

# Cross-Modal Interactive Perception Network with Mamba for Lung Tumor Segmentation in PET-CT Images

Jie Mei<sup>1,2\*</sup>, Chenyu Lin<sup>2\*</sup>, Yu Qiu<sup>3†</sup>, Yaonan Wang<sup>1</sup>, Hui Zhang<sup>1</sup>, Ziyang Wang<sup>4</sup>, Dong Dai<sup>4</sup>

<sup>1</sup>Hunan University <sup>2</sup>Nankai University <sup>3</sup>Hunan Normal University <sup>4</sup>Tianjin Medical University Cancer Institute and Hospital  
 {jiemei, yaonan, zhanghui1983}@hnu.edu.cn, {chenyulin, yqiu}@mail.nankai.edu.cn  
 {wangziyang, dongdai}@tmu.edu.cn

## Abstract

Lung cancer is a leading cause of cancer-related deaths globally. PET-CT is crucial for imaging lung tumors, providing essential metabolic and anatomical information, while it faces challenges such as poor image quality, motion artifacts, and complex tumor morphology. Deep learning-based models are expected to address these problems, however, existing small-scale and private datasets limit significant performance improvements for these methods. Hence, we introduce a large-scale PET-CT lung tumor segmentation dataset, termed PCLT20K, which comprises 21,930 pairs of PET-CT images from 605 patients. Furthermore, we propose a cross-modal interactive perception network with Mamba (CIPA) for lung tumor segmentation in PET-CT images. Specifically, we design a channel-wise rectification module (CRM) that implements a channel state space block across multi-modal features to learn correlated representations and helps filter out modality-specific noise. A dynamic cross-modality interaction module (DCIM) is designed to effectively integrate position and context information, which employs PET images to learn regional position information and serves as a bridge to assist in modeling the relationships between local features of CT images. Extensive experiments on a comprehensive benchmark demonstrate the effectiveness of our CIPA compared to the current state-of-the-art segmentation methods. We hope our research can provide more exploration opportunities for medical image segmentation. The dataset and code are available at <https://github.com/mj129/CIPA>.

## 1. Introduction

Lung cancer remains one of the most significant causes of cancer-related mortality globally. Positron Emission Tomography-Computed Tomography (PET-CT) has become a cornerstone in the imaging and evaluation of lung

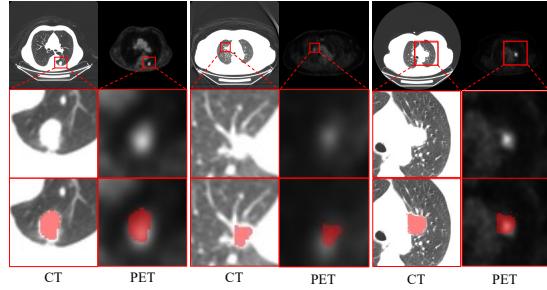


Figure 1. Some examples of the PET-CT images and the corresponding tumor picked from the PCLT20K dataset. The metabolic data from PET enhances sensitivity to lesion location, while the anatomical details from CT help precise localization and morphological characterization.

tumors, providing crucial metabolic and anatomical information [37]. The accurate segmentation of lung tumors in PET-CT images is essential for effective diagnosis, treatment planning, and monitoring of therapeutic responses. However, this task faces substantial challenges due to the complex nature of lung tumors and the variability in PET-CT images, such as poor image quality, motion artifacts or noise, complex morphology, irregular shape, and unclear tumor boundaries. Deep learning-based segmentation models are expected to address these problems, which can significantly improve segmentation accuracy, reduce the workload of radiologists, and lead to consistent assessments.

However, these models are usually data-hungry, and their performance gains mainly benefit from the large amounts of accurately labeled training data. Although some annotated datasets have been built for model training and validation, they are usually small-scale [4, 6, 15, 16, 33, 37, 43, 44]. The two largest publicly reported datasets of lung only comprise 6,985 PET-CT pairs from 126 patients [37] and 3,905 pairs from 290 patients [33], respectively. What's worse, almost all of these datasets are private, severely hindering further advancements in this field. We summarize the recent PET-CT tumor segmentation datasets in Table 1. Although

\*Equal contribution. †Corresponding author.

these datasets have supported the development of some segmentation models [6], their limited scale and restricted accessibility significantly constrain the models’ applicability and robustness in broader clinical scenarios. Consequently, the foremost challenge for PET-CT lung tumor segmentation lies in the absence of a publicly available, large-scale dataset. This limitation arises from stringent privacy concerns associated with PET-CT data and the need for substantial clinical expertise to ensure accurate annotation.

To address these limitations, we first establish a comprehensive, large-scale PET-CT lung tumor segmentation dataset, named **PCLT20K**. PCLT20K consists of 21,930 pairs of PET-CT images collected from 605 unique patients. All PET-CT image pairs are manually annotated with the pixel-level masks of tumors. To ensure the quality and reliability of these annotations, a rigorous three-stage annotation process is conducted with all annotators being experienced clinicians. Representative examples of PET-CT images and their corresponding tumor labels are shown in Fig. 1, illustrating the significant challenges of PET-CT lung tumor segmentation: 1) the low contrast, low spatial resolution, and substantial noise can obscure the tumor appearances, especially their boundaries; 2) the shapes and structures of tumors are irregular and complex; 3) tumors vary considerably in size, location, and appearance, with some tumors being very small or situated in intricate anatomical regions; and 4) some lung tumors grow infiltratively, diffusely invading surrounding tissues. These factors are still challenging for most existing segmentation algorithms.

To achieve high-performance lung tumor segmentation from PET-CT images, we propose a cross-modal interactive perception network with Mamba (named **CIPA**) for lung tumor segmentation in PET-CT images. Specifically, a channel-wise rectification module (CRM) is proposed by implementing a channel state space block across multi-modal features, which enhances shared representation learning and helps to filter out modality-specific noise by emphasizing the correlated features between the two modalities. Additionally, considering that PET images often reveal potential lesion areas while CT images present detailed texture information of the lungs, we design a dynamic cross-modality interaction module (DCIM) to effectively integrate position and context information. DCIM employs the interactive region Mamba block and local Mamba block to enable effective cross-modality fusion by dynamically interacting PET and CT information, allowing for a more comprehensive and synergistic feature representation.

