# TDPSR-Net: Topographic Distance Priors and a Spatial Regularization Network for Enhanced PET Segmentation

Lin Yang, Mingxiang Wu, Meiyun Wang, Yaping Wu, Chuanli Cheng, Chao Zou, Raymond Chan, Hairong Zheng, *Senior Member, IEEE*, Dong Liang, *Senior Member, IEEE*, Zhanli Hu, *Senior Member, IEEE*, Zhi-Feng Pang, Xue-Cheng Tai, Na Zhang, *Member, IEEE*

***Abstract*—Positron emission tomography (PET) provides functional information by capturing tracer uptake and is widely used for disease assessment. Accurate segmentation of regions of interest is essential for quantitative analysis and clinical decision-making. However, PET images often exhibit low spatial resolution, high noise, and blurred boundaries due to partial volume effects, which hampers precise delineation. To address this, we propose TDPSR-Net, a PET image segmentation network that integrates topographic distance priors (TDP) and spatial regularization techniques. Our method automatically generates marker points for computing topographic distances, and the network jointly extracts features from both PET images and the resulting TDP maps. To enhance spatial coherence and boundary consistency, we introduce a Soft Threshold Dynamics of Sigmoid (STD-Sigmoid) layer that imposes spatial regularization on the network output, and we further establish a theoretical connection between the proposed formulation and a Potts-type model. We evaluate TDPSR-Net on multiple datasets, including liver, hippocampus, and lung cancer tumor segmentation, and the results demonstrate consistently high accuracy and robustness across diverse datasets, highlighting the potential of TDPSR-Net for a wide range of clinical applications.**

***Index Terms*—PET Segmentation, Topographic Distance, Spatial Regularization, Soft Threshold Dynamic Method**

This work was supported by the National Key Research and Development Program of China (2024YFE0202400), the National Natural Science Foundation of China (82372038, 12326607, 12471398), the Shenzhen Excellent Technological Innovation Talent Training Project of China (RCJC20200714114436080), the Guangdong Basic and Applied Basic Research Foundation of China (2023B1515120007, 2024B1515040018), the Shenzhen Science and Technology Program of China (JCYJ20220818101804009 and KJZD20240903101307010), the Natural Science Foundation of Henan Province (232300420108), the Science and Technology Major Project of Henan Province (221100310200), the HKRGC GRF grants (CityU1101120, CityU11309922), the CRF grant (C1013-21GF), the HKITF MHKJFS Grant (MHP05422, LU105824), Key Project of Henan Province Medical Science and Technology Project (SBGJ202502010), and the Key Laboratory for Magnetic Resonance and Multimodality Imaging of Guangdong Province (2023B1212060052).

L. Yang is with the Lauterbur Research Center for Biomedical Imaging, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with University of Chinese Academy of Sciences, Beijing 101408, China(e-mail: l.yang2@siat.ac.cn).

C. Cheng, C. Zou, H. Zheng and N. Zhang are with the Lauterbur Research Center for Biomedical Imaging, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: cl.cheng@siat.ac.cn, chao.zou@siat.ac.cn, hr.zheng@siat.ac.cn, na.zhang@siat.ac.cn).

M. Wu, Z.-F. Pang is with the College of Mathematics and Statistics, Henan University, Kaifeng 475004, China (e-mail: mingxiangwu2022@163.com, zhifengpang@henu.edu.cn).

M. Wang and Y. Wu is with the Department of Medical Imaging, Henan Provincial People's Hospital & People's Hospital of Zhengzhou University, Zhengzhou 450003, China (e-mail: mywang@zzu.edu.cn, ypwu@ha.edu.cn).

R. Chan is with Lingnan University, Hong Kong, China and with Hong Kong Centre for Cerebro-Cardiovascular HealthEngineering, Hong Kong, China (e-mail: raymond.chan@ln.edu.hk).

D. Liang and Z. Hu are with Research Center for Medical AI, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: dong.liang@siat.ac.cn, zl.hu@siat.ac.cn).

X.-C. Tai is with the Norwegian Research Centre, Nygardstangen, NO-5838 Bergen, Norway (e-mail: xtai@norceresearch.no, xuechengtai@gmail.com).

L. Yang, M. Wu and M. Wang contributed equally to this work.

Corresponding authors: Z.-F. Pang, X.-C. Tai and N. Zhang.

## I. Introduction

POSITRON emission tomography (PET) detects $\gamma$ rays produced by positron-electron annihilation to reconstruct the distribution of radioactive tracers in the body. Since pathological tissues often exhibit elevated metabolic rates, PET provides high sensitivity for early lesion detection. Clinically, PET imaging is commonly combined with computed tomography (CT) or magnetic resonance imaging (MRI) to provide both functional and anatomical information. PET has been widely applied in the diagnosis and treatment planning of tumors [1], neurological disorders [2], as well as inflammatory and infectious diseases [3]. Accurate segmentation of tumors, organs, or functional regions is essential for quantitative analysis, disease characterization, and clinical decision-making. However, limitations in imaging physics and detector performance result in low spatial resolution and pronounced noise in PET [4]. Moreover, the partial volume effect (PVE) [5] blurs boundaries between adjacent structures, further hindering their separation. These factors pose substantial challenges for robust PET segmentation.

Traditional PET segmentation methods typically model and segment targets using region-based or pixel-level features [6] (e.g., intensity values [7], gradients [8], and statistical descriptors [9]). Variational energy–based segmentation models [10], [11] leverage these features as similarity measures to construct a data-fidelity term, and incorporate a regularization term to promote spatial smoothness and structural consistency. A key advantage is the explicit formulation of regularization and prior assumptions (e.g., contour smoothness or intensity statistics), which yields clear mathematical interpretability. However, PET images often exhibit high noise and weak boundaries, making traditional models prone to produce stable and accurate delineations.

Deep learning–based methods have substantially advanced medical image segmentation in recent years [12], [13]. Since

the introduction of fully convolutional networks (FCNs) [14], encoder–decoder architectures have become the dominant framework, with representative methods including UNet and its numerous variants [15]–[17]. In parallel, Transformer-based or hybrid architectures have been introduced to enhance global context modeling capability, such as SwinUNet [18], and SwinUNETR [19]. However, when confronted with the inherent challenges of PET imaging, such as low spatial resolution, high noise levels, and boundary ambiguity caused by partial volume effects, segmentation methods that rely solely on data-driven feature learning are prone to producing irregular contours and small spurious artifacts. To enhance the model's ability to perceive organ-specific intensity thresholds and shapes, Gao et al. [20] introduced textual prompts and disentangled organ features. Zhang et al. [21] proposed a cross-prompting confident learning (CPCL) strategy combined with an uncertainty-guided self-rectification process to learn representations for promptable segmentation. However, these models are primarily evaluated on organ segmentation tasks. To handle irregularly shaped tissues such as tumors, incorporating explicit spatial or geometric priors is crucial for stabilizing predicted boundaries and improving structural consistency.

Existing studies often adopt multimodal segmentation strategies by incorporating anatomical information from CT or MR images to assist PET segmentation. For example, Xiang et al. [22] proposed a modality-specific segmentation network (MoSNet) for lung tumor segmentation using PET/CT. Tang et al. [23] proposed a 3D whole-brain PET/MR segmentation network with a cross-fusion mechanism that jointly exploits functional and structural information. A substantial body of research [24], [25] has demonstrated the feasibility and effectiveness of multimodal approaches for PET image segmentation. Despite their demonstrated effectiveness, multimodal PET segmentation still faces potential limitations. In PET/CT and PET/MR imaging, inherent inter-modality discrepancies and motion-induced misregistration can compromise the reliability of fused inputs [26], [27]. Furthermore, low-dose protocols or prolonged scans may reduce anatomical fidelity, thereby limiting the contribution of complementary modalities. For radiation-sensitive populations (e.g., pediatric and elderly patients), low-dose CT is often used in PET/CT, which may degrade CT image quality and blur anatomical details. Consequently, reducing the reliance on multimodal inputs and fully exploiting the structural information inherently contained in single-modality PET images for accurate segmentation remains of significant clinical importance.

Introducing explicit structural priors into segmentation models has been shown to improve accuracy and robustness [28]. Such priors include connectivity priors [29], topological constraints [30], star-shaped priors [31], convexity constraints [32], and distance-based information [33]. In addition, boundary-length regularization from active-contour models has been integrated into neural networks as an explicit prior [31]. These advances suggest that explicitly modeling geometric and spatial priors complements data-driven learning, particularly under low signal-to-noise ratio (SNR) conditions. In PET segmentation, severe noise and low contrast often render object boundaries weak or ambiguous, leading to fragmented predictions and irregular shapes. Therefore, a key requirement is to introduce spatial guidance that can propagate globally while remaining robust to noise. Topographic distance priors (TDP) address this need by propagating distance information from sparse annotations under a topographic metric, thereby preserving global structure even when boundaries are indistinct. Meanwhile, boundary-length regularization enforces boundary consistency and suppresses common noise-induced artifacts (e.g., jagged edges) in PET. In our previous work [34], we investigated the use of topographic distance priors for PET segmentation. However, constructing such priors by solving the Eikonal equation relied on clinicians to manually select marker points, which increases annotation cost and limits the practicality of the topographic distance prior.

To address these limitations, we propose TDPSR-Net, a unified framework that integrates topographic distance priors with spatial regularization for PET segmentation. The main contributions are summarized as follows:

- We propose an Eikonal-equation–based approach to compute TDP maps from PET images without manual marker selection and incorporate them as additional network inputs. These priors emphasize target regions while effectively suppressing background noise. PET images and TDP maps are processed through a dual-encoder UNet to extract complementary features.
- A variational interpretation of the sigmoid function is provided, and spatial regularization is embedded via soft threshold dynamics (STD), yielding a differentiable STD-Sigmoid layer. The proposed layer is further shown to be equivalent to the Potts model under appropriate parameter settings. By mapping logits to probabilities while explicitly enforcing spatial consistency, STD-Sigmoid promotes boundary coherence and mitigates noise-induced artifacts.
- A Global Context Scalar Estimator (GCSE) is introduced to predict the scalar parameters of STD-Sigmoid in a content-adaptive manner, enabling automatic adaptation across datasets and imaging conditions.

## II. Method

In this section, we first explain the details in obtaining TDPs and the derivation of the STD-Sigmoid function, followed by an explanation of the proposed method.

### A. Topographic Distance Priors

*1) Topographic Distance Function:* The topographic distance measures the shortest total elevation difference between two points along a surface. Let $I(\mathrm{x})$ be a image that is a twice continuously differentiable real-valued function defined on the image domain $\Omega$. For a two-dimensional image, $\Omega \subset \mathbb{R}^2$; for a three-dimensional image, $\Omega \subset \mathbb{R}^3$. The topographic distance between any two points $\mathrm{x}, \mathrm{y} \in \Omega$ is defined as [35]:

$$d(\mathrm{x}, \mathrm{y}) = \inf_{\gamma \in \mathcal{P}(\mathrm{x},\mathrm{y})} \int_0^1 \|\nabla I(\gamma(s))\| \mathrm{d}s. \tag{1}$$

where $\mathcal{P}(\mathrm{x}, \mathrm{y})$ denotes the set of all feasible paths from point x to y, and the parametric curve $\gamma(s)$ satisfies $\gamma(0) = \mathrm{x}$ and

$\gamma(1) = \mathrm{y}$. The notation $\|\cdot\|$ denotes the Euclidean norm and $\nabla$ denotes the spatial gradient operator.

