Review Article

# Multi-modal tumor segmentation methods based on deep learning: a narrative review

**Hengzhi Xue[1], Yudong Yao[2,3], Yueyang Teng[1,4]**

[1]College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China; [2]Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ, USA; [3]Research Institute for Medical and Biological Engineering, Ningbo University, Ningbo, China; [4]Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Shenyang, China

*Contributions:* (I) Conception and design: Y Yao, Y Teng; (II) Administrative support: Y Yao, Y Teng; (III) Provision of study materials or patients: Y Yao, Y Teng; (IV) Collection and assembly of data: H Xue; (V) Data analysis and interpretation: H Xue; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

*Correspondence to:* Yueyang Teng, PhD. Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Shenyang 110169, China; College of Medicine and Biological Information Engineering, Northeastern University, 195 Innovation Road, Hunnan New District, Shenyang 110016, China. Email: tengyy@bmie.neu.edu.cn.

**Background and Objective:** Automatic tumor segmentation is a critical component in clinical diagnosis and treatment. Although single-modal imaging provides useful information, multi-modal imaging provides a more comprehensive understanding of the tumor. Multi-modal tumor segmentation has been an essential topic in medical image processing. With the remarkable performance of deep learning (DL) methods in medical image analysis, multi-modal tumor segmentation based on DL has attracted significant attention. This study aimed to provide an overview of recent DL-based multi-modal tumor segmentation methods.

**Methods:** In in the PubMed and Google Scholar databases, the keywords "multi-modal", "deep learning", and "tumor segmentation" were used to systematically search English articles in the past 5 years. The date range was from 1 January 2018 to 1 June 2023. A total of 78 English articles were reviewed.

**Key Content and Findings:** We introduce public datasets, evaluation methods, and multi-modal data processing. We also summarize common DL network structures, techniques, and multi-modal image fusion methods used in different tumor segmentation tasks. Finally, we conclude this study by presenting perspectives for future research.

**Conclusions:** In multi-modal tumor segmentation tasks, DL technique is a powerful method. With the fusion methods of different modal data, the DL framework can effectively use the characteristics of different modal data to improve the accuracy of tumor segmentation.

**Keywords:** Multi-modal image; tumor segmentation; fusion methods; review

Submitted Jun 07, 2023. Accepted for publication Oct 19, 2023. Published online Jan 02, 2024.
doi: 10.21037/qims-23-818
**View this article at:** https://dx.doi.org/10.21037/qims-23-818

## Introduction

Tumors pose a significant threat to human health and well-being, with gliomas in the brain, squamous cell tumors in the head and neck, melanomas in the skin, and systemic lymphomas among the prominent examples (1). Advanced medical imaging techniques, including magnetic resonance imaging (MRI), computed tomography (CT), and positron emission tomography (PET), play a crucial role in tumor staging, localization, diagnosis, and treatment planning. Accurate segmentation of tumor regions from medical images is critical to these processes. However, images obtained from a single modality may have limitations in

**Table 1** The search strategy summary

| Items | Specification |
|---|---|
| Date of search | 03/03/2023–06/01/2023 |
| Databases and other sources searched | PubMed and Google Scholar |
| Search terms used | Use "multi-modal", "deep learning", and "tumor segmentation" as keywords to search |
| Timeframe | 2018–2023 |
| Inclusion and exclusion criteria | Published English journals were selected, excluding conferences and non-English papers. Papers containing "multi-modal", "deep learning", and "tumor segmentation" were selected, otherwise they were excluded |
| Selection process | The literature selection was conducted independently by H.X. and Y.T. |

accurately depicting the tumor shape, making multi-modal imaging advantageous in tumor segmentation tasks (2,3). For instance, in the segmentation of brain gliomas, the four MRI sequences T1, T2, T1ce, and fluid-attenuated inversion recovery (FLAIR) provide complementary information on the tumor shape and other lesion structures in the brain (4,5). Similarly, in head and neck tumor segmentation, PET and CT images can offer additional information on the tumor's location and contour (6).

Manual tumor segmentation is a common technique used in computer-aided diagnosis (CAD) systems, but it has limitations due to the subjectivity of doctors' experience, which may lead to deviation, as well as time-consuming and labor-intensive processes (7). Therefore, accurate automatic segmentation is essential. In recent years, deep learning (DL) techniques, for example, convolutional neural networks (CNN), have been widely used in multi-modal tumor segmentation tasks for various body parts, including the brain (8), head and neck (9), and lungs (10). The fundamental idea behind these techniques is to learn tumor features from training data and automatically segment tumors in unknown data, which can reduce the cost of manual segmentation and improve segmentation accuracy. Multi-modal DL-based tumor segmentation algorithms have emerged as a prominent trend and have attracted increasing attention for achieving accurate segmentation of tumors.

This study provides a comprehensive overview of DL algorithms for multi-modal tumor segmentation, including public datasets, evaluation methods, segmentation networks, common techniques, and evaluation indicators analysis under various multi-modal data fusion methods. The benchmark dataset from the Open Challenge (https://grand-challenge.org) can validate the tumor segmentation performance under different multi-modal data fusion methods, and researchers can either apply these methods to specific task datasets or innovate based on them. Additionally, publicly available datasets such as BraTS (11) for head glioma, HECKTOR (12,13) for head and neck squamous cell carcinoma (HNSCC), and autoPET (14) for lymphoma, melanoma, or lung cancer, provide valuable resources for researchers to develop and evaluate their segmentation algorithms.

The discussion of this paper is divided into four main parts. The first part introduces multi-modal datasets and evaluation methods. The second part describes data processing. In the third part, we conduct a retrospective analysis of multi-modal data from the perspective of DL structures, covering preprocessing, network structures, fusion strategies, loss functions, and post-processing methods. The fourth part analyzes the tumor location retrospectively, focusing on brain gliomas, HNSCC, and systemic lymphomas. We present this article in accordance with the Narrative Review reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-23-818/rc).

## Methods

We performed a search of the PubMed and Google Scholar databases for existing research work on multi-modal tumor segmentation. The keywords used were "multi-modal", "deep learning", and "tumor segmentation". The search time comprised research conducted in the past 5 years, and the specific time was from 1 January 2018 to 1 June 2023. The search strategy is shown in *Table 1*. In the end, a total of 78 English articles were reviewed. After searching the articles, we analyzed and discussed the applied public

 

**Table 2** Multi-modal medical image tumor segmentation datasets, including MRI multi-sequence data set and PET/CT dual-modal dataset

| Image modality | Type of tumor | Dataset | Year | Number of cases | | |
|---|---|---|---|---|---|---|
| | | | | Training set | Validation set | Test set |
| MRI sequence (T1, T2, T1ce, FLAIR) | Brain tumor | BraTS2013 | 2013 | 30 | N/A | 35 |
| | | BraTS2014 | 2014 | 200 | N/A | 38 |
| | | BraTS2015 (11) | 2015 | 274 | N/A | 110 |
| | | BraTS2016 (15) | 2016 | 274 | N/A | 191 |
| | | BraTS2017 (16) | 2017 | 285 | 46 | 146 |
| | | BraTS2018 (17) | 2018 | 285 | 66 | 191 |
| | | BraTS2019 (18) | 2019 | 335 | 125 | 166 |
| | | BraTS2020 (19) | 2020 | 369 | 125 | 166 |
| | | BraTS2021 (20) | 2021 | 1,251 | 219 | 530 |
| PET/CT | Head and neck tumor | HNC | 2017 | 250 | N/A | N/A |
| | | HECKTOR (12) | 2020 | 201 | N/A | 53 |
| | | HECKTOR21 (13) | 2021 | 224 | N/A | 101 |
| | Lung cancer | Lung-PET-CT-Dx (21) | 2020 | 355 | N/A | N/A |
| | Soft tissue sarcoma | STS (22) | 2015 | 51 | N/A | N/A |
| | Malignant melanoma, lymphoma, or lung cancer | autoPET (14) | 2022 | 1,014 | N/A | 150 |

MRI, magnetic resonance imaging; PET, positron emission tomography; CT, computed tomography; FLAIR, fluid-attenuated inversion recovery; N/A, not applicable.

dataset, evaluation metrics, data processing, DL networks and technology, and multi-modal tumor segmentation methods according to the work content.

## Discussion

### *Datasets and evaluation methods*

#### Public dataset

This study presents a summary of publicly available datasets commonly used for multi-modal tumor segmentation tasks. The BraTS Challenge provides MRI datasets with multiple sequences, which include T1, T2, T1ce, and FLAIR. Additionally, the HECKTOR and autoPET challenges provide PET/CT multi-modal datasets. Other public datasets, such as Lung-PET-CT-Dx and Soft Tissue Sarcoma (STS), are also available on The Cancer Imaging Archive (TCIA) website (https://www.cancerimagingarchive.net). *Table 2* summarizes the key features of these publicly available datasets.

