and Grant 20DZ2220400. (Corresponding author: Yun Gu.)

X. You, J. Yang and Y. Gu are with the Institute of Image Processing and Pattern Recognition, and also with the Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai 200240, China. (Email: {sjtu_youxin, jieyang, geron762}@sjtu.edu.cn)

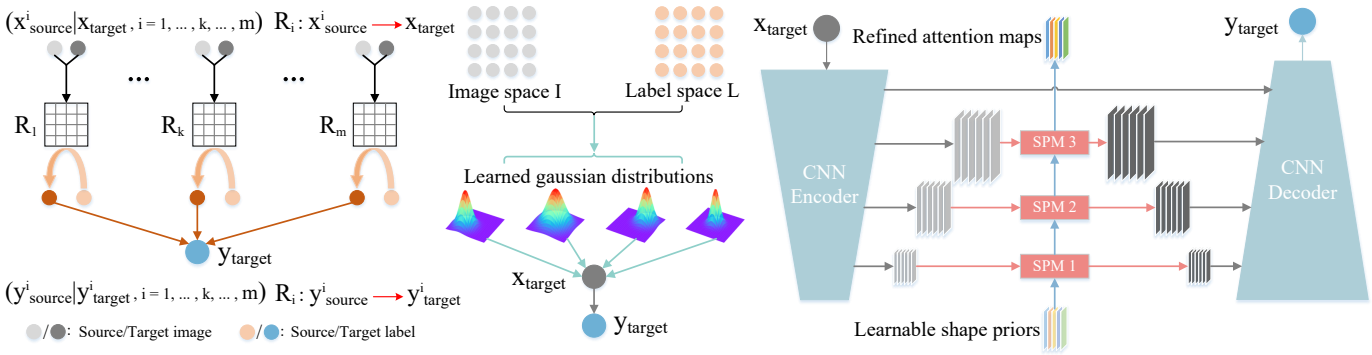
J. He is with Shanghai AI Laboratory, Shanghai 200240, China. (Email: hejunjun@pjlab.org.cn)

A. Limitations

However, current UNet-based models suffer from the following limitations for medical image segmentation. ① CNNs bear limited receptive fields due to intrinsic properties of convolutional kernels, which cannot exploit long-range and global spatial relations between organs or tissues, then they fail to achieve fine shape representations. Thus, some implicit attention modules [17], [24], [41], [52] are employed to enlarge models' receptive fields, and they are termed implicit shape models in our work. We will talk about them comprehensively in Section I-B. ② Segmentation masks are primarily based on the training of the final learnable prototype [59] interpreted by the segmentation head. Specifically, considering a segmentation task with N semantic classes, N class-wise prototypes are learned for pixel-wise classification. Only one learnable prototype is learned for each class, which employs limited representation abilities, thus insufficient to describe rich intra-class variance. Under this circumstance, UNet-based models meet a major challenge to extract precise shape information of organs or tumors. Specific loss functions [18], [37] are designed to integrate explicit shape priors or anatomical constraints to segmentation frameworks instead of Dice loss or cross-entropy loss, which can extract sufficient structure information related to the regions of interest, including shapes and topology. However, these loss functions are task-specific and cannot be easily extended on different datasets. Moreover, explicit shape models [22], [29] are proposed to enhance models' capacities for shape representations, with shape priors as an additional input. More thorough descriptions can be referred to in Section I-C.

B. Implicit Shape Models

To address the limitation on restricted receptive fields of CNNs, previous works try to introduce implicit anatomical shape priors to the U-shape structures, which are termed implicit shape models. These shape priors can be injected into



(a) Atlas-based model (b) Statistical shape model: GMM (c) Ours

Fig. 1. The comparison between different segmentation paradigms with explicit shape priors. (a) Atlas-based models employ m ground truths from source datasets as shape priors, and construct the transformation matrix $\mathcal{R}$ between source and target images. Then $\mathcal{R}$ is applied to shape priors for achieving segmentation masks. (b) Gaussian Mixture Model (GMM) gathers N (the number of segmentation classes) learnable Gaussian distributions as shape priors to deal with the per-pixel classification. (c) Our model adopts N-channel learnable shape priors as additional inputs to boost the segmentation performance of deep neural networks.

network structures via implicit attention modules. Theoretically, attention modules $\mathcal{M}$ [10], [17], [24], [41], [52] are introduced to strengthen deep features generated from encoders $\mathcal{E}$, driving them more focused on foreground regions with specific shapes. Then decoders $\mathcal{D}$ merge enhanced features to attain more precise masks $\mathcal{Y}$ of corresponding images $\mathcal{X}$. This process can be described as follows:

$$\mathcal{Y} = \mathcal{D}(\mathcal{M}(\mathcal{E}(\mathcal{X}))) \tag{1}$$

More specifically, the attention modules can be divided into two categories. The first type belongs to convolution-based attention modules. BB-UNet [17] enhanced skipped features via bounding box (BB) filters generated before training. Though BB filters can provide shape information of specific organs, obtaining BB filters requires manual interventions. Besides, Attention UNet [41] employs attention gates (AGs) to enhance salient features beneficial for specific tasks. AGs can also suppress redundant feature activation from irrelevant regions, which bear shape priors to some extent. SE blocks [24] adaptively recalibrate channel-wise feature responses by modeling inter-dependencies between channels. CBAM [52] adopted the channel attention module and spatial attention module to boost representation abilities on the shape of specific regions. However, stacking convolution-based attention modules cannot efficiently broaden effective receptive fields [36]. And they still show limited competence to model long-term dependency.

Different from that, the second type of attention module is based on the self-attention mechanism [16], which provides a feasible way of modeling global contexts via query, key, and value vectors. Many Transformer-based models with various types of self-attention [10], [20], [48], [49], [58] are proposed to model the long-range dependency of medical images. TransUNet [10] combines 2D UNet with a pre-trained Vision Transformer (ViT) to solve volumetric image segmentation by stacking each slice's prediction. SwinUNETR [48] adopts the shifted window-based attention to extract features of 3D patches, then merge multi-scale encoded features via residual convolutional blocks to attain final masks. However, unlike CNNs, these Transformer-based models require large data resources for training, which fail to simply and finely learn inductive bias such as shape prior information inside data sources [54].

C. Explicit Shape Models

To relieve the heavy dependence on the training of the final learnable prototype, prior methods attempt to introduce additional shape information into the segmentation framework, which we call explicit shape priors. Different from implicit shape priors mentioned above, explicit shape priors show strong interpretability, which presents a rough localization for the regions of interest (ROIs). These works can be divided into two categories, including atlas-based models [29], and statistical shape models, represented by Gaussian Mixture Model (GMM) [47].

Specifically, the first paradigm is based on the atlas, whose essence is the label propagation via registration transform between source and target images [6]. Then after applying this transform to source ground truths (GTs), we can attain GTs of target images. Obviously, nonrigid registration cannot perfectly deal with the segmentation task due to limited data sources and imaging noise [1]. Thus, a more feasible solution is to build matching relations in a non-local way. Besides, it is beneficial to achieve a more refined segmentation mask by adopting the weighted combination of a group of candidates from source images, which are called registration bases. The segmentation masks of registration bases serve as shape priors to promote the segmentation of target images. The whole model can be described by Eq (2):

$$\mathcal{Y}_{test} = \sum_{i=1}^{m} \omega_i \times \mathcal{T}(\mathcal{Y}_b^i; \mathcal{R}(\mathcal{X}_b^i \Rightarrow \mathcal{X}_{test})), \mathcal{X}_b^i \in \mathcal{X}_{train}, \mathcal{Y}_b^i \in \mathcal{Y}_{train} \tag{2}$$

where $\mathcal{X}_{train}$ and $\mathcal{Y}_{train}$, $\mathcal{X}_{test}$ and $\mathcal{Y}_{test}$ represent images and GTs from the training and testing data sources, $\mathcal{X}_b$ and $\mathcal{Y}_b$ represent the group of registration bases containing m pairs of source images and GTs, ω_i is the weighting coefficient. Besides, $\mathcal{R}$ refers to the registration transform between $\mathcal{X}_b^i$ and $\mathcal{X}_{test}$. And $\mathcal{T}$ is the transformation matrix, which applies the registration transform to each GT of registration bases $\mathcal{Y}_b^i$.

For atlas-based models, there exists a large computational cost during inference. Furthermore, the choice of registration

bases is significant to the model's robustness. Detailedly, the base vector should cover the distribution property of the whole dataset. However, adopting fixed template shapes cannot cover all biological objects due to the existence of shape variations. Thus, gathering a number of shape priors in a statistical way from training datasets is essential to boost models' segmentation performance and robustness.

The second segmentation paradigm is statistical shape models [22]. A representative method is the Gaussian Mixture Model (GMM) [42], which completes a consecutive mapping from image space I to label space L via a group of learnable Gaussian distributions. And these Gaussian distributions can be regarded as explicit shape priors of the dataset. During training, the Expectation Maximization (EM) algorithm is iteratively implemented to update the learnable Gaussian distributions and segmentation masks [32], [47]. In the inference process, we utilize learned shape priors as independent kernels. The whole model is illustrated by the following equation:

$$\mathcal{Y}_{test} = \underset{i=1,\ldots,N}{\arg\max}\, \mathcal{G}(\mathcal{X}_{test}; \mathcal{K}_i), \mathcal{K}_i \in \mathcal{K}(\mathcal{X}_{train}, \mathcal{Y}_{train}) \quad (3)$$

Where $\mathcal{K}$ means learned Gaussian distributions generated from the training process, $\mathcal{K}_i$ refers to each element in $\mathcal{K}$, N is the number of Gaussian kernels, which is also the number of semantic classes. $\mathcal{G}$ is a mapping function by distributing n Gaussian probability values to each pixel, each value generated from kernel $\mathcal{K}_i$. However, GMM is still sensitive to noises and dynamic backgrounds. Besides, the initial setting of the EM algorithm is crucial to the final solution [42].

Some other related works expand the impacts of statistical shape models. The point distribution model (PDM) [15] was devised to represent the mean geometry of a shape and some statistical modes of geometric variation inferred from a training set of shapes. Active shape model (ASM) [14] is a statistical model of the object shape that iteratively deforms to fit an example of the object in a new image. The shapes are constrained by the point distribution model to vary in ways seen in a training set of labeled examples. However, ASM only uses shape constraints and does not take advantage of all the available information. Thus, the active appearance Model (AAM) [13] was proposed for matching a statistical model of object shape and appearance to a new image. Specifically, this algorithm uses the difference between the current estimate of appearance & shape and the target image to drive the optimization process. Besides, constrained local models [55] combine a global shape model with local texture models for the delineation of every landmark point. This method is effective in modeling the shape deformation in local regions. Nevertheless, these explicit shape models cannot be embedded into deep segmentation networks, which still suffer from poor generalization abilities for unseen datasets.