To summarize, our main contributions are as follows:

- We propose the first publicly available large-scale dataset for researching the task of lung tumor segmentation from PET-CT images, named **PCLT20K**, which contains 21,930 pairs of PET-CT images from 605 patients with high-quality pixel-level annotations.

Dataset	Location	#Cases	#Pairs	Public
HECKTOR [24]	Head & Neck	845	-	✓
MFCNet [32]	Pancreas	93	-	✗
AutoPET [7]	Whole Body	900	-	✓
STS [30]	Lung	51	2,409	✓
Co-Segmentation [43]	Lung	32	-	✗
DFCN [44]	Lung	60	-	✗
Co-Learning [15]	Lung	50	855	✗
FVM-PET [16]	Lung	36	-	✗
NSCLC [6]	Lung	50	876	✗
MoSNet [37]	Lung	126	6,985	✗
Dual-Modality [33]	Lung	290	3,905	✗
Joint-Level-Set [4]	Lung	20	-	✗
PCLT20K (Ours)	Lung	605	21,930	✓

Table 1. Summary of PET-CT tumor segmentation datasets.

- We propose a cross-modal interactive perception network with Mamba, named **CIPA**, which designs a channel-wise rectification module and a dynamic cross-modality interaction module to effectively interact with the correlation information between multi-modal data.
- Based on the PCLT20K dataset, we establish a comprehensive benchmark to facilitate future research. Extensive experiments validate that our CIPA can achieve state-of-the-art performance.

## 2. Related Work

### 2.1. Tumor Segmentation in PET-CT Images

The availability of high-quality datasets is crucial for developing and evaluating segmentation algorithms [26]. In recent years, some PET-CT tumor segmentation datasets have been proposed successively, covering organs such as the lung [1, 15], pancreas [32], head and neck [24], and so on [7]. However, as summarized in Table 1, almost all existing lung tumor segmentation datasets are small-scale and private [4, 6, 33, 37], which limits the further development of the field. Despite these limitations, several lesion segmentation algorithms based on these datasets have been proposed [25]. Convolutional neural networks (CNNs), especially fully convolutional networks (FCNs), have been widely applied to tumor segmentation from PET-CT images and remarkably boosted the segmentation performance [15, 23]. Among these methods, UNet and its variants demonstrated superior accuracy on PET-CT datasets thanks to their enhanced feature extraction capabilities [37]. More recently, some works have employed vision Transformers for PET-CT image segmentation [42]. These models leverage self-attention mechanisms to capture global contextual information, outperforming traditional FCN-based models, particularly in cases with complex tumor morphologies. Besides, there are efforts dedicated to using multi-modal fusion techniques to further improve PET-CT segmentation accuracy [32]. Nevertheless, the small scale and privacy of current PET-CT datasets limit the reproducibility of these

algorithms and impede the advancement of this field.

## 2.2. Mamba in Medical Image Segmentation

The emergence of Mamba [8] has brought new possibilities to the field of deep learning-based medical image segmentation. Traditional CNN-based models are inherently limited in their ability to capture long-range dependencies. Transformers ensure long-range modeling through self-attention mechanisms, while they do so at the cost of quadratic time complexity. Mamba addresses this issue by introducing a Selective State Space Model (SSM) that achieves long-range interaction with linear time complexity, which achieves the current optimal performance [8]. U-shaped Mamba networks have become the mainstream architectures, such as U-Mamba [22], SegMamba [38], and Swin-UMamba [19]. Huang *et al.* [11] introduced a novel local scanning strategy that divides images into distinct windows, effectively capturing local dependencies while maintaining a global perspective. Mamba-based models have also been extended to lightweight [18], semi-supervised [21], and weakly-supervised [36] medical image segmentation field. Despite their advancements, Mamba struggles to effectively explore regional correlations and complementary information between multi-modal images due to its 2D selective scanning mechanism, thereby restricting the performance of tumor segmentation in PET-CT images.

## 3. PCLT20K Dataset

### 3.1. Data Collection

The PET-CT images of PCLT20K are collected from the department of molecular imaging and medicine at a top-tier hospital. The patients' examination dates range from June 2016 to April 2020. Besides, the PET-CT images are acquired using a GE Discovery Elite PET/CT scanner (GE Medical Systems). Prior to PET/CT imaging, patients fast for approximately six hours with a serum glucose level maintained below 11.1 mmol/L. Images are acquired 50 to 60 minutes after the injection of 4.2 MBq/kg  $^{18}\text{F}$ -FDG. A spiral CT scan (80 mAs, 120 kVp, and 5-mm slice thickness) is performed for precise anatomical localization and attenuation correction, and a PET emission scan in 3D mode is followed from the distal femur to the top of the skull. The voxel size of CT images is  $0.98 \times 0.98 \times 2.8$  mm. PET images are reconstructed using an ordered-subset expectation maximization (OSEM) iterative algorithm to achieve a final pixel size of  $3.6 \times 3.6 \times 3.3$  mm. It is worth noting that during data collection at the hospital, all private information is removed, including hospital names, doctor names, patient names, and other identifying details. Additionally, images with foreign object interference and motion artifacts are excluded to ensure the quality of the collected images.

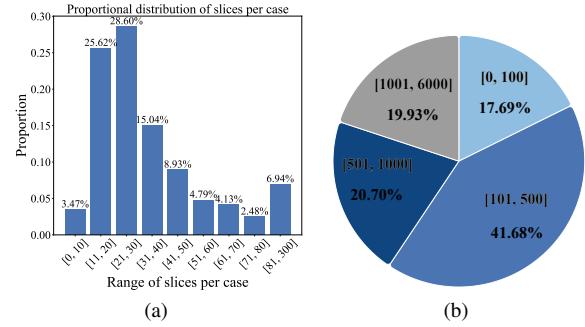


Figure 2. (a) Proportional statistics of the number of slices per case. (b) Statistics of the pixel counts in tumor regions.

### 3.2. Data Annotation

To guarantee the accuracy and reliability of tumor region annotations, we employ a three-stage annotation process conducted by experienced physicians. The first annotation stage occurs when doctors diagnose the patients in the hospital. During this stage, doctors analyze CT and PET images to generate examination reports that include the approximate location, appearance, and type of the tumor. Referring to these medical reports, physicians in the second stage conduct a detailed annotation for each case on a slice-by-slice basis, marking all the pixels in the tumor regions. In the third annotation stage, another doctor reviews and refines the annotations to form the final pixel-level annotations of the tumor regions. If there are any discrepancies between the annotations from the second and third stages, the doctors will discuss and re-annotate the tumor regions together. Note that in the second and third stages, doctors select the tumor parenchyma area as the lesion area for annotation. Due to the uneven growth of the tumor and its invasion into surrounding tissues, the annotated lesion area may be discontinuous.