#### MRI sequence dataset

The BraTS challenge contains four MRI sequences T1, T1ce, T2, and FLAIR sequences. T1 sequences provide basic tissue structure and can be used to identify morphological features of tumors (4). T1ce sequences can reflect the fat content of cells (4). The location and size of the lesion can be clearly reflected in the T2 sequences (4). Compared with the T2 sequences, FLAIR sequences can better show the surrounding conditions of the tumor site, and clearly show the edema area (5). BraTS2013 contains the data of 65 glioma patients, 30 people as training data, and 35 people as testing data. Some 14 of these patients had low-grade glioma (LGG) and 51 had high-grade glioma (HGG). The BraTS2014 training set contains the data of 200 patients, and the testing set contains the data of 38 patients. On the basis of BraTS2014, the BraTS2015 training set contains the data of 220 cases of HGG and 54 cases of LGG, and the testing set contains 110 cases of unknown grade data, with a total of 384 cases of data. The BraTS2016 dataset and the 2015 dataset have the same

 

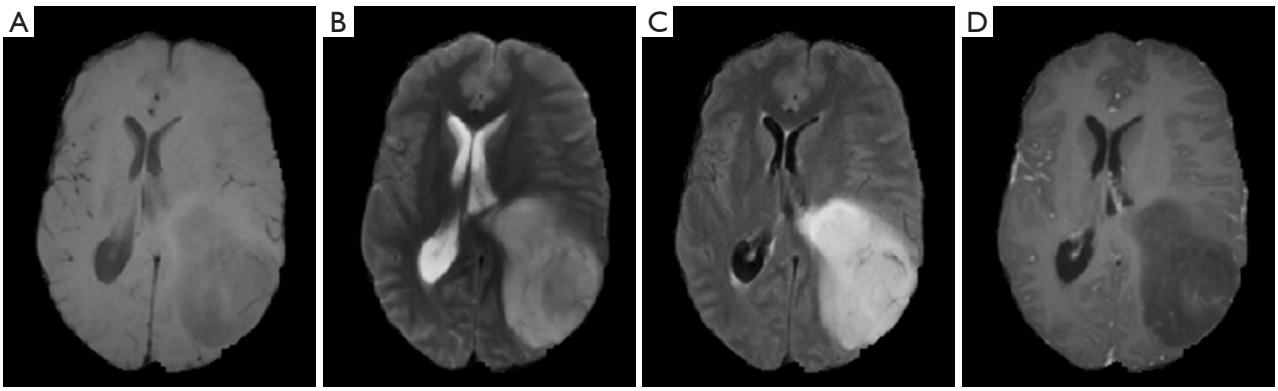


**Figure 1** MRI sequence images of a brain tumor patient in the BraTS2021 (20) database. (A) T1. (B) T2. (C) FLAIR. (D) T1ce. MRI, magnetic resonance imaging; FLAIR, fluid-attenuated inversion recovery.

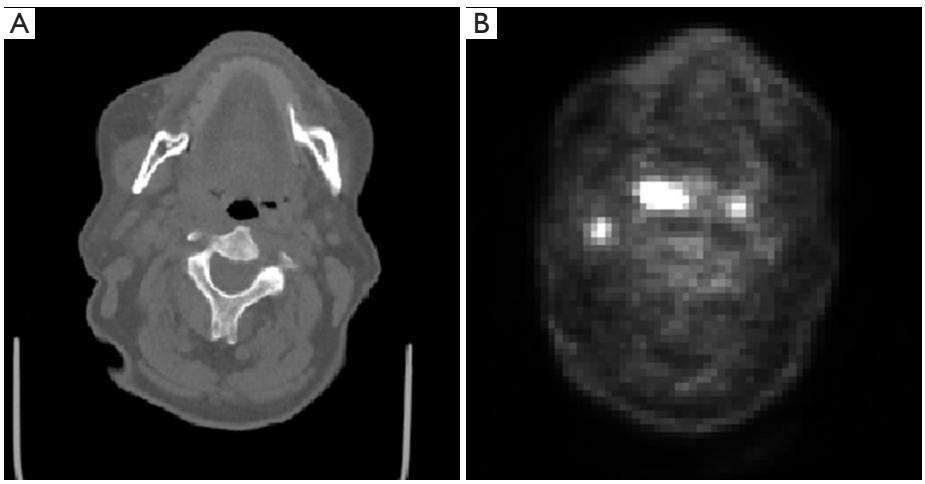


**Figure 2** PET/CT images of a patient with squamous cell carcinoma in the HECKTOR database. (A) CT. (B) PET. PET, positron emission tomography; CT, computed tomography.

training set, but the testing set has 191 cases. BraTS2017 contains 210 cases of HGG and 75 cases of LGG, and the validation set and testing set contain data of 46 cases and 146 cases of unknown grade, respectively. BraTS2018 training set and BraTS2017 have the same training set, and the verification set and testing set contain data of 66 cases and 191 cases, respectively. BraTS2019 contains 335 cases of data, including 259 cases of HGG and 76 cases of LGG. The verification set and testing set contain data of 125 cases and 166 cases, respectively. BraTS2020 contains data of 369 cases, including 293 cases of HGG and 76 cases of LGG. The BraTS2020 dataset and the BraTS2019 dataset have the same verification set and testing set. BraTS2021 contains multi-institution and multi-parameter MRI datasets, including 1,251 training sets, 219 validation sets, and 530 test data. *Figure 1* shows an image of a patient in the BraTS2021 dataset.

### PET/CT dataset

The HNC dataset, obtained from TCIA, comprises of 250 patients (192 males and 58 females) who have been diagnosed with HNSCC. The data format used for storage is Digital Imaging and Communication in Medicine (DICOM). In HECKTOR2020, a dataset with 254 cases was created based on the HNC dataset, consisting of 201 cases for training and 53 cases for testing. HECKTOR2021 further expanded this dataset by adding 71 more patients, totaling 325 patient data. Among them, 224 cases were used for

 

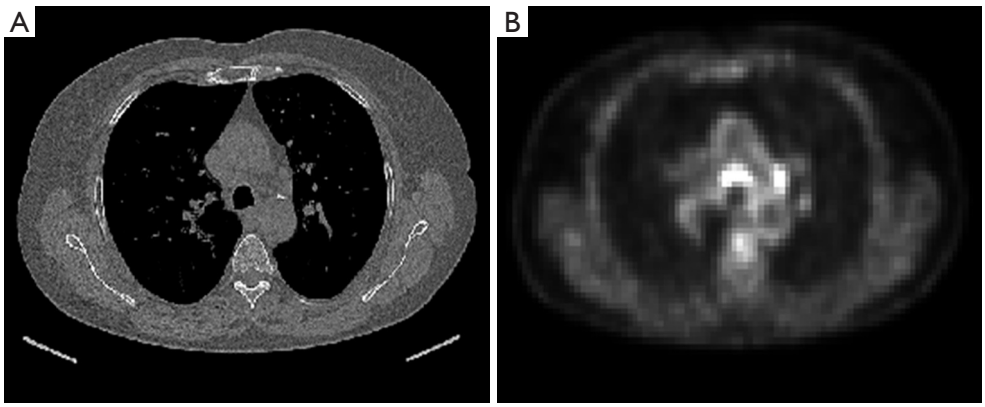


**Figure 3** PET/CT images of a patient with lung cancer in the Lung-PET-CT-Dx database. (A) CT. (B) PET. PET, positron emission tomography; CT, computed tomography.

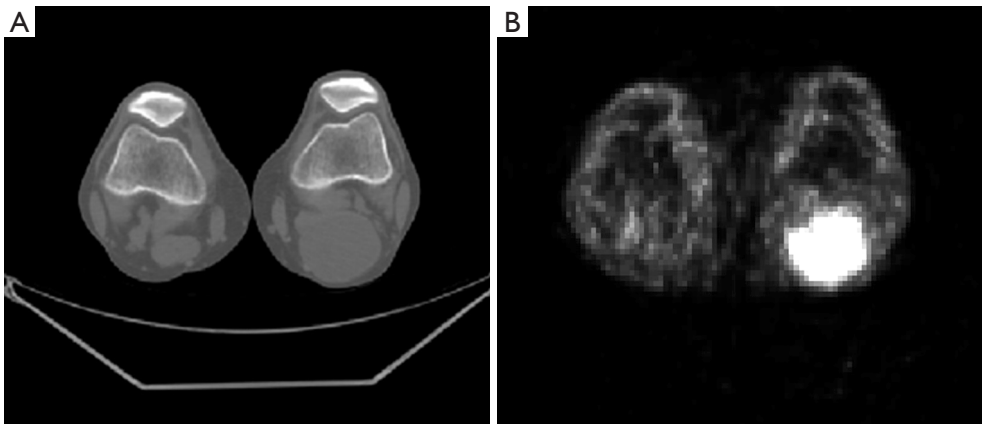


**Figure 4** PET/CT images of a patient with soft tissue sarcoma of extremities in the STS database. (A) CT. (B) PET. PET, positron emission tomography; CT, computed tomography.

training, and 101 cases for testing. *Figure 2* presents a PET/CT image of a patient from the HECKTOR2020 challenge dataset. The size of CT in the data set is 512×512, and the axial resolution is 0.9766×0.9766, whereas that of the corresponding PET is 128×128, and the axial resolution is 3.516×3.516. Therefore, in order to avoid the difficult problem of segmentation caused by low resolution, for some methods (23-26) in data preprocessing, the CT and PET volume are resampled to an isotropic 1×1×1 $mm^3$ voxel spacing using trilinear interpolation.

The Lung-PET-CT-Dx dataset, sourced from TCIA, comprises DICOM-formatted data from 355 patients diagnosed with lung cancer. The CT images in the dataset are of size 512×512, whereas the corresponding PET images are of size 200×200. *Figure 3* depicts a PET/CT image of a patient from the Lung-PET-CT-Dx dataset.

The STS dataset, sourced from TCIA, comprises data from 51 patients diagnosed with soft tissue sarcoma of extremities, stored in DICOM format. Each CT image is of size 512×512, whereas the corresponding PET image is of size 128×128. *Figure 4* presents a PET/CT image from a patient included in the STS dataset.

