H-DenseFormer: An Efficient Hybrid Densely Connected Transformer for Multimodal Tumor Segmentation

Jun Shi[1], Hongyu Kan[1], Shulan Ruan[1], Ziqi Zhu[1], Minfan Zhao[1], Liang Qiao[1], Zhaohui Wang[1], Hong An[1(✉)], and Xudong Xue[2]

[1] University of Science and Technology of China, Hefei, China
shijun18@mail.ustc.edu.cn, han@ustc.edu.cn
[2] Hubei Cancer Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Abstract. Recently, deep learning methods have been widely used for tumor segmentation of multimodal medical images with promising results. However, most existing methods are limited by insufficient representational ability, specific modality number and high computational complexity. In this paper, we propose a hybrid densely connected network for tumor segmentation, named **H-DenseFormer**, which combines the representational power of the Convolutional Neural Network (CNN) and the Transformer structures. Specifically, H-DenseFormer integrates a Transformer-based Multi-path Parallel Embedding (**MPE**) module that can take an arbitrary number of modalities as input to extract the fusion features from different modalities. Then, the multimodal fusion features are delivered to different levels of the encoder to enhance multimodal learning representation. Besides, we design a lightweight Densely Connected Transformer (**DCT**) block to replace the standard Transformer block, thus significantly reducing computational complexity. We conduct extensive experiments on two public multimodal datasets, HECKTOR21 and PI-CAI22. The experimental results show that our proposed method outperforms the existing state-of-the-art methods while having lower computational complexity. The source code is available at https://github.com/shijun18/H-DenseFormer.

Keywords: Tumor segmentation · Multimodal medical image · Transformer · Deep learning

1 Introduction

Accurate tumor segmentation from medical images is essential for quantitative assessment of cancer progression and preoperative treatment planning [3]. Tumor tissues usually present different features in different imaging modalities. For example, Computed Tomography (CT) and Positron Emission Tomography

J. Shi and H. Kan contributed equally. This study was supported by the Fundamental Research Funds for the Central Universities (No. YD2150002001).

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 692–702, 2023.
https://doi.org/10.1007/978-3-031-43901-8_66

(PET) are beneficial to represent morphological and metabolic information of tumors, respectively. In clinical practice, multimodal registered images, such as PET-CT images and Magnetic Resonance (MR) images with different sequences, are often utilized to delineate tumors to improve accuracy. However, manual delineation is time-consuming and error-prone, with a low inter-professional agreement [12]. These have prompted the demand for intelligent applications that can automatically segment tumors from multimodal images to optimize clinical procedures.

Recently, multimodal tumor segmentation has attracted the interest of many researchers. With the emergence of multimodal datasets (e.g., BRATS [25] and HECKTOR [1]), various deep-learning-based multimodal image segmentation methods have been proposed [3,10,13,27,29,31]. Overall, large efforts have been made on effectively fusing image features of different modalities to improve segmentation accuracy. According to the way of feature fusion, the existing methods can be roughly divided into three categories [15,36]: *input-level fusion*, *decision-level fusion*, and *layer-level fusion*. As a typical approach, input-level fusion [8,20,26,31,34] refers to concatenating multimodal images in the channel dimension as network input during the data processing or augmentation stage. This approach is suitable for most existing end-to-end models [6,32], such as U-Net [28] and U-Net++ [37]. However, the shallow fusion entangles the low-level features from different modalities, preventing the effective extraction of high-level semantics and resulting in limited performance gains. In contrast, [35] and [21] propose a solution based on decision-level fusion. The core idea is to train an independent segmentation network for each data modality and fuse the results in a specific way. These approaches can bring much extra computation at the same time, as the number of networks is positively correlated with the number of modalities. As a compromise alternative, layer-level fusion methods such as HyperDense-Net [10] advocate the cross-fusion of the multimodal features in the middle layer of the network.