We finally obtain 21,930 2D slice pairs of PET-CT images from 605 unique patients with high-quality pixel-level annotations. We randomly split them into a training set and a testing set at an 8:2 ratio, naming them PCLT20K-TR (17,416 pairs) and PCLT20K-TE (4514 pairs), respectively. The data splitting is conducted at the patient level.

In the PCLT20K dataset, PET-CT images undergo pre-processing before being input into the network. Based on doctors' experience, the Hounsfield Unit (HU) values of the CT images are clipped to the range [-1200, -200] and then normalized to [0, 255]. The PET images are normalized by converting them to Standard Uptake Value (SUV) [29], which are then scaled to [0, 255]. All PET-CT slices in PCLT20K are resized to the same resolution of  $512 \times 512$  to ensure they share the same coordinate space, which is a standard procedure for analyzing PET-CT data [14, 37].

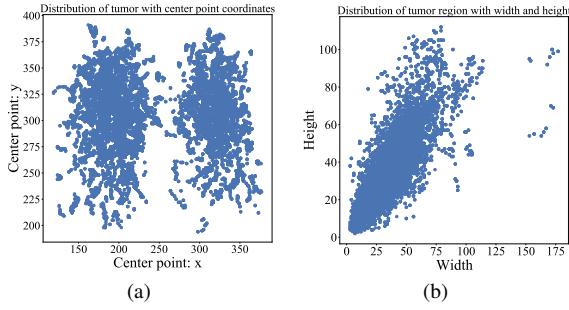


Figure 3. Distribution of tumor center points and sizes. (a) Distribution of tumor in terms of center point coordinates. (b) Distribution of tumor in terms of width and height.

### 3.3. Data Statistics

We conduct statistical analyses to present the distribution of tumor sizes and locations in the PCLT20K dataset. The proportional statistics of the number of slices per case are shown in Fig. 2(a), 72.73% of the tumors have fewer than 40 slices, while only 13.55% of the tumors exhibit more than 60 slices. Notably, PCLT20K demonstrates considerable variability, with the minimum number of tumor-containing slices being 4 and the maximum reaching 278. The size distribution of tumors, as visualized in Fig. 2(b), reveals that a majority of tumors occupy relatively limited spatial regions within the slices. In detail, 59.37% of tumors have an area smaller than 500 pixels. The number of tumors whose sizes are over 1,000 pixels only accounts for 19.93%, reflecting the distribution’s skew toward smaller lesions within the PCLT20K dataset. Among all lesions, the smallest tumor only has 11 pixels, while the largest occupies an area of 5,830 pixels. Furthermore, we plot the locations of center points of tumors in Fig. 3(a). Obviously, all tumors are randomly distributed on the lung regions without bias, indicating the universal property of PCLT20K. Besides, we show the distribution of tumor sizes using height and width in Fig. 3(b). This representation highlights that the tumors within the PCLT20K are generally small, reaffirming the dataset’s value in representing typical clinical scenarios where tumors may be of minimal size.

## 4. Proposed Method

### 4.1. Preliminaries

The recent State Space Models (SSMs) [9, 10, 28] represent a class of sequence-to-sequence modeling systems characterized by constant dynamics over time, a property also known as linear time-invariant (LTI). It maps a 1D sequential input  $x(t) \in \mathbb{R} \rightarrow y(t) \in \mathbb{R}$  through an implicit latent state  $h(t) \in \mathbb{R}^N$ , which effectively capture the inherent dynamics of systems. An SSM is defined by four parameters

$(\Delta, A, B, C)$  with the following operations:

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t, y_t = Ch_t, \quad (1)$$

where  $(\bar{A}, \bar{B})$  represent the zero-order hold (ZOH) [10] discretization of  $(A, B)$  using the transformations  $\bar{A} = \exp(\Delta A)$  and  $\bar{B} = (\Delta A)^{-1}(\exp(A) - I) \cdot \Delta B$ . This discretization facilitates efficient parallelizable training through a global convolution approach:

$$y = x * \bar{K}, \bar{K} = \left( C\bar{B}, C\bar{A}\bar{B}, \dots, C\bar{A}^{L-1}\bar{B} \right), \quad (2)$$

where  $*$  represents the convolution operation,  $\bar{K} \in \mathbb{R}^L$  is the SSM kernel. Although discretization enhances computational efficiency, the parameters  $(\Delta, \bar{A}, \bar{B}, C)$  in the SSM are data-independent and time-invariant, limiting the expressiveness of the hidden state to compress seen information. Selective SSM (or Mamba) [8] introduces data-dependent parameters  $(B, C, \Delta)$  that effectively select relevant information in  $x_t$ , which vary with the input through liner projections:  $B_t = \text{Linear}_B(x_t)$ ,  $C_t = \text{Linear}_C(x_t)$ ,  $\Delta_t = \text{SoftPlus}(\text{Linear}_\Delta(x_t))$ .

Through hardware-aware optimizations, the selective SSM achieves linear computation and memory complexity with respect to sequence length, effectively compressing essential context across the entire input sequence. In visual tasks, VMamba introduced the 2D Selective Scan (SS2D) [45]. This method preserves the integrity of 2D image structures by scanning four directed feature sequences, each independently processed within an S6 (*i.e.*, Selective Scan Space State Sequential Model) block before being integrated to create a cohesive 2D feature map.

### 4.2. Network Structure

As illustrated in Fig. 4, our proposed method comprises two parallel branches for extracting features from PET- and CT-modal inputs, a channel-wise feature rectification module, and a dynamic cross-modality interaction module, which form an architecture entirely composed of state space models. During the encoding phase, each branch sequentially processes the input through four Visual State Space (VSS) blocks with downsampling operations to extract multi-level image features. The two encoder branches share weights to optimize computational efficiency. At each stage of the dual-branch architecture, a channel-wise rectification module is introduced between both branches to refine features based on two modalities. Additionally, cross-modal feature fusion is facilitated at each stage using rectified features from distinct sources through our dynamic cross-modality interaction module. In the decoding phase, fused features are further processed using Channel-Aware Visual State Space (CVSS) blocks proposed in [31] with upsampling operations. Finally, the resulting features are fed into a classifier to generate lesion segmentation.