The autoPET challenge provides a training set of PET/CT data for patients diagnosed with melanoma, lymphoma, or lung cancer. The dataset includes a total of 1,014 studies (900 cases) obtained at the University Hospital Tübingen and stored in DICOM, NIfTI, and HDF5 formats on TCIA. *Figure 5* displays a PET/CT image of a patient in the autoPET dataset.

### Basic evaluation metrics

Tumor segmentation involves pixel-wise classification of

 

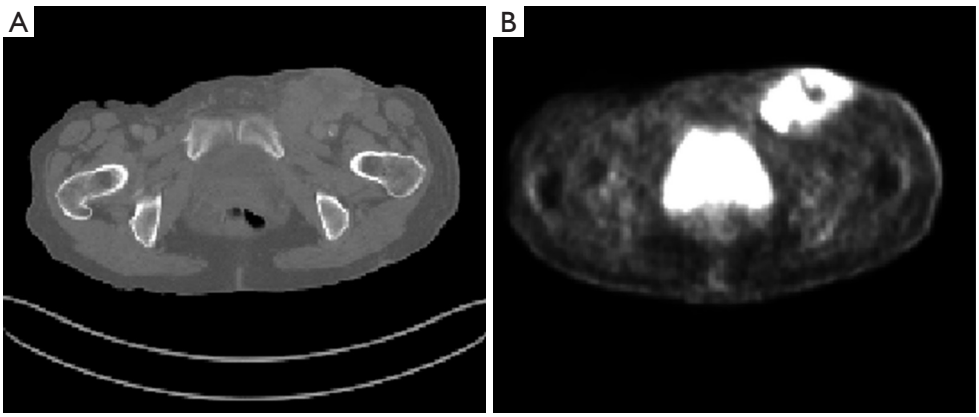


**Figure 5** PET/CT images of a patient with melanoma in the autoPET database. (A) CT. (B) PET. PET, positron emission tomography; CT, computed tomography.

**Table 3** Confusion matrix

| Confusion matrix | Ground truth | |
|---|---|---|
| | Positive | Negative |
| Positive | TP | FP |
| Negative | FN | TN |

TP, true positive; FP, false positive; FN, false negative; TN, true negative.

the image, and the accuracy of the segmentation can be evaluated using a confusion matrix, which indicates the relationship between the segmentation result and the ground truth. *Table 3* provides a summary of the confusion matrix.

In tumor segmentation, the confusion matrix serves as a representation of the relationship between the segmentation results and the ground truth. The basic evaluation indicators that can be derived from the confusion matrix include accuracy rate (ACC), positive predictive value (PPV), sensitivity, known as true positive rate (TPR), specificity, known as true negative rate (TNR), Jaccard similarity coefficient (JC), dice similarity coefficient (DSC), F1 score, and the harmonic mean value of sensitivity (HMSD). ACC indicates the percentage of correctly predicted results among all samples. PPV reflects the proportion of all predicted positive samples that are actually positive. TPR signifies the proportion of actual positive samples that are predicted as positive. TNR represents the ability to identify negative samples, which is the proportion of negative samples that are correctly identified. The calculation formulas for these indicators are presented in *Table 4*.

**Table 4** Evaluation index and its calculation formula

| Assessment | Formula |
|---|---|
| ACC | $\frac{TP+TN}{TP+TN+FP+FN}$ |
| PPV (precision) | $\frac{TP}{TP+FP}$ |
| TPR (sensitivity) | $\frac{TP}{TP+FN}$ |
| TNR (specificity) | $\frac{TN}{FP+TN}$ |
| IoU (JC) | $\frac{TP}{TP+FP+FN}$ |
| DSC | $\frac{2TP}{2TP+FP+FN}$ |
| F1 score | $\frac{2\times PPV\times TPR}{PPV+TPR}$ |
| HMSD | $\frac{2}{1/DSC+1/TPR}$ |

ACC, accuracy; TP, true positive; TN, true negative; FP, false positive; FN, false negative; TPR, true positive rate; TNR, true negative rate; IoU, intersection over union; JC, Jaccard similarity coefficient; DSC, dice similarity coefficient; PPV, positive predictive value; HMSD, harmonic mean value of sensitivity.

### Evaluation metrics based on segmentation boundaries

The evaluation of the boundary of the segmentation result can be achieved using two distance metrics: mean surface distance (MSD), also known as average symmetric surface distance (ASSD), and Hausdorff distance (HD). These two

 

metrics are more sensitive to the segmentation boundary. Let $P$ denote the predicted tumor, $G$ denote the ground truth tumor, and $P_C$ and $G_C$ denote the corresponding contours. The calculation formulas for MSD and HD are as follows:

$$MSD(P_C,G_C)=\frac{1}{2}\left(\frac{1}{|P_C|}\sum_{p\in P_C}\min_{g\in G_C}d(p,g)+\frac{1}{|G_C|}\sum_{g\in G_C}\min_{p\in P_C}d(g,p)\right) \quad [1]$$

$$HD(P_C,G_C)=\max\left\{\max_{p\in P_C}\min_{g\in G_C}d(p,g),\max_{g\in G_C}\min_{p\in P_C}d(g,p)\right\} \quad [2]$$

where $d(p,g)$ represents the Euclidean distance between point $p$ and point $g$, $p\in p_c, g\in G_c$. $|p_c|$ and $|G_c|$ represent the total number of pixels on the predicted and ground truth contours, respectively.

**Evaluation metrics based on segmentation volume**

For the task of tumor segmentation, the performance of segmentation can be evaluated by measuring the segmented volume. The two evaluation indicators commonly used are the relative volume difference (RVD) and the absolute volume difference (AVD). The calculation formulas for RVD and AVD are given below:

$$RVD(P,G)=\frac{|P|-|G|}{G} \quad [3]$$

$$AVD(P,G)=|P-G| \quad [4]$$

where $P$ is the predicted tumor volume, and $G$ is the actual tumor volume.

### *Data processing*

This section will discuss the preprocessing and post-processing steps applied to the raw data in preparation for input to the segmentation network. This includes standardizing the data dimensions of different modalities and normalizing pixel intensity values prior to input. In cases where data are insufficient, data augmentation techniques may be employed. Additionally, post-processing techniques may be applied to refine the preliminary segmentation results generated by the network and improve performance.

**Pre-processing**

Data preprocessing is a crucial step in the segmentation task that can enhance the network model's performance. Multi-modal image registration, in particular, is a crucial step in data preprocessing for multi-modal tumor segmentation (27). It involves aligning different imaging modalities, such as MRI sequences and PET/CT scans, to a common spatial reference frame. By performing image registration, the inherent spatial discrepancies between modalities are minimized, ensuring accurate fusion and correlation of tumor information across diverse imaging sources (28). This alignment facilitates the creation of a comprehensive and integrated view of the tumor, enhancing the segmentation process and enabling better visualization and analysis of the tumor's spatial distribution and characteristics (29).

Data size is also an important consideration, and in some experiments, the BraTS dataset's original size of 240×240×155 is cropped or resized to 128×128×128 to reduce storage requirements (30-34). In the HECKTOR dataset, where the axial size of PET and CT images is different, some methods resample PET/CT images to 1×1×1 mm in each direction and crop them to 144×144×144 size (23,24,35). For the STS dataset, where the PET axial size is 128×128, and the CT axial size is 256×256, some works (36,37) resize CT to match PET size to 128×128, or linearly interpolate PET to CT size (38).

Pixel value processing, such as bias field correction, intensity normalization, and intensity shift and scale, is also an essential pre-processing step (19,39). For example, some methods use the N4ITK method to correct the BraTS dataset's bias field (8,31,40), and then the pixel values are normalized (2,16,41) using z-score normalization. For the HECKTOR dataset, some methods intercept the Hounsfield unit (HU) value of CT images to eliminate irrelevant information (25,42,43), and normalize the intercepted HU value to [0, 1] or [–1, 1] (9,25,44). For PET images, it is typically converted into standard uptake values (SUVs) (45), and then z-score normalization is applied.

Finally, data volume augmentation is used to improve the model's robustness and avoid overfitting. Data augmentation methods including flip, rotations, random noise, scaling, elastic deformations, and the mixed data (mix-up) method (46,47) are commonly used in PET/CT tumor segmentation tasks. The mix-up (48) technique can increase data diversity, which generates randomly weighted combined image pairs based on training data pairs. For example, the generated mixed training data pair ($x$,$y$) can be weighted from the original training data pair ($x_1$,$y_1$), ($x_2$,$y_2$) and can be expressed as:

$$x=\lambda x_1+(1-\lambda)x_2 \quad [5]$$

$$y=\lambda y_1+(1-\lambda)y_2 \quad [6]$$

Where $\lambda\sim\text{Beta}(\alpha=0.2)$, means that the probability of data enhancement is 0.2.