In addition to the progress on the fusion of multimodal features, improving the model representation ability is also an effective way to boost segmentation performance. In the past few years, Transformer structure [11,24,30], centered on the multi-head attention mechanism, has been introduced to multimodal image segmentation tasks. Extensive studies [2,4,14,16] have shown that the Transformer can effectively model global context to enhance semantic representations and facilitate pixel-level prediction. Wang et al. [31] proposed TransBTS, a form of input-level fusion with a U-like structure, to segment brain tumors from multimodal MR images. TransBTS employs the Transformer as a bottleneck layer to wrap the features generated by the encoder, outperforming the traditional end-to-end models. Saeed et al. [29] adopted a similar structure in which the Transformer serves as the encoder rather than a wrapper, also achieving promising performance. Other works like [9] and [33], which combine the Transformer with the multimodal feature fusion approaches mentioned above, further demonstrate the potential of this idea for multimodal tumor segmentation.

Although remarkable performance has been accomplished with these efforts, there still exist several challenges to be resolved. Most existing methods are either limited to specific modality numbers due to the design of asymmetric connections or suffer from large computational complexity because of the huge amount of model parameters. Therefore, how to improve model ability while ensuring computational efficiency is the main focus of this paper.

To this end, we propose an efficient multimodal tumor segmentation solution named Hybrid Densely Connected Network (**H-DenseFormer**). First, our method leverages Transformer to enhance the global contextual information of different modalities. Second, H-DenseFormer integrates a Transformer-based Multi-path Parallel Embedding (**MPE**) module, which can extract and fuse multimodal image features as a complement to naive input-level fusion structure. Specifically, MPE assigns an independent encoding path to each modality, then merges the semantic features of all paths and feeds them to the encoder of the segmentation network. This decouples the feature representations of different modalities while relaxing the input constraint on the specific number of modalities. Finally, we design a lightweight, Densely Connected Transformer (**DCT**) module to replace the standard Transformer to ensure performance and computational efficiency. Extensive experimental results on two publicly available datasets demonstrate the effectiveness of our proposed method.

2 Method

2.1 Overall Architecture of H-DenseFormer

Figure 1 illustrates the overall architecture of our method. H-DenseFormer comprises a Multi-path Parallel Embedding (MPE) module and a U-shaped segmentation backbone network in form of input-level fusion. The former serves

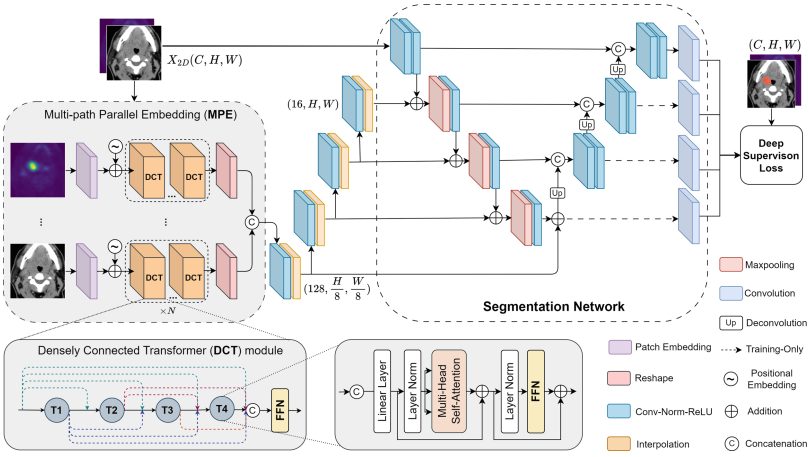

Fig. 1. Overall architecture of our proposed H-DenseFormer.

as the auxiliary extractor of multimodal fusion features, while the latter is used to generate predictions. Specifically, given a multimodal image input $\mathbf{X}_{3D} \in \mathbb{R}^{C\times H\times W\times D}$ or $\mathbf{X}_{2D} \in \mathbb{R}^{C\times H\times W}$ with a spatial resolution of $H \times W$, the depth dimension of D (number of slices) and C channels (number of modalities), we first utilize MPE to extract and fuse multimodal image features. Then, the obtained features are progressively upsampled and delivered to the encoder of the segmentation network to enhance the semantic representation. Finally, the segmentation network generates multi-scale outputs, which are used to calculate deep supervision loss as the optimization target.