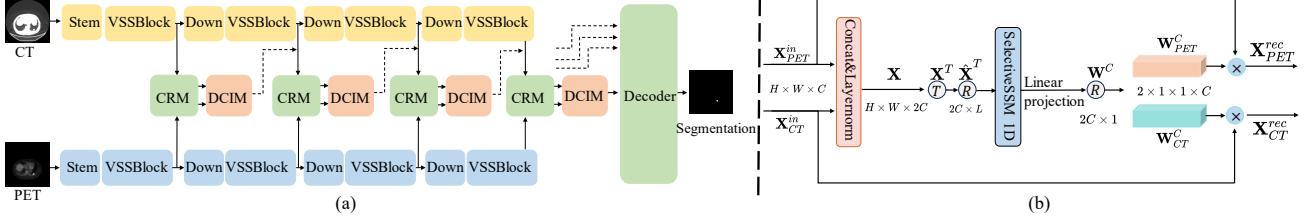


Figure 4. (a) Overall architecture of our proposed cross-modal interactive perception network with Mamba (CIPA) for lung tumor segmentation in PET-CT images. CIPA consists of: (1) a channel-wise rectification module (CRM) to learn shared representations; (2) a dynamic cross-modality interaction module (DCIM) to integrate position and context information effectively. (b) Illustration of CRM.

### 4.3. Channel-wise Rectification Module

One limitation of the existing Mamba-based network is that, although the global context is captured by scanning image feature maps, the channel information, especially across different modalities, is usually overlooked. To address this, we propose a channel-wise rectification module (CRM) that integrates multi-modal features to learn shared representations and enhance correlated features between the two modalities. As shown in Fig. 4(b), given two input features  $\mathbf{X}_{PET}^{in}$  and  $\mathbf{X}_{CT}^{in}$  from the PET and CT modalities, we first concatenate them along the channel dimension and apply layer normalization to obtain the feature  $\mathbf{X} \in \mathbb{R}^{H \times W \times 2C}$ . We transpose  $\mathbf{X}$  to  $\mathbf{X}^T \in \mathbb{R}^{2C \times H \times W}$  and flatten it to  $\hat{\mathbf{X}}^T \in \mathbb{R}^{2C \times L}$ , where  $L = H \times W$ . This can be interpreted as using flattened feature pixels as the channel representation. We then apply the 1D selective SSM by:

$$\hat{\mathbf{X}}^T = \text{SelectiveSSM}(\hat{\mathbf{X}}^T). \quad (3)$$

This operation effectively mixes and memorizes channel-wise correlation features by scanning the channel mapping representations of the two modalities. Next, we employ linear projection  $\mathcal{F}$ , followed by a sigmoid function  $\sigma$  to obtain weights  $\mathbf{W}^C \in \mathbb{R}^{2C}$  for each channel within both modalities. These weights reflect the importance of each channel for the two modalities, which are split into  $\mathbf{W}_{PET}^C$  and  $\mathbf{W}_{CT}^C$  correspond to the PET and CT channels:

$$\mathbf{W}_{PET}^C, \mathbf{W}_{CT}^C = \mathcal{F}_{\text{split}}\left(\sigma\left(\mathcal{F}(\hat{\mathbf{X}}^T)\right)\right). \quad (4)$$

Finally, we apply channel-wise rectification by scaling the original input features with the computed weights:

$$\mathbf{X}_{PET}^{rec} = \mathbf{W}_{PET}^C \circledast \mathbf{X}_{PET}^{in}, \mathbf{X}_{CT}^{rec} = \mathbf{W}_{CT}^C \circledast \mathbf{X}_{CT}^{in}, \quad (5)$$

where  $\circledast$  denotes channel-wise multiplication.

By modeling dependencies among channels from both PET and CT modalities, CRM effectively integrates multi-modal information to enhance shared representation learning. This rectification helps to filter out modality-specific noise by emphasizing the most informative and correlated features between the two modalities.

### 4.4. Dynamic Cross-Modality Interaction Module

For lung PET-CT images, PET scans typically exhibit SUV values and appear brighter visually due to increased metabolic activity. However, they suffer from relatively low resolution and often present blurred tumor boundaries. Conversely, CT scans offer higher spatial resolution and provide detailed structural information about the tumor, while the contrast between the lesions and background regions is lower. By focusing on metabolic and morphological tissue characteristics, PET and CT images can provide complementary and synergistic information for accurate tumor region identification. However, the current VMamba [45] cannot effectively combine this complementary regional information due to its 2D selective scan mechanism. To address this limitation, we propose a dynamic cross-modality interaction module (DCIM) to effectively integrate position and context information between PET and CT images, as shown in Fig. 5.

Given two rectified features  $\mathbf{X}_{PET}^{rec}$  and  $\mathbf{X}_{CT}^{rec}$ , we first construct a convolutional stem composed of stacked  $3 \times 3$  convolutions to process them and produce the PET feature  $\mathbf{X}_{PET} \in \mathbb{R}^{\frac{H}{r} \times \frac{W}{r} \times D}$  and CT feature  $\mathbf{X}_{CT} \in \mathbb{R}^{H \times W \times C}$ .  $r$  is the resolution of the regional patch in the PET feature,  $D$  and  $C$  are the feature dimensions for PET and CT, respectively. We can obtain PET regional tokens:

$$\mathcal{X}_{PET} = [X_{PET}^1, X_{PET}^2, \dots, X_{PET}^n] \in \mathbb{R}^{n \times D}, \quad (6)$$

where  $n = \frac{H}{r} \times \frac{W}{r}$  and each  $X_{PET}^i \in \mathbb{R}^D$  corresponds to the PET feature in the  $i$ -th region.

To facilitate dynamic interactions, we first split the CT feature into  $n$  regional patches:  $\mathcal{X}_{CT} = [X_{CT}^1, X_{CT}^2, \dots, X_{CT}^n] \in \mathbb{R}^{n \times r \times r \times C}$ . Each regional patch of the CT feature is further divided into  $m$  local patches, i.e., a regional patch is composed of a sequence of local tokens:

$$X_{CT}^i = [x_{CT}^{i,1}, x_{CT}^{i,2}, \dots, x_{CT}^{i,m}], \quad (7)$$

where  $x_{CT}^{i,j} \in \mathbb{R}^C$  is the  $j$ -th local token of the  $i$ -th regional patches, and  $j = 1, 2, \dots, m$ .