**Post-processing**

After training a neural network, the resulting prediction

 

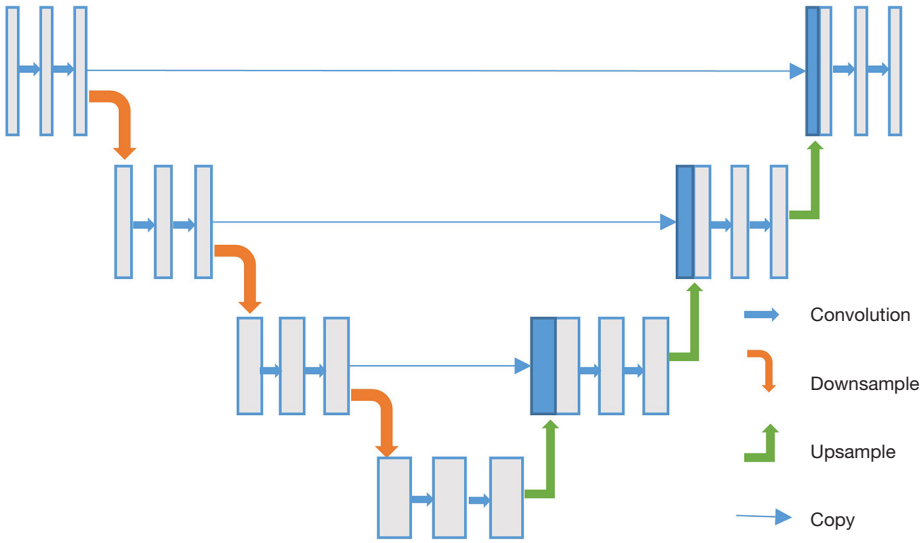


**Figure 6** U-Net network structure with skip connections for tumor segmentation. The blue box indicates copied feature maps, while the white box represents a multi-channel feature map.

may require further refinement using post-processing methods to improve segmentation accuracy. These methods commonly involve the use of conditional random field (CRF), removal of small connected domains and outliers (49), and setting thresholds for the number of pixels. For instance, Kamnitsas *et al.* (50,51) utilized CRF to eliminate false positives and remove connected domains that contain fewer than 250 pixels. In the case of glioma-enhanced region segmentation, if the number of pixels in the enhanced region is below a predefined threshold, necrotic regions can be substituted instead (52,53).

### *DL network and technology*

This section provides a comprehensive overview of the DL network architectures and techniques frequently employed in multi-modal tumor segmentation tasks. Among the commonly used network structures, CNNs and generative adversarial networks (GANs) are prominent. Moreover, several techniques are utilized to enhance the segmentation performance, such as attention mechanisms and uncertainty analysis.

#### CNN-based methods

CNNs have been widely used in computer vision applications, such as image classification (54,55), segmentation (56), and detection (57). The basic components of a CNN include the convolutional layer, pooling layer, activation layer, and fully connected layer. In 2014, Google proposed a network structure called GoogleNet (58), which included the Inception module. In multi-modal tumor segmentation tasks (40,59,60), the Inception module has also been used as a component of the structural framework. For example, Qayyum *et al.* (60) used the Inception structure as the encoding part in the HNSCC segmentation task. They extracted features of different levels to avoid gradient disappearance and reduce computational costs, resulting in improved segmentation performance. In addition to CNNs, the fully connected layer can be removed to obtain the fully convolutional network (FCN) (61), which is a milestone in the field of image segmentation. FCN is often used in multi-modal tumor segmentation (53,62-65), and Sun *et al.* (53) constructed a brain glioma segmentation framework based on FCN. Their framework is a multi-path structure that includes convolutional layers, pooling layers, dilated convolutional layers, deconvolutional layers, and activation function components. Similarly, the U-Net (56) network follows the encoder-decoder network structure and has become the most popular model in the field of medical image segmentation. Its structure is shown in *Figure 6*, and it consists of multi-level convolutions, including downsampling and upsampling operations. To solve the problem of feature disappearance in the downsampling process, skip connections are designed to fuse the features of the encoding part during the upsampling process. Some works (10,30,52,66-71) have performed multi-modal tumor

 

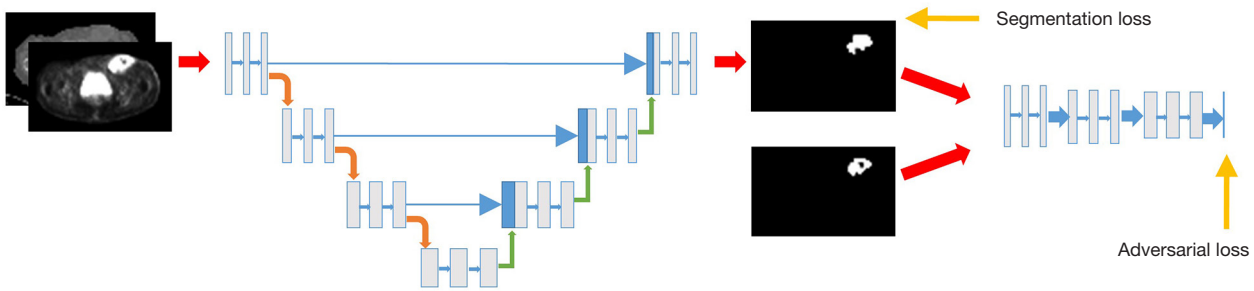


**Figure 7** GAN network structure for tumor segmentation. GAN, generative adversarial network.

segmentation based on the U-Net framework and achieved satisfactory segmentation results. Raza *et al.* (30) designed the encoder part as a residual network based on the U-Net framework to solve the problem of gradient disappearance. They also designed skip connections between residual and convolutional blocks to speed up the training of the network.

### GAN-based methods

GANs (72) are composed of two competing neural networks: a generator and a discriminator. The generator is responsible for generating fake data, while the discriminator is used to differentiate real and fake data. In the context of multi-modal tumor segmentation tasks, the generator can be directly utilized to generate segmentation results (73). A typical GAN network structure is illustrated in *Figure* 7, where the generator is a U-Net network, and the discriminator is a classification network. The predicted result and the real label are fed into the discriminator network to distinguish whether the input is the predicted result or the actual label. In a study by Huang *et al.* (73), transformer technology was employed in the generator against the network framework in the nasopharyngeal carcinoma PET/CT dataset, whereas the discriminator was used to constrain the final segmentation results and achieved good performance. GAN's generators can also be utilized to generate image data to improve the feature extraction capabilities for specific modalities. Zhang *et al.* (74) proposed to use two discriminators, where PET/CT images are encoded separately and the mode-specific decoding branch and the fusion branch of the two modalities are sent to the discriminator to train the network. This approach is used to extract features of a specific modality. Moreover, Xiang *et al.* (75) added a modality discriminator based on Zhang's work to determine whether the fusion branch and two specific decoding branches belong to PET or CT modality.

### Attention methods

In multi-modal tumor segmentation tasks, it is crucial for the network to automatically determine which feature information is the most informative. Therefore, the attention mechanism has been widely used in this area (9,18,24,69,76-83). Feature channel attention and feature space attention are two common types of attention methods. To enhance the tumor region information and suppress the normal physiological regions in PET images, spatial attention modules have been introduced by Diao *et al.* (84) and Fu *et al.* (3). In MRI multi-sequence data, Zhou *et al.* (31,32,34) applied channel and spatial attention mechanisms for feature fusion after extracting features from four MRI sequence images to improve the model's feature extraction ability.

In order to enhance the feature extraction and representation capabilities of the network, Hu *et al.* (85) proposed the squeeze-and-excitation network (SENet). Compared with traditional weight transfer methods, such as transferring feature maps to the next layer, SENet establishes interdependence between channels and adaptively corrects feature strengths among channels through a global loss function. Some methods (25,42,43,60,86-88) integrate SENet into their own network to improve feature extraction abilities. For example, Qayyum *et al.* (60,86) embedded SENet into the Inception coding part to obtain useful features, whereas Yuan *et al.* (42,43) and Yan *et al.* (87) incorporated SENet into the ResNet structure to improve tumor segmentation performance. Yousefirizi *et al.* (25) integrated SENet into the nnU-Net network to improve head and neck tumor segmentation results. Furthermore, Yao *et al.* (88) combined SENet and Dense to enhance feature extraction in PET/MR pancreatic tumors and improve segmentation accuracy.

In the field of natural language processing, transformer technology introduced a self-attention mechanism to connect

long-sequence long-distance contexts. This concept was later extended to image segmentation. Liu *et al.* (89) proposed the Swin Transformer technology that employs sliding window technology to achieve parameter reduction and global information modeling. Similarly, Cao *et al.* (90) developed a U-Net shaped transformer for medical image segmentation, which outperformed CNNs. Several studies built on these foundations to address specific segmentation tasks. For example, Zhu *et al.* (91) and Li *et al.* (92) employed Swin Transformers for semantic segmentation. Likewise, Liang *et al.* (93), Huang *et al.* (73), and Cai *et al.* (94) used U-Net shaped transformers for multi-modal tumor segmentation. Yue *et al.* (95) designed a U-shaped structure with a global attention transformer and a local attention transformer for esophageal squamous cell carcinoma data, which led to improved segmentation outcomes.

**Uncertainty mechanism**

Quantifying network segmentation uncertainty is a crucial aspect of medical image segmentation (37,77,96-98). To describe the uncertainty of the segmentation task, De Biase *et al.* (77) utilized generative probability maps. Diao *et al.* (37) proposed the evidence loss function to express the uncertainty of the PET and CT output results. The single-modal segmentation outcomes were eventually merged using evidence fusion, whereby the network becomes simpler and segmentation performance is enhanced. Hu *et al.* (97) implemented uncertainty analysis using the Monte Carlo dropout technique. It proposed uncertainty criteria based on lesion regions (impacting DSC and sensitivity) and background regions (impacting specificity) to improve performance. Huang *et al.* (98) added an evidence layer to the feature space to calculate the uncertainty of each voxel and improve segmentation accuracy.

### *Multi-modal tumor segmentation methods*

In recent years, a variety of DL-based methods have been proposed to tackle the problem of multi-modal tumor segmentation. Specifically, various methods have been proposed for brain glioma, HNSCC, and whole-body tumor segmentation.