According to definition (1), if the integrand $\|\nabla I(\gamma(s))\|$ is replaced by a constant, the topographic distance degenerates into the Euclidean distance. Unlike the Euclidean distance, which reflects only the geometric spatial separation between points, the topographic distance integrates the rate of intensity variation along the connecting path $\gamma(s)$, thereby capturing the cumulative intensity change between the two points. Specifically, if $\mathrm{x}$ and $\mathrm{y}$ lie within the same homogeneous region, $d(\mathrm{x}, \mathrm{y})$ is small. If the path $\gamma(s)$ crosses an edge or a region with a sharp intensity change, $d(\mathrm{x}, \mathrm{y})$ increases significantly.

In image segmentation tasks, effective feature extraction of the region of interest (ROI) is crucial for improving segmentation accuracy. By selecting a set of marker points $\mathcal{M}$ within the ROI, the topographic distance can characterize the structural dissimilarity between the marker points and other pixels, thereby significantly enhancing the discriminability of the target region. For any point $\mathrm{x} \in \Omega$, its topographic distance to marker set $\mathcal{M}$ is defined as

$$D(\mathrm{x}) = \min_{\bar{\mathrm{x}} \in \mathcal{M}} d(\mathrm{x}, \bar{\mathrm{x}}). \tag{2}$$

According to [36], $D(\mathrm{x})$ can be expressed as the unique viscosity solution of the Eikonal equation:

$$\begin{cases} \|\nabla D(\mathrm{x})\| = \|\nabla I(\mathrm{x})\|, & \text{if } \mathrm{x} \notin \mathcal{M}, \\ D(\mathrm{x}) = 0, & \text{if } \mathrm{x} \in \mathcal{M}, \end{cases} \tag{3}$$

which is a boundary value problem of a first-order nonlinear partial differential equation.

In the topographic distance, the original gradient $\|\nabla I(\mathrm{x})\|$ is highly sensitive to noise, count fluctuations, and boundary blurring, making it challenging to apply to PET images. To address this issue, we modify the gradient field. First, low-pass filtering is applied to the image using a Gaussian kernel $G_\sigma(\mathrm{x})$ with standard deviation $\sigma$, which suppresses Poisson noise while preserving large-scale edges. Then, gradient truncation is performed. This operation forces weak gradients to zero, ensuring that the metric only responds to regions of significant intensity variation (such as boundaries), thereby preventing the accumulation of false distances along noisy paths. Specifically, equation (3) is modified as

$$\begin{cases} \|\nabla D_K(\mathrm{x})\| = h(\mathrm{x}), & \text{if } \mathrm{x} \notin \mathcal{M}, \\ D_K(\mathrm{x}) = 0, & \text{if } \mathrm{x} \in \mathcal{M}, \end{cases} \tag{4}$$

where $h(\mathrm{x})$ is defined by

$$h(\mathrm{x}) = \begin{cases} 0, & \text{if } \|\nabla G_\sigma(\mathrm{x}) * I(\mathrm{x})\| < K, \\ \|\nabla G_\sigma(\mathrm{x}) * I(\mathrm{x})\|, & \text{otherwise.} \end{cases} \tag{5}$$

and $*$ denotes the convolution operation. The equation (4) can be solved efficiently using the fast sweeping Method [37].

To intuitively demonstrate the properties of the topographic distance, two sets of experiments are presented in Fig. 1. For the simple synthetic image shown in Fig. 1(a), direct thresholding of the topographic distance $D(\mathrm{x})$ yields an accurate segmentation result, as illustrated in Fig. 1(b). However, when the image contains fine structures and is heavily contaminated by noise (Fig. 1(c)), the original gradient field $\|\nabla I(\mathrm{x})\|$ becomes severely distorted, causing $D(\mathrm{x})$ to fail in reflecting the true topographic distance between the marker points and the remaining pixels, as illustrated in Fig. 1(d). After introducing the corrected gradient $h(\mathrm{x})$, the resulting robust topographic distance effectively suppresses noise interference and accurately captures the image structure within the target region, as shown in Fig. 1(e).

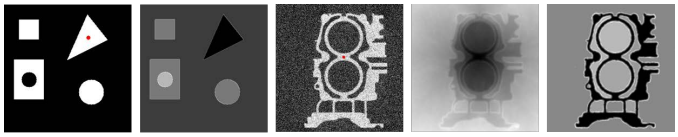


Fig. 1. Examples of topographic distance for two groups of synthetic images. (a) Simple synthetic image; (b) Corresponding original topographic distance $D(\mathrm{x})$ for (a); (c) Noise-contaminated synthetic image; (d) Corresponding original topographic distance $D(\mathrm{x})$ for (c); (e) Corrected topographic distance $D_K(\mathrm{x})$ for (c) with truncation threshold $K = 0.05$. (Red dots indicate the marker points.)

*2) Marker Points Generation Strategy:* The structural information characterized by topographic distance is inherently dependent on the selection of marker points. In segmentation tasks, the primary focus lies on the topological structure within the ROI, rather than the background. However, the manual selection of internal markers within the ROI (particularly in tumor regions within PET images) is not only labor-intensive but also highly contingent upon the clinical expertise of radiologists. To achieve a fully automated segmentation workflow and eliminate user intervention, we propose a learning-based approach for the automatic generation of the marker set $\mathcal{M}$.

Specifically, we adopt a segmentation network based on the standard UNet [15] architecture to generate an initial probability map of the target region. This network adopts the classic encoder-decoder structure. The encoder consists of four convolutional blocks, each comprising two $3 \times 3$ convolutions (stride 1, padding 1) followed by a ReLU activation function. Downsampling is performed via $2 \times 2$ max pooling (stride 2), with the number of feature channels doubling after each pooling operation. The decoder symmetrically comprises four stages. Each stage initiates with upsampling via a $2 \times 2$ transposed convolution (stride 2). Subsequently, skip connections concatenate features from the corresponding encoder layers along the channel dimension, followed by two consecutive $3 \times 3$ convolutions with ReLU activation to refine spatial details. Finally, a $1 \times 1$ convolution restores the output to the original image resolution, and a Sigmoid activation function maps the results to the probability interval $[0, 1]$. The network is trained using the cross-entropy loss function, with the Adam optimizer employed for parameter updates. After obtaining the initial probability map, a binary mask is first generated by applying a threshold (e.g., 0.5). Subsequently, morphological erosion is performed to suppress boundary uncertainty and extract the target core regions. The contours of each connected component are then identified, and their areas are computed and ranked. According to the number of target regions, the centroids of the top-ranked connected components are selected

as marker points $\mathcal{M}$. Based on these marker points, the topographic distance map $D_K(\mathrm{x})$ is computed and concatenated with the original PET image as a combined input. The overall workflow is illustrated in Fig. 2.

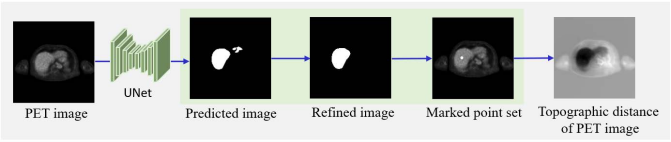


Fig. 2. Demonstration of the marked point set generation strategy.

### B. Spatially Regularized Sigmoid Function

*1) Potts Model:* The Potts model is a widely used image segmentation model that is based on an energy functional [38]. It obtains the desired segmentation results by solving the following optimization problem [10]:

$$\begin{cases} \min\limits_{\Omega_1,\Omega_2} \int_{\Omega_1} o_1(\mathrm{x})\mathrm{dx} + \int_{\Omega_2} o_2(\mathrm{x})\mathrm{dx} + \lambda\left|\partial\Omega_1\right|, \\ \Omega_1 \cup \Omega_2 = \Omega, \Omega_1 \cap \Omega_2 = \varnothing. \end{cases} \tag{6}$$

where $o_1(\mathrm{x})$ and $o_2(\mathrm{x})$ extract the grayscale information of the foreground and background of the image. Specifically, when $o_1(\mathrm{x}) = (I(\mathrm{x}) - c_1)^2$ and $o_2(\mathrm{x}) = (I(\mathrm{x}) - c_2)^2$, the Potts model degenerates into the classical Chan-Vese model [11], where $c_1$ and $c_2$ represent the mean grayscale values of the input image $I(\mathrm{x})$ in the subregion $\Omega_1$ and $\Omega_2$, respectively. The last term is a regularization term, penalizing the length of the boundary, denoted by $|\partial\Omega|$, while the parameter $\lambda > 0$ controls the extent of the penalty.

By introducing indicator functions

$$u(\mathrm{x}) = \begin{cases} 1, & \mathrm{x} \in \Omega_1 \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

the boundary length term in (6) can be replaced by threshold dynamic regularization [39]

$$|\partial\Omega| \approx \sqrt{\frac{\pi}{\tau}} \int_\Omega u(\mathrm{x}) \left(G_\tau(\mathrm{x}) * (1 - u(\mathrm{x}))\right) \mathrm{dx}, \tag{8}$$

where $G_\tau(\mathrm{x}) = \frac{1}{2\pi\tau^2} e^{-\frac{\|\mathrm{x}\|^2}{2\tau^2}}$ is the Gaussian kernel. It has been shown that as $\tau \to 0$, this threshold dynamic regularization term $\Gamma$-converges to boundary length. Then, the Potts model (6) can be equivalently expressed as:

$$\min_{u\in[0,1]} \langle o_1 - o_2, u\rangle + \lambda\sqrt{\frac{\pi}{\tau}}\langle u, G_\tau * (1-u)\rangle, \tag{9}$$

where $\langle\cdot,\cdot\rangle$ represents the inner product. Consequently, we have $\langle o_1 - o_2, u\rangle = \int_\Omega (o_1(\mathrm{x}) - o_2(\mathrm{x}))\, u(\mathrm{x})\mathrm{dx}$ and $\langle u, G_\tau * (1-u)\rangle = \int_\Omega u(\mathrm{x}) \left(G_\tau(\mathrm{x}) * (1 - u(\mathrm{x}))\right) \mathrm{dx}$.

*2) Variational Explanation for Sigmoid Function:* To embed spatial regularization into neural networks, Liu [31] and Zhang [40] started from the Potts model and developed a variational interpretation of the Softmax activation, leading to the soft threshold dynamics (STD) formulation. Although STD has demonstrated strong performance in multi-class segmentation, it is only formally similar to the Potts model and does not explicitly model interactions among different feature classes.