2.2 Multi-path Parallel Embedding

Many methods [5,10,15] have proved that decoupling the feature representation of different modalities facilitates the extraction of high-quality multimodal features. Inspired by this, we design a Multip-path Parallel Embedding (MPE) module to enhance the representational ability of the network. As shown in Fig. 1, each modality has an independent encoding path consisting of a patch embedding module, stacked Densely Connected Transformer (DCT) modules, and a reshape operation. The independence of the different paths allows MPE to handle an **arbitrary** number of input modalities. Besides, the introduction of the Transformer provides the ability to model global contextual information. Given the input X_{3D}, after convolutional embedding and tokenization, the obtained feature of the i-th path is $\mathbf{F}_i \in \mathbb{R}^{l\times \frac{H}{p}\times \frac{W}{p}\times \frac{D}{p}}$, where $i \in [1, 2, ..., C]$, $p = 16$ and $l = 128$ denote the path size and embedding feature length respectively. First, we concatenate the features of all modalities and entangle them using a convolution operation. Then, interpolation upsampling is performed to obtain the multimodal fusion feature $\mathbf{F}_{out} \in \mathbb{R}^{k\times \frac{H}{8}\times \frac{W}{8}\times \frac{D}{8}}$, where $k = 128$ refers to the channel dimension. Finally, $\mathbf{F}_{out}$ is progressively upsampled to multiple scales and delivered to different encoder stages to enhance the learned representation.

2.3 Densely Connected Transformer

Standard Transformer structures [11] typically consist of dense linear layers with a computational complexity proportional to the feature dimension. Therefore, integrating the Transformer could lead to a mass of additional computation and memory requirements. Shortening the feature length can effectively reduce computation, but it also weakens the representation capability meanwhile. To address this problem, we propose the Densely Connected Transformer (DCT) module inspired by DenseNet [17] to balance computational cost and representation capability. Figure 1 details the DCT module, which consists of **four** Transformer layers and a feedforward layer. Each Transformer layer has a linear projection layer that reduces the input feature dimension to $g = 32$ to save computation. Different Transformer layers are connected densely to preserve representational power with lower feature dimensions. The feedforward layer at the end generates the fusion features of the different layers. Specifically, the output $\mathbf{z}_j$ of the j-th ($j \in [1, 2, 3, 4]$) Transformer layer can be calculated by:

$$\tilde{\mathbf{z}}_{j-1} = p(cat([\mathbf{z}_0; \mathbf{z}_1; ...; \mathbf{z}_{j-1}])), \tag{1}$$

$$\tilde{\mathbf{z}}_j = att(norm(\tilde{\mathbf{z}}_{j-1})) + \tilde{\mathbf{z}}_{j-1}, \tag{2}$$

$$\mathbf{z}_j = f(norm(\tilde{\mathbf{z}}_j)), \tag{3}$$

where $\mathbf{z}_0$ represents the original input, $cat(\cdot)$ and $p(\cdot)$ denote the concatenation operator and the linear layer, respectively. The $norm(\cdot)$, $att(\cdot)$, $f(\cdot)$ are the regular layer normalization, multi-head self-attention mechanism, and feedforward layer. The output of DCT is $\mathbf{z}_{out} = f(cat([\mathbf{z}_0; \mathbf{z}_1; ...; \mathbf{z}_4]))$. Table 1 shows that the stacked DCT has lower parameters and computational complexity than a standard Transformer structure with the same number of layers.

Table 1. Comparison of the computational complexity between the standard 12-layer Transformer structure and the stacked 3 (=12/4) DCT modules.