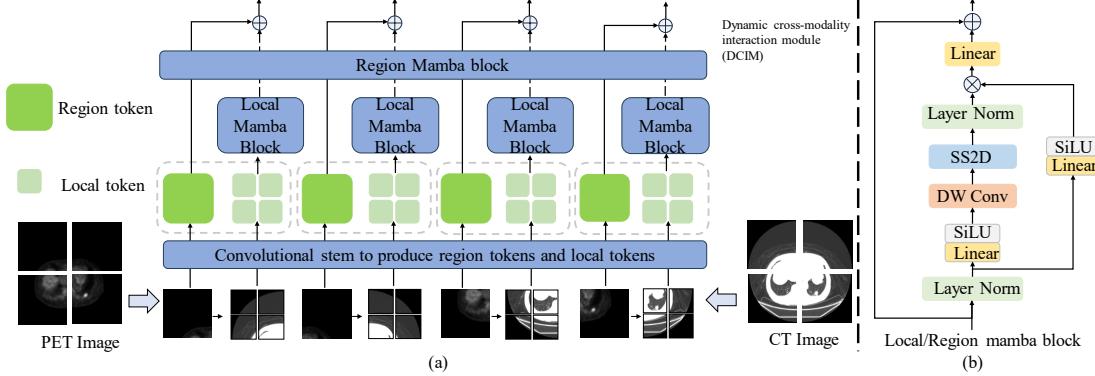


Figure 5. (a) Illustration of dynamic cross-modality interaction module (DCIM), which mainly includes a convolutional stem, local Mamba block, and region Mamba block. The black dashed arrows indicate bypassing region Mamba blocks. (b) The structure of Mamba block.

The core of our DCIM is the dynamic interaction between the position information in PET regional tokens and the structural information in CT local tokens. Specifically, a region Mamba block is designed to process the regional position information inherent in PET features, while a local Mamba block is proposed to model the intricate dependencies among fine-grained local CT features within each regional patch. The region Mamba block first involves all regional tokens to learn the global position information efficiently as follows:

$$\hat{\mathcal{X}}_{PET} = \mathcal{X}_{PET} + \text{Mamba}_{region} (\text{LN} (\mathcal{X}_{PET})) , \quad (8)$$

where LN denotes layer normalization. Then the local Mamba block models the local dependencies among the local tokens within each regional patch:

$$\hat{\mathcal{X}}_{CT}^i = \mathcal{X}_{CT}^i + \text{Mamba}_{local} (\text{LN} (\mathcal{X}_{CT}^i)) . \quad (9)$$

The two kinds of Mamba blocks enhance the global and local information of the tokens, respectively. To enable the high-contrast PET images to guide the high-resolution CT images, we integrate the PET regional tokens with the CT local tokens within the same region. Specifically, we add the PET regional token to each of the CT local tokens in the corresponding region:

$$\begin{aligned} \hat{\mathcal{X}}_{PET} &= [\hat{\mathcal{X}}_{PET}^1, \hat{\mathcal{X}}_{PET}^2, \dots, \hat{\mathcal{X}}_{PET}^n] , \\ \hat{\mathcal{X}}_{CT}^i &= [\hat{x}_{CT}^{i,1}, \hat{x}_{CT}^{i,2}, \dots, \hat{x}_{CT}^{i,m}] , \\ \hat{\mathcal{X}}_{fuse}^i &= [\hat{x}_{CT}^{i,1} + \hat{\mathcal{X}}_{PET}^i, \hat{x}_{CT}^{i,2} + \hat{\mathcal{X}}_{PET}^i, \dots, \hat{x}_{CT}^{i,m} + \hat{\mathcal{X}}_{PET}^i] . \end{aligned} \quad (10)$$

The region Mamba block focuses on extracting high-level information from regional tokens and serves as a bridge to transfer position information of lesion areas between local tokens. In this case, each local token can gather global positional information while paying close attention to its immediate neighbors. The DCIM thus enables effective cross-modality fusion by dynamically interacting PET

and CT information. It facilitates the model to leverage the complementary strengths of PET (e.g., lesion localization due to high contrast) and CT (e.g., structural details due to high resolution), allowing for a more comprehensive and synergistic feature representation that transcends traditional multiscale fusion methods.

## 5. Experiments

### 5.1. Implementation Details

In this work, we implement our CIPA with the representative PyTorch platform on a PC with 4 RTX 4090 GPUs (24 GB). We adopt the following data enhancement: flipping horizontally or vertically, and random cropping in the size range of [0.7, 0.9]. We use the AdamW optimizer [20] with an initial learning rate of  $6e^{-5}$  and the cosine annealing strategy. The proposed method is trained for 50 epochs on our PCLT20K dataset with a batch size of 16. We utilize the pre-trained model on ImageNet-1K [27] provided by VMamba [45] for our encoder.

We adopt four widely used metrics to evaluate the segment performance of all models, including the Intersection over Union (IoU), accuracy, F1-score, and HD95 (95% Hausdorff Distance).

### 5.2. Comparison with Other Methods

#### 5.2.1. Quantitative Comparison

As listed in Table 2, we compare CIPA with 12 state-of-the-art segmentation models. The comparisons of the number of parameters and floating point operations (FLOPs) are also presented. To ensure fairness, all methods are trained on the PCLT20K-TR dataset and tested on the PCLT20K-TE dataset. Notably, CIPA achieves the best results across all four evaluation metrics on the PCLT20K dataset. The model complexity of our CIPA is competitive compared to other methods. Besides, we also conduct experiments using different backbone. CIPA with Mamba as the backbone