D. Contributions

To address those two limitations in the meantime, we incorporate learnable explicit shape priors to enhance shape representations of UNet-based models. In the field of object detection, DETR [8] introduced a set of learnable object queries, then reasons about the relations of the objects and the global image context to directly output the final predictions. Motivated by this design, we devise the learnable shape prior, which is indeed an N-channel vector, and each channel contains rich shape information for the specific class of regions. Shape priors are generated based on self-attention, which endow the segmentation model with global receptive fields. Meanwhile, learnable shape priors can boost encoded deep features with richer shape information, then drive networks to generate better masks, which can ease the heavy dependence on the learnable prototype. Also, encoded features will contribute to iterative updates of shape priors. Based on this theory, we propose the shape prior module (SPM) consisting of the self-update block (SUB) and cross-update block (CUB).

Firstly, SUB is devised to generate global shape priors of specific datasets. Based on the self-attention mechanism [16], learnable shape priors are globalized to model the inter-class relations by calculating the similarity between each channel pair of shape priors. Here each channel in shape priors corresponds to shape representations of the specific anatomical structure. Thus, SUB plays an essential role in modeling the long-range dependency of datasets, which alleviates the drawback of convolution-based implicit attention modules.

Secondly, the structure of SUB is deficient in reductive bias [54] to model refined shape representations of anatomies. That is why CUB is designed to characterize local shape priors. Convolutional features from the encoder reveal strong abilities to localize discriminative structures [60]. Hence, the interaction between convolutional features and global shape priors will output local priors with finer shape representations, which mitigates the convergence challenge of self-attention-based implicit shape enhancement modules. Meanwhile, shape priors can enrich convolutional features with abundant global contexts, including texture and structural information.

Thirdly, learnable shape priors are explicitly incorporated into segmentation models for better performance with good interpretability as illustrated by Fig.1. And they are capable of relieving the heavy reliance on prototype learning. In comparison with other explicit shape models, our proposed SPM presents stronger generalization abilities on different datasets. Specifically, learnable N-channel shape priors are more robust than the atlas selected from training sets due to the fact that fixed templates cannot cover the distribution property of the whole dataset. Besides, deep segmentation networks with SPM are less sensitive to background noise and the initial solution in the optimization process compared to statistic shape models (SSM) including GMM, ASM, etc.

Lastly, to demonstrate the efficacy of the proposed SPM, we carry out evaluations on three public datasets, containing BraTS 2020, ACDC, and VerSe 2019. Here SPM is plugged into the position of skip connections as shown in Fig.1. And our model achieves state-of-the-art (SOTA) segmentation performance on these challenging datasets. Specifically, under the productive guidance of explicit shape priors, our model achieves an absolute gain of 1.06% in the term of Dice score on VerSe 2019, a significant decrease of $0.92mm$ in the term of HD_{95} on BraTS 2020, a finer connectivity for the shape of the myocardial region. Besides, due to its plug-and-play property, we probe into the generalization ability on other

networks, including classic CNNs and recent Transformer-based models. Our contributions are briefly summarized as follows:

1) We conduct a thorough comparative analysis of three types of segmentation paradigms with explicit shape priors, consisting of Atlas-based models, Statistical shape models represented by GMM and learnable shape priors plugged into deep segmentation networks.

2) We propose a novel shape prior module (SPM), comprised of the self-update and cross-update blocks. And they will generate global contexts and local shape priors with finer shape representations of anatomical structures.

3) The proposed module plugged into deep segmentation networks alleviates the limitations on networks' limited receptive fields and prototype learning. The enhanced segmentation model achieves SOTA performance on BraTS 2020, VerSe 2019, and ACDC.

4) SPM is a plug-and-play structure, which brings a significant boost to shape representations of classic CNNs and Transformer backbones.

II. Methodology

A. Unified Framework for Explicit Shape Models

As shown in Fig.1, we mainly discuss three types of segmentation paradigms, which can provide explicit shape priors. These paradigms can be unified as follows:

$$\mathcal{O} = \mathcal{D}(\mathcal{P}(\mathcal{I}; \mathcal{S})) \tag{4}$$

where $\mathcal{I}$ represents testing images as the input of the segmentation framework, and $\mathcal{O}$ refers to the model's outputs. $\mathcal{S}$ denotes explicit shape priors generated with different manners, which are used for enhancing the segmentation performance. $\mathcal{P}$ refers to the process of model prediction with joint inputs. $\mathcal{D}$ means the one-hot decoding on the generated N-channel prediction, and N is the number of segmentation classes.

In this work, the proposed paradigm introduces learnable explicit shape priors $\mathcal{S}$ to U-shape neural networks. Specifically, $\mathcal{S}$ is utilized as inputs of networks combined with images. The outputs of networks are predicted masks and attention maps generated by S. Then channels of attention maps can provide rich shape information of the ground truth region. The explicit-shape-prior model can be depicted as follows:

$$\mathcal{Y}_{test}, attention = \mathcal{F}(\mathcal{X}_{test}, \mathcal{S}(\mathcal{X}_{train}, \mathcal{Y}_{train})) \tag{5}$$

where $\mathcal{F}$ represents the forward propagation during inference, $\mathcal{S}$ stands for consecutive shape priors constructing the mapping between image space I and label space L. Here $\mathcal{S}$ is updated in the training process as the image-GT pair varies. Once training is finished, learnable shape priors are fixed, which can dynamically generate refined shape priors as input patches vary in the inference process. And refined shape priors serve as attention maps, which can localize regions of interest, and suppress background areas. Furthermore, a small portion of inaccurate ground truths will not affect the learning for $\mathcal{S}$ significantly, revealing the robustness of our proposed paradigm.

B. Shape Prior Module

Overview. As depicted in Fig.1, our proposed model is a hierarchical U-shape network, which consists of a ResNet-like encoder, a Resblock-based [48] decoder and the shape prior module (SPM). And SPM is a plug-and-play module, which can be flexibly plugged into other network structures to improve segmentation performance. In the sections below, we will give a detailed description of SPM, including the motivation for this module, the detailed structure, and the functions.

In order to get rid of the dependency on the final learnable prototype, we propose to introduce explicit shape priors to UNet-based networks, which will exert anatomical shape constraints for each class to enhance the representation abilities of networks. Motivated by DETR [8], we devise n (the number of segmentation classes) learnable prototypes, the analogy to object queries in the Transformer decoder of DETR. As shown in Fig.2, inputs of SPM are original skipped features F_o and original shape priors S_o, which are refined as enhanced skipped features F_e and enhanced shape priors S_e. Specifically, learnable shape priors will generate refined attention maps with sufficient shape information under the guidance of convolutional encoded features. In the meantime, encoded features will generate more accurate segmentation masks via shape priors. Different from DETR, SPM will interact with multi-scale features, not just features from the bottleneck of the encoder. Thus, hierarchical encoded features before skip connections will be equipped with richer shape information via SPM. Enhanced shape priors are made up of two components, global and local shape priors, generated from the self-update block and cross-update block respectively. We will give a more elaborate description of these two blocks.

Self-Update Block: Modeling long-range dependency. On the ground that we aim to introduce explicit shape priors which can localize the target regions, the size of shape priors $\mathcal{S}_o$ is $N \times$ spatial dimension. N refers to the number of classes, and spatial dimension is related to the patch size. To alleviate the drawback of limited receptive fields, the long-range dependency inside shape priors is considered in this work. Specifically, the self-update block (SUB) is proposed to model relations between inter-classes and generate global shape priors with interactions between N channels. Motivated by the self-attention mechanism of Vision Transformer (ViT) [16], the affinity map of self-attention S_{map} between N classes is constructed by the Eq (6), which describes the similarity and dependency relationship between each channel of shape priors.

$$S_{map} = Softmax(\frac{Q_s(\mathcal{S}_o) \times K_s(\mathcal{S}_o)^T}{\sqrt{N}}) \tag{6}$$

where Q_s and K_s represent convolutional transforms which project $\mathcal{S}_o$ into the query and key vector, T is the transpose operator, and the dimension for S_{map} is $N \times N$. The vanilla self-attention module shows quadratic computational complexity, which poses an obstacle to dense prediction tasks. Thus, many related works [9] attempt to reduce the computational cost of the self-attention module for faster convergence and less memory consumption. Here we set the spatial dimension of $\mathcal{S}_o$ as $h \times w \times l$, $\frac{1}{16}$ ratio of the patch size $H \times W \times L$.