Feature Dimension	Resolution	Transformer		Stacked DCT (×3)	
		GFLOPs ↓	Params ↓	GFLOPs ↓	Params ↓
256	(512,512)	6.837	6.382M	**2.671**	**1.435M**
512	(512,512)	26.256	25.347M	**3.544**	**2.290M**

2.4 Segmentation Backbone Network

The H-DenseFormer adopts a U-shaped encoder-decoder structure as its backbone. As shown in Fig. 1, the encoder extracts features and reduces their resolution progressively. To preserve more details, we set the maximum downsampling factor to 8. The multi-level multimodal features from MPE are fused in a bitwise addition way to enrich the semantic information. The decoder is used to restore the resolution of the features, consisting of deconvolutional and convolutional layers with skip connections to the encoder. In particular, we employ Deep Supervision (**DS**) loss to improve convergence, which means that the multiscale output of the decoder is involved in the final loss computation.

Deep Supervision Loss. During training, the decoder has four outputs; for example, the i-th output of 2D H-DenseFormer is $\mathbf{O}^i \in \mathbb{R}^{c \times \frac{H}{2^i} \times \frac{W}{2^i}}$, where $i \in [0, 1, 2, 3]$, and $c = 2$ (tumor and background) represents the number of segmentation classes. To mitigate the pixel imbalance problem, we use a combined loss of Focal loss [23] and Dice loss as the optimization target, defined as follows:

$$\zeta_{FD} = 1 - \frac{2\sum_{t=1}^{N} p_t q_t}{\sum_{t=1}^{N} p_t + q_t} + \frac{1}{N}\sum_{t=1}^{N} -(1-p_t)^{\gamma} log(p_t), \tag{4}$$

where N refers to the total number of pixels, p_t and q_t denote the predicted probability and ground truth of the t-th pixel, respectively, and $r = 2$ is the modulation factor. Thus, DS loss can be calculated as follows:

$$\zeta_{\mathbf{DS}} = \sum \alpha_i \cdot \zeta_{FD}(\mathbf{O}^i, \mathbf{G}^i), \alpha_i = 2^{-i}. \tag{5}$$

where $\mathbf{G}^i$ represents the ground truth after resizing and has the same size as $\mathbf{O}^i$. α is a weighting factor to control the proportion of loss corresponding to the output at different scales. This approach can improve the convergence speed and performance of the network.

3 Experiments

3.1 Dataset and Metrics

To validate the effectiveness of our proposed method, we performed extensive experiments on **HECKTOR21** [1] and **PI-CAI22**[1]. HECKTOR21 is a dual-modality dataset for head and neck tumor segmentation, containing 224 PET-CT image pairs. Each PET-CT pair is registered and cropped to a fixed size of (144,144,144). PI-CAI22 provides multimodal MR images of 220 patients with prostate cancer, including T2-Weighted imaging (T2W), high b-value Diffusion-Weighted imaging (DWI), and Apparent Diffusion Coefficient (ADC) maps. After standard resampling and center cropping, all images have a size of (24,384,384). We randomly select 180 samples for each dataset as the training set and the rest as the independent test set (44 cases for HECKTOR21 and 40 cases for PI-CAI22). Specifically, the training set is further randomly divided into five folds for cross-validation. For quantitative analysis, we use the Dice Similarity Coefficient (**DSC**), the Jaccard Index (**JI**), and the 95% Hausdorff Distance (**HD95**) as evaluation metrics for segmentation performance. A better segmentation will have a smaller HD95 and larger values for DSC and JI. We also conduct holistic t-tests of the overall performance for our method and all baseline models with the two-tailed $p < 0.05$.

3.2 Implementation Details

We use Pytorch to implement our proposed method and the baselines. For a fair comparison, all models are trained from scratch using two NVIDIA A100 GPUs and all comparison methods are implemented with open-source codes, following their original configurations. In particular, we evaluate the 3D and 2D H-DenseFormer on HECKTOR21 and PI-CAI22, respectively. During the training phase, the Adam optimizer is employed to minimize the loss with an initial learning rate of 10^{-3} and a weight decay of 10^{-4}. We use the PolyLR strategy [19] to control the learning rate change. We also use an early stopping strategy with a tolerance of 30 epochs to find the best model within 100 epochs. Online data augmentation, including random rotation and flipping, is performed to alleviate the overfitting problem.

[1] https://pi-cai.grand-challenge.org/.