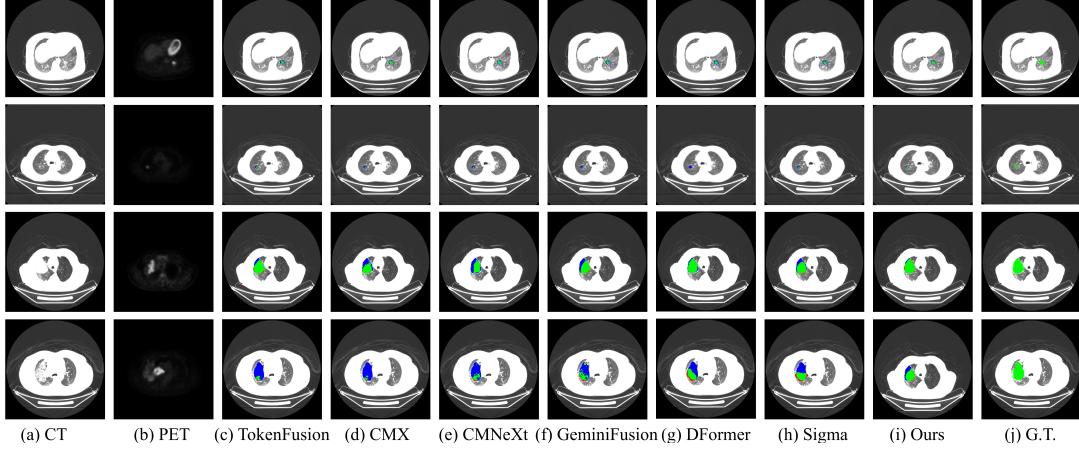


Figure 6. Qualitative comparison of our CIPA with other segmentation methods on the PCLT20K dataset. Green: true positive, red: false positive, blue: false negative.

Method	IoU ↑ (%)	F1 ↑ (%)	Acc ↑ (%)	HD95 ↓	Flops (G)	Params (M)
TokenFusion [34]	61.21	75.94	84.52	25.52	92.22	45.91
CEN [35]	61.28	75.99	85.16	30.33	524.75	118.11
CMX [40]	61.64	76.27	83.99	24.68	65.43	66.56
CMNeXt [41]	63.12	77.39	85.77	24.95	119.49	58.68
AsymFormer [5]	50.03	66.69	76.52	30.40	34.20	33.05
DFormer [39]	61.09	75.85	83.83	29.57	14.83	39.01
Sigma [31]	63.26	77.50	85.37	22.56	76.18	48.28
GeminiFusion [13]	62.49	76.91	85.09	20.70	129.92	75.54
nnU-Net [12]	61.44	76.11	85.61	22.94	—	—
Swin-Unet [2]	36.41	53.38	75.48	124.68	62.25	54.29
TransUNet [3]	51.72	68.18	78.01	102.97	258.58	186.46
nnSAM [17]	62.93	77.25	85.24	22.49	—	—
CIPA (ResNet-50)	53.54	70.63	81.76	38.97	77.77	58.04
CIPA (Swin-T)	62.82	77.17	86.84	20.24	92.48	121.22
CIPA (ours)	<b>63.81</b>	<b>77.91</b>	<b>89.01</b>	<b>17.74</b>	76.18	54.57

Table 2. Comparison between our proposed CIPA and other methods on the testing set of our PCLT20K.

Method	IoU (%)	F1 (%)	Acc (%)
TokenFusion [34]	42.26	59.41	75.52
CEN [35]	42.35	59.50	74.97
CMX [40]	51.47	67.96	79.66
CMNeXt [41]	55.75	71.59	81.13
AsymFormer [5]	53.07	69.34	80.75
DFormer [39]	55.99	71.78	82.10
Sigma [31]	58.14	73.53	85.34
GeminiFusion [13]	59.02	74.23	84.80
CIPA (ours)	<b>60.33</b>	<b>75.26</b>	<b>86.03</b>

Table 3. Comparison between our proposed CIPA and other methods on the testing set of STS dataset [30].

No.	CRM	DCIM	IoU	F1	Acc	Params (M)
1			62.12	76.64	84.95	33.92
2	✓		62.70	77.08	86.09	36.32
3		✓	63.48	77.66	86.77	52.17
4	✓	✓	<b>63.81</b>	<b>77.91</b>	<b>89.01</b>	54.57

Table 4. Ablation for the proposed CRM and DCIM (%).

achieves superior performance with fewer parameters compared to CIPA with ResNet-50 and with Swin-T. These superior results validate the effectiveness of CIPA for lung tumor segmentation in PET-CT images.

### 5.2.2. Qualitative Comparison

As illustrated in Fig. 6, we present the qualitative comparison of CIPA and segmentation methods in several examples. Overall, our method can accurately locate and segment complete lung tumors with well-preserved edge details. Additionally, CIPA produces fewer false-positive lesions, which can effectively reduce the burden on doctors during further screening. We provide additional quantitative comparison examples in the supplementary material.

### 5.2.3. Qualitative Comparison on STS dataset

We also conduct comparative experiments on the publicly available PET-CT tumor segmentation dataset, STS [30].

The STS dataset consists of 2,409 pairs of PET-CT images collected from 51 patients, which are split into a training set with 1,941 pairs and a test set with 468 pairs of images. Besides, all compared methods use their default optimal experiment settings. As listed in Table 3, our CIPA also achieves the best results on the STS dataset. The state-of-the-art results on our PCLT20K and STS datasets further validate the performance of our proposed CIPA.

## 5.3. Ablation Study

### 5.3.1. Effectiveness of Key Modules

As shown in Table 4, we conduct ablation experiments for the proposed CRM and DCIM in CIPA on PCLT20K dataset. By sequentially adding the CRM and DCIM modules to the baseline model based on shared Mamba, the

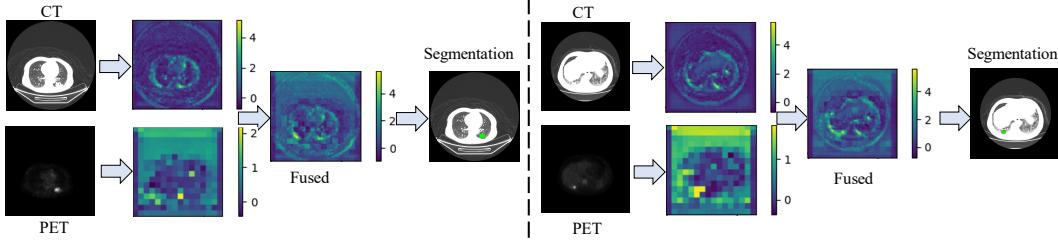


Figure 7. Illustration of the process of DCIM with two examples. Each example illustrates the feature maps of the PET image processed by the region Mamba block, the feature maps of the CT image processed by the local Mamba block, and the integrated features resulting from the fusion of these two sets of information.