In this subsection, we derive a variational formulation of the Sigmoid activation and show its mathematical equivalence to the Potts model. For binary segmentation tasks, the Sigmoid activation function is typically employed to map the foreground feature $o_1(\mathrm{x})$ to the probability interval $[0, 1]$, formulated as:

$$\mathcal{S}\left(o_1(\mathrm{x})\right) = \frac{1}{e^{-o_1(\mathrm{x})} + 1}.$$

During the inference stage, the segmentation result can be interpreted as a pointwise comparison of the foreground feature $o_1(\mathrm{x})$ with a threshold function $k(\mathrm{x})$, followed by binarization:

$$u(\mathrm{x}) = \begin{cases} 1, & o_1(\mathrm{x}) - k(\mathrm{x}) > 0 \\ 0, & \text{others}. \end{cases} \tag{10}$$

The threshold is typically set as $k(\mathrm{x}) \equiv \frac{1}{2}\max\limits_{\mathrm{x}\in\Omega} o_1(\mathrm{x})$, which is equivalent to classifying pixel x as foreground if its foreground probability $\mathcal{S}\left(o_1(\mathrm{x})\right) > \frac{1}{2}$, and as background otherwise. As demonstrated in [31], equation (7) is equivalent to the optimal solution to the k-means maximization problem, given by $\int_\Omega \max\left\{o_1(\mathrm{x}) - k(\mathrm{x}), 0\right\}\mathrm{dx}$.

To enable the backpropagation of the max function in deep learning frameworks, we introduce its smooth approximation—the log-sum-exp function:

$$E_\varepsilon(o_1) = \int_\Omega \varepsilon ln(e^{\frac{o_1(\mathrm{x}) - k(\mathrm{x})}{\varepsilon}} + 1)\mathrm{dx}. \tag{11}$$

The $E_\varepsilon(o_1)$ is smooth and convex and it is not difficult to verify $\lim\limits_{\varepsilon\to 0} E_\varepsilon(o_1) = \int_\Omega \max\left\{o_1(\mathrm{x}) - k(\mathrm{x}), 0\right\}\mathrm{dx}$.

According to the definition of the Fenchel–Legendre transform, the Fenchel–Legendre transform of $E_\varepsilon(o_1)$ is

$$\begin{aligned} E_\varepsilon^*(u) &:= \max_{o_1}\{\langle u, o_1\rangle - E_\varepsilon(o_1)\} \\ &= \begin{cases} \langle k, u\rangle + \varepsilon\left(\langle u, \ln u\rangle + \langle 1-u, \ln(1-u)\rangle\right), & u(\mathrm{x}) \in [0,1], \\ +\infty, & \text{otherwise}. \end{cases} \end{aligned}$$

and twice Fenchel–Legendre transformation of $E_\varepsilon(o_1)$ is

$$\begin{aligned} E_\varepsilon^{**}(o_1) &:= \max_u\{\langle k - o_1, u\rangle - \varepsilon\left(\langle u, \ln u\rangle + \langle 1-u, \ln(1-u)\rangle\right)\}, \\ &= \min_u\{\langle o_1 - k, u\rangle + \varepsilon\left(\langle u, \ln u\rangle + \langle 1-u, \ln(1-u)\rangle\right)\}. \end{aligned}$$

Since $E_\varepsilon(o_1)$ is a strictly convex function, the Fenchel–Legendre theory yields the biconjugate identity: $E_\varepsilon(o_1) = E_\varepsilon^{**}(o_1)$. Therefore, the Sigmoid function can be approximated as the solution of the following variational problem:

$$\min_{u\in[0,1]} \mathcal{F}(u), \tag{12}$$

where

$$\mathcal{F}(u) = \langle o_1 - k, u\rangle + \varepsilon\left(\langle u, \ln u\rangle + \langle 1-u, \ln(1-u)\rangle\right),$$

and $\varepsilon$ is a balancing parameter. The optimal solution $\hat{u}$ of problem (12) can be readily derived in the following Sigmoid form:

$$\hat{u}(\mathrm{x}) = \frac{1}{e^{-\frac{o_1(\mathrm{x})-k(\mathrm{x})}{\varepsilon}}+1} = \mathcal{S}\left(\frac{o_1(\mathrm{x})-k(\mathrm{x})}{\varepsilon}\right).$$

In the specific case when $\varepsilon = 1$ and $k = 0$, $\mathcal{F}(u)$ reduces to the classical Sigmoid function. Furthermore, in the limiting case when $\varepsilon \to 0$ and $k = o_2$ where $o_2(\mathrm{x}) := \max_{\mathrm{x}\in\Omega}\{o_1(\mathrm{x})\} - o_1(\mathrm{x})$ defined as background feature, $\mathcal{F}(u)$ degenerates into the Potts model comprising solely the data fidelity term.

*3) Soft Threshold Dynamics of Sigmoid (STD-Sigmoid):* The standard Sigmoid operation performs pointwise normalization only on the foreground features, without incorporating any spatial structure or contextual information from the background region. The variational formulation of (12) naturally allows the embedding of spatial regularization terms. By introducing threshold dynamic regularization, the following variational problem is obtained:

$$\min_{u\in[0,1]} \mathcal{F}(u) + \lambda\langle u, G_\tau * (1-u)\rangle. \tag{13}$$

Since the term $\langle u, G_\tau * (1-u)\rangle$ is nonconvex in $u$, we employ a linearization strategy to replace this concave component to facilitate optimization. This leads to the following iterative scheme:

$$\begin{aligned} u^{t+1} &= \operatorname*{argmin}_{u\in[0,1]} \mathcal{F}(u) + \left\langle p^{t+1}, u - u^t \right\rangle \\ &= \mathcal{S}\left(\frac{o_1(\mathrm{x}) - k(\mathrm{x}) - p^{t+1}}{\varepsilon}\right), \quad t \in \{0,1,2,\cdots T\}. \end{aligned} \tag{14}$$

where $p^{t+1} = \lambda G_\tau * (1-2u^t)$ denotes the gradient of the last term evaluated at the solution $u^t$ from the $t$-th iteration.

Equation (14) embeds the threshold dynamic regularization term into the Sigmoid function (STD-sigmoid). By setting $k = o_2$, this formulation further yields an equivalent Sigmoid representation of the Potts model. Compared with the standard Sigmoid function, the proposed STD-Sigmoid serves two key purposes:

1) It rescales the feature $o_1$ into the $[0, 1]$ range, making it compatible with the subsequent loss computation;
2) The segmentation is further refined by enforcing spatial consistency between foreground and background features, as prescribed by the Potts model, while the threshold dynamic regularization term enhances boundary coherence.

The overall STD-Sigmoid procedure is summarized in Algorithm 1.

**Algorithm 1** STD-Sigmoid function

**Input:** Foreground features $o_1$, parameters $\lambda$ and $\varepsilon$.
**Output:** Segmentation result $u$.

$u^0 \leftarrow \mathcal{S}\left(\frac{-o_1(\mathbf{x})}{\varepsilon}\right)$.
**for** $t = 1, 2, \cdots, T$ **do**
  $p^{t+1} \leftarrow \lambda G_\tau * (1 - 2u^t)$.
  Compute the STD-Sigmoid solution via (14).
**end for**
**return** Segmentation result $u = u^T$.

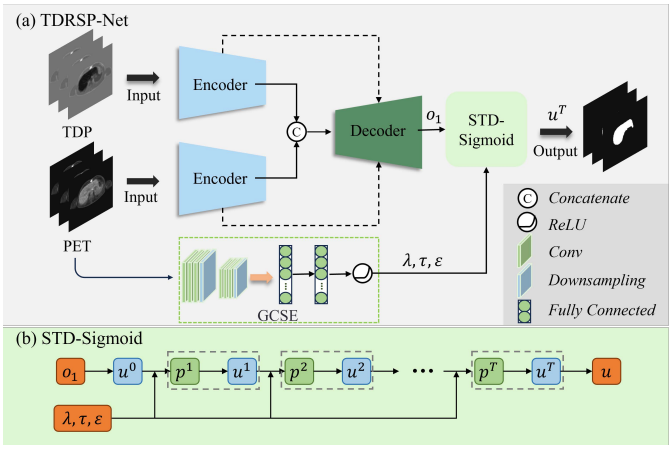


Fig. 3. (a) The network architecture of TDRSP-Net. It consists of two weight-shared encoders, a decoder, an STD-Sigmoid layer, and an image-driven global context scalar estimator (GCSE). (b) The flowchart of the STD-Sigmoid module.

## C. Model and Framework

We propose topographic distance priors and spatial regularization network (TDPSR-Net) to enhance PET image segmentation. The overall architecture is illustrated in Fig. 3. The network consists of two weight-shared encoders, a decoder, an STD-Sigmoid layer, and an image-driven global context scalar estimator (GCSE). The inputs are paired PET image $I(\mathrm{x})$ and its corresponding TDP map $D_K(\mathrm{x})$. The encoders independently extract PET features and distance-related features, which are then concatenated and passed to the decoder. Through skip connections, the decoder progressively restores the spatial resolution while integrating PET and topographic-distance–based priors at each stage to achieve deep feature fusion. The resulting features are subsequently fed into the STD-Sigmoid layer for further refinement. In addition, the GCSE adaptively estimates the hyperparameters of the STD-Sigmoid based on the input PET image.

*1) Encoder:* In the encoding phase, two encoders simultaneously extract features from the PET and TDP. Each encoder consists of 4 downsampling layers and 1 bridging layer. The downsampling layers comprise convolutional blocks and max-pooling layers, which capture multi-scale features. Each convolutional block consists of 3×3 convolutions with stride 1, instance normalization, and ReLU activation; the bridging layer further refines the downsampled features using dual convolutions. Once feature extraction is complete, the PET and TDP features are concatenated and passed to the decoder. The two encoders share parameters, ensuring that the introduction of the TDP input does not increase the overall parameter count.

*2) Decoder:* Similar to the encoder, the decoder consists of 4 upsampling layers, each comprising transposed convolutions and convolutional blocks, gradually recovering the original

spatial resolution. At each scale, PET and TDP features from the two encoders are concatenated and fed to the corresponding decoder stage by skip connections. The decoder produces a feature map with the same spatial resolution as the input image. This feature map is then transformed into a probability map through the STD-Sigmoid, where the spatial regularization term is simultaneously incorporated.

*3) Global Context Scalar Estimator:* In the conventional Potts model, the scalar parameter $\lambda$ must be manually tuned for each image to balance the data fidelity and regularization terms, which is both labor-intensive and time-consuming. In contrast, the proposed STD-Sigmoid involves three scalar parameters ($\lambda$, $\tau$, and $\varepsilon$), making manual selection for each image even more challenging. Under the deep learning paradigm, fixing these parameters to constant values limits the model's ability to adapt to diverse image characteristics. To address this limitation, we introduce a Global Context Scalar Estimator (GCSE), which adaptively learns the optimal scalar parameters based on the content of each input image. Taking the estimation of $\lambda$ as an example, it is mathematically expressed as

$$\lambda = f_\theta(I), \qquad f_\theta : \Omega \to \mathbb{R}^+. \tag{15}$$

where $f_\theta$ is a deep nonlinear mapping function that can be trained end-to-end. Similarly, $\tau$ and $\varepsilon$ are learned using two independent estimators, respectively. Specifically, $f_\theta$ employs a cascade of multi-scale convolutional downsampling units to progressively enlarge the effective receptive field. The input first passes through two such units, each comprising $7 \times 7$, $5 \times 5$, $3 \times 3$ convolutions followed by downsampling. This hierarchical convolutional design enables the joint capture of global semantic context and fine-grained structural details. The resulting high-dimensional representation is subsequently compressed into a single scalar through a two-stage fully connected transformation with ReLU activation, thereby allowing adaptive balancing of the associated regularization terms.