3.3 Overall Performance

Table 2. Comparison with existing methods on independent test set. We show the mean ± std (standard deviation) scores of averaged over the 5 folds.

Methods (Year)	Params↓	GFLOPs↓	DSC(%) ↑	HD95(mm) ↓	JI(%) ↑
HECKTOR21, two modalities (CT and PET)					
3D U-Net (2016) [7]	12.95M	629.07	68.8 ± 1.4	14.9 ± 2.2	58.0 ± 1.4
UNETR (2022) [16]	95.76M	282.19	59.6 ± 2.5	23.7 ± 3.4	48.2 ± 2.6
Iantsen et al. (2021) [18]	38.66M	1119.75	72.4 ± 0.8	9.6 ± 1.0	60.5 ± 1.1
TransBTS (2021) [31]	30.62M	372.80	64.8 ± 1.0	20.9 ± 3.9	52.9 ± 1.2
3D H-DenseFormer	**3.64M**	**242.96**	**73.9 ± 0.5**	**8.1 ± 0.6**	**62.5 ± 0.5**
PI-CAI22, three modalities (T2W, DWI and ADC)					
Deeplabv3+ (2018) [6]	12.33M	**10.35**	47.4 ± 1.9	48.4 ± 14.3	35.4 ± 1.7
U-Net++ (2019) [37]	15.97M	36.08	49.7 ± 3.9	38.5 ± 6.7	36.9 ± 3.3
ITUNet (2022) [22]	18.13M	32.67	42.1 ± 2.3	67.6 ± 10.3	31.3 ± 1.6
Transunet (2021) [4]	93.23M	72.62	44.8 ± 3.0	59.3 ± 14.8	33.2 ± 2.5
2D H-DenseFormer	**4.25M**	31.46	**49.9 ± 1.2**	**35.9 ± 8.2**	**37.1 ± 1.2**

Table 2 compares the performance and computational complexity of our proposed method with the existing state-of-the-art methods on the independent **test** sets. For HECKTOR21, 3D H-DenseFormer achieves a DSC of 73.9%, HD95 of 8.1mm, and JI of 62.5%, which is a significant improvement ($p < 0.01$) over 3D U-Net [7], UNETR [16], and TransBTS [31]. It is worth noting that the performance of hybrid models such as UNETR is not as good as expected, even worse than 3D U-Net, perhaps due to the small size of the dataset. Moreover, compared to the champion solution of HECKTOR20 proposed by Iantsen et al. [18], our method has higher accuracy and about **10** and **5** times lower amount of network parameters and computational cost, respectively. For PI-CAI22, the 2D variant of H-DenseFormer also outperforms existing methods ($p < 0.05$), achieving a DSC of 49.9%, HD95 of 35.9 mm, and JI of 37.1%. Overall, H-DenseFormer reaches an effective balance of performance and computational cost compared to existing CNNs and hybrid structures. For qualitative analysis, we show a visual comparison of the different methods. It is evident from Fig. 2 that our approach can describe tumor contours more accurately while providing better segmentation accuracy for small-volume targets. These results further demonstrate the effectiveness of our proposed method in multimodal tumor segmentation tasks.

3.4 Parameter Sensitivity and Ablation Study

Impact of DCT Depth. As illustrated in Table 3, the network performance varies with the change in DCT depth. H-DenseFormer achieves the best performance at the DCT depth of **6**. An interesting finding is that although the depth

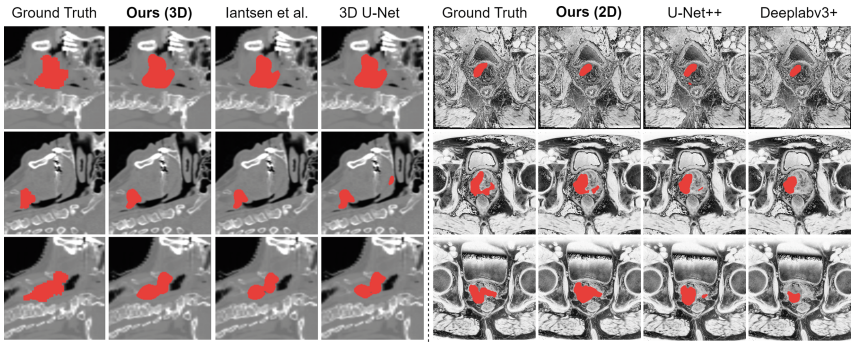

Fig. 2. Visualizations of different models on HECKTOR21 (left) and PI-CAI22 (right).