No.	Locals in region	IoU	F1	Acc	Params (M)
1	$2 \times 2$	62.87	77.20	87.59	54.92
2	$4 \times 4$	<b>63.81</b>	<b>77.91</b>	89.01	54.57
3	$8 \times 8$	63.07	77.35	<b>89.16</b>	54.82

Table 5. Ablation for the number of local patches in the corresponding region patch in DCIM (%).

No.	Operation	IoU	F1	Acc	Params (M)
1	$Region_{PET} \& Local_{PET}$	42.06	59.22	80.05	54.57
2	$Region_{CT} \& Local_{CT}$	61.37	76.06	85.28	54.57
3	$Region_{CT} \& Local_{PET}$	60.63	75.49	86.54	54.57
4	$Local_{CT}$	62.65	77.04	87.68	48.48
5	$Region_{PET}$	62.26	76.76	86.46	48.43
6	DCIM	<b>63.81</b>	<b>77.91</b>	<b>89.01</b>	54.57

Table 6. Ablation for the various operations in DCIM (%).

performance across all three evaluation metrics improves. Adding both modules simultaneously achieved the best performance. These results validate the effectiveness of these two modules for tumor segmentation in PET-CT images.

### 5.3.2. Effectiveness of Configurations in DCIM

As listed in Table 5, we report the results for the number of local patches in the corresponding region patch in the DCIM module. It can be observed that when a PET image's regional patch contains  $4 \times 4$  local patches from the CT image, both the model's complexity and segmentation performance show advantages. Besides, we also conduct ablation studies for the various operations of DCIM in Table 6. Compared to the five configurations of applying both region Mamba block and local Mamba block to either PET or CT image (No.1 and No.2), applying a region Mamba block to the CT image while using a local Mamba block for the PET image (No.3), applying a local Mamba block only to the CT image and directly adding the PET images (No.4), applying a region Mamba block only to the PET image and directly adding the CT images (No.5), the default configuration of DCIM (applying region Mamba block to the PET image while using local Mamba block for the CT image) achieves the best performance.

### 5.4. Visualization of Intermediate Feature

To facilitate a deeper comprehension of how our proposed CIPA leverages the DCIM module to enhance the structural features within tumor areas, we visualize the features of PET image, CT image, and the features after their interaction and fusion within the DCIM, as depicted in Fig. 7. The results reveal that these integrated features exhibit a more pronounced delineation of lung structures and lesion areas compared to individual modalities. This improved structural clarity is advantageous for the accurate segmentation of lung tumors, thereby facilitating a more reliable basis for downstream diagnostic and therapeutic in lung oncology.

## 6. Conclusion

This paper presents a comprehensive study on lung tumor segmentation from PET-CT images. We first provide a large-scale and publicly available PET-CT lung tumor segmentation dataset, PCLT20K, which contains 21,930 pairs of PET-CT images. Furthermore, we propose a cross-modal interactive perception network with Mamba (CIPA), which explores the innovative application of State Space Models in PET-CT tumor segmentation for the first time. A channel-wise rectification module (CRM) and a dynamic cross-modality interaction module (DCIM) are designed to effectively interact and leverage the correlation information between the multi-modal data. Extensive experiments demonstrate that CIPA achieves state-of-the-art performance on our PCLT20K and the publicly available STS datasets. This study not only establishes a benchmark for lung tumor segmentation but also provides a solid foundation for the development of advanced multi-modal models in medical image analysis. In the future, we plan to expand our PCLT20K dataset to support additional tasks such as PET-CT lung tumor detection and multi-modal image fusion.