The network parameters $\theta$ of the GCSE are optimized jointly with the segmentation network using the same loss function. Let the loss function be defined as $\mathcal{L} := \mathcal{L}(\hat{u}, u_{gt})$ where $\hat{u}$ denotes the output of the STD-Sigmoid layer and $u_{gt}$ is the ground truth. The parameters $\theta$ are updated according to

$$\theta \leftarrow \theta - \Lambda \frac{\partial \mathcal{L}}{\partial \theta}, \tag{16}$$

where $\Lambda$ is the learning rate determined by the optimizer (e.g., SGD or Adam). The partial derivatives $\frac{\partial \mathcal{L}}{\partial \theta}$ can be computed using the chain rule. By optimizing the network parameters $\theta$ during the training phase, the GCSE is able to adaptively predict the scalar parameter $\lambda, \tau, \varepsilon$ for each input image during testing.

## III. Experiments and Results

### A. Clinical Datasets

*1) Liver Dataset:* The liver dataset used in this study was collected from Henan Provincial People's Hospital via a PET/CT scanner, and it includes 108 subjects. All the subjects fasted for at least 6 hours before receiving an injection of the $^{18}$F-FDG. The study was approved by the ethics committees of Henan Provincial People's Hospital and Zhengzhou University People's Hospital, and written informed consent was obtained from all participants before enrollment. PET voxel size was $4.06\times4.06\times3\text{mm}^3$, and CT voxel size was $0.98\times0.98\times3\text{mm}^3$. After PET–CT registration and resampling, all volumes were standardized to a matrix size of $160 \times 160 \times 160$. Only axial slices containing the liver were retained, yielding a total of 5,671 slices. The dataset was split at the patient level into training, validation, and test sets, comprising 70, 8, and 30 patients, respectively.

*2) Hippocampus Dataset:* The hippocampus dataset was collected at Henan Provincial People's Hospital and includes 92 subjects. All subjects fasted for at least 6 h before $^{18}$F-FDG injection, followed by brain imaging using a PET/MRI scanner. The original T1-weighted MR images had a matrix size of $345 \times 384 \times 264$ and a voxel size of $0.67 \times 0.67 \times 0.67$ mm$^3$. The original PET images had a matrix size of $192\times 192 \times 227$ with a voxel size of $0.86 \times 0.86 \times 1.00\text{mm}^3$. PET and MR images were preregistered and resampled to a final matrix size of $256 \times 256 \times 256$. In our experiments, coronal slices containing the hippocampus were selected, resulting in 3,648 slices. The dataset was split at the patient level into training, validation, and test sets, comprising 54, 8, and 22 patients, respectively.

*3) PCLT20K Dataset [41]:* The PCLT20K dataset was acquired using the GE Discovery Elite PET/CT scanner (GE Healthcare), with data from 605 patients diagnosed with lung tumors. Prior to scanning, patients fasted for approximately 6 hours, with blood glucose levels maintained below 11.1 mmol/L. After injection of 4.2 MBq/kg $^{18}$F-FDG, data collection took place 50–60 minutes later. A helical CT scan (80 mAs, 120 kVp, 5-mm slice thickness) was first performed for anatomical localization and attenuation correction, followed by a 3D PET emission scan from the distal femur to the cranial apex. The CT voxel size was $0.98 \times 0.98 \times 2.8$ mm$^3$, while the PET images were reconstructed using OSEM iterative reconstruction, yielding a final voxel size of $3.6 \times 3.6 \times 3.3$ mm$^3$. The PET images were first converted into standardized uptake value (SUV) maps, followed by registration with the corresponding CT images and resampling to a uniform matrix size of $512 \times 512$. The dataset was split into training, validation, and test sets, comprising 410, 15, and 180 patients, respectively.

### B. Implementation Details

In the experiments, all input images were intensity-normalized before being fed into the network. During training, random flipping was employed as a data augmentation strategy to enhance the model's robustness. For the Liver and Hippocampus datasets, we trained the model on full 2D slices at fixed in-plane resolutions, with input sizes of $160 \times 160$ and $256 \times 256$, respectively. For the PCLT20K dataset, we randomly cropped 3D patches of size $512 \times 512 \times 32$. For 2D settings, we adopted 2D convolutional layers; for 3D settings, we replaced them with 3D convolutions while keeping the same architecture. The network was optimized using a binary cross-entropy loss function and the Adam optimizer, with an

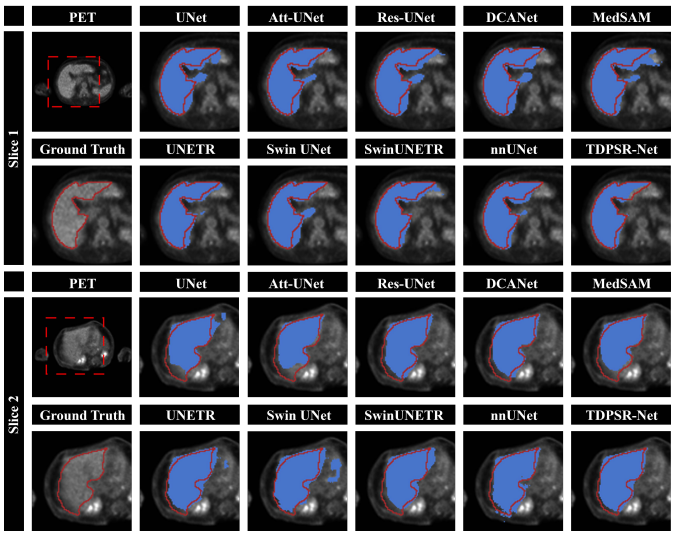


Fig. 4. Visualization of Results from Different Methods on the Liver Dataset. (The red dashed box represents the zoomed-in region, the blue area corresponds to the prediction results of different methods, and the red curve denotes the true target boundary.)

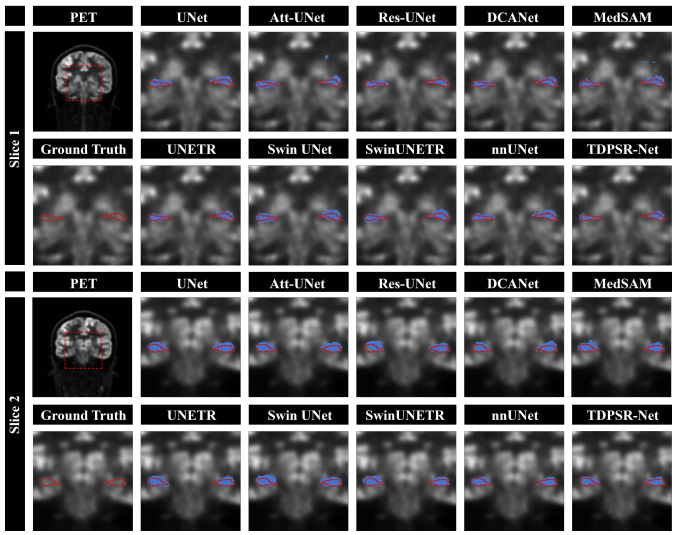


Fig. 5. Visualization of Results from Different Methods on the hippocampus Dataset. (The red dashed box represents the zoomed-in region, the blue area corresponds to the prediction results of different methods, and the red curve denotes the true target boundary.)

initial learning rate of $1 \times 10^{-4}$. The batch size was set to 16 for 2D and 8 for 3D, and the network was trained for 200 epochs. All experiments were conducted on a server equipped with two NVIDIA RTX 4090 GPUs.

### C. Evaluation Metrics

In this study, we use four evaluation metrics to assess the segmentation results, including the Dice similarity coefficient (DSC), Jaccard index (JAC), precision, and average symmetric surface distance (ASSD). The formulas for these evaluation metrics are defined as follows:

$$\mathrm{DSC} = \frac{2 \times |S_g \cap S_p|}{|S_g| + |S_p|}, \qquad \mathrm{JAC} = \frac{|S_g \cap S_p|}{|S_g \cup S_p|},$$

$$\mathrm{Precision} = \frac{|S_g \cap S_p|}{|S_p|}, \qquad \mathrm{ASSD} = \frac{H(g,p)}{|\partial S_p| + |\partial S_g|},$$

where $H(g,p) = \sum_{p \in \partial S_p} \min_{g \in \partial S_g} d_E(p,g) + \sum_{g \in \partial S_g} \min_{p \in \partial S_p} d_E(g,p)$. In this context, $S_g$ represents the ground-truth target region, and $S_p$ represents the predicted target region. The sets $\partial S_g$ and $\partial S_p$ denote the boundary points of $S_g$ and $S_p$, respectively. $d_E(p,g)$ denotes the Euclidean distance between point $p$ and point $g$.

### D. Results

TDPSR-Net was compared with nine representative classical and state-of-the-art methods, including UNet [15], Att-UNet [16], Res-UNet [17], DCANet [42], MedSAM [43], UNETR [44], Swin UNet [18], SwinUNETR [19], and nnUNet [45]. For all comparative methods, the hyperparameter configurations were strictly adopted from the respective original papers to ensure a fair and reproducible comparison. Furthermore, all baseline models were trained from scratch on our dataset without using any pretrained weights to eliminate potential biases.

*1) Organ Segmentation:* Fig. 4 presents a visual comparison of the liver dataset. Due to the blurred boundaries in PET images, the target region is easily confused with adjacent tissues. Models such as UNet, Att-UNet, Res-UNet, DCANet, MedSAM, and Swin UNet exhibit significant mis-segmentation, incorrectly classifying surrounding tissues as the foreground. Although UNETR, SwinUNETR, and nnUNet show some improvements, visible deviations from the ground truth remain. In contrast, TDPSR-Net demonstrates a closer alignment with the true target boundaries.

To further assess generalization, comparative experiments were conducted on the hippocampus dataset. This dataset features small target volumes and bilateral structures, making the segmentation task particularly challenging. The visualization results in Fig. 5 demonstrate that although all methods can roughly locate the hippocampus, over-segmentation or under-segmentation occurs in areas with blurred boundaries. By contrast, TDPSR-Net produces contours that better match the reference annotations, improving boundary accuracy.