**Brain tumor segmentation**

This section will focus on the segmentation of brain glioma using the BraTS dataset. We present an overview of the current state-of-the-art DL methods for multi-modal medical image segmentation in the BraTS challenge, which are summarized in *Table 5*. The methods can be categorized into three types based on multi-modal image fusion: input-level fusion using input data, feature-level fusion based on feature extraction, and result-level fusion based on segmentation results. The table provides details on the data preprocessing technique, segmentation network structure, dataset used, as well as Dice results for whole tumor, tumor core, and enhanced tumor.

The input-level fusion method involves merging the four input MRI sequences (T1, T2, T1ce, and FLAIR) into the network model after fusion at the channel layer. This method is commonly used for tumor segmentation. For instance, Liu *et al.* (19) proposed a multi-task model for segmenting tumors in the BraTS dataset and added a variational autoencoder (VAE) to the segmentation network to reconstruct the input data as an auxiliary task. The reconstruction part extraction helped in segmenting the multi-modal feature extraction capabilities of the encoded part of the network, leading to improved segmentation performance. Similarly, Cai *et al.* (94) fused multiple sequences of the BraTS dataset at the input-level. They used the U-Net network as the framework for segmentation and employed convolution and transformer blocks as the components of the encoding and decoding, thereby achieving comprehensive learning of global and local information.

Feature-level fusion is also a common approach for multi-modal tumor segmentation. After extracting features from T1, T2, T1c, and FLAIR sequences, the fusion is performed at the feature level. Zhu *et al.* (91) proposed a segmentation model for the BraTS dataset, consisting of three modules: segmentation, edge detection, and feature fusion. In the segmentation module, T1, T2, T1c, and FLAIR sequence data are fused at the input level, whereas in the edge detection module, T1ce and FLAIR sequences are fused at the input level as special features. The feature fusion module fuses the intermediate features of the segmentation module and the edge detection module in a feature-level manner, enhancing segmentation performance. Zhou *et al.* (34) used independent encoders for four different image sequences to learn specific features. They employed cross-connections to learn information of related modalities, and an attention fusion module to fuse the extracted features.

In addition, there are tumor segmentation methods based on result-level fusion, such as fusion by majority voting and weighted averaging. Ding *et al.* (100) adopted a dynamic decision fusion method to integrate the results from

 

**Table 5** Summary of the deep learning approaches for BraTS challenge segmentation

| Fusion method | Study | Network | Pre-processing | Database | Result DSC (whole/core/enhanced) |
|---|---|---|---|---|---|
| Input level | (39) | 3D U-Net | Normalization | BraTS18** | 0.8839/0.8154/0.7664 |
| | | VAE | Data augmentation | | |
| | (19)* | 3D U-Net | Normalization | BraTS20 | 0.8900/0.8300/0.8100 |
| | | 3D ResNet | Crop | | |
| | (8) | 2D CNN | Bias field correction | BraTS13** | 0.8800/0.8300/0.7700 |
| | | | Normalization | BraTS15 | 0.7800/0.6500/0.7500 |
| | (16) | 3D U-Net | Normalization | BraTS15 | 0.8500/0.7400/0.6100 |
| | | | Data augmentation | BraTS17 | 0.8960/0.7970/0.7320 |
| | (51)* | 3D CNN | Normalization | BraTS15** | 0.8490/0.6670/0.6340 |
| | | CRF | | | |
| | (53)* | 3D FCN | Crop | BraTS18 | 0.9000/0.7900/0.7700 |
| | | | Normalization | BraTS19 | 0.8900/0.7800/0.7600 |
| | (99)* | 3D U-Net | N/A | BraTS15 | 0.8700/0.7500/0.6400 |
| | | 3D ResNet | | BraTS17 | 0.9040/0.8280/0.7780 |
| | (94)* | 3D Transformer | N/A | BraTS18 | 0.7160/0.7610/0.8740 |
| | | | | BraTS21 | 0.8400/0.8740/0.9110 |
| | (15) | 3D FCN | Bias field correction | BraTS13 | – |
| | | CRF | Normalization | BraTS15 | 0.8600/0.7300/0.6200 |
| | | RNN | | BraTS16 | 0.8400/0.7300/0.6200 |
| Feature level | (34) | 3D U-Net | Crop | BraTS18 | 0.8560/0.8700/0.7940 |
| | | | Bias field correction | | |
| | | | Normalization | | |
| | (49) | 3D U-Net | Crop and resize | BraTS18 | 0.8820/0.7860/0.6940 |
| | | | Bias field correction | BraTS19 | 0.8970/0.7750/0.7060 |
| | | | Normalization | | |
| | (18) | 2D U-Net | Crop | BraTS19 | 0.9267/0.8947/0.8354 |
| | (91)* | 2D Transformer | Normalization | BraTS18 | 0.9089/0.8796/0.8194 |
| | | | | BraTS19 | 0.9158/0.8924/0.8384 |
| | | | | BraTS20 | 0.9103/0.8822/0.8461 |
| Result level | (50)* | 3D CNN | Bias field correction | BraTS17** | 0.8860/0.7850/0.7290 |
| | | 3D FCN | Normalization | | |
| | | 3D U-Net | | | |
| | (100) | 2D FCN | Normalization | BraTS15 | 0.8500/0.7100/0.6100 |
| | | 2D U-Net | | BraTS18 | 0.8300/0.7360/0.7120 |

*, the method has public code. **, the first result of the challenge. DSC, dice similarity coefficient; VAE, variational autoencoder; CNN, convolutional neural network; CRF, conditional random field; FCN, fully convolutional network; N/A, not applicable; RNN, recurrent neural network.

 

**Table 6** Summary of the deep learning approaches for HECKTOR challenge segmentation

| Fusion method | Study | Network | Pre-processing | Database | Result DSC/HD95 |
|---|---|---|---|---|---|
| Input level | (24)* | 3D U-Net | Normalization<br>Clip | HECKTOR20 | 0.7530/3.28 |
| | (35)* | 3D U-Net | Normalization | HECKTOR20 | 0.7590/– |
| | (42) | 3D ResNet | Normalization<br>Crop<br>Data augmentation | HECKTOR20 | 0.7318/– |
| | (43) | 3D ResNet | Normalization<br>Crop<br>Data augmentation | HECKTOR21 | 0.7608/3.27 |
| | (25) | 3D U-Net | Normalization<br>Clip | HECKTOR21 | 0.7700/3.01 |
| | (60)* | 3D Inception<br>3D ResNet | Data augmentation | HECKTOR21 | 0.8110/5.75 |
| | (26) | 3D U-Net | Normalization | HECKTOR21 | 0.7681/3.15 |
| | (86)* | 3D Inception | Normalization<br>Data augmentation | HECKTOR21<br>HECKTOR22 | 0.8240/–<br>0.7540/– |
| Feature level | (9) | 3D U-Net | Normalization<br>Data augmentation | HECKTOR21 | 0.7367/3.27 |
| | (92) | 3D Transformer | N/A | HECKTOR21 | 0.7690/– |
| | (101) | 2D U-Net<br>3D U-Net | N/A | HECKTOR21 | 0.8104/3.42 |

*, the method has public code. DSC, dice similarity coefficient; HD95, Hausdorff distance 95%; N/A, not applicable.

multiple views to improve segmentation performance. First, the 3D data is sliced into 2D images from axial, sagittal, and coronal directions, and the 2D segmentation results from the three directions are fused to obtain the final integrated segmentation result.

### Head and neck tumor segmentation

This section discusses the segmentation method for HNSCC based on the HECKTOR dataset and summarizes a range of state-of-the-art networks based on multi-modal segmentation, as shown in *Table 6*. The networks are categorized into input-level fusion and feature-level fusion, and the table includes information on the network structure, preprocessing method, dataset used, and common segmentation result indicators. Qayyum *et al.* validated the HECKTOR21 and HECKTOR22 datasets in their work (60,86). They fused PET and CT 2-modality images at the input-level and then performed segmentation using a U-shaped segmentation network. The encoding part of the network employed a 3D inception structure with a 3D squeeze and excitation module, whereas the decoding part was constructed based on ResNet. This approach helped to calibrate channel features and integrate coarse and fine features for accurate tumor segmentation.

In the context of the HECKTOR dataset, multi-modal data can also be segmented by fusing features extracted from different modalities in a feature-level manner. Lee *et al.* (9) proposed a dual-path cross-attention U-network that encodes PET and CT modalities in two separate paths in the encoding layer. The channel attention module is used in each layer of encoding to extract key features. The modality-specific features are then aggregated using 1×1×1

 

**Table 7** Summary of the deep learning approaches for whole body tumor segmentation

| Fusion method | Study | Network | Pre-processing | Database | Result DSC |
|---|---|---|---|---|---|
| Input level | (103) | U-Net | Normalization | Private data | 0.8609 |
| | | | Clip | | |
| Feature level | (104) | 3D U-Net | Crop | Private data | 0.8585 |
| | | | Random rotate | | |
| | (105) | 3D U-Net | Crop | autoPET | 0.6450 |
| | | | Random rotate | | |
| | (106) | 3D U-Net | Crop | Private data | 0.8690 |
| | | | Random rotate | | |
| Result level | (107) | 2D ResU-Net | Rescale | Private data | 0.6664 |
| | | 3D ResU-Net | | | |

DSC, dice similarity coefficient.

convolution and sent to the decoder for final segmentation. Ahmad *et al.* (101) used the attention mechanism to fuse PET modal data features in the decoding part after extracting features from CT, which improved the segmentation performance of the network. Zhu *et al.* (102) used the cruciform structure extracted from PET images as additional information. Then, three different encoders performed feature extraction on cruciform structure images, PET images, and CT images. The extracted features were fused in a feature-level manner to extract tumor structure and boundary information.