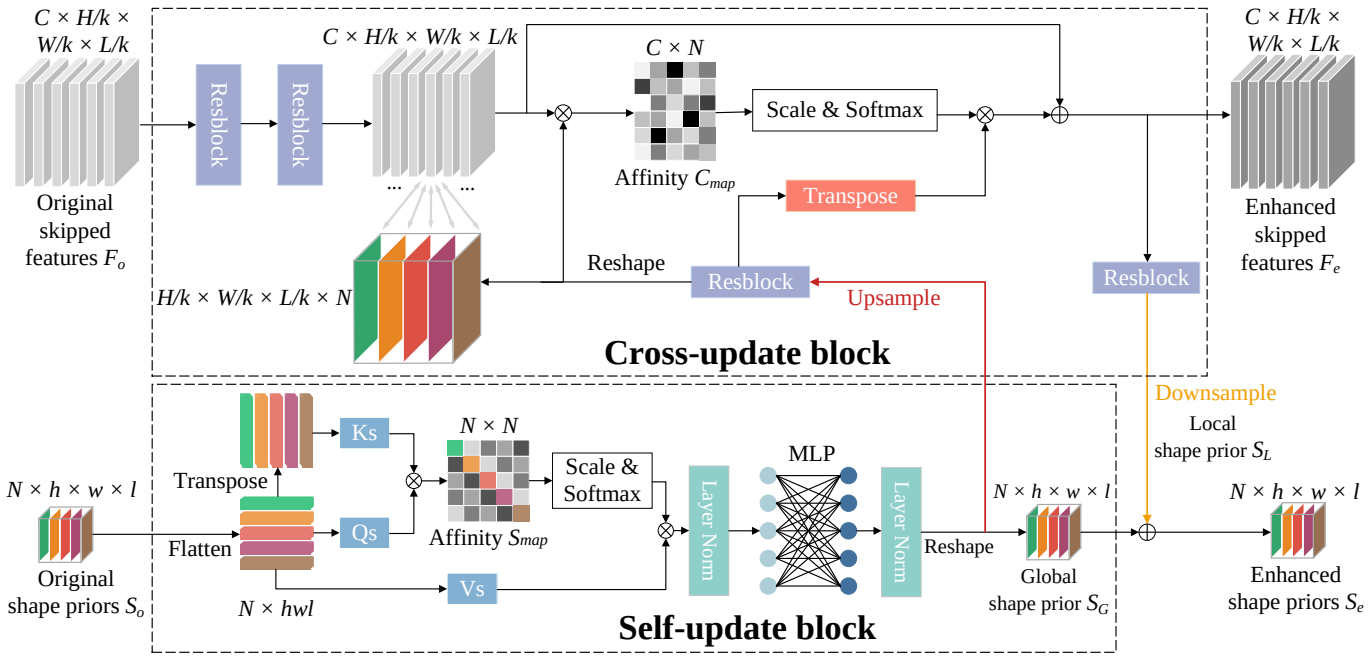


Fig. 2. Illustration of the shape prior module (SPM). SPM consists of the self-update block (SUB) and cross-update block (CUB). Specifically, SUB aims to extract global shape priors via a linear attention module. CUB introduces global shape priors as additional guidance to refine original skipped features $\boldsymbol{F_o}$. Also, the convolutional feature guides the generation of local shape priors with finer shape information. Here $\boldsymbol{H} \times \boldsymbol{W} \times \boldsymbol{L}$ is the input patch size (k = 2, 4, 8), and $\boldsymbol{h} \times \boldsymbol{w} \times \boldsymbol{l}$ is $\frac{1}{16}$ of the patch size.

Besides, $\mathcal{S}_o$ bears N tokens, which means the self-attention module in SUB is a linear attention module. And Eq (6) shows $\mathcal{O}(N^2 \times HWL)$ computational complexity.

After that, the weighted sum of S_{map} and value vector of $\mathcal{S}_o$, noted as $V_s(\mathcal{S}_o)$, are adopted to obtain global shape priors. This process also requires $\mathcal{O}(N^2 \times HWL)$ computational costs. To further model the long-range dependency inside learnable shape priors, multi-layer perceptron (MLP) and layer normalization (LN) are employed. The detailed process can be illustrated as follows:

$$\mathcal{S}' = LN(S_{map} \times V_s(\mathcal{S}_o)) + \mathcal{S}_o \tag{7}$$

$$\mathcal{S}_\mathcal{G} = LN(MLP(\mathcal{S}')) + \mathcal{S}' \tag{8}$$

where V_s represents the convolutional transform, $\times$ means the process of matrix multiplication, $\mathcal{S}_\mathcal{G}$ is global shape priors. Detailedly, $\mathcal{S}_\mathcal{G}$ can build the long-term dependency inside $\mathcal{S}_o$, which contains global contexts of sampled input patches, including coarse shape and positional representations combined with sufficient texture information for global regions.

Cross-Update Block: Modeling local shape priors. To relieve the dependence on the learnable prototype, we attempt to introduce explicit shape priors to boost the representation abilities for shape information. However, the structure of SUB falls lack of inductive bias [54] to model local visual structures and localize objects with various scales. As a result of that, global shape priors do not have precise shape and contour information. Further, models have to learn intrinsic inductive bias from large amounts of data for a longer training time. To address this limitation, we propose the cross-update block (CUB). Motivated by the fact that convolutional kernels intrinsically bear the inductive bias of locality and scale invariance, CUB based on convolution injects inductive bias to SPM for local shape priors with finer shape information. Moreover, based on the fact that convolutional features from the encoder have remarkable potentials to localize discriminative regions [60], we attempt to interact original skipped features F_o from the backbone with shape priors $\mathcal{S}_o$ as demonstrated in Fig.2.

Specifically, we calculate the similarity map between features F_o and shape priors $\mathcal{S}_o$. Here the dimension of F_o is $C \times \frac{H}{k} \times \frac{W}{k} \times \frac{L}{k} (k = 2, 4, 8)$, and C represents the channel number of features. However, F_o and $\mathcal{S}_o$ bear different scales, which makes it difficult to fuse two elements. Thus, we firstly upsample $\mathcal{S}_o$ to the same resolution as F_o, then integrate them based on the cross attention mechanism [8]. The detailed computational process is illustrated as Eq (9):

$$C_{map} = Softmax(\frac{Q_c(F_o) \times K_c(Upsample(\mathcal{S}_\mathcal{G}))^T}{\sqrt{N}}) \tag{9}$$

where C_{map} means the affinity map in the cross attention stage, Q_c and K_c represent convolutional transforms which project F_o and $\mathcal{S}_o$ into the query and key vector. C_{map} is a $C \times N$ matrix, which evaluates the relations between C-channel feature maps F_o and N-channel shape priors. The specific channel of convolutional feature maps F_o correlates with specific channels of shape priors. After that, C_{map} acts on transformed global shape priors $\mathcal{S}_\mathcal{G}$ to refine F_o, with more accurate shape characteristics and rich global textures.

$$F_e = C_{map} \times V_c(Upsample(\mathcal{S}_\mathcal{G})) + F_o \tag{10}$$

where V_c refers to convolutional transforms projecting $\mathcal{S}_o$ into the value vector, F_e represents enhanced skipped features. At the same time, local shape priors $\mathcal{S}_\mathcal{L}$ are generated from downsampled F_e, which bear the property to model local

TABLE I
COMPARISON WITH OTHER MODELS ON VERSE 2019. (CERV: CERVICAL VERTEBRAE, THOR: THORACIC VERTEBRAE, LUMB: LUMBAR VERTEBRAE, MEAN: THE AVERAGE EVALUATION METRIC OF ALL CASES, MEDIAN: THE MEDIAN EVALUATION METRIC OF ALL CASES.)

Method	Dice score (%) ↑					HD_{95} (mm) ↓					Params(M)	FLOPs(G)
	Cervical	Thoracic	Lumbar	Mean	Median	Cervical	Thoracic	Lumbar	Mean	Median		
3D UNet [12]	83.10	78.37	70.88	81.28	87.54	4.04	8.75	11.56	7.97	4.84	16.49	520.98
VNet [39]	86.32	87.78	73.45	85.57	92.14	2.82	4.77	10.29	5.24	2.27	45.73	957.88
nnUNet [28]	87.81	**88.80**	**74.96**	86.59	92.62	2.94	3.74	**8.39**	5.21	2.60	30.90	502.80
TransUNet (3D) [10]	85.49	82.67	73.88	83.53	88.06	2.56	4.90	9.38	5.43	3.34	146.68	682.96
CoTr [53]	81.48	79.68	68.83	80.59	85.72	5.57	10.01	17.42	11.34	7.77	48.53	507.68
UNeXt [49]	77.00	86.73	71.06	83.36	88.39	4.57	3.98	12.24	5.82	3.51	4.02	12.17
maskformer [11]	87.14	84.59	69.72	82.20	90.27	2.38	5.80	25.50	7.55	2.79	64.40	943.48
EG-Trans3DUNet [56]	83.67	82.41	74.11	86.01	91.12	3.38	4.06	9.87	5.32	3.18	161.89	748.20
Verteformer [57]	87.25	88.76	72.73	86.54	90.74	2.50	3.62	10.90	4.93	2.52	330.65	336.45
Swin UNETR [48]	89.30	81.43	73.36	83.46	88.91	2.21	7.79	10.09	7.66	5.22	62.19	732.18
Ours	**89.87**	88.69	74.15	**87.65**	**93.43**	**2.01**	**3.15**	9.08	**4.36**	**1.81**	46.55	457.63

visual structures (edges or corners).

$$\mathcal{S}_{\mathcal{L}} = Downsample(Conv(F_e)) \tag{11}$$

$$\mathcal{S}_e = \mathcal{S}_{\mathcal{L}} + \mathcal{S}_{\mathcal{G}} \tag{12}$$

In conclusion, original shape priors are enhanced with global and local characteristics [46]. Global shape priors can model the inter-class relations, which bear coarse shape priors with sufficient global texture information based on the self-attention block. Local shape priors show finer shape information via the introduction of inductive bias based on convolution. Besides, original skipped features are further enhanced via the interaction with global shape priors, which will promote generating features with discriminative shape representations and global contexts, then acquire more precise predicted masks.

III. EXPERIMENT

A. Datasets

In this work, we conduct experiments on three public datasets for segmentation including the Brain Tumor Segmentation (BraTS) 2020 challenge [2], [3], [38], the Large Scale Vertebrae Segmentation Challenge (VerSe 2019) [44] and the Automatic Cardiac Diagnosis Challenge (ACDC) [4].

BraTS 2020: This MRI dataset contains 369 training cases, 125 validation cases, and 166 testing cases. Each case bears the same volume size $155\times240\times240$ and the same voxel space $1\times1\times1\ mm^3$. Besides, each sample consists of four modality inputs, which are T1, T1-weighted, T2-weighted, and T2-FLAIR. The segmentation ground truth contains four classes, label 0 for background, label 1 for non-enhancing tumor core (NET), label 2 for peritumoral edema (ED), and label 4 for GD-enhancing tumor (ET). And the final evaluation metrics are Dice scores [39] and 95% Hausdorff distance HD_{95} [27] on three regions, ET region (label 4), tumor core (TC, including label 1 and 4), the whole tumor (WT, containing label 1, 2 and 4). Furthermore, we introduce the average Dice score and HD_{95} for an average evaluation of three regions.

VerSe 2019: This CT dataset is composed of 80 training cases, 40 validation cases, and 40 testing cases. There are 26 segmentation classes, including label 0 for the background and label 1-25 for 25 vertebrae. Of all 25 vertebrae, label 1-7 represents cervical vertebrae, label 8-19 for thoracic vertebrae and label 20-25 for lumbar vertebrae. Different samples show different field of views (FOVs), which means they may have different kinds of vertebrae. Here we select Dice scores and HD_{95} for cervical, thoracic, and lumbar. Besides, mean and median values for all testing cases are also reported.

ACDC: This dataset involves 100 MRI scans from 100 patients. The target ROIs are the left ventricle (LV), right ventricle (RV), and myocardium (Myo). And we follow the data split setting of nnFormer [58], with 70 training cases, 10 validation cases, and 20 testing cases.