## 7. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62303173, Grant 62403256, and Grant 62027810.

## References

- [1] Pam Angelus and J Kirby. A large-scale ct and pet/ct dataset for lung cancer diagnosis (lung-pet-ct-dx). *Cancer Imaging Archive*, 2020. 2
- [2] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Eur. Conf. Comput. Vis.*, pages 205–218. Springer, 2022. 7
- [3] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint:2102.04306*, 2021. 7
- [4] Zhe Chen, Nan Qiu, Hui Feng, and Dongfang Dai. Pet-ct image co-segmentation of lung tumor using joint level set model. *Comput. Electr. Eng.*, 105:108545, 2023. 1, 2
- [5] Siqi Du, Weixi Wang, Renzhong Guo, Ruisheng Wang, and Shengjun Tang. Asymformer: Asymmetrical cross-modal representation learning for mobile platform real-time rgb-d semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 7608–7615, 2024. 7
- [6] Xiaohang Fu, Lei Bi, Ashnil Kumar, Michael Fulham, and Jinman Kim. Multimodal spatial attention module for targeting multimodal pet-ct lung tumor segmentation. *IEEE J. Biomed. Health Inform.*, 25(9):3507–3516, 2021. 1, 2
- [7] Sergios Gatidis, Marcel Fröh, Matthias Fabritius, Sijing Gu, Konstantin Nikolaou, Christian La Fougère, Jin Ye, Junjun He, Yige Peng, Lei Bi, et al. The autopet challenge: towards fully automated lesion segmentation in oncologic pet/ct imaging. *Research Square*, 2023. 2
- [8] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 3, 4
- [9] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Adv. Neural Inform. Process. Syst.*, 34:572–585, 2021. 4
- [10] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *Int. Conf. Learn. Represent.*, 2022. 4
- [11] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*, 2024. 3
- [12] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 7
- [13] Ding Jia, Jianyuan Guo, Kai Han, Han Wu, Chao Zhang, Chang Xu, and Xinghao Chen. Geminifusion: Efficient pixel-wise multimodal fusion for vision transformer. In *Int. Conf. Mach. Learn.*, 2024. 7
- [14] Ashnil Kumar, Michael Fulham, Dagan Feng, and Jinman Kim. Co-learning feature fusion maps from pet-ct images of lung cancer. *IEEE Trans. Med. Image.*, 39(1):204–217, 2019. 3
- [15] Ashnil Kumar, Michael Fulham, Dagan Feng, and Jinman Kim. Co-learning feature fusion maps from pet-ct images of lung cancer. *IEEE Trans. Med. Image.*, 39(1):204–217, 2020. 1, 2
- [16] Laquan Li, Xiangming Zhao, Wei Lu, and Shan Tan. Deep learning for variational multimodality tumor segmentation in pet/ct. *Neurocomputing*, 392:277–295, 2020. 1, 2
- [17] Yunxiang Li, Bowen Jing, Xiang Feng, Zihan Li, Yongbo He, Jing Wang, and You Zhang. nnSam: Plug-and-play segment anything model improves nnunet performance. *arXiv preprint arXiv:2309.16967*, 2023. 7
- [18] Weibin Liao, Yinghao Zhu, Xinyuan Wang, Cehngwei Pan, Yasha Wang, and Liantao Ma. Lightm-unet: Mamba assists in lightweight unet for medical image segmentation. *arXiv preprint arXiv:2403.05246*, 2024. 3
- [19] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Yizhou Yu, Yong Liang, Guangming Shi, Shaotong Zhang, Hairong Zheng, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. *arXiv preprint arXiv:2402.03302*, 2024. 3
- [20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [21] Chao Ma and Ziyang Wang. Semi-mamba-unet: Pixel-level contrastive and pixel-level cross-supervised visual mamba-based unet for semi-supervised medical image segmentation. *arXiv e-prints*, pages arXiv–2402, 2024. 3
- [22] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. 3
- [23] Jie Mei, Ming-Ming Cheng, Gang Xu, Lan-Ruo Wan, and Huan Zhang. Sanet: A slice-aware network for pulmonary nodule detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(8):4374–4387, 2021. 2
- [24] Valentin Oreiller, Vincent Andreczyk, Mario Jreige, Sarah Bougħdad, Hesham Elhalawani, Joel Castelli, Martin Vallières, Simeng Zhu, Juanying Xie, Ying Peng, et al. Head and neck tumor segmentation in pet/ct: the hecktor challenge. *Med. Image Anal.*, 77:102336, 2022. 2
- [25] Yu Qiu and Jing Xu. Delving into universal lesion segmentation: Method, dataset, and benchmark. In *Eur. Conf. Comput. Vis.*, pages 485–503. Springer, 2022. 2
- [26] Yu Qiu, Yun Liu, Shijie Li, and Jing Xu. Miniseg: an extremely minimum network based on lightweight multiscale learning for efficient covid-19 segmentation. *IEEE Trans. Neur. Net. Learn. Syst.*, 2022. 2
- [27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115: 211–252, 2015. 6
- [28] Jimmy TH Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *Int. Conf. Learn. Represent.*, 2023. 4
- [29] Joseph A Thie. Understanding the standardized uptake value, its methods, and implications for usage. *J. Nucl. Med.*, 45(9): 1431–1434, 2004. 3

- [30] Martin Vallières, Carolyn R Freeman, Sonia R Skamene, and Issam El Naqa. A radiomics model from joint fdg-pet and mri texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys. Med. Biol.*, 60(14):5471, 2015. [2](#), [7](#)
- [31] Zifu Wan, Yuhao Wang, Silong Yong, Pingping Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Sigma: Siamese mamba network for multi-modal semantic segmentation. *arXiv preprint arXiv:2404.04256*, 2024. [4](#), [7](#)
- [32] Fei Wang, Chao Cheng, Weiwei Cao, Zhongyi Wu, Heng Wang, Wenting Wei, Zhuangzhi Yan, and Zhaobang Liu. Mfcnet: A multi-modal fusion and calibration networks for 3d pancreas tumor segmentation on pet-ct images. *Comput. Biol. Med.*, 155:106657, 2023. [2](#)
- [33] Siqiu Wang, Rebecca Mahon, Elisabeth Weiss, Nuzhat Jan, Ross James Taylor, Philip Reed McDonagh, Bridget Quinn, and Lulin Yuan. Automated lung cancer segmentation using a pet and ct dual-modality deep learning neural network. *Int. J. Radiat. Oncol. Biol. Phys.*, 115(2):529–539, 2023. [1](#), [2](#)
- [34] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12186–12195, 2022. [7](#)
- [35] Yikai Wang, Fuchun Sun, Wenbing Huang, Fengxiang He, and Dacheng Tao. Channel exchanging networks for multi-modal and multitask dense image prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5):5481–5496, 2022. [7](#)
- [36] Ziyang Wang and Chao Ma. Weak-mamba-unet: Visual mamba makes cnn and vit work better for scribble-based medical image segmentation. *arXiv preprint arXiv:2402.10887*, 2024. [3](#)
- [37] Dehui Xiang, Bin Zhang, Yuxuan Lu, and Shengming Deng. Modality-specific segmentation network for lung tumor segmentation in pet-ct images. *IEEE J. Biomed. Health Inform.*, 27(3):1237–1248, 2022. [1](#), [2](#), [3](#)
- [38] Zhaochu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv preprint arXiv:2401.13560*, 2024. [3](#)
- [39] Bowen Yin, Xuying Zhang, Zhong-Yu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. In *Int. Conf. Learn. Represent.*, 2024. [7](#)
- [40] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Trans. Intell. Transport. Syst.*, 2023. [7](#)
- [41] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1136–1147, 2023. [7](#)
- [42] Wenjie Zhao, Zhenxing Huang, Si Tang, Wenbo Li, Yunlong Gao, Yingying Hu, Wei Fan, Chuanli Cheng, Yongfeng Yang, Hairong Zheng, et al. Mmca-net: A multimodal cross attention transformer network for nasopharyngeal carcinoma tumor segmentation based on a total-body pet/ct system. *IEEE J. Biomed. Health Inform.*, 2024. [2](#)
- [43] Zisha Zhong, Yusung Kim, Leixin Zhou, Kristin Plichta, Bryan Allen, John Buatti, and Xiaodong Wu. 3d fully convolutional networks for co-segmentation of tumors on pet-ct images. In *IEEE ISBI*, pages 228–231. IEEE, 2018. [1](#), [2](#)
- [44] Zisha Zhong, Yusung Kim, Kristin Plichta, Bryan G Allen, Leixin Zhou, John Buatti, and Xiaodong Wu. Simultaneous cosegmentation of tumors in pet-ct images using deep fully convolutional networks. *Medical physics*, 46(2):619–633, 2019. [1](#), [2](#)
- [45] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Int. Conf. Mach. Learn.*, 2024. [4](#), [5](#), [6](#)