Table 3. Parameter sensitivity analysis on DCT depth.

DCT Depth	Params↓	GFLOPs↓	DSC (%) ↑	HD95 (mm) ↓	JI (%) ↑
3	3.25M	242.38	73.5 ± 1.4	8.4 ± 0.7	62.2 ± 1.6
6	3.64M	242.96	**73.9 ± 0.5**	**8.1 ± 0.6**	**62.5 ± 0.5**
9	4.03M	243.55	72.7 ± 1.2	8.7 ± 0.6	61.2 ± 1.3

of the DCT has increased from 3 to 9, the performance does not improve or even worsen. We suspect that the reason is over-fitting due to over-parameterization. Therefore, choosing a proper DCT depth is crucial to improve accuracy.

Impact of Different Modules. The above results demonstrate the superiority of our method, but it is unclear which module plays a more critical role in performance improvement. Therefore, we perform ablation experiments on MPE, DCT and DS loss. Specifically, w/o MPE refers to keeping one embedding path, w/o DCT signifies using a standard 12-layer Transformer, and w/o DS loss denotes removing the deep supervision mechanism. As shown in Table 4, the performance decreases with varying degrees when removing them separately, which means all the modules are critical for H-DenseFormer. We can observe that DCT has a greater impact on overall performance than the others, further demonstrating its effectiveness. In particular, the degradation after removing the MPE also con-

Table 4. Ablation study of 3D H-DenseFormer, w/o denotes without.

Method	DSC (%) ↑	HD95 (mm) ↓	JI (%) ↑
3D H-DenseFormer **w/o MPE**	72.1 ± 0.8	10.8 ± 1.1	60.4 ± 0.8
3D H-DenseFormer **w/o DCT**	70.7 ± 1.8	11.9 ± 1.9	58.6 ± 2.1
3D H-DenseFormer **w/o DS loss**	72.2 ± 0.9	10.2 ± 1.0	60.1 ± 1.2
3D H-DenseFormer	**73.9 ± 0.5**	**8.1 ± 0.6**	**62.5 ± 0.5**

firms that decoupling the feature expression of different modalities helps obtain higher-quality multimodal features and improve segmentation performance.

4 Conclusion

In this paper, we proposed an efficient hybrid model (H-DenseFormer) that combines Transformer and CNN for multimodal tumor segmentation. Concretely, a Multi-path Parallel Embedding module and a Densely Connected Transformer block were developed and integrated to balance accuracy and computational complexity. Extensive experimental results demonstrated the effectiveness and superiority of our proposed H-DenseFormer. In future work, we will extend our method to more tasks and explore more efficient multimodal feature fusion methods to further improve computational efficiency and segmentation performance.

References

1. Andrearczyk, V., et al.: Overview of the HECKTOR challenge at MICCAI 2020: automatic head and neck tumor segmentation in PET/CT. In: Andrearczyk, V., Oreiller, V., Depeursinge, A. (eds.) HECKTOR 2020. LNCS, vol. 12603, pp. 1–21. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67194-5_1
2. Cao, H., et al.: Swin-UNet: UNet-like pure transformer for medical image segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) Computer Vision – ECCV 2022 Workshops. ECCV 2022, Part III. LNCS, vol. 13803, pp. 205–218. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25066-8_9
3. Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.-A.: Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: Shen, D., et al. (eds.) MICCAI 2019, Part III 22. LNCS, vol. 11766, pp. 447–456. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_50
4. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Chen, L., Wu, Y., DSouza, A.M., Abidin, A.Z., Wismüller, A., Xu, C.: MRI tumor segmentation with densely connected 3D CNN. In: Medical Imaging 2018: Image Processing, vol. 10574, pp. 357–364. SPIE (2018)
6. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with Atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 833–851. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_49
7. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
8. 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**, 4940593 (2018)