Table I reports the mean $\pm$ standard deviation of the four metrics for all methods on the liver and hippocampus datasets. Overall, pure CNN-based approaches (UNet, Att-UNet, Res-UNet, DCANet) tend to underperform Transformer-based or hybrid methods (e.g. UNETR, SwinUNETR). For liver segmentation, Res-UNet achieved the lowest DSC (83.91%), followed by DCANet (84.42%). SwinUNETR and nnUNet achieved competitive results; TDPSR-Net further improved DSC by 0.38% and 0.47% over these methods, respectively. A similar trend was observed on the hippocampus dataset: Res-UNet and DCANet again showed the worst and second-worst performance, with DSC of 71.73% and 73.05%, respectively. Among the competing methods, UNETR achieved the highest DSC (75.34%), but it still underperformed TDPSR-Net. Notably, although MedSAM achieved the best ASSD, this advantage may partly arise from the use of box prompts

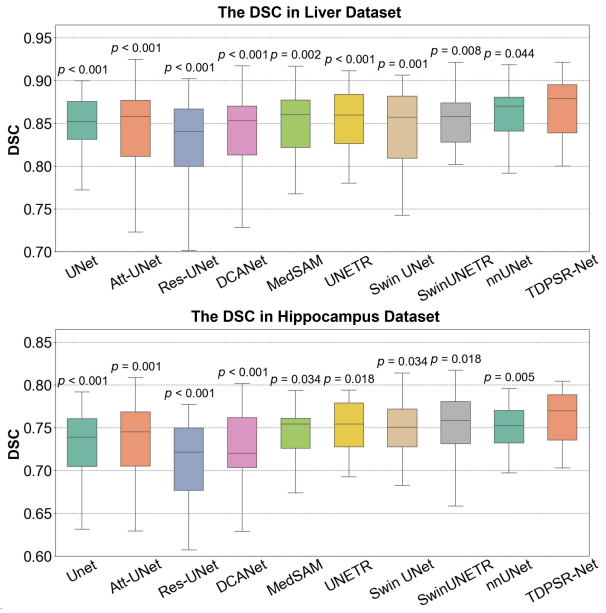


Fig. 6. Boxplot comparison of DSC across methods. Paired $t$-tests were performed to compare TDPSR-Net with each competing method with Holm–Bonferroni correction to adjust the resulting $p$-values for multiple comparisons; all $p$-values were below 0.05.

during inference. In our experiments, bounding boxes were generated from the ground-truth labels to ensure complete coverage of the target region. This strategy constrains the predicted boundary within a predefined spatial range, thereby reducing surface distance errors. Nevertheless, even without considering this inherent advantage of MedSAM, the proposed method still achieved superior ASSD performance compared with the remaining approaches.

To intuitively illustrate the performance differences, Fig. 6 presents the boxplots of the DSC values for all methods on the liver and hippocampus datasets. TDPSR-Net achieved the highest median DSC and a relatively narrow interquartile range (IQR), indicating improved stability. Moreover, paired $t$-tests were conducted to compare TDPSR-Net with each competing method, and the resulting $p$-values were adjusted for multiple comparisons using the Holm–Bonferroni correction to control the family-wise error rate (FWER) at 0.05. All Holm-adjusted $p$-values remained below 0.05, confirming that the performance improvements are statistically significant.

*2) Tumor Segmentation:* Fig. 7 shows qualitative results for lung tumor segmentation on the PCLT20K dataset. For relatively large lesions, Att-UNet and DCANet were noticeably affected by high-uptake regions around the tumor, leading to evident false positives. UNet and Res-UNet exhibited pronounced under-segmentation. Although UNETR and nnUNet delineated the main tumor region more completely, they still underperformed TDPSR-Net in capturing fine boundary details. For small lesions, UNet, Att-UNet, Res-UNet, and DCANet all suffered from varying degrees of under-segmentation. Swin UNet showed notable limitations and failed to identify tumor regions effectively. By contrast, MedSAM, UNETR, SwinUNETR, and nnUNet achieved relatively satisfactory performance, whereas TDPSR-Net further

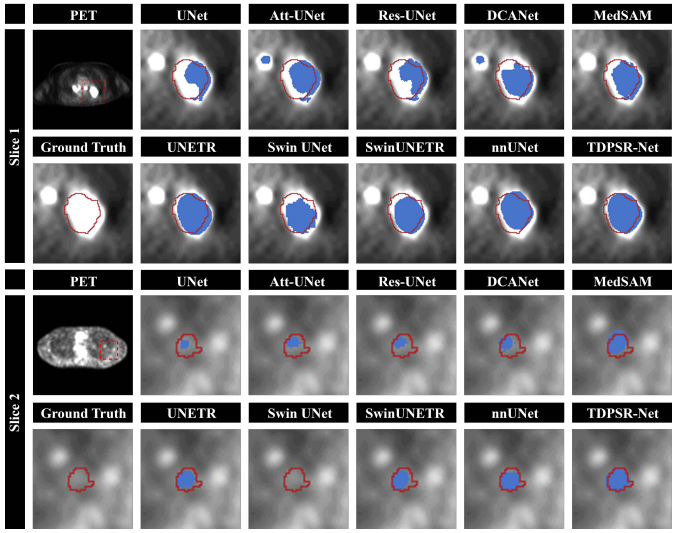


Fig. 7. Visualization of Results from Different Methods on the PCLT20K Dataset. (The red dashed box represents the zoomed-in region, the blue area corresponds to the prediction results of different methods, and the red curve denotes the true target boundary.)

improved lesion localization and boundary consistency.

Table II summarizes quantitative segmentation performance on PCLT20K and compares model efficiency in terms of parameter count, floating-point operations (FLOPs), and inference time. Overall, conventional CNN-based architectures (UNet, Att-UNet, Res-UNet, and DCANet) struggled with PET lung tumor imaging challenges, including heterogeneous tracer uptake and blurred lesion boundaries, with mean DSC values below 59% for all these methods. Although Res-UNet achieved the lowest computational cost (1.62M parameters and 74.87G FLOPs), this efficiency came at the expense of segmentation accuracy. UNETR, SwinUNETR, and nnUNet demonstrated notable performance improvements. However, these gains were accompanied by increased model complexity and longer inference times. With model complexity and inference time comparable to SwinUNETR, TDPSR-Net further improved DSC by 0.31%, indicating a more favorable trade-off between accuracy and efficiency.

### *E. Ablation Experiments*

*1) Impact of Marker Prediction Accuracy:* The marker points for topographic distance were generated by the first-stage UNet. However, due to the limited segmentation accuracy of the UNet, these markers did not always lie precisely within the true target regions. To evaluate the effect of marker localization on final segmentation, we calculated the proportion of markers that fell within the ground-truth regions. Table III summarizes the overall localization accuracy (Acc.) of the marker points across the three datasets and further compares the segmentation performance under correct and failed marker localization. For the hippocampus dataset, because each image contained two target structures, a sample was counted as correct only when both targets were localized; otherwise, it was treated as a failure. To further assess the robustness of the proposed method, we compared TDPSR-Net with the

TABLE I
QUANTITATIVE COMPARISON ON LIVER AND HIPPOCAMPUS DATASETS. ALL SEGMENTATION METRICS ARE REPORTED AS MEAN ± STANDARD DEVIATION, AND THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

| Methods | Liver | | | | Hippocampus | | | |
|---|---|---|---|---|---|---|---|---|
| | DSC(%)↑ | JAC(%)↑ | Prec.(%)↑ | ASSD(mm)↓ | DSC(%)↑ | JAC(%)↑ | Prec.(%)↑ | ASSD(mm)↓ |
| UNet [15] | 84.72±14.05 | 75.53±17.08 | 84.27±17.58 | 0.65±2.08 | 73.68±19.64 | 61.17±18.91 | 72.74±20.28 | 1.10±3.87 |
| Att-UNet [16] | 84.53±17.29 | 76.04±18.91 | 85.91±18.12 | 0.50±1.16 | 73.39±20.65 | 61.05±19.87 | 71.92±21.05 | 1.40±4.72 |
| Res-UNet [17] | 83.91±17.20 | 75.14±19.05 | 83.52±19.58 | 0.66±1.87 | 71.73±21.74 | 59.31±20.61 | 71.62±22.21 | 1.10±3.69 |
| DCANet [42] | 84.42±16.93 | 75.70±18.82 | 84.21±18.14 | 0.51±1.37 | 73.05±21.18 | 60.82±20.24 | 73.63±20.65 | 1.18±4.16 |
| MedSAM [43] | 85.06±15.85 | 76.43±18.61 | 80.82±19.85 | 0.57±1.16 | 74.47±16.63 | 61.45±16.73 | 71.91±18.93 | **0.80±1.41** |
| UNETR [44] | 85.41±14.43 | 76.91±17.34 | 85.83±16.21 | 0.50±1.12 | 75.34±16.91 | 62.81±17.21 | 73.74±19.52 | 0.82±1.49 |
| Swin UNet [18] | 85.02±15.78 | 76.31±18.42 | 85.14±17.95 | 0.56±1.18 | 74.14±17.23 | 61.53±17.52 | 72.34±19.83 | 0.95±2.10 |
| SwinUNETR [19] | 86.14±11.58 | 77.71±14.72 | 87.13±15.12 | 0.44±1.33 | 74.91±15.00 | 62.13±16.24 | 73.71±14.80 | 0.86±1.18 |
| nnUNet [45] | 86.05±13.52 | 77.52±17.54 | 87.32±16.51 | 0.44±1.11 | 75.21±17.01 | 62.61±17.31 | 73.42±19.64 | 0.83±1.65 |
| **TDPSR-Net** | **86.52±12.64** | **78.01±15.52** | **87.93±16.21** | **0.42±0.91** | **75.63±16.23** | **63.21±16.91** | **74.15±17.72** | 0.82±1.52 |