Implementation Details. The proposed model is implemented with PyTorch 1.8.0 and trained on 2 NVIDIA Telsa V100, with a batch size of 2 in each GPU. For the BraTS 2020 dataset, all models are trained with the AdamW [35] optimizer for 2000 epochs, with a warm-up cosine scheduler for the first 50 epochs. The initial learning rate is set as $8e$-4 with $1e$-5 weight decay. And the size of cropped patches is $128\times128\times128$. We do not utilize complicated data augmentations like previous works [28], [58]. Instead, we adopt strategies of random mirror flipping, random rotation, random intensity shift and scale. For the VerSe 2019 dataset, we train all models for 1000 epochs. All preprocessed cases are cropped with a patch size of $128\times160\times96$. Random rotation between $[-15°, 15°]$ and random flipping along the XOZ or YOZ plane are employed for the data diversity. For the ACDC dataset, models are trained for 1500 epochs and the patch size is set as $20\times256\times256$. Similarly, we augment the cardiac data with random rotation and random flipping. And for both VerSe 2019 and ACDC, we choose the Adamw optimizer with the initial learning rate set as $5e$-4 and the cosine warm-up strategy for 50 epochs during training. Following the setup in [28], we choose a sum of Dice loss and cross-entropy loss for model training.

B. Experimental Results

Brain tumor segmentation: Table II illustrates the quantitative segmentation performance of our proposed model compared with other CNNs and Transformer-based models. It can be figured out that our model shows absolute superiority on the Dice score and HD_{95} of all three regions. Compared with TransBTSV2 [30], our model achieves higher Dice scores of 0.07%, 0.52%, 0.85% and a lower HD_{95} of $0.46mm$, $0.35mm$, $0.48mm$ on the ET, WT and TC region. This improvement results from the fact that enhanced shape priors

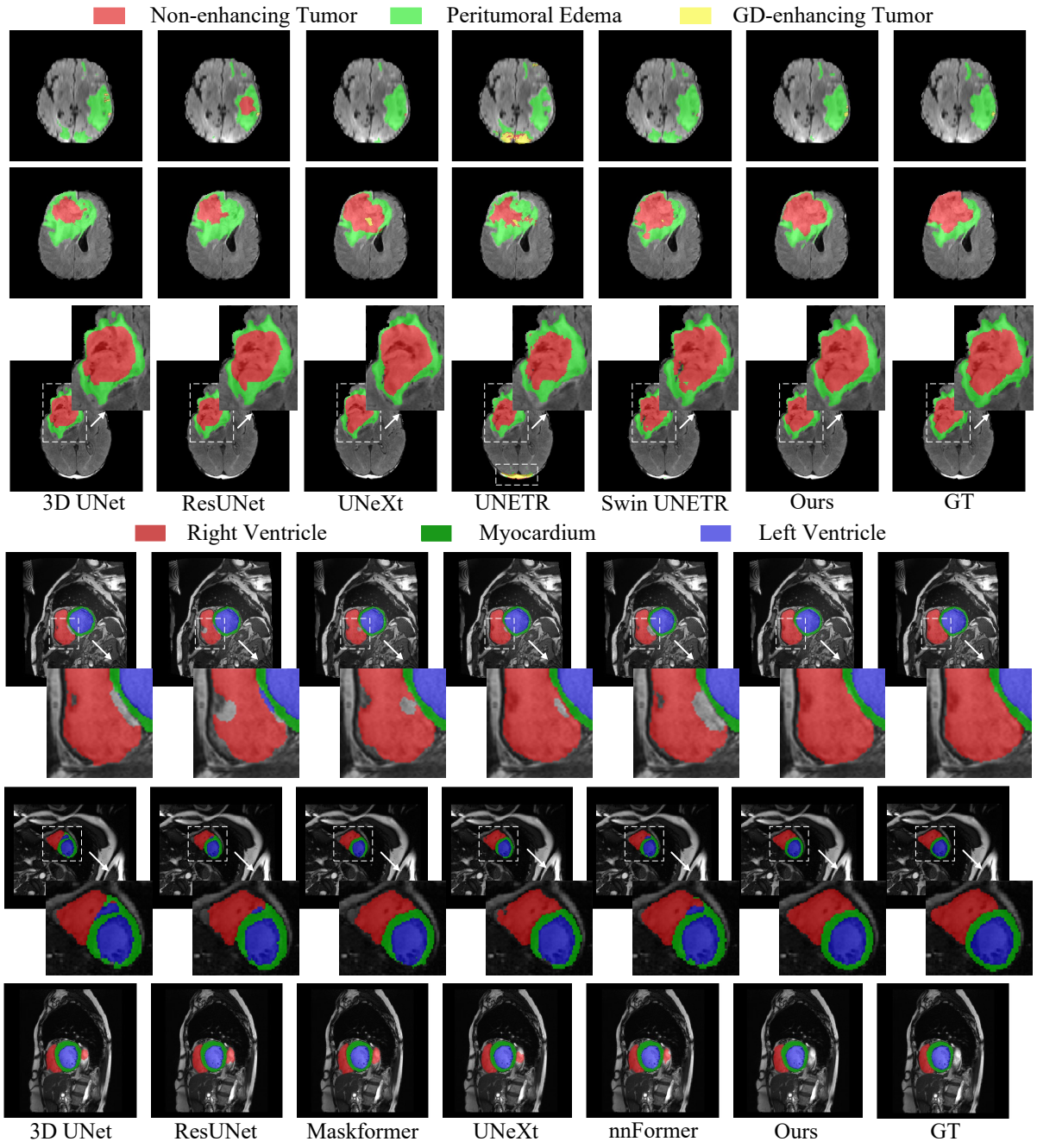


Fig. 3. Predicted masks of different models on BraTS 2020 and ACDC. Each column refers to a segmentation result of a network model.

serve as anatomical priors to be injected into networks, which eases the dependence on the final learnable prototype. Compared with nnUNet [28], the proposed model outperforms that powerful baseline on the HD_{95} metric, with a significant drop of $2.71mm$ and $0.92mm$ on the ET and the whole foreground regions. This performance superiority indicates that the shape prior module (SPM) can enlarge models' receptive fields. Furthermore, our model significantly surpasses other Transformer-based baselines, revealing a more reasonable inductive bias for unseen data. It is worth mentioning that there is a significant improvement in the metrics for WT and TC regions. We claim that our model with refined shape priors is aimed at enhancing shape representations for region WT and TC because they bear a relatively fixed shape in contrast to region ET. We will prove this viewpoint in the ablation study on the generalization abilities of SPM with different network structures.

Vertebrae segmentation: To further evaluate the performance of our proposed model, we conduct experiments on VerSe 2019. Table I presents the segmentation performance on the hidden test dataset. Our model outperforms the powerful nnUNet on the metric of cervical, and thoracic vertebrae. Specifically, there is 2.06%, 1.06% Dice score increases and $0.93mm$, $0.85mm$ HD_{95} decreases for cervical and the whole spine. Besides, different cases have different field of views (FoVs), which bring difficulty for models to identify the last several vertebrae. And nnUNet is superior to other models

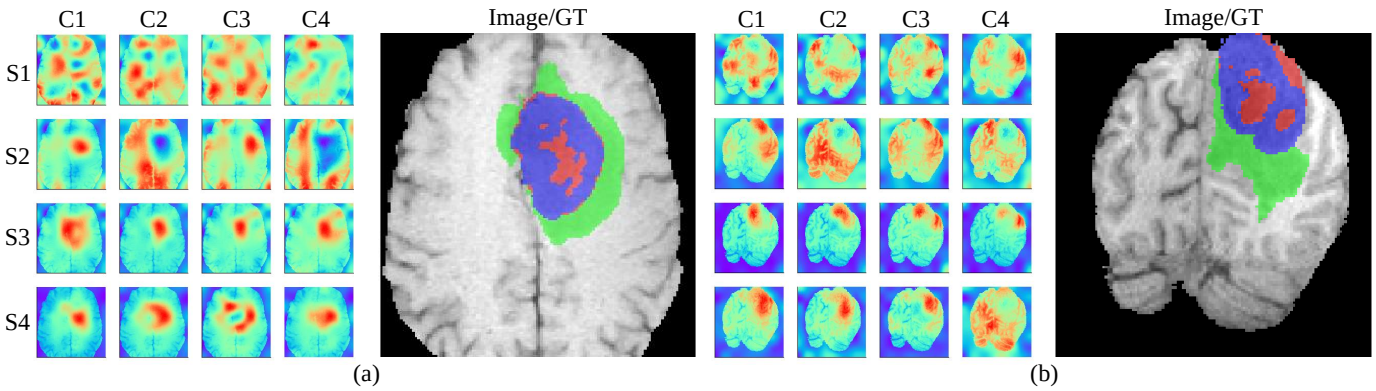


Fig. 4. Different channels (from Channel 1 to Channel 4) and different stages (from Stage 1 to Stage 4) of generated shape priors on two cases of BraTS 2020. C_i and S_i (i=0, 1, 2, 3) represent the respective channels and stages.

TABLE II

COMPARISON WITH OTHER MODELS ON BRATS 2020. (ET: THE GD-ENHANCING TUMOR, WT: THE WHOLE TUMOR, TC: THE TUMOR CORE, MEAN: THE AVERAGE EVALUATION METRIC OF THREE REGIONS.)

Method	Dice score (%) ↑				HD_{95} (mm) ↓				Params(M)	FLOPs(G)
	ET	WT	TC	Mean	ET	WT	TC	Mean		
3D UNet [12]	77.85	90.41	83.26	83.84	17.94	4.90	5.77	9.53	16.47	516.71
Liu et al. [34]	76.37	88.23	80.12	81.57	21.39	6.68	6.49	11.52	-	-
Vu et al. [50]	77.17	90.55	82.67	83.46	27.04	4.99	8.63	13.55	-	-
Nguyen et al. [40]	78.43	89.99	84.22	84.21	24.02	5.68	9.57	13.09	-	-
ResUNet [21]	78.64	90.48	85.18	84.77	17.77	6.56	5.46	9.93	17.16	334.31
TransUNet [10]	78.42	89.46	78.37	82.08	12.85	5.97	12.84	10.55	105.18	1035.52
Swin UNETR [48]	79.61	89.51	84.69	84.60	14.61	11.18	6.10	10.63	62.19	790.61
UNeXt [49]	76.49	88.70	81.37	82.19	16.61	4.98	11.50	11.04	4.02	14.92
TransBTS [51]	78.73	90.09	81.73	83.52	17.95	4.96	9.77	10.89	32.99	412.97
TransBTSV2 [30]	79.63	90.56	84.50	84.90	12.52	4.27	5.56	7.45	15.30	320.54
nnUNet [28]	79.37	91.05	85.11	85.18	14.79	**3.65**	5.36	7.94	30.43	534.36
Ours	**79.70**	**91.08**	**85.35**	**85.38**	**12.06**	3.92	**5.08**	**7.02**	43.53	438.23

TABLE III

COMPARISON WITH OTHER MODELS ON ACDC. METRIC: DICE SCORES (%). (RV: RIGHT VENTRICLE, MYO: MYOCARDIUM, LV: LEFT VENTRICLE, MEAN: THE AVERAGE EVALUATION METRIC OF ALL REGIONS.)