### Whole body tumor segmentation

This section presents a detailed review and analysis of lymphoma segmentation, including a summary of state-of-the-art networks based on multi-modal segmentation. These networks are classified into input-level fusion, feature-level fusion, and result-level fusion, and are presented in *Table* 7 along with the network structure, preprocessing method, data set, and common segmentation result indicators. As an example of input-level fusion, Shi *et al.* (103) used the CycleGAN network to generate metabolic anomaly appearance (MAA) with whole-body PET/CT data. MAA was directly fused with PET and CT at the channel level in an input-level manner, and the network was trained to improve lymphoma segmentation performance. Wang *et al.* (104-106) used a threshold method to obtain high fluorodeoxyglucose (FDG) uptake sites (sFHU) based on PET images as additional prior information. Different from the study by Shi *et al.* (103), Wang *et al.*'s studies (104-106) performed feature extraction on sFHU, PET, and CT images separately, and then fused the extraction features in a feature-level manner to obtain the segmentation results. Result-level fusion also yielded good segmentation results, as demonstrated by Hu *et al.* (107), who fused the original PET/CT image, the preliminary results of multi-angle 2D segmentation, and the preliminary results of 3D segmentation to perform lymphoma segmentation.

## Conclusions

In this paper, the importance of using multi-modal data for accurate tumor segmentation in medical images is highlighted. Various DL-based methods for multi-modal tumor segmentation are discussed, along with commonly used datasets and evaluation methods, multi-modal data processing techniques, DL network structures, and data fusion methods. Commonly used CNN networks such as FCN (53,62-65) and U-Net (10,30,52,66-71), as well as GAN-based segmentation methods (73-75), are summarized. Attention mechanisms such as channel and spatial attention (84), SE module (60,86), and transformer attention (89) are also discussed. Uncertainty analysis techniques (37,77,96-98) are used to quantify segmentation uncertainty and improve segmentation results. *Tables* 5-7 summarize the segmentation methods for BraTS, HECKTOR dataset, and whole-body tumor dataset, respectively, including the fusion methods of different modal data, the network structure used, the data processing method, and the segmentation results.

 

Despite the remarkable progress made by CNN and GAN networks, as well as attention and uncertainty techniques, in multi-modal tumor segmentation, there is still a need for further improvement in existing methods. There are several research directions that can be explored in the future. Firstly, most existing methods fuse multi-modal data directly at the input level, feature level, or result level, without taking into consideration whether the features provided by each modality data are helpful for tumor segmentation. Therefore, there is a need for uncertainty analysis and quantification of the extracted features to further improve segmentation performance. Secondly, in whole-body tumor segmentation tasks, such as lymphoma, obtaining large and accurate annotations is challenging. Therefore, it is a challenging and practical direction to explore unsupervised or weakly supervised methods for segmenting lymphoma in the absence of or with inaccurate annotations.

Automatic tumor segmentation holds paramount significance in clinical diagnosis and treatment, and the use of multi-modal images allows for a more comprehensive representation of tumor characteristics. Thus, this paper aimed to review the advancements made in multi-modal tumor segmentation based on DL over the past 5 years. The research work was scrutinized in terms of the utilized datasets, evaluation metrics, data preprocessing techniques, DL networks and technology, and multi-modal tumor segmentation methods. Additionally, the paper includes a thorough analysis of different modality fusion approaches concerning tumor segmentation in various contexts. Moreover, the paper identifies the urgent challenges that demand attention in future research, such as uncertainty analysis of extracted features and the exploration of unsupervised learning techniques in scenarios with limited precise annotations.

## Acknowledgments

*Funding:* This work was supported by the Natural Science 556 Foundation of Liaoning Province (No. 2022-MS-114).

## Footnote

*Reporting Checklist:* The authors have completed the Narrative Review reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-23-818/rc

*Conflicts of Interest:* All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-818/coif). All authors report that this work was supported by the Natural Science 556 Foundation of Liaoning Province (No. 2022-MS-114). The authors have no other conflicts of interest to declare.

*Ethical Statement:* The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.



## References

1. Burstein HJ, Krilov L, Aragon-Ching JB, Baxter NN, Chiorean EG, Chow WA, et al. Clinical Cancer Advances 2017: Annual Report on Progress Against Cancer From the American Society of Clinical Oncology. J Clin Oncol 2017;35:1341-67.
2. Yang Q, Guo X, Chen Z, Woo PYM, Yuan Y. D(2)-Net: Dual Disentanglement Network for Brain Tumor Segmentation With Missing Modalities. IEEE Trans Med Imaging 2022;41:2953-64.
3. Fu X, Bi L, Kumar A, Fulham M, Kim J. Multimodal Spatial Attention Module for Targeting Multimodal PET-CT Lung Tumor Segmentation. IEEE J Biomed Health Inform 2021;25:3507-16.
4. Katti G, Ara SA, Shireen A. Magnetic resonance imaging (MRI)–A review. International Journal of Dental Clinics 2011;3:65-70.
5. Stall B, Zach L, Ning H, Ondos J, Arora B, Shankavaram U, Miller RW, Citrin D, Camphausen K. Comparison of T2 and FLAIR imaging for target delineation in high grade gliomas. Radiat Oncol 2010;5:5.
6. Ma, J, Yang X. Combining CNN and Hybrid Active Contours for Head and Neck Tumor Segmentation in CT and PET Images. In: Andrearczyk V, Oreiller V. Depeursinge A, editors. Head and Neck Tumor

 

Segmentation. HECKTOR 2020. Lecture Notes in Computer Science, vol 12603. Cham: Springer. 2021. doi: 10.1007/978-3-030-67194-5_7.
7. Dubey RB, Hanmandlu M, Vasikarla S. Evaluation of Three Methods for MRI Brain Tumor Segmentation. 2011 Eighth International Conference on Information Technology: New Generations, Las Vegas, NV, USA; 2011:494-9.
8. Pereira S, Pinto A, Alves V, Silva CA. Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. IEEE Trans Med Imaging 2016;35:1240-51.
9. Lee J, Kang J, Shin EY, Kim RE, Lee M. Dual-Path Connected CNN for Tumor Segmentation of Combined PET-CT Images and Application to Survival Risk Prediction. In: Andrearczyk V, Oreiller V, Hatt M, Depeursinge A. editors. Head and Neck Tumor Segmentation and Outcome Prediction. HECKTOR 2021. Lecture Notes in Computer Science, vol 13209. Cham: Springer; 2022:248-56.
10. Protonotarios NE, Katsamenis I, Sykiotis S, Dikaios N, Kastis GA, Chatziioannou SN, Metaxas M, Doulamis N, Doulamis A. A few-shot U-Net deep learning model for lung cancer lesion segmentation via PET/CT imaging. Biomed Phys Eng Express 2022.
11. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans Med Imaging 2015;34:1993-2024.
12. Oreiller V, Andrearczyk V, Jreige M, Boughdad S, Elhalawani H, Castelli J, et al. Head and neck tumor segmentation in PET/CT: The HECKTOR challenge. Med Image Anal 2022;77:102336.
13. Andrearczyk V, Oreiller V, Hatt M, Depeursinge A. Head and Neck Tumor Segmentation and Outcome Prediction. Second Challenge, HECKTOR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings. Springer Nature; 2022.
14. Gatidis S, Hepp T, Früh M, La Fougère C, Nikolaou K, Pfannenberg C, Schölkopf B, Küstner T, Cyran C, Rubin D. A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions. Sci Data 2022;9:601.
15. Zhao X, Wu Y, Song G, Li Z, Zhang Y, Fan Y. A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. Med Image Anal 2018;43:98-111.
16. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH, editors. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3; Springer; 2018.
17. Weninger L, Rippel O, Koppers S, Merhof D, editors. Segmentation of brain tumors and patient survival prediction: Methods for the brats 2018 challenge. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4; Springer;2019.
18. Fang F, Yao Y, Zhou T, Xie G, Lu J. Self-Supervised Multi-Modal Hybrid Fusion Network for Brain Tumor Segmentation. IEEE J Biomed Health Inform 2022;26:5310-20.
19. Liu Y, Mu F, Shi Y, Chen X. Sf-net: A multi-task model for brain tumor segmentation in multimodal mri via image fusion. IEEE Signal Process Lett 2022;29:1799-803.
20. Baid U, Ghodasara S, Mohan S, Bilello M, Calabrese E, Colak E, Farahani K, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. 2021. Available online: https://arxiv.org/abs/2107.02314
21. Li P, Wang S, Li T, Lu J, HuangFu Y, Wang D. A large-scale CT and PET/CT dataset for lung cancer diagnosis [dataset]. The Cancer Imaging Archive 2020. doi: 10.7937/TCIA.2020.NNC2-0461
22. Vallières M, Freeman CR, Skamene SR, El Naqa I. 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 2015;60:5471-96.
23. Andrearczyk V, Oreiller V, Vallières M, Castelli J, Elhalawani H, Jreige M, Boughdad S, Prior JO, Depeursinge A. Automatic segmentation of head and neck tumors and nodal metastases in PET-CT scans. Proceedings of the Third Conference on Medical Imaging with Deep Learning, Proceedings of Machine Learning Research. 2020;121:33-43.
24. Bourigault E, McGowan DR, Mehranian A, Papież BW. Multimodal PET/CT tumour segmentation and prediction of progression-free survival using a full-scale UNet with attention. Head and Neck Tumor Segmentation and Outcome Prediction: Second Challenge, HECKTOR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings. Springer; 2022:189-201. Available online: https://arxiv.org/abs/2111.03848
25. Yousefirizi F, Janzen I, Dubljevic N, Liu YE, Hill C,