TABLE II
QUANTITATIVE COMPARISON OF SEGMENTATION PERFORMANCE AND MODEL COMPLEXITY (PARAMETERS, FLOPS, AND INFERENCE TIME) ON THE PCLT20K DATASET. ALL SEGMENTATION METRICS ARE REPORTED AS MEAN ± STANDARD DEVIATION, AND THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

| Methods | PCLT20K | | | | | | |
|---|---|---|---|---|---|---|---|
| | DSC(%)↑ | JAC(%)↑ | Prec.(%)↑ | ASSD(mm)↓ | Parameters(M) | FLOPs(G) | Inferencing time(ms) |
| UNet [15] | 57.15±25.63 | 47.26±20.57 | 64.96±27.57 | 5.51±6.41 | 1.78 | 115.13 | **244.02** |
| Att-UNet [16] | 58.07±24.47 | 47.38±20.10 | 67.65±26.45 | 4.92±4.61 | 28.72 | 1636.78 | 1191.30 |
| Res-UNet [17] | 57.56±25.29 | 47.12±20.48 | 67.23±27.14 | 5.30±5.19 | **1.62** | **74.87** | 374.68 |
| DCANet [42] | 57.85±16.93 | 47.40±18.82 | 65.82±18.14 | 5.62±7.02 | 4.13 | 3814.13 | 4712.12 |
| MedSAM [43] | 58.54±24.86 | 47.82±20.08 | 68.26±26.95 | 5.07±4.99 | 93.73 | 743.98 | 2021.74 |
| UNETR [44] | 60.20±23.26 | 48.84±18.80 | 68.20±26.13 | 4.94±4.35 | 92.61 | 391.28 | 1378.11 |
| Swin UNet [18] | 44.30±18.38 | 38.95±12.70 | 60.50±33.13 | 7.09±6.19 | 48.81 | 270.32 | 1420.62 |
| SwinUNETR [19] | 62.13±22.64 | 50.58±18.53 | 68.29±24.46 | **4.82±5.03** | 61.98 | 1568.41 | 2735.16 |
| nnUNet [45] | 61.73±24.35 | 50.50±19.94 | 68.48±26.71 | 4.96±4.44 | 18.46 | 1193.98 | 1814.67 |
| **TDPSR-Net** | **62.44±23.99** | **51.05±19.76** | **68.91±26.89** | 4.89±6.13 | 77.59 | 1675.15 | 1638.45 |

second-best method (SBM) on each dataset in the Table III, namely SwinUNETR for the liver dataset, UNETR for the hippocampus dataset, and SwinUNETR for the PCLT20K dataset. When the marker localization was correct, TDPSR-Net and the SBM achieved closely matched performance across the three datasets. In contrast, when marker localization failed, incorporating TDP still enabled TDPSR-Net to achieve improved segmentation performance. In particular, under failed marker localization, TDPSR-Net improved DSC over the SBM by 3.16%, 0.09%, and 3.36% on the liver, hippocampus, and PCLT20K datasets, respectively, while also achieving lower ASSD on all three datasets. Fig. 8 presents a representative example in which the marker point deviates from the target. Although the initial marker lay outside the object, the topographic distance still captured the target topology and provided a meaningful shape prior. Compared with the original PET image, the distance information encoded in TDP improved the discriminability between regions. The proposed TDPSR-Net effectively leveraged the TDP to substantially improve the final segmentation, reducing false positives and refining boundary delineation, compared with the coarse segmentation generated by the UNet. These results indicated that TDPSR-Net did not rely entirely on marker accuracy; even when the preliminary segmentation failed to place valid markers, TDP still provided meaningful global structural information.

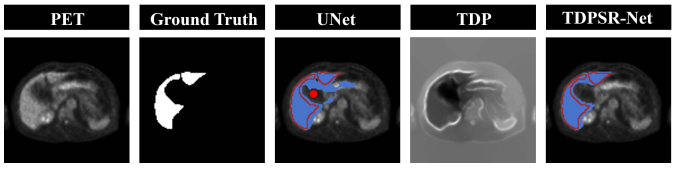


Fig. 8. A segmentation example with incorrectly predicted marked point. (The blue area corresponds to the prediction results of different methods, the red curve denotes the true target boundary, and the red dot represents marked point.)

*2) Impact of STD-Sigmoid:* The behavior of the STD-Sigmoid was primarily governed by the number of iterations $T$. To examine its impact on segmentation performance, we gradually varied $T$ from 0 to 20 on the liver dataset and evaluated the corresponding results. When $T = 0$, the network employed the conventional Sigmoid activation. Fig.9 presented the quantitative performance as a function of the iteration number $T$. Compared with $T = 0$, DSC improved noticeably at $T = 2$. With larger $T$, both DSC and ASSD gradually converged, indicating that performance became stable with respect to $T$. Based on these observations, we set $T = 20$ in experiments to ensure convergence.

In addition, we replaced the proposed STD-Sigmoid with commonly used activation functions in segmentation, includ-

TABLE III
THE IMPACT OF MARKER LOCALIZATION ACCURACY ON SEGMENTATION PERFORMANCE. CORRECT MARKER LOCALIZATION INDICATES THAT THE PREDICTED MARKER POINTS FALL WITHIN THE GROUND-TRUTH TARGET REGION; FAILED MARKER LOCALIZATION INDICATES THAT AT LEAST ONE PREDICTED MARKER POINT LIES OUTSIDE THE TARGET REGION. AND SBM DENOTES THE SECOND BEST METHOD ON EACH DATASET, NAMELY SWINUNETR FOR LIVER, UNETR FOR HIPPOCAMPUS, AND SWINUNETR FOR PCLT20K.

| Dataset | Acc. | Correct Marker Localization | | | | | | Failed Marker Localization | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | UNet | | SBM | | TDPSR-Net | | UNet | | SBM | | TDPSR-Net | |
| | | DSC↑ | ASSD↓ | DSC↑ | ASSD↓ | DSC ↑ | ASSD ↓ | DSC↑ | ASSD↓ | DSC ↑ | ASSD ↓ | DSC ↑ | ASSD ↓ |
| Liver | 96.28% | 85.45 | 0.64 | 86.66 | 0.42 | **86.93** | **0.41** | 65.83 | 0.95 | 72.77 | 0.91 | **75.93** | **0.81** |
| Hippocampus | 85.52% | 76.79 | 0.77 | 77.46 | 0.62 | **77.78** | **0.63** | 55.33 | 3.05 | 62.85 | 2.01 | **62.94** | **1.98** |
| PCLT20K | 73.55% | 59.84 | 4.23 | **65.93** | **3.48** | 65.91 | 3.58 | 49.68 | 9.10 | 51.56 | 8.60 | **54.92** | **8.54** |

TABLE IV
QUANTITATIVE METRICS AND COMPUTATIONAL COMPLEXITY OF DIFFERENT ACTIVATION STRATEGIES ON THE LIVER DATASET.

| Activate Function | DSC↑ | JAC↑ | Pred.↑ | ASSD↓ | FLOPs | Inferencing Time |
|---|---|---|---|---|---|---|
| Sigmoid | 85.64 | 76.84 | 86.36 | 0.47 | **177.79** | **4.63** |
| Softmax | 84.90 | 76.08 | 84.42 | 0.56 | 177.81 | 4.92 |
| STD [31] | 85.05 | 76.35 | 84.98 | 0.50 | 177.81 | 5.84 |
| STD-Sigmoid | **86.52** | **78.01** | **87.93** | **0.42** | 191.57 | 7.21 |

ing Sigmoid, Softmax, and STD (a variational formulation of Softmax [31]). Table IV reported the quantitative segmentation performance and computational efficiency of different activation functions. Notably, Softmax and STD had identical FLOPs, indicating that the variational formulation did not introduce noticeable computational overhead. By contrast, the increased FLOPs of STD-Sigmoid relative to Sigmoid mainly stemmed from the GCSE forward computation, rather than the activation operation itself. Despite a modest increase in inference time, STD-Sigmoid yielded substantial improvements in DSC and other metrics, demonstrating a favorable trade-off between accuracy and computational cost.

Fig. 9 further illustrated the loss convergence under different activation strategies during training. As can be observed, STD-Sigmoid exhibits competitive convergence compared with Sigmoid, Softmax, and STD and reaches a lower loss level in the later stage of training. Although STD also shows a rapid initial descent, it stabilizes at a higher loss level under the same experimental setting. Notably, although direct comparison of loss values across different activation functions is subject to limitations because they alter the optimization landscape, the final segmentation performance reported in Table IV still supports the effectiveness of the STD-Sigmoid configuration.

*3) Impact of Network Architecture:* To analyze the architecture and the contribution of each module, we designed a series of network variants, as shown in Fig. 10. Model (a) was the standard UNet and serves as the baseline. Starting from model (a), we separately incorporated TDP and STD-Sigmoid, yielding model (b) (UNet-TDP) and model (c) (UNet-STD). By simultaneously introducing both TDP and STD-Sigmoid, we obtain model (d), referred to as UNet-TDP-STD. Model (e) further extends the architecture by adopting a dual-branch encoder, where PET and TDP are processed by two encoder branches with shared weights. The extracted features were concatenated at the encoder output and then fed into a shared

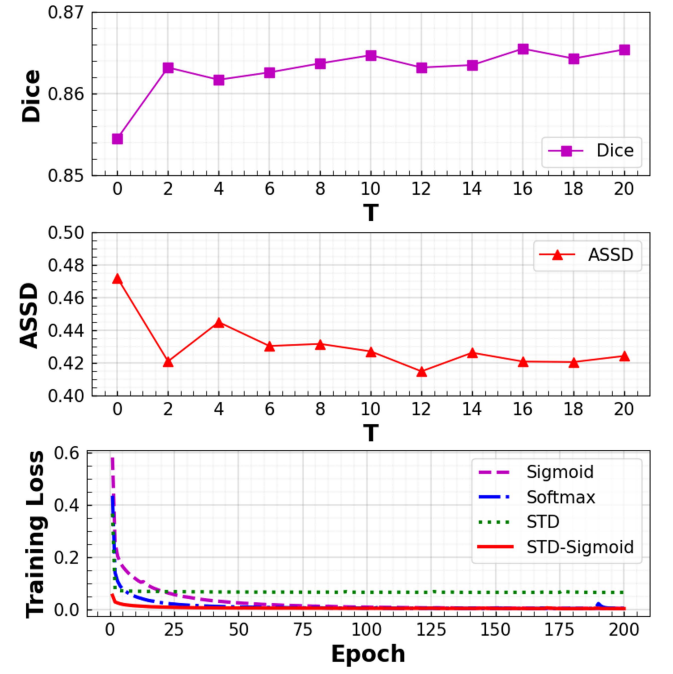


Fig. 9. Impact of the iteration number $T$ in STD-Sigmoid on Dice and ASSD, along with the loss convergence curves under different activation strategies during training.

decoder, enabling more effective feature fusion. Building on model (e), model (f) replaced the output-layer activation with STD-Sigmoid, resulting in the proposed TDPSR-Net. For TDPSR-Net, the parameters mainly consisted of three components: the encoder, the decoder, and the GCSE module. Since the two encoder branches shared parameters, the total number of encoder parameters was identical to that of a single encoder branch. However, concatenating the features from the two branches before the decoder increases the channel dimensionality of the decoder input, thereby substantially increasing the number of decoder parameters compared with the single-branch setting. To ensure a fair comparison, we considered two parameter-matched evaluation scenarios. First, under the condition of identical encoder parameter counts, models (a)–(d) were denoted as UNet-E, UNet-TDP-E, UNet-STD-E, and UNet-TDP-STD-E, respectively. Second, under the condition of identical decoder parameter counts, models (a)–(d) were denoted as UNet-D, UNet-TDP-D, UNet-STD-

TABLE V
QUANTITATIVE COMPARISON OF DIFFERENT NETWORK FRAMEWORKS UNDER EQUAL ENCODER-PARAMETER AND EQUAL DECODER-PARAMETER CONFIGURATIONS.

| Model | Parameters (M) | | | DSC↑ | JAC↑ | Prec.↑ | ASSD↓ |
|---|---|---|---|---|---|---|---|
| | Total | Encoder | Decoder | | | | |
| UNet-E | 8.63 | 4.71 | 3.92 | 84.70 | 75.48 | 84.19 | 0.65 |
| UNet-TDP-E | 8.63 | 4.71 | 3.92 | 85.44 | 76.58 | 84.79 | 0.52 |
| UNet-STD-E | 29.26 | 4.71 | 3.92 | 85.23 | 76.18 | 84.53 | 0.51 |
| UNet-TDP-STD-E | 29.26 | 4.71 | 3.92 | 85.71 | 76.89 | 86.09 | 0.46 |
| UNet-D | 34.51 | 18.74 | 15.67 | 85.42 | 76.73 | 86.47 | 0.54 |
| UNet-TDP-D | 34.51 | 18.74 | 15.67 | 85.55 | 76.76 | 86.32 | 0.48 |
| UNet-STD-D | 55.14 | 18.74 | 15.67 | 85.59 | 76.82 | 86.00 | 0.51 |
| UNet-TDP-STD-D | 55.14 | 18.74 | 15.67 | 85.80 | 77.00 | 86.38 | 0.47 |
| YNet-TDP | 20.38 | 4.71 | 15.67 | 85.64 | 76.84 | 86.36 | 0.47 |
| **TDPSR-Net** | 41.01 | 4.71 | 15.67 | **86.52** | **78.01** | **87.93** | **0.42** |

D, and UNet-TDP-STD-D, respectively.