Method	RV	Myo	LV	Mean	Params (M)	FLOPs (G)
R50 Att-UNet [41]	87.58	79.20	93.47	86.75	107.60	639.09
TransUNet [10]	88.86	84.54	95.73	89.71	105.50	643.20
Swin-UNet [7]	88.55	85.62	95.83	90.00	27.17	118.00
UNETR [20]	85.29	86.52	94.02	88.61	93.10	195.62
MISSFormer [26]	86.36	85.75	91.59	87.90	42.46	187.80
Maskformer [11]	87.89	87.34	94.92	90.05	45.87	712.69
nnUNet [28]	90.24	89.24	95.36	91.61	30.43	471.86
nnFormer [58]	90.94	**89.58**	95.65	92.06	147.11	291.76
Ours	**92.28**	88.13	**96.61**	**92.34**	41.86	257.90

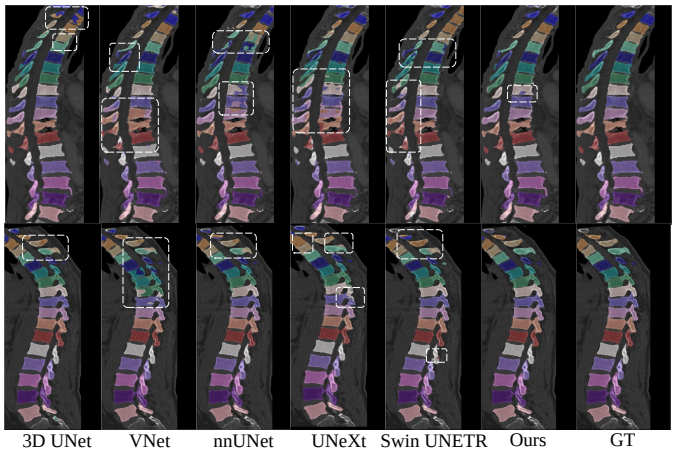


Fig. 5. Predicted masks of different models on VerSe 2019.

including ours on the ability to localize and segment lumbar vertebrae (Label 20-25). Furthermore, in contrast with other models, we achieve the highest median Dice score 93.43% and lowest HD_{95} $1.81mm$, which means that our model can boost the overall performance of testing cases. Due to the additional shape guidance, segmentation performance will not be restricted by the limited generalization ability of learnable prototypes. Fig.5 illustrates that our model presents better visualization results, with more consistent predictions in a singular vertebra. This phenomenon indicates that learnable explicit shape priors for vertebrae are truly effective in the refinement of predicted masks.

Automated cardiac segmentation: We also conduct quantitative and qualitative experiments on the ACDC dataset. As shown in Table III, our model outperforms the previous SOTA model nnFormer [58] on the evaluation metrics for the RV and LV regions. Specifically, the Dice scores of RV and LV reach 92.28% and 96.61%, with 1.34% and 0.96% higher than that of nnFormer. Fig.3 demonstrates that our model outputs more accurate segmentation masks, particularly in the RV and LV regions. That phenomenon proves that SPM can boost shape representations for anatomical structures with relatively fixed shapes. However, the segmentation performance for Myo is lower than that of nnUNet [28] and nnFormer. We argue that networks will be more focused on larger regions and ignore smaller regions due to the label imbalance between RV, LV and Myo [37]. Besides, this MRI dataset shows a large voxel space, which will aggravate the effect of label imbalance. And the unique and effective resampling strategy of nnUNet and nnFormer will improve the imbalanced distribution of the myocardium tissue, which results in stronger attention of models on this region. Thus, nnUNet and nnFormer achieve a higher Dice score on the Myo region.

C. Visualizations of shape priors and skipped features

In this section, we probe into the qualitative results of shape priors and skipped features. In fact, they are mutually enhanced. We first discuss the impact of skipped features on explicit shape priors. As mentioned in section II-A, explicit shape priors are iteratively updated under the guidance of skipped convolutional features, then optimized shape priors will activate regions of interest. We visualize two cases from the BraTS 2020 dataset in Fig.4. Case (a) illustrates generated explicit shape priors from different stages. Specifically, shape

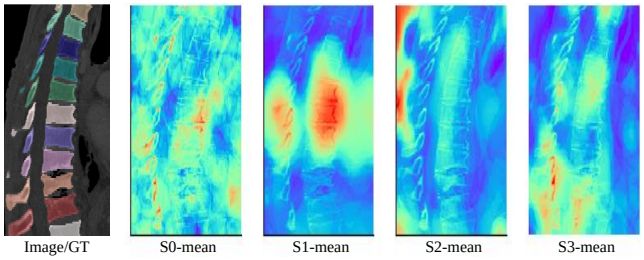


Fig. 6. Mean shape priors on Channel 1 to Channel 26 of different stages (from Stage 0 to Stage 3) on two cases of VerSe 2019. S_i-mean (i=0, 1, 2, 3) means average shape priors on channels of the i_{th} stage.

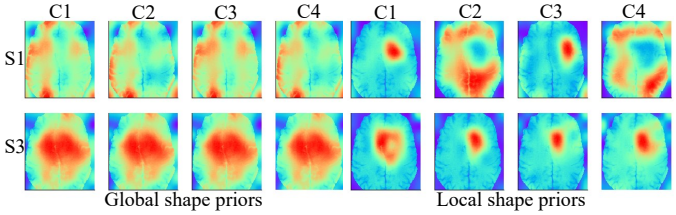


Fig. 7. Different channels and different stages of global and local shape priors on case (a) in Fig.4. C_i and S_i (i=0, 1, 2, 3) represent the respective channels and stages.

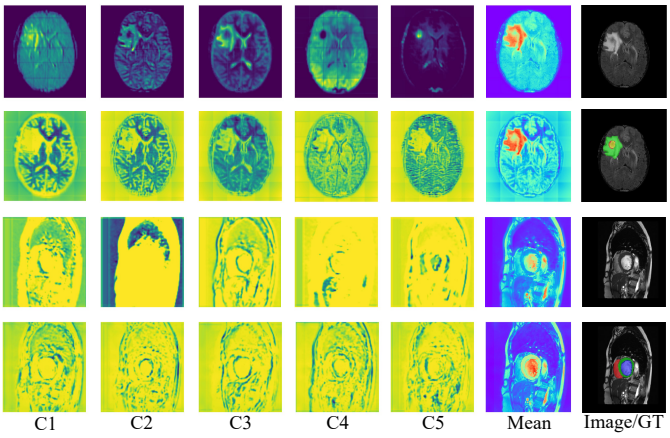


Fig. 8. Feature visualizations of baseline and baseline with SPM on BraTS 2020 and ACDC. C_i (i=1, 2, 3, 4, 5) represents the respective channel. 'Mean' refers to mean features across all 32 channels.

priors consist of N-channel attention maps, in which N refers to the number of segmentation classes, and each row represents shape priors from each stage. We can figure out that shape priors reveal more accurate activation maps for the ground truth region as the top-to-down process. In particular, wrongly activated regions in the first stage will be suppressed in the second and third stages of SPM. Here in our visualization results, there exists a phenomenon called the reverse activation [19], which means that all regions except the ground truth area are activated. A canonical example is visualized in the last stage and last channel of shape priors in case (b). We claim that this phenomenon results from the global shape priors, which bring global contexts and sufficient texture information for the whole region, even including regions from the background. In essence, it is simple to locate the ROIs via reverse attention, in which ROIs are highlighted with distinct contours. From this point of view, reverse activation is similar to positive activation.

Further, we decompose shape priors into two components, global and local shape priors generated from SUB and CUB respectively. We visualize these two components of case (a) in Fig.7. Due to the self-attention module [16], global shape priors bear globalized receptive fields, containing contexts and textures. However, the structure of SUB lacks inductive bias to model local visual structures. Here we can discover that global shape priors are responsible for a coarse localization for the region of ground truths. And local shape priors generated from CUB can provide finer shape information for the ROIs via the introduction of convolutional kernels, which bear the inductive bias of locality.

We then thoroughly analyze the impact of shape priors on skipped features with the direct comparison between original skipped features F_o and enhanced skipped features F_e. In detail, specific channels of F_o and F_e are selected from all 32 channels for a qualitative visualization. As shown in Fig.8, features in the tumor region are enhanced and some voxels which are not activated before, are highlighted after processing by SPM. Besides, via the introduction of global shape priors, skipped features are enriched with sufficient texture information for the whole region. We also explain the process of feature refinement with a cardiac CT case as shown in Fig.8. The channel-average skipped features are refined with more attention on the LV and Myo regions.

D. Ablation Studies

1) Plug-and-play: To prove the plug-and-play characteristic of SPM, we detailedly carry out ablation studies on the generalization ability of SPM on different network structures. Here we choose CNNs, Transformer-based and MLP-based models, including 3D UNet [12], ResUNet [21], UNETR [20], Swin UNETR [48] and UNeXt [49]. For evaluations on the BraTS 2020 dataset, we report the segmentation performance on 41 split validation cases during training. For the ACDC dataset, each MRI scan consists of thick slices, which is not applicable to the input size of UNETR and Swin UNETR. Thus, we only report quantitative results on 3D UNet, ResUNet and UNeXt.