 

MacAulay C, Rahmim A. Segmentation and risk score prediction of head and neck cancers in PET/CT volumes with 3D U-Net and Cox proportional hazard neural networks. Head and Neck Tumor Segmentation and Outcome Prediction: Second Challenge, HECKTOR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 2022:236-247. 2022 Feb 16. Available online: https://arxiv.org/abs/2202.07823

26. Wang G, Huang Z, Shen H, Hu Z. The head and neck tumor segmentation in PET/CT based on multi-channel attention network. Head and Neck Tumor Segmentation and Outcome Prediction: Second Challenge, HECKTOR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings. Springer; 2022:68-74.
27. Islam KT, Wijewickrema S, O'Leary S. A deep learning based framework for the registration of three dimensional multi-modal medical images of the head. Sci Rep 2021;11:1860.
28. Zitová B, Flusser J. Image registration methods: a survey. Image Vision Comput 2003;21:977-1000.
29. Boveiri HR, Khayami R, Javidan R, Mehdizadeh A. Medical image registration using deep neural networks: a comprehensive review. Comput Electr Eng 2020;87:106767.
30. Raza R, Bajwa UI, Mehmood Y, Anwar MW, Jamal MH. dResU-Net: 3D deep residual U-Net based brain tumor segmentation from multimodal MRI. Biomed Signal Process Control 2023;79:103861.
31. Zhou T, Ruan S, Guo Y, Canu S. A multi-modality fusion network based on attention mechanism for brain tumor segmentation. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 2020:377-80.
32. Zhou T, Canu S, Vera P, Ruan S. Latent Correlation Representation Learning for Brain Tumor Segmentation With Missing MRI Modalities. IEEE Trans Image Process 2021;30:4263-74.
33. Zhou T, Vera P, Canu S, Ruan S. Missing data imputation via conditional generator and correlation learning for multimodal brain tumor segmentation. Pattern Recognit Lett 2022;158:125-32.
34. Zhou T. Modality-level cross-connection and attentional feature fusion based deep neural network for multi-modal brain tumor segmentation. Biomed Signal Process Control 2023;81:104524.
35. Iantsen A, Visvikis D, Hatt M. Squeeze-and-Excitation Normalization for Automated Delineation of Head and Neck Primary Tumors in Combined PET and CT Images. In: Andrearczyk V, Oreiller V, Depeursinge A, editors. Head and Neck Tumor Segmentation. HECKTOR 2020. Lecture Notes in Computer Science, vol 12603. Cham: Springer. 2021. doi: 10.1007/978-3-030-67194-5_4.
36. Luo S, Jiang H, Wang M. C(2)BA-UNet: A context-coordination multi-atlas boundary-aware UNet-like method for PET/CT images based tumor segmentation. Comput Med Imaging Graph 2023;103:102159.
37. Diao Z, Jiang H, Han XH, Yao YD, Shi T. EFNet: evidence fusion network for tumor segmentation from PET-CT volumes. Phys Med Biol 2021.
38. Xu G, Cao H, Udupa JK, Tong Y, Torigian DA. DiSegNet: A deep dilated convolutional encoder-decoder architecture for lymph node segmentation on PET/CT images. Comput Med Imaging Graph 2021;88:101851.
39. Myronenko A. 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization. In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T, editors. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2018. Lecture Notes in Computer Science, vol 11384. Cham: Springer; 2019.
40. Chen W, Liu B, Peng S, Sun J, Qiao X. S3D-UNet: separable 3D U-Net for brain tumor segmentation. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4; Springer; 2019.
41. Wang G, Li W, Ourselin S, Vercauteren T. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3 2018:178-190. 2017 Dec 15. Available online: https://arxiv.org/abs/1709.00382
42. Yuan Y. Automatic Head and Neck Tumor Segmentation in PET/CT with Scale Attention Network. In: Andrearczyk V, Oreiller V, Depeursinge A, editors. Head and Neck Tumor Segmentation. HECKTOR 2020. Lecture Notes in Computer Science, vol 12603. Cham: Springer; doi: 10.1007/978-3-030-67194-5_5.
43. Yuan Y, Adabi S, Wang X. Automatic head and neck tumor segmentation and progression free survival analysis on PET/CT images. In: Andrearczyk V, Oreiller V, Hatt M, Depeursinge A, editors. Head and Neck Tumor Segmentation and Outcome Prediction. HECKTOR 2021. Lecture Notes in Computer Science, vol 13209.

 

Cham: Springer; 2022:179-88.
44. Guo Z, Guo N, Gong K, Zhong S, Li Q. Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. Phys Med Biol 2019;64:205015.
45. Xue Z, Li P, Zhang L, Lu X, Zhu G, Shen P, Ali Shah SA, Bennamoun M. Multi-Modal Co-Learning for Liver Lesion Segmentation on PET-CT Images. IEEE Trans Med Imaging 2021;40:3531-42.
46. Huang Z, Zou S, Wang G, Chen Z, Shen H, Wang H, Zhang N, Zhang L, Yang F, Wang H, Liang D, Niu T, Zhu X, Hu Z. ISA-Net: Improved spatial attention network for PET-CT tumor segmentation. Comput Methods Programs Biomed 2022;226:107129.
47. Wang M, Jiang H, Shi T, Wang Z, Guo J, Lu G, Wang Y, Yao YD. PSR-Nets: Deep neural networks with prior shift regularization for PET/CT based automatic, accurate, and calibrated whole-body lymphoma segmentation. Comput Biol Med 2022;151:106215.
48. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond empirical risk minimization. 2018 Apr 27. Available online: https://arxiv.org/abs/1710.09412
49. Zhang D, Huang G, Zhang Q, Han J, Han J, Wang Y, Yu Y. Exploring Task Structure for Brain Tumor Segmentation from Multi-modality MR Images. IEEE Trans Image Process 2020. [Epub ahead of print]. doi: 10.1109/TIP.2020.3023609.
50. Kamnitsas K, Bai W, Ferrante E, McDonagh S, Sinclair M, Pawlowski N, Rajchl M, Lee M, Kainz B, Rueckert D, editors. Ensembles of multiple models and architectures for robust brain tumour segmentation. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3; Springer: 2018.
51. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal 2017;36:61-78.
52. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH. No new-net. In: International MICCAI brainlesion workshop. Granada: Springer; 2018:234-44.
53. Sun J, Peng Y, Guo Y, Li D. Segmentation of the multimodal brain tumor image used the multi-pathway architecture method based on 3D FCN. Neurocomputing 2021;423:34-45.
54. Jmour N, Zayen S, Abdelkrim A. Convolutional neural networks for image classification. 2018 International Conference on Advanced Systems and Electric Technologies (IC_ASET), Hammamet, Tunisia, 2018:397-402.
55. Rawat W, Wang Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Comput 2017;29:2352-449.
56. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. 2015 May 18. Available online: https://arxiv.org/abs/1505.04597
57. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell 2017;39:1137-49.
58. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. 2014 Sep 17. Available online: https://arxiv.org/abs/1409.4842
59. Wang F, Cheng C, Cao W, Wu Z, Wang H, Wei W, Yan Z, Liu Z. MFCNet: A multi-modal fusion and calibration networks for 3D pancreas tumor segmentation on PET-CT images. Comput Biol Med 2023;155:106657.
60. Qayyum A, Benzinou A, Razzak I, Mazher M, Nguyen TT, Puig D, Vafaee F. 3D-IncNet: Head and Neck (H&N) Primary Tumors Segmentation and Survival Prediction. IEEE J Biomed Health Inform 2022. doi: 10.1109/JBHI.2022.3219445.
61. Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:640-51.
62. Li L, Zhao X, Lu W, Tan S. Deep Learning for Variational Multimodality Tumor Segmentation in PET/CT. Neurocomputing (Amst) 2020;392:277-95.
63. Bi L, Fulham M, Li N, Liu Q, Song S, Dagan Feng D, Kim J. Recurrent feature fusion learning for multi-modality pet-ct tumor segmentation. Comput Methods Programs Biomed 2021;203:106043.
64. Zhao X, Li L, Lu W, Tan S. Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network. Phys Med Biol 2018;64:015011.
65. Cui S, Mao L, Jiang J, Liu C, Xiong S. Automatic Semantic Segmentation of Brain Gliomas from MRI Images Using a Deep Cascaded Neural Network. J Healthc Eng 2018;2018:4940593.
66. Naser MA, van Dijk LV, He R, Wahid KA, Fuller CD. Tumor Segmentation in Patients with Head and Neck Cancers Using Deep Learning Based-on Multi-modality

 