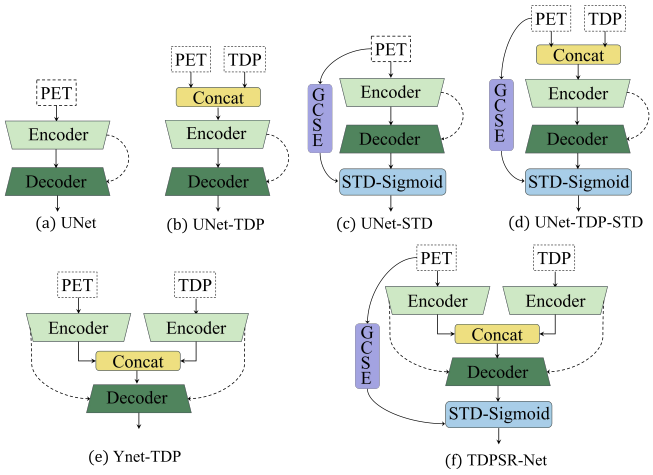


Fig. 10. The structures of the network variants used for ablation studies.

Table V summarized the quantitative results achieved by different model variants. Under both encoder- and decoder-parameter-matched settings, models (b) and (c) consistently outperformed the baseline model (a), indicating that each module contributed to performance gains. Model (d) further surpassed models (b) and (c) across all metrics, indicating that jointly integrating both modules was more beneficial than using either module alone. Model (e) achieved better performance than model (b), supporting the dual-branch encoder design for feature representation. Notably, despite using fewer encoder parameters, TDPSR-Net still outperformed UNet-STD-D and UNet-TDP-STD-D, indicating that separately encoding heterogeneous features led to more discriminative representations and improved segmentation.

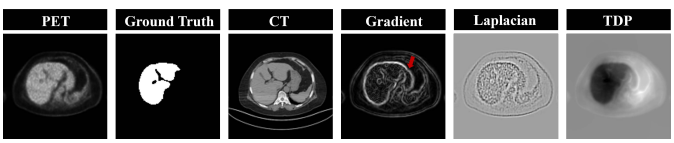


Fig. 11. Examples of PET, CT, gradient magnitude, laplacian, and topographic distance.

*4) Impact of TDP Inputs:* In TDPSR-Net, prior information was constructed to assist PET segmentation. To examine the impact of using TDP as an input prior, we replaced TDP with other priors, including gradient magnitude $\|\nabla I\|$, the Laplacian operator $\Delta I$, and an additional modality (CT). Moreover, to investigate the feasibility of TDP in a multimodal setting, we performed a PET+CT+TDP segmentation experiment in which PET, CT, and TDP were jointly used as inputs. In this setting, three encoders with shared weights were employed to process the three inputs independently.

Table VI reports the corresponding results. Using gradient magnitude or Laplacian features yielded inferior performance compared with using TDP. Fig. 11 presents representative examples of different features. It can be observed that gradient-based and Laplacian-based representations mainly emphasize local boundary responses and lack target-specific guidance. In addition, due to the blurred boundaries and low spatial resolution of PET images, gradient computation is easily disturbed by weak or spurious edges highlighted by the red arrows in Fig. 11. As a result, incorporating gradient information may even degrade performance and thus negatively impact the model. In contrast, TDP was generated from marker points, which facilitated target localization while providing global topological information. Replacing TDP with CT led to a marked performance improvement, mainly because CT provided clearer anatomical delineation. TDP served as a prior tailored to low-resolution and boundary-ambiguous PET conditions. Notably, when PET, CT, and TDP were jointly used, the DSC was higher than that obtained with PET and CT alone. A paired statistical test further showed that this improvement was statistically significant ($p = 0.008$). This result indicated that TDP remained beneficial and feasible in multimodal segmentation.

TABLE VI
ABLATION STUDY ON THE EFFECT OF DIFFERENT INPUT ON THE LIVER DATASET.

| Input | DSC↑ | JAC↑ | Pred.↑ | ASSD↓ |
|---|---|---|---|---|
| CT | 89.84 | 84.85 | 89.43 | 0.21 |
| PET + Gradient Magnitude | 84.22 | 71.87 | 85.45 | 0.78 |
| PET + Laplacian | 85.13 | 74.03 | 86.26 | 0.54 |
| PET + CT | 90.42 | 86.68 | 90.57 | 0.14 |
| PET + CT + TDP | **90.84** | **86.79** | **90.71** | **0.11** |
| PET + TDP | 86.52 | 78.01 | 87.93 | 0.42 |

*5) Impact of Hyperparameters in the Topographic Distance:* In TDP, two hyperparameters, $K$ and $\sigma$, were specified. In this section, we examined their effects on TDP qualitatively. Specifically, we fixed $\sigma = 1$ and varied $K$ in $[0, 0.01, 0.03, 0.05, 0.08, 0.1]$ to analyze its influence on the resulting TDP. For visualization, Fig. 12 showed pseudo-color maps of TDP under different $K$ settings. When $K \leq 0.05$, variations in $K$ had only a marginal impact on TDP. Although TDP generally captured the overall target shape, gradient accumulation caused by grayscale inhomogeneity in PET led to larger distance values in regions farther from the marker points. When $K = 0.08$ and $K = 0.1$, visual differences between the corresponding TDP maps were negligible. However, at $K = 0.08$, background regions exhibited relatively larger distance values, resulting in more pronounced contrast

between target and background. This contrast enhancement improved target distinguishability.

To examine the effect of $\sigma$, we fixed $K = 0.08$ and varied $\sigma$ in $[0, 0.3, 0.5, 0.7, 0.9, 1]$, where $\sigma = 0$ indicated that no Gaussian filtering was applied. Fig. 13 illustrated the influence of $\sigma$ on TDP. Without Gaussian filtering or with small $\sigma$, gradient accumulation due to grayscale inhomogeneity prevented TDP from accurately delineating target boundaries. As $\sigma$ increased, TDP values within the target became progressively more uniform, yielding clearer boundary representation and the best visual quality at $\sigma = 1$. Based on these observations, we selected $K = 0.08$ and $\sigma = 1$ for subsequent experiments, as this configuration yielded the most favorable TDP representation.

## IV. Discussion and Conclusion

In this work, we propose a novel PET image segmentation framework, termed TDPSR-Net, which integrates TDP and spatial regularization into a unified network to address the intrinsic challenges of PET imaging, including low spatial resolution, noise, and metabolic heterogeneity. By incorporating these components, the proposed framework improves both the accuracy and stability of target region segmentation in PET images.

To mitigate boundary ambiguity commonly observed in PET images, we introduce TDP features generated from marker points as an additional network input to explicitly model the global topological structure of the target region. By encoding topological information, the network is able to capture target regions with varying sizes and structural complexities. The marker points required for TDP construction are obtained automatically using a UNet [15]. Notably, the predicted marker points are not always located within the target region. Nevertheless, experimental results demonstrate that even when the marker points deviate from the target region, the resulting TDP can still capture the global topological structure of the target and provide an effective shape prior for the subsequent segmentation network.

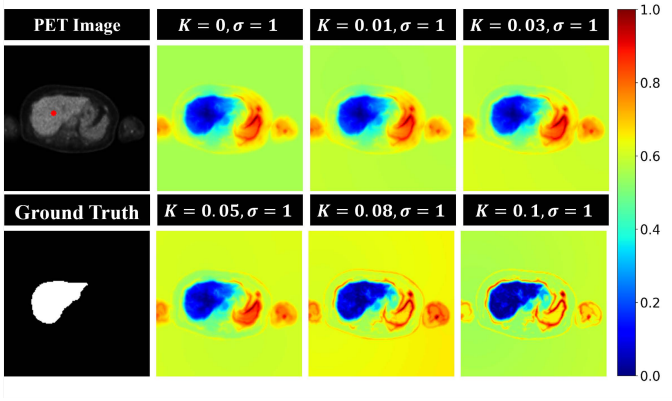


Fig. 12. The visual effects of the topographic distance prior under different parameter K settings.

The conventional Sigmoid activation is reinterpreted from a variational perspective and augmented with an explicit spatial regularization term, yielding a formulation that establishes a theoretical connection to the Potts model [10]. Compared with the standard Sigmoid function, this formulation introduces spatial consistency constraints during probability map generation, which effectively suppress image noise and improve the smoothness and coherence of segmentation boundaries. Moreover, a Global Context Scalar Estimator (GCSE) is designed to predict the STD-Sigmoid scalar parameters from the input image. This mechanism enables automatic parameter estimation directly from the image data, thereby eliminating the need for manual parameter tuning required by traditional approaches.

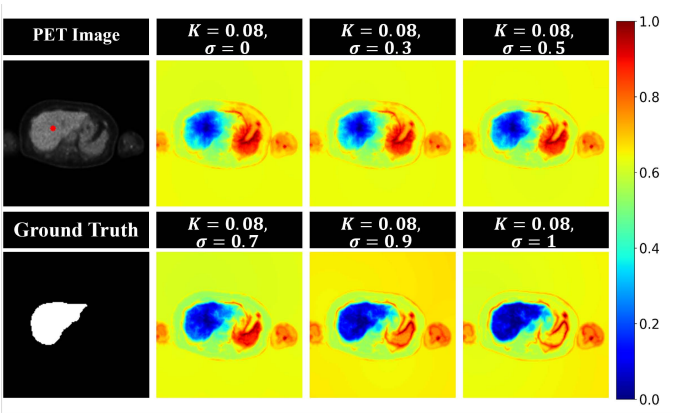


Fig. 13. The visual effects of the topographic distance prior under different parameter $\sigma$ settings, where $\sigma = 0$ indicates no Gaussian filter is applied.

Experiments are conducted on multiple datasets to evaluate the proposed method. The results demonstrate that TDPSR-Net consistently achieves strong segmentation performance across organs and tumors with different sizes and scales. Ablation studies further support the effectiveness of the proposed components for PET segmentation. In addition, the feasibility of incorporating TDP into a PET/CT multimodal segmentation setting is investigated. The results indicate that TDP remains beneficial even when anatomical information from CT is available to guide PET segmentation. This benefit is attributed to explicit encoding of global target topology, suggesting that TDP complements other modalities by providing a robust shape prior for PET segmentation.