According to Table IV, it can be observed that SPM can boost the segmentation performance of different networks. Specifically, SPM can bring an increase on ResUNet [21], with a Dice score increase of 0.77%, 0.51%, 1.48% on the ET, WT, TC region. And combined with SPM, 3D UNet [12]shows an improvement on the metric of HD_{95}, with a decrease of $0.43mm$, $0.67mm$, $0.07mm$ on region ET, WT, TC. Furthermore, SPM also upgrades the segmentation performance of Transformer-based models. SPM brings a Dice score increase of 1.27% and 0.97% on the WT region for UNETR [20] and Swin UNETR [48], which reveals the potential of SPM to significantly enhance the representation ability for regions with relatively regular shapes. And we can also explain this phenomenon from the perspective of inductive bias. With the introduction of shape priors, we introduce a strong inductive bias to Transformer-based models, which will relieve the requirements for a huge amount of datasets and accelerate the convergence of Transformers. Besides, SPM can improve the segmentation performance of the enhanced tumors, which

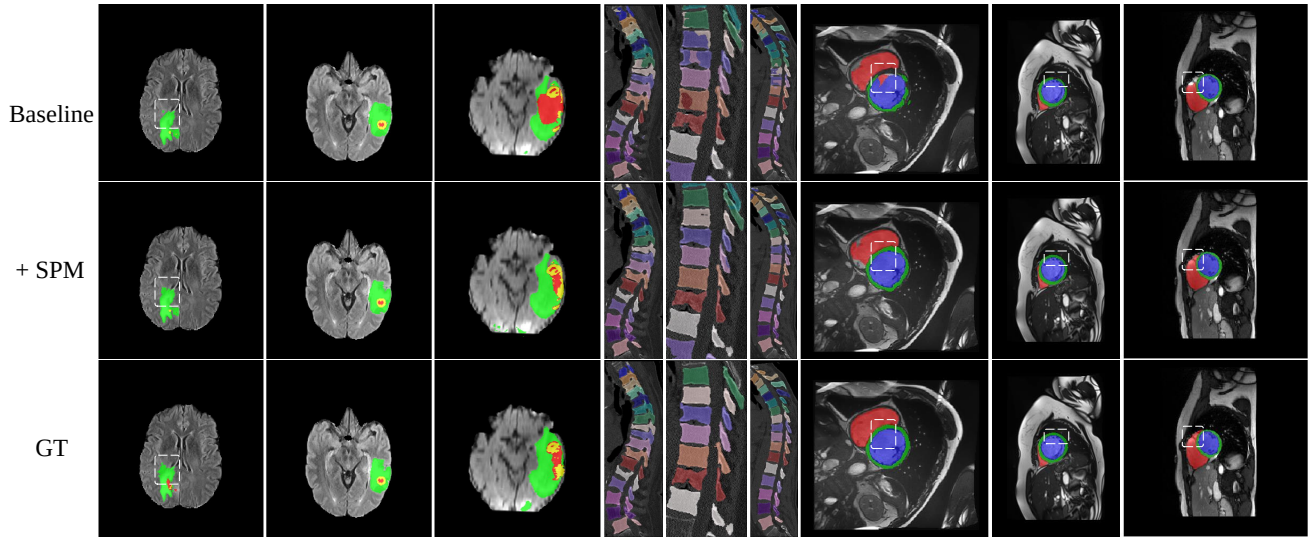


Fig. 9. Qualitative comparisons of baseline and baseline with SPM on BraTS 2020, VerSe 2019 and ACDC.

TABLE IV
ABLATION STUDY ABOUT THE PLUG-AND-PLAY CHARACTERISTIC AND STRUCTURAL COMPOSITION OF SPM ON THE BRATS 2020, VERSE 2019 AND ACDC DATASETS.

BraTS

Method	Dice score (%) ↑				HD_{95} (mm) ↓			
	ET	WT	TC	Mean	ET	WT	TC	Mean
3D UNet [12]	82.38	91.89	90.24	88.16	4.20	4.93	3.35	4.16
+ SPM	**82.86**	**92.42**	**90.46**	**88.58**	**3.77**	**4.26**	**3.28**	**3.77**
ResUNet [21]	83.33	92.36	89.47	88.38	4.09	6.38	4.09	4.85
+ SPM	**84.26**	**92.70**	**91.05**	**89.34**	**4.04**	**3.30**	**3.56**	**3.63**
UNETR [20]	79.84	88.81	83.22	83.96	8.95	17.99	14.02	13.65
+ SPM	**80.25**	**90.08**	**85.69**	**85.34**	**8.41**	**15.16**	**10.65**	**11.41**
UNeXt [49]	79.72	91.61	89.32	86.88	4.06	4.64	4.46	4.39
+ SPM	**80.61**	**91.89**	**89.75**	**87.42**	**3.40**	**4.11**	**4.39**	**3.97**
Swin UNETR [48]	83.51	91.95	90.20	88.55	5.95	8.99	**4.93**	6.63
+ SPM	**83.68**	**92.92**	**90.36**	**88.99**	**4.83**	**6.73**	6.39	**5.98**
nnUNet [28]	83.42	92.56	90.87	88.95	3.78	6.17	4.29	4.75
+ SPM	**84.47**	**93.27**	**91.10**	**89.61**	**3.77**	**3.60**	**3.16**	**3.51**
Baseline	83.33	92.36	89.47	88.38	4.09	6.38	4.09	4.85
+ CUB	84.19	92.47	90.40	89.02	**3.86**	3.52	4.09	3.82
+ SUB & CUB	**84.26**	**92.70**	**91.05**	**89.34**	4.04	**3.30**	**3.56**	**3.63**

VerSe 2019

Method	Dice score (%) ↑				HD_{95} (mm) ↓			
	Cerv	Thor	Lumb	Mean	Cerv	Thor	Lumb	Mean
3D UNet [12]	83.10	78.37	70.88	81.28	4.04	8.75	11.56	7.97
+ SPM	**85.30**	**84.79**	**72.28**	**84.16**	**3.14**	**7.17**	**10.46**	**7.38**
ResUNet [21]	88.59	84.50	73.08	84.98	2.66	5.34	9.97	5.49
+ SPM	**89.87**	**88.69**	**74.15**	**87.65**	**2.01**	**3.15**	**9.08**	**4.36**
UNETR [20]	72.10	69.33	**67.07**	73.86	9.29	13.53	18.11	13.42
+ SPM	**79.95**	**74.62**	66.32	**75.86**	**5.34**	**9.41**	**17.46**	**11.08**
UNeXt [49]	77.00	**86.73**	71.06	83.36	4.57	**3.98**	12.24	5.82
+ SPM	**85.35**	86.18	**73.15**	**84.62**	**2.61**	4.22	**8.86**	**5.15**
Swin UNETR [48]	**89.30**	81.43	73.36	83.46	**2.21**	7.79	10.09	7.66
+ SPM	81.82	**85.58**	**73.61**	**84.69**	4.32	**4.45**	**10.07**	**6.04**
Baseline	88.59	84.50	73.08	84.98	2.66	5.34	9.97	5.49
+ CUB	82.48	85.63	74.02	84.46	4.09	5.06	9.47	5.95
+ SUB & CUB	**89.87**	**88.69**	**74.15**	**87.65**	**2.01**	**3.15**	**9.08**	**4.36**

ACDC

Method	Dice score (%) ↑				HD_{95} (mm) ↓			
	RV	Myo	LV	Mean	RV	Myo	LV	Mean
3D UNet [12]	91.13	85.75	95.95	90.94	1.46	1.17	**1.07**	1.24
+ SPM	**91.78**	**86.73**	**96.29**	**91.60**	**1.45**	**1.12**	1.09	**1.22**
ResUNet [21]	91.66	85.26	95.72	90.88	1.47	1.21	**1.09**	1.26
+ SPM	**92.28**	**88.13**	**96.61**	**92.34**	**1.43**	**1.10**	**1.09**	**1.21**
UNeXt [49]	91.85	86.91	96.34	91.70	1.47	**1.09**	**1.05**	1.21
+ SPM	**92.14**	**87.48**	**96.58**	**92.07**	**1.37**	1.13	**1.05**	**1.18**
nnUNet [28]	90.24	89.24	95.36	91.61	1.53	1.12	1.17	1.27
+ SPM	**91.52**	**89.67**	**95.94**	**92.38**	**1.43**	**1.09**	**1.10**	**1.21**
Baseline	91.66	85.26	95.72	90.88	1.47	1.21	**1.09**	1.26
+ CUB	91.60	87.69	**96.61**	91.97	1.46	1.15	1.10	1.24
+ SUB & CUB	**92.28**	**88.13**	**96.61**	**92.34**	**1.43**	**1.10**	**1.09**	**1.21**

bear various and irregular shapes. This phenomenon can be explained by the fact that global shape priors inject global texture information to skipped features as shown in Fig.8, and the context information is effective to improve models' representation abilities for enhanced tumors.

As shown in Table IV, a significant improvement is obtained on the segmentation performance of the ACDC dataset after employing the proposed SPM to the baseline model. Particularly, the Dice score for the myocardium region increases by 0.98%, 2.87%, 0.57% on 3D UNet [12], ResUNet [21] and UNeXt [49] respectively. We carry out a visualization comparison between ResUNet and ResUNet with SPM. Fig.9 reveals that SPM can refine segmentation masks with a roughly circular shape, which benefits from the learnable shape priors. Besides, false positive predictions of RV introduce inconsistency to the segmentation result of LV, which will be suppressed by anatomical shape priors learned from training datasets.

Furthermore, we conduct comprehensive experiments on VerSe 2019 to evaluate the efficacy of SPM when plugged into CNNs and Transformer-based structures. As shown in Table IV, SPM brings considerable improvements on the Dice score and 95% Hausdorff distance of cervical, thoracic, and lumbar vertebrae. For ResUNet, the introduction of shape priors brings remarkable gains on the segmentation performance of cervical, thoracic, lumbar vertebrae, with 1.28%, 4.19%, 1.07% Dice score increases and $0.65mm$, $2.19mm$, $0.89mm$ HD_{95} decreases. And for Swin UNETR [48], a Transformer-based model, substantial improvements have been achieved on the Dice score and HD_{95} of thoracic vertebrae ($\uparrow 4.15\%$, $\downarrow 3.34mm$). However, SPM will degrade the segmentation performance of cervical vertebrae, with the Dice score decreasing by 7.48%. We argue that there are 220 cervical, 884 thoracic, 621 lumbar vertebrae in the VerSe 2019 dataset [44], in which cervical vertebrae make up a small proportion of all vertebrae. Therefore, there might be a risk of biased learning, where explicit shape priors focus on the leaning of thoracic and lumbar vertebrae. Even if a segmentation degradation in the cervical region, average evaluation metrics for the whole spline is still improved, with a 1.23% Dice score increase

TABLE V
QUANTITATIVE COMPARISONS BETWEEN SPM AND OTHER ATTENTION MODULES ON BRATS 2020, VERSE 2019 AND ACDC (THE PARAMETERS AND FLOPS ARE CALCULATED BASED ON MODELS DEVISED FOR BRATS 2020)

Method	BraTS 2020		VerSe 2019		ACDC		Params(M)	FLOPs(T)
	Dice ↑	HD_{95} ↓	Dice ↑	HD_{95} ↓	Dice ↑	HD_{95} ↓		
Baseline + attention gate [41]	88.77	4.54	85.47	5.55	91.67	1.23	44.21	456.98
Baseline + residual attention gate	88.84	4.02	86.40	5.29	91.91	1.24	44.21	456.98
Baseline + SE [24]	88.87	4.09	86.89	5.06	91.33	1.32	20.24	354.34
Baseline + CBAM [52]	88.65	3.83	85.00	5.52	91.91	1.27	22.21	386.52
Baseline + SPM	**89.34**	**3.63**	**87.65**	**4.36**	**92.34**	**1.21**	43.53	438.23

and a $1.62mm$ HD_{95} decrease. Besides, Fig.9 shows more consistent predictions in each vertebra, which is a strong proof that shape priors can enhance models' representation abilities via the introduction of shape constraints.