PET/CT Images. In Head and Neck Tumor Segmentation: First Challenge, HECKTOR 2020, Held in Conjunction with MICCAI 2020, Proceedings 1 2021; Lima, Peru: Springer International Publishing; 2020;85-98.
67. Murugesan GK, Mccrumb D, Brunner E, Kumar J, Soni R, Grigorash V, Chang A, VanOss J, Moore S. Automatic Whole Body FDG PET/CT Lesion Segmentation using Residual UNet and Adaptive Ensemble. bioRxiv 2023. doi: 10.1101/2023.02.06.525233
68. Zhong Z, Kim Y, Zhou L, Plichta K, Allen B, Buatti J, Wu X. 3D fully convolutional networks for co-segmentation of tumors on PET-CT Images. Proc IEEE Int Symp Biomed Imaging 2018;2018:228-31.
69. Li J, Chen H, Li Y, Peng Y, Sun J, Pan P. Cross-modality synthesis aiding lung tumor segmentation on multi-modal MRI images. Biomed Signal Process Control 2022;76:103655.
70. Rahimpour M, Bertels J, Radwan A, Vandermeulen H, Sunaert S, Vandermeulen D, Maes F, Goffin K, Koole M. Cross-Modal Distillation to Improve MRI-Based Brain Tumor Segmentation With Missing MRI Sequences. IEEE Trans Biomed Eng 2022;69:2153-64.
71. Jemaa S, Fredrickson J, Carano RAD, Nielsen T, de Crespigny A, Bengtsson T. Tumor Segmentation and Feature Extraction from Whole-Body FDG-PET/CT Using Cascaded 2D and 3D Convolutional Neural Networks. J Digit Imaging 2020;33:888-94.
72. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative Adversarial Networks. 2014 Jun 10. Available online: https://arxiv.org/abs/1406.2661
73. Huang Z, Tang S, Chen Z, Wang G, Shen H, Zhou Y, Wang H, Fan W, Liang D, Hu Y, Hu Z. TG-Net: Combining transformer and GAN for nasopharyngeal carcinoma tumor segmentation based on total-body uEXPLORER PET/CT scanner. Comput Biol Med 2022;148:105869.
74. Zhang X, Zhang B, Deng S, Meng Q, Chen X, Xiang D. Cross modality fusion for modality-specific lung tumor segmentation in PET-CT images. Phys Med Biol 2022.
75. Xiang D, Zhang B, Lu Y, Deng S. Modality-Specific Segmentation Network for Lung Tumor Segmentation in PET-CT Images. IEEE J Biomed Health Inform 2023;27:1237-48.
76. Zhang Y, Lu Y, Chen W, Chang Y, Gu H, Yu B. MSMANet: A multi-scale mesh aggregation network for brain tumor segmentation. Appl Soft Comput 2021;110:107733.
77. De Biase A, Sijtsema NM, van Dijk LV, Langendijk JA, van Ooijen PMA. Deep learning aided oropharyngeal cancer segmentation with adaptive thresholding for predicted tumor probability in FDG PET and CT images. Phys Med Biol 2023.
78. Chen S, Li A, Chen J, Zhang X, Jiang C, Xu J. Hybrid Attention Fusion Segmentation Network for Diffuse Large B-cell Lymphoma in PET-CT. 2022 14th International Conference on Wireless Communications and Signal Processing (WCSP), Nanjing, China. 2022:72-6.
79. Zhou Y, Jiang H, Diao Z, Tong G, Luan Q, Li Y, Li X. MRLA-Net: A tumor segmentation network embedded with a multiple receptive-field lesion attention module in PET-CT images. Comput Biol Med 2023;153:106538.
80. Kumar A, Fulham M, Feng D, Kim J. Co-Learning Feature Fusion Maps from PET-CT Images of Lung Cancer. IEEE Trans Med Imaging 2019. [Epub ahead of print]. doi: 10.1109/TMI.2019.2923601.
81. Zhang J, Jiang H, Shi T. ASE-Net: A tumor segmentation method based on image pseudo enhancement and adaptive-scale attention supervision module. Comput Biol Med 2023;152:106363.
82. Zhang G, Shen X, Zhang YD, Luo Y, Luo J, Zhu D, Yang H, Wang W, Zhao B, Lu J. Cross-Modal Prostate Cancer Segmentation via Self-Attention Distillation. IEEE J Biomed Health Inform 2022;26:5298-309.
83. Matkovic LA, Wang T, Lei Y, Akin-Akintayo OO, Abiodun Ojo OA, Akintayo AA, Roper J, Bradley JD, Liu T, Schuster DM, Yang X. Prostate and dominant intraprostatic lesion segmentation on PET/CT using cascaded regional-net. Phys Med Biol 2021.
84. Diao Z, Jiang H, Shi T. A spatial squeeze and multimodal feature fusion attention network for multiple tumor segmentation from PET–CT Volumes. Eng Appl Artif Intell 2023;121:105955.
85. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA. 2018:7132-41.
86. Qayyum A, Mazher M, Khan T, Razzak I. Semi-supervised 3D-InceptionNet for segmentation and survival prediction of head and neck primary cancers. Eng Appl Artif Intell 2023;117:105590.
87. Yan C, Ding J, Zhang H, Tong K, Hua B, Shi S. SEResU-Net for Multimodal Brain Tumor Segmentation. IEEE Access 2022;10:117033-44.
88. Yao Y, Chen Y, Gou S, Chen S, Zhang X, Tong N. Auto-segmentation of pancreatic tumor in multi-modal image

 

using transferred DSMask R-CNN network. Biomed Signal Process Control 2023;83:104583.
89. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision. 2021. doi: 10.1109/ICCV48922.2021.00986
90. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M. Swin-unet: Unet-like pure transformer for medical image segmentation. Available online: https://arxiv.org/pdf/2105.05537.pdf
91. Zhu Z, He X, Qi G, Li Y, Cong B, Liu Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. Inf Fusion 2023;91:376-87.
92. Li GY, Chen J, Jang SI, Gong K, Li Q. SwinCross: Cross-modal Swin transformer for head-and-neck tumor segmentation in PET/CT images. Med Phys 2023. [Epub ahead of print]. doi: 10.1002/mp.16703.
93. Liang J, Yang C, Zeng M, Wang X. TransConver: transformer and convolution parallel network for developing automatic brain tumor segmentation in MRI images. Quant Imaging Med Surg 2022;12:2397-415.
94. Cai Y, Long Y, Han Z, Liu M, Zheng Y, Yang W, Chen L. Swin Unet3D: a three-dimensional medical image segmentation network combining vision transformer and convolution. BMC Med Inform Decis Mak 2023;23:33.
95. Yue Y, Li N, Zhang G, Zhu Z, Liu X, Song S, Ta D. Automatic segmentation of esophageal gross tumor volume in (18)F-FDG PET/CT images via GloD-LoATUNet. Comput Methods Programs Biomed 2023;229:107266.
96. Huang L, Ruan S. Application of belief functions to medical image segmentation: A review. Inf Fusion 2023;91:737-56.
97. Hu X, Guo R, Chen J, Li H, Waldmannstetter D, Zhao Y, Li B, Shi K, Menze B. Coarse-to-Fine Adversarial Networks and Zone-Based Uncertainty Analysis for NK/T-Cell Lymphoma Segmentation in CT/PET Images. IEEE J Biomed Health Inform 2020;24:2599-608.
98. Huang L, Ruan S, Decazes P, Denœux T. Lymphoma segmentation from 3D PET-CT images using a deep evidential network. Int J Approximate Reasoning 2022;149:39-60.
99. Zhou C, Ding C, Lu Z, Wang X, Tao D. One-Pass Multi-task Convolutional Neural Networks for Efficient Brain Tumor Segmentation. In: Frangi A, Schnabel J, Davatzikos C, Alberola-López C, Fichtinger G, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018. Lecture Notes in Computer Science, vol 11072. Cham: Springer; 2018. doi: 10.1007/978-3-030-00931-1_73
100. Ding Y, Zheng W, Geng J, Qin Z, Choo KR, Qin Z, Hou X. MVFusFra: A Multi-View Dynamic Fusion Framework for Multimodal Brain Tumor Segmentation. IEEE J Biomed Health Inform 2022;26:1570-81.
101. Ahmad I, Xia Y, Cui H, Islam ZU. AATSN: Anatomy Aware Tumor Segmentation Network for PET-CT volumes and images using a lightweight fusion-attention mechanism. Comput Biol Med 2023;157:106748.
102. Zhu X, Jiang H, Diao Z. CGBO-Net: Cruciform structure guided and boundary-optimized lymphoma segmentation network. Comput Biol Med 2023;153:106534.
103. Shi T, Jiang H, Wang M, Diao Z, Zhang G, Yao YD. Metabolic Anomaly Appearance Aware U-Net for Automatic Lymphoma Segmentation in Whole-Body PET/CT Scans. IEEE J Biomed Health Inform 2023;27:2465-76.
104. Wang M, Jiang H, Shi T, Yao YD. HD-RDS-UNet: Leveraging Spatial-Temporal Correlation Between the Decoder Feature Maps for Lymphoma Segmentation. IEEE J Biomed Health Inform 2022;26:1116-27.
105. Wang M, Jiang H, Shi T, Yao YD. SCL-Net: Structured Collaborative Learning for PET/CT Based Tumor Segmentation. IEEE J Biomed Health Inform 2022. [Epub ahead of print]. doi: 10.1109/JBHI.2022.3226475.
106. Wang M, Jiang H. Memory-Net: Coupling feature maps extraction and hierarchical feature maps reuse for efficient and effective PET/CT multi-modality image-based tumor segmentation. Knowl Based Syst 2023;265:110399.
107. Hu H, Shen L, Zhou T, Decazes P, Vera P, Ruan S. Lymphoma segmentation in PET images based on multi-view and Conv3D fusion strategy. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA. 2020:1197-200.

**Cite this article as:** Xue H, Yao Y, Teng Y. Multi-modal tumor segmentation methods based on deep learning: a narrative review. Quant Imaging Med Surg 2024;14(1):1122-1140. doi: 10.21037/qims-23-818

 