From a clinical perspective, the proposed PET segmentation framework holds potential practical value and broad applicability. In routine practice, PET is typically performed alongside CT or MRI to provide anatomical reference and support correction procedures for quantitative PET [46]. However, the reliability of these complementary modalities could be compromised by patient tolerance and acquisition-time constraints [26]. For radiation-sensitive populations (e.g., pediatric and elderly patients), CT acquisition could be adjusted or limited to reduce dose [47]. Similarly, MRI acquisition is often constrained by long scan times and the requirement for patient cooperation. Patient motion during prolonged MRI examinations may introduce motion-related artifacts, thereby degrading image reliability and limiting its effectiveness as a complementary modality [48]. In this context, effectively leveraging the structural and functional information inherently

present in PET images to enhance segmentation performance carries important clinical implications.

However, the current framework still has some limitations. In particular, the method adopts a relatively cumbersome two-stage design, in which the marker points used to construct TDP are generated by a first-stage UNet. Consequently, inaccurate marker localization may degrade the quality of the derived TDP maps and introduce error propagation to the subsequent segmentation stage. In future work, we will explore an end-to-end framework to jointly optimize marker generation and segmentation for improved robustness and a more streamlined pipeline. In addition, future work will investigate latent structural and functional characteristics in PET and extend the framework to whole-body and multi-target segmentation. In addition, integrating topographic priors with multimodal frameworks will be explored to leverage complementary CT/MRI information when available while maintaining robustness across diverse clinical settings.

## V. Acknowledgement

All authors declare that they have no known conflicts of interest in terms of competing financial interests or personal relationships that could have an influence or are relevant to the work reported in this paper.

## References

[1] B. Foster, U. Bagci, A. Mansoor, Z. Xu, and D. J. Mollura, "A review on segmentation of positron emission tomography images," *Computers in Biology and Medicine*, vol. 50, pp. 76–96, 2014.

[2] Y. Onishi, T. Isobe, M. Ito, F. Hashimoto, T. Omura, and E. Yoshikawa, "Performance evaluation of dedicated brain PET scanner with motion correction system," *Annals of Nuclear Medicine*, vol. 36, no. 8, pp. 746–755, 2022.

[3] L. Vass, M. Fisk, S. Lee, F. J. Wilson, J. Cheriyan, and I. Wilkinson, "Advances in PET to assess pulmonary inflammation: a systematic review," *European Journal of Radiology*, vol. 130, p. 109182, 2020.

[4] R. L. Wahl, H. Jacene, Y. Kasamon, and M. A. Lodge, "From RECIST to percist: evolving considerations for PET response criteria in solid tumors," *Journal of Nuclear Medicine*, vol. 50, no. Suppl 1, pp. 122S–150S, 2009.

[5] M. Soret, S. L. Bacharach, and I. Buvat, "Partial-volume effect in PET tumor imaging," *Journal of Nuclear Medicine*, vol. 48, no. 6, pp. 932–945, 2007.

[6] M. Hatt, J. A. Lee, C. R. Schmidtlein, I. E. Naqa, C. Caldwell, E. De Bernardi, W. Lu, S. Das, X. Geets, V. Gregoire *et al.*, "Classification and evaluation strategies of auto-segmentation approaches for PET: Report of aapm task group no. 211," *Medical Physics*, vol. 44, no. 6, pp. e1–e42, 2017.

[7] B. Berthon, C. Marshall, M. Evans, and E. Spezi, "Evaluation of advanced automatic PET segmentation methods using nonspherical thin-wall inserts," *Medical Physics*, vol. 41, no. 2, p. 022502, 2014.

[8] X. Geets, J. A. Lee, A. Bol, M. Lonneux, and V. Grégoire, "A gradient-based method for segmenting FDG-PET images: methodology and validation," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 34, no. 9, pp. 1427–1438, 2007.

[9] M. Hatt, C. C. Le Rest, A. Turzo, C. Roux, and D. Visvikis, "A fuzzy locally adaptive bayesian segmentation approach for volume determination in PET," *IEEE Transactions on Medical Imaging*, vol. 28, no. 6, pp. 881–893, 2009.

[10] X. Tai, L. Li, and E. Bae, "The potts model with different piecewise constant representations and fast algorithms: a survey," in *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging: Mathematical Imaging and Vision*. Springer, 2023, pp. 1–41.

[11] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.

[12] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of u-net," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[13] P.-H. Conze, G. Andrade-Miranda, V. K. Singh, V. Jaouen, and D. Visvikis, "Current and emerging trends in medical image segmentation with deep learning," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 7, no. 6, pp. 545–569, 2023.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[16] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: learning where to look for the pancreas," in *Medical Imaging with Deep Learning*, 2018.

[17] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[18] 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 *Proceedings of the European Conference on Computer Vision Workshops(ECCVW)*, 2022.

[19] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *International MICCAI Brainlesion Workshop*. Springer, 2022, pp. 272–284.

[20] Y. Gao, Z. Huang, Y. Wu, W. Li, M. Wen, W. Zhao, Q. Yang, C. Cheng, X. Yang, Y. Yang *et al.*, "Dual-prompt-enhanced multiorgan segmentation model for total-body pet images," *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2025.

[21] Y. Zhang, L. Xue, W. Zhang, L. Li, Y. Liu, C. Jiang, Y. Cheng, and Y. Qi, "Seganypet: Universal promptable segmentation from positron emission tomography images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 21 107–21 116.

[22] D. Xiang, B. Zhang, Y. Lu, and S. Deng, "Modality-specific segmentation network for lung tumor segmentation in PET-CT images," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1237–1248, 2022.

[23] H. Tang, Z. Huang, W. Li, Y. Wu, J. Yuan, Y. Yang, Y. Zhang, J. Qin, H. Zheng, D. Liang *et al.*, "Automatic brain segmentation for pet/MR dual-modal images through a cross-fusion mechanism," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[24] I. Mecheter, L. Alic, M. Abbod, A. Amira, and J. Ji, "MR image-based attenuation correction of brain PET imaging: review of literature on machine learning approaches for segmentation," *Journal of Digital Imaging*, vol. 33, no. 5, pp. 1224–1241, 2020.

[25] F. Yousefirizi, I. S. Klyuzhin, J. H. O, S. Harsini, X. Tie, I. Shiri, M. Shin, C. Lee, S. Y. Cho, T. J. Bradshaw *et al.*, "Tmtv-net: fully automated total metabolic tumor volume segmentation in lymphoma PET/CT images—a multi-center generalizability analysis," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 51, no. 7, pp. 1937–1954, 2024.

[26] A. Z. Kyme and R. R. Fulton, "Motion estimation and correction in SPECT, PET and CT," *Physics in Medicine & Biology*, vol. 66, no. 18, p. 18TR02, 2021.

[27] C. Catana, "Motion correction options in PET/MRI," in *Seminars in nuclear medicine*, vol. 45, no. 3. Elsevier, 2015, pp. 212–223.

[28] L. Yang, D. Shao, C. Cheng, C. Zou, Z. Huang, H. Zheng, D. Liang, Z.-F. Pang, X.-C. Tai, and Z. Hu, "An automatic 3D PET tumor segmentation framework assisted by geodesic sequences," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[29] S. Nowozin and C. H. Lampert, "Global connectivity potentials for random field models," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 818–825.

[30] C. Le Guyader and L. A. Vese, "Self-repelling snakes for topology-preserving segmentation models," *IEEE Transactions on Image Processing*, vol. 17, no. 5, pp. 767–779, 2008.

[31] J. Liu, X. Wang, and X.-C. Tai, "Deep convolutional neural networks with spatial regularization, volume and star-shape priors for image segmentation," *Journal of Mathematical Imaging and Vision*, vol. 64, no. 6, pp. 625–645, 2022.

[32] S. Luo, X.-C. Tai, and R. Glowinski, "Convex object (s) characterization and segmentation using level set function," *Journal of Mathematical Imaging and Vision*, vol. 64, no. 1, pp. 68–88, 2022.
[33] M. Roberts, K. Chen, and K. L. Irion, "A convex geodesic selective model for image segmentation," *Journal of Mathematical Imaging and Vision*, vol. 61, no. 4, pp. 482–503, 2019.
[34] L. Yang, D. Shao, Z. Huang, H. Zheng, D. Liang, Z.-F. Pang, X.-C. Tai, and Z. Hu, "Constructing prior-aided pet tumor segmentation using eikonal equation," 2024.
[35] J. B. Roerdink and A. Meijster, "The watershed transform: Definitions, algorithms and parallelization strategies," *Fundamenta Informaticae*, vol. 41, no. 1-2, pp. 187–228, 2000.
[36] L. D. Cohen and T. Deschamps, "Segmentation of 3D tubular objects with adaptive front propagation and minimal tree extraction for 3D medical imaging," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 10, no. 4, pp. 289–305, 2007.
[37] H. Zhao, "A fast sweeping method for eikonal equations," *Mathematics of Computation*, vol. 74, no. 250, pp. 603–627, 2005.
[38] X.-C. Tai, H. Liu, and R. Chan, "PottsMGNet: A Mathematical Explanation of Encoder-Decoder Based Neural Networks," *SIAM Journal on Imaging Sciences*, vol. 17, no. 1, pp. 540–594, 2023.
[39] M. Miranda Jr, D. Pallara, F. Paronetto, and M. Preunkert, "Short-time heat flow and functions of bounded variation in $\mathbf{R}^N$," in *Annales de la Faculté des sciences de Toulouse: Mathématiques*, vol. 16, 2007, pp. 125–145.
[40] J. Zhang and W. Guo, "A new regularization for deep learning-based segmentation of images with fine structures and low contrast," *Sensors*, vol. 23, no. 4, p. 1887, 2023.
[41] J. Mei, C. Lin, Y. Qiu, Y. Wang, H. Zhang, Z. Wang, and D. Dai, "Cross-modal interactive perception network with mamba for lung tumor segmentation in PET-CT images," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
[42] G. C. Ates, P. Mohan, and E. Celik, "Dual cross-attention for medical image segmentation," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 107139, 2023.
[43] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.
[44] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.
[45] 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.
[46] J. Ren, J. G. Eriksen, J. Nijkamp, and S. S. Korreman, "Comparing different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation," *Acta Oncologica*, vol. 60, no. 11, pp. 1399–1406, 2021.
[47] N. A. Bebbington, K. B. Christensen, L. L. Østergård, and P. C. Holdgaard, "Ultra-low-dose ct for attenuation correction: dose savings and effect on pet quantification for protocols with and without tin filter," *EJNMMI physics*, vol. 10, no. 1, p. 66, 2023.
[48] V. R. Tripathi, M. N. Tibdewal, and R. Mishra, "A survey on motion artifact correction in magnetic resonance imaging for improved diagnostics," *SN Computer Science*, vol. 5, no. 3, p. 281, 2024.