9. Dobko, M., Kolinko, D.I., Viniavskyi, O., Yelisieiev, Y.: Combining CNNs with transformer for multimodal 3D MRI brain tumor segmentation. In: Crimi, A., Bakas, S. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, 27 September 2021, Revised Selected Papers, Part II, pp. 232–241. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-09002-8_21
10. Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ayed, I.B.: HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. IEEE Trans. Med. Imag. **38**(5), 1116–1126 (2018)
11. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Foster, B., Bagci, U., Mansoor, A., Xu, Z., Mollura, D.J.: A review on segmentation of positron emission tomography images. Comput. Bio. Med. **50**, 76–96 (2014)
13. 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. **25**(9), 3507–3516 (2021)
14. Gao, Y., Zhou, M., Metaxas, D.N.: UTNet: a hybrid transformer architecture for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021, Part III. LNCS, vol. 12903, pp. 61–71. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_6
15. Guo, Z., Li, X., Huang, H., Guo, N., Li, Q.: Deep learning-based image segmentation on multimodal medical imaging. IEEE Trans. Radiat. Plasma Med. Sci. **3**(2), 162–169 (2019)
16. Hatamizadeh, A., et al.: UNETR: transformers for 3D medical image segmentation. In: WACV 2022 Proceedings, pp. 574–584 (2022)
17. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR 2017 Proceedings, pp. 4700–4708 (2017)
18. 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. (eds.) HECKTOR 2020. LNCS, vol. 12603, pp. 37–43. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67194-5_4
19. Isensee, F., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
20. Kamnitsas, K., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. **36**, 61–78 (2017)
21. Kamnitsas, K., et al.: Ensembles of multiple models and architectures for robust brain tumour segmentation. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 450–462. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_38
22. Kan, H., et al.: ITUnet: Integration of transformers and UNet for organs-at-risk segmentation. In: EMBC 2022, pp. 2123–2127. IEEE (2022)
23. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV 2017 Proceedings, pp. 2980–2988 (2017)
24. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: ICCV 2021 Proceedings, pp. 10012–10022 (2021)
25. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imag. **34**(10), 1993–2024 (2014)

26. Pereira, S., Pinto, A., Alves, V., Silva, C.A.: Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans. Med. Imag. **35**(5), 1240–1251 (2016)
27. Rodríguez Colmeiro, R.G., Verrastro, C.A., Grosges, T.: Multimodal brain tumor segmentation using 3D convolutional networks. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 226–240. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_20
28. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015, Part III 18. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
29. Saeed, N., Sobirov, I., Al Majzoub, R., Yaqub, M.: TMSS: an end-to-end transformer-based multimodal network for segmentation and survival prediction. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022, Part VII. LNCS, vol. 13437, pp. 319–329. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_31
30. Vaswani, A., et al.: Attention is all you need. In: NIPS 2017, vol. 30 (2017)
31. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: TransBTS: multimodal brain tumor segmentation using transformer. In: de Bruijne, M., et al. (eds.) MICCAI 2021, Part I 24. LNCS, vol. 12901, pp. 109–119. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_11
32. Xiao, X., Lian, S., Luo, Z., Li, S.: Weighted Res-UNet for high-quality retina vessel segmentation. In: ITME 2018, pp. 327–331. IEEE (2018)
33. Zhang, Y., et al.: mmFormer: multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022 Proceedings, Part V, vol. 13435, pp. 107–117. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_11
34. Zhao, X., et al.: A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. Med. Image Anal. **43**, 98–111 (2018)
35. Zhong, Z., et al.: 3D fully convolutional networks for co-segmentation of tumors on PET-CT images. In: ISBI 2018, pp. 228–231. IEEE (2018)
36. Zhou, T., Ruan, S., Canu, S.: A review: deep learning for medical image segmentation using multi-modality fusion. Array **3**, 100004 (2019)
37. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imag. **39**(6), 1856–1867 (2019)