2) Structural Ablations of SPM: SPM is composed of SUB and CUB, which play a different role in the process of enhancing skipped features and refining explicit shape priors. Thus, we further research on the effectiveness of these two components. According to Table IV, the introduction of the individual CUB brings a significant improvement on the baseline model, with a 0.64% ↑ Dice score increase and a $1.03mm$ ↓ HD_{95} decrease on the BraTS 2020, a 1.09% ↑ average Dice score increase on ACDC. And this is a strong proof that CUB can boost the shape representation for local GT regions. When SUB is applied to the structure of SPM, there is a further performance increase on these two datasets. However, for the VerSe 2019 dataset, removing SUB from SPM will sharply degrade the segmentation performance, even lower than that of the baseline as shown in Table IV. We give an explanation that global contexts from SUB are essential for the identification of vertebrae due to the existence of long-range dependency contained in the longitudinal axis of the spines. Besides, after introducing CUB to the baseline model, the segmentation performance for cervical vertebrae declines significantly while thoracic and lumbar vertebrae are finely segmented, which might result from the biased ratio between three kinds of vertebrae [44].

E. Discussions

1) Comparison with other implicit shape models: Here SPM is intrinsically a type of attention module to enhance skipped features with luxurious shape information. Thus, we conduct quantitative experiments on the performance comparison with other classic attention modules, popularly employed in the field of medical image segmentation. As shown in Table V, our proposed attention module can achieve better improvements compared with other attention modules on the three public datasets. Although the parameters and FLOPs of SPM are heavier than those of SE [24] and CBAM [52] block, there exist significant improvements on the average Dice score due to the fact that our module introduces additional shape information generated from learnable shape priors. Specifically, SPM outperforms SE block on VerSe 2019 and ACDC with 0.76% and 1.01% Dice score increases, shows superior to CBAM block on BraTS2020 and VerSe 2019 with a 0.69% and 2.65% Dice score increase. Apart from that, the attention gate from attention UNet [41] bears the same form for outputs, with a refined skipped feature and an attention map for the target region. Thus, we give a qualitative visualization for attention maps between the attention gate and SPM. As illustrated by Fig.10, shape priors generated from SPM show stronger localization abilities than attention maps from the attention gate. More detailedly, the latter attention can cover regions of the whole CT spine, which proves that SPM can model global dependency due to the existence of SUB.

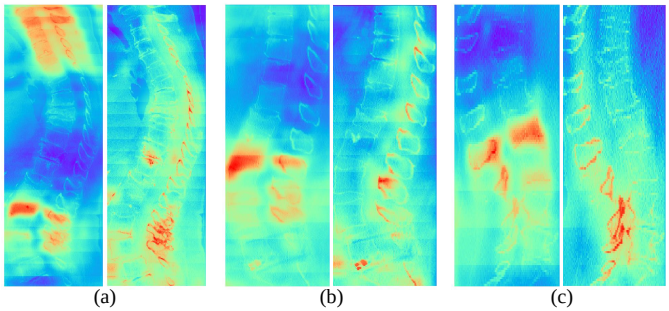


Fig. 10. Attention map comparisons between attention gate and SPM. For (a)-(c) cases, the left and right columns represent attention maps from the attention gate and SPM respectively.

TABLE VI
QUANTITATIVE COMPARISONS WITH OTHER ATLAS-BASED MODELS AND STATISTICAL SHAPE MODELS ON BRATS 2020, VERSE 2019 AND ACDC. DAFWM: DIFFERENTIABLE ATLAS FEATURE WARPING MODULE.

BraTS

Method	Dice score (%) ↑				HD_{95} (mm) ↓			
	ET	WT	TC	Mean	ET	WT	TC	Mean
ResUNet [21]	78.64	90.48	85.18	84.77	17.77	6.56	5.46	9.93
Deep Atlas Prior [25]	78.19	90.04	82.36	83.53	16.39	5.21	8.87	10.16
DAFWM [33]	78.39	87.60	83.47	83.15	20.69	8.61	5.81	11.70
Deep GMM [32]	56.71	69.06	74.63	66.80	45.49	15.74	11.12	24.12
DeepSSM [5]	78.99	90.44	85.11	84.84	17.66	7.80	5.38	10.28
Ours	**79.70**	**91.08**	**85.35**	**85.38**	**12.06**	**3.92**	**5.08**	**7.02**

VerSe 2019

Method	Dice score (%) ↑				HD_{95} (mm) ↓			
	Cerv	Thor	Lumb	Mean	Cerv	Thor	Lumb	Mean
ResUNet [21]	88.59	84.50	73.08	84.98	2.66	5.34	9.97	5.49
Deep Atlas Prior [25]	87.53	87.92	73.58	85.53	2.52	4.91	9.70	5.18
DAFWM [33]	84.19	81.41	72.56	83.31	3.93	6.70	10.59	7.10
DeepSSM [5]	87.54	87.58	73.76	85.85	2.85	4.25	9.16	4.94
Ours	**89.87**	**88.69**	**74.15**	**87.65**	**2.01**	**3.15**	**9.08**	**4.36**

ACDC

Method	Dice score (%) ↑				HD_{95} (mm) ↓			
	RV	Myo	LV	Mean	RV	Myo	LV	Mean
ResUNet [21]	91.66	85.26	95.72	90.88	1.47	1.21	**1.09**	1.26
Deep Atlas Prior [25]	90.99	86.15	96.14	91.09	1.59	1.25	1.12	1.28
DAFWM [33]	91.15	86.56	95.86	91.19	1.54	1.17	**1.09**	1.26
Deep GMM [32]	88.79	34.50	93.79	72.36	3.81	23.64	1.30	9.58
DeepSSM [5]	91.44	86.88	96.45	91.59	1.52	1.22	1.15	1.30
Ours	**92.28**	**88.13**	**96.61**	**92.34**	**1.43**	**1.10**	**1.09**	**1.21**

2) Comparisons with other explicit shape models: In this part, extensive comparative experiments are carried out between SPM and other atlas-based & statistical-based models on BraTS 2020, VerSe 2019 and ACDC. The backbone of all methodologies is set as ResUNet [21]. And we also present how SPM can theoretically overcome the drawbacks of other explicit shape models.

Atlas-based models: We adopt two recent deep atlas-based methods [25], [33] for comparisons. Deep atlas prior (DAP) [25] is integrated into deep segmentation networks through the prior loss, which contains prior location and shape information of organs. Thus, DAP can boost the segmentation performance on ACDC and VerSe 2019, in which anatomies show a relatively fixed shape and location. As revealed by Table VI, there exist 0.89% and 3.42% Dice increases for the Myo and thoracic vertebral region. Additionally, a differentiable

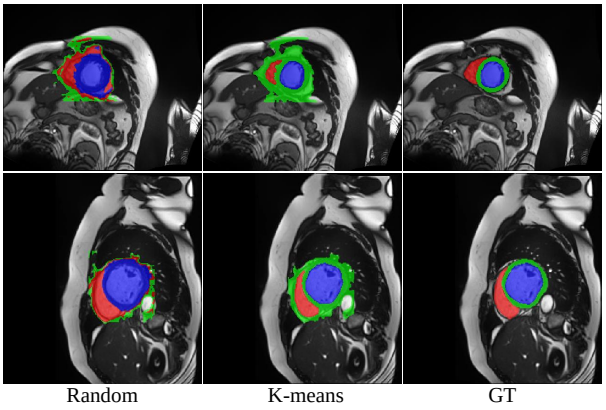


Fig. 11. Predictions on ACDC by deep GMM under two settings of initialized seed points: Random vs. K-means.

atlas feature warping module (DAFWM) [33] is devised to establish feature-level atlas-target correspondence. Here two pairs of source images/GTs are chosen as atlas for modeling training. However, the atlas cannot cover the label distribution properties of the whole dataset. For vertebral data, different CT scans bear different field-of-views, and different scans suffer from a **label mismatch**. As a result, DAFWM induces 3.09% and 1.67% Dice decreases for evaluations on thoracic vertebrae and the whole spine. Besides the label property, datasets like BraTS characterize a large **shape variance**, which poses a challenge for segmenting tumor regions. Table VI illustrates a 1.62% average Dice decrease and a $1.77mm$ average HD_{95} increase. Indeed, DAP is encountered with the same obstacle of shape variance, with a $2.82\% \downarrow$ Dice score and $3.41mm \uparrow$ HD_{95} on TC.

Compared with atlas-based models aforementioned, SPM supplies models with N-channel explicit shape priors, which can be regarded as learnable atlas corresponding to different segmentation classes. During training, shape priors are iteratively updated under the coarse shape guidance from encoded features [60]. Thus, learnable priors are adaptive to different datasets with various shapes, alleviating the challenge of label mismatch and shape variance.

Statistical-based models: Here deep GMM [32] and DeepSSM [5] serve as two baselines. Specifically, deep GMM calculates N Gaussian distributions via the Expectation Maximization (EM) algorithm based on decoded features from ResUNet. As depicted by Fig.11, GMM is sensitive to initial seed points [42]. Specifically, different predicted masks are achieved under settings of the random initialization and K-means initialization. Therefore, we choose a unified K-means initialization mode, with 3 times of initializations (best results are kept). Another baseline is DeepSSM, which is aimed at implementing data augmentations via the Principal Component Analysis (PCA) algorithm.

However, SSMs mentioned above are afflicted by **input noise**. Since different regions of backgrounds represent distinct feature representations, deep GMM tends to classify some background voxels as foregrounds. As shown in Fig.12, this model wrongly activates background areas for BraTS/-VerSe/ACDC, resulting in a huge amount of outliers. Consequently, the HD_{95} value is significantly increased compared with ResUNet, with the average HD_{95} increasing by $14.19mm$ and $8.32mm$ for BraTS and ACDC. Meanwhile, linear transforms [5] generated from PCA suffer from CT/MRI imaging noise, which leads to inaccurate transformed masks. That is why DeepSSM cannot improve the evaluation metric of HD_{95} as illustrated in Table VI. Moreover, SSMs fail to address the **label imbalance** issue. For VerSe, the ratio and volume of cervical vertebrae are smaller than those of thoracic and lumbar. Thus, the main component transform of DeepSSM will damage original shape properties of cervical vertebrae, causing a $1.05\% \downarrow$ Dice score. As revealed by Fig.12, DeepSSM neglects the prediction of several cervical vertebrae. As for deep GMM, when one mixture set has insufficient points, estimating the covariance matrices becomes difficult, and the algorithm is known to diverge [47]. This phenomenon is the same with BraTS and ACDC.

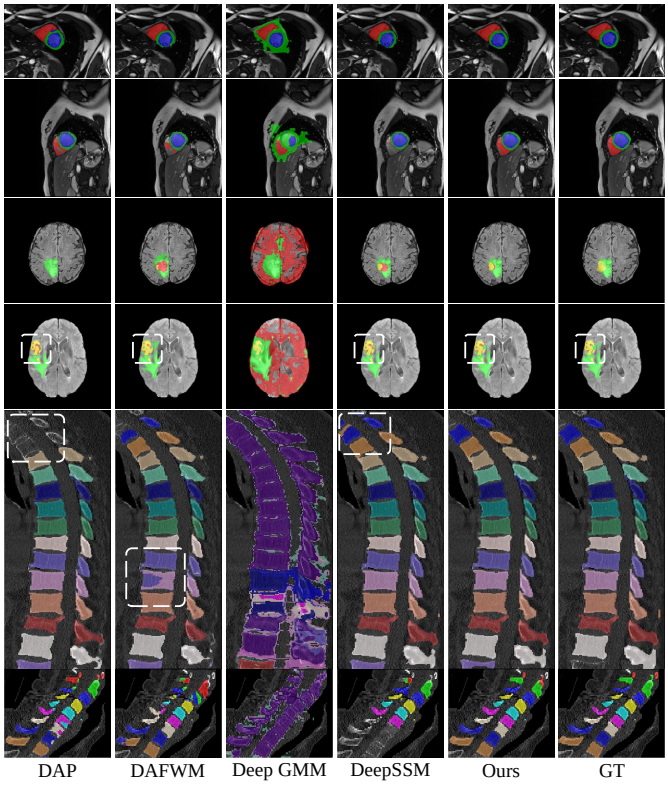


Fig. 12. Qualitative comparisons between our model and other atlas-based models & SSMs on BraTS 2020, VerSe 2019 and ACDC. DAP: deep atlas prior, DAFWM: differentiable atlas feature warping module.

In contrast, shape priors of SPM are randomly initialized for model training, demonstrating the strong robustness compared with deep GMM. Due to the fact that multi-scale skipped features are enhanced with the interaction of global shape priors, refined features bear luxurious shape and texture information as revealed in Fig.8. Hence, decoded features will suppress imaging noise from backgrounds. For the challenge of label imbalance, well-trained shape priors can provide additional shape guidance for regions with smaller data ratios or volumes, boosting models' representation abilities for anatomies with various sizes.

3) Limitations of SPM: Quantitative and qualitative experimental results on BraTS 2020, VerSe 2019 and ACDC validate the effectiveness of SPM. Besides, as shown in Fig.4, explicit

shape priors show the potential for providing sufficient shape information to guide the segmentation task. However, there exist two drawbacks for the structure of SPM currently. The first drawback is related to the phenomenon called repetitive activation, which means that different channels of explicit shape priors tend to activate the same region. Here the affinity map S_{map} in the self-update block is employed to describe inter-relations between different channels of shape priors. And we expect to attain shape priors, in which each channel shows discriminative shape priors. Due to the fact that no constraint conditions are imposed to guide the learning of affinity map S_{map}, there still exists dependency relations between channels of global shape priors, which results in the repetitive activation in the last stage of shape priors. Furthermore, the second drawback is the occasional degradation of shape priors. On the ground that global priors generated from SUB contain global contexts and textures for the whole region, the last-stage shape priors of some cases show less accurate shape information than those in the second stage. As a result, the design for SUB will be a potential direction that needs to be further researched.

IV. Conclusion

In this paper, we detailedly discuss three types of segmentation models with shape priors, which consist of atlas-based models, statistical-based models and UNet-based models. To enhance the interpretability of shape priors on UNet-based models, we proposed a shape prior module (SPM), which could explicitly introduce shape priors to promote the segmentation performance on different datasets. And our model achieves state-of-the-art performance on the datasets of BraTS 2020, VerSe 2019 and ACDC. Furthermore, according to quantitative and qualitative experimental results, SPM shows a good generalization ability on different backbones, which can serve as a plug-and-play structure.

References

[1] M. Bach Cuadra, V. Duay, and J.-P. Thiran. Atlas-based segmentation. *Handbook of Biomedical Imaging: Methodologies and Clinical Research*, pages 221–244, 2015.

[2] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

[3] S. Bakas *et al.* Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

[4] 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*, 37(11):2514–2525, 2018.

[5] R. Bhalodia, S. Elhabian, J. Adams, W. Tao, L. Kavan, and R. Whitaker. Deepssm: A blueprint for image-to-shape deep learning models. *Medical Image Analysis*, 91:103034, 2024.

[6] M. Cabezas, A. Oliver, X. Lladó, J. Freixenet, and M. B. Cuadra. A review of atlas-based segmentation for magnetic resonance brain images. *Computer methods and programs in biomedicine*, 104(3):e158–e177, 2011.

[7] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *ECCV*, pages 205–218. Springer, 2022.

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020.

[9] C.-F. R. Chen et al. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *ICCV*, pages 357–366, 2021.

[10] 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.

[11] B. Cheng, A. Schwing, and A. Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

[12] Ö. Çiçek et al. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *MICCAI*, pages 424–432. Springer, 2016.

[13] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[14] T. F. Cootes et al. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[15] T. F. Cootes and C. J. Taylor. Combining point distribution models with shape models based on finite element analysis. *Image and Vision Computing*, 13(5):403–409, 1995.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[17] R. El Jurdi, C. Petitjean, P. Honeine, and F. Abdallah. Bb-unet: U-net with bounding box prior. *IEEE JSTSP*, 14(6):1189–1198, 2020.

[18] R. El Jurdi *et al.* High-level prior-based loss functions for medical image segmentation: A survey. *CVIU*, 210:103248, 2021.

[19] D.-P. Fan et al. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, pages 263–273. Springer, 2020.

[20] A. Hatamizadeh *et al.* Unetr: Transformers for 3d medical image segmentation. In *WACV*, pages 574–584, 2022.

[21] 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*, pages 770–778, 2016.

[22] T. Heimann and H.-P. Meinzer. Statistical shape models for 3d medical image segmentation: a review. *MedIA*, 13(4):543–563, 2009.

[23] M. H. Hesamian, W. Jia, X. He, and P. Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32:582–596, 2019.

[24] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[25] H. Huang, H. Zheng, L. Lin, M. Cai, H. Hu, Q. Zhang, Q. Chen, Y. Iwamoto, X. Han, Y.-W. Chen, et al. Medical image segmentation with deep atlas prior. *IEEE TMI*, 40(12):3519–3530, 2021.

[26] X. Huang *et al.* Missformer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*, 2021.

[27] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993.

[28] F. Isensee *et al.* nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[29] H. Kalinic. Atlas-based image segmentation: A survey. *Croatian Scientific Bibliography*, pages 1–7, 2009.

[30] J. Li et al. Transbtsv2: Wider instead of deeper transformer for medical image segmentation. *arXiv preprint arXiv:2201.12785*, 2022.

[31] J. Li et al. Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. *Medical image analysis*, 85:102762, 2023.

[32] C. Liang, W. Wang, J. Miao, and Y. Yang. Gmmseg: Gaussian mixture based generative semantic segmentation models. *Advances in Neural Information Processing Systems*, 35:31360–31375, 2022.

[33] H. Liu, D. Nie, J. Yang, J. Wang, and Z. Tang. A new multi-atlas based deep learning segmentation framework with differentiable atlas feature warping. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[34] C. Liu *et al.* Brain tumor segmentation network using attention-based fusion and spatial relationship constraint. In *BrainLes 2020*, pages 219–229. Springer, 2021.

[35] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[36] W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. *NeurIPS*, 29, 2016.

[37] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A. L. Martel. Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71:102035, 2021.

[38] B. H. Menze et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE TMI*, 34(10):1993–2024, 2014.

[39] F. Milletari *et al.* 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.
[40] H. T. Nguyen *et al.* Enhancing mri brain tumor segmentation with an additional classification network. In *BrainLes 2020*, pages 503–513. Springer, 2021.
[41] O. Oktay et al. Attention u-net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning*, 2022.
[42] D. A. Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.
[43] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.
[44] A. Sekuboyina et al. Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images. *MedIA*, 73:102166, 2021.
[45] D. Shen et al. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.
[46] Y. Su et al. Yolo-logo: A transformer-based yolo segmentation model for breast mass detection and segmentation in digital mammograms. *CMPB*, 221:106903, 2022.
[47] Y.-W. Tai, J. Jia, and C.-K. Tang. Local color transfer via probabilistic segmentation by expectation-maximization. In *CVPR*, volume 1, pages 747–754. IEEE, 2005.
[48] Y. Tang et al. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *CVPR*, pages 20730–20740, 2022.
[49] J. M. J. Valanarasu and V. M. Patel. Unext: Mlp-based rapid medical image segmentation network. In *MICCAI*, pages 23–33. Springer, 2022.
[50] M. H. Vu, T. Nyholm, and T. Löfstedt. Multi-decoder networks with multi-denoising inputs for tumor segmentation. In *BrainLes 2020*, pages 412–423. Springer, 2021.
[51] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li. Transbts: Multimodal brain tumor segmentation using transformer. In *MICCAI*, pages 109–119, 2021.
[52] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
[53] Y. Xie et al. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *MICCAI*, pages 171–180. Springer, 2021.
[54] Y. Xu, Q. Zhang, J. Zhang, and D. Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *NeurIPS*, 34, 2021.
[55] P. Yan et al. Global structure constrained local shape prior estimation for medical image segmentation. *Computer Vision and Image Understanding*, 117(9):1017–1026, 2013.
[56] X. You, Y. Gu, et al. Eg-trans3dunet: a single-staged transformer-based model for accurate vertebrae segmentation from spinal ct images. In *ISBI*, pages 1–5. IEEE, 2022.
[57] X. You, Y. Gu, et al. Verteformer: A single-staged transformer network for vertebrae segmentation from ct images with arbitrary field of views. *Medical Physics*, 50(10):6296–6318, 2023.
[58] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, and Y. Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 2023.
[59] T. Zhou, W. Wang, et al. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2582–2593, 2022.
[60] B. Zhou *et al.* Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.