3D-TransUNet for Brain Metastases Segmentation in the BraTS2023 Challenge

Siwei Yang[1], Xianhang Li[1], Jieru Mei[2], Jieneng Chen[2], Cihang Xie[1], and Yuyin Zhou[1]

[1] University of California, Santa Cruz
[2] The Johns Hopkins University

Abstract. Segmenting brain tumors is complex due to their diverse appearances and scales. Brain metastases, the most common type of brain tumor, are a frequent complication of cancer. Therefore, an effective segmentation model for brain metastases must adeptly capture local intricacies to delineate small tumor regions while also integrating global context to understand broader scan features. The TransUNet model, which combines Transformer self-attention with U-Net's localized information, emerges as a promising solution for this task. In this report, we address brain metastases segmentation by training the 3D-TransUNet [6] model on the Brain Tumor Segmentation (BraTS-METS) 2023 challenge dataset. Specifically, we explored two architectural configurations: the **Encoder-only 3D-TransUNet**, employing Transformers solely in the encoder, and the **Decoder-only 3D-TransUNet**, utilizing Transformers exclusively in the decoder. For Encoder-only 3D-TransUNet, we note that Masked-Autoencoder pre-training is required for a better initialization of the Transformer Encoder and thus accelerates the training process.
We identify that the Decoder-only 3D-TransUNet model should offer enhanced efficacy in the segmentation of brain metastases, as indicated by our 5-fold cross-validation on the training set[3]. However, our use of the Encoder-only 3D-TransUNet model already yield notable results, with an average lesion-wise Dice score of 59.8% on the test set, securing second place in the BraTS-METS 2023 challenge.

Keywords: Brain Tumor Segmentation · Transformer

1 Introduction

Tumors, with their subtle intensity variations compared to surrounding tissues, often pose difficulties, as evidenced by inconsistencies in even expert-driven manual annotations[8,24,9,13,1]. Additionally, the wide variance in tumor appearances and dimensions across patients challenges the efficacy of traditional shape and location models[15,19]. Brain metastases, which are brain tumors that originate from primary cancers elsewhere in the body, represent the most prevalent

[3] The code and models are available at https://github.com/Beckschen/3D-TransUNet

malignant tumors in the central nervous system. With an annual incidence of 24 per 100,000 individuals [11,3,17], brain metastases outnumber the occurrence of all primary brain cancers combined.

In terms of segmentation methods, Convolutional Neural Networks (CNNs), especially Fully Convolutional Networks (FCNs)[14], have established their prominence. Among various architectures, the u-shaped architecture, popularly known as U-Net [20], stands out for its symmetrical encoder-decoder framework and skip-connections, excelling at preserving image intricacies. However, these methods often struggle with modeling long-range dependencies due to convolution's inherent locality. To address this, researchers have turned to Transformers, which rely solely on attention mechanisms, showcasing success in capturing global contexts [22]. An example is TransUNet [5], a hybrid CNN-Transformer model, seamlessly blending localized convolution's efficiency with global attention's comprehension. Anchored in the encoder-decoder paradigm, this innovation leverages and elevates both paradigms, promising a new frontier in segmentation precision.

This report aims to validate the performance of 3D-TransUNet [6] on the segmentation of brain metastases in the BraTS 2023 challenge. 3D-TransUNet has two opted self-attention modules: 1) A *Transformer encoder*, which tokenizes image patches extracted from CNN feature maps to capture extensive global contexts using transformer blocks, and 2) A *Transformer decoder*, which innovatively redefines the process of medical image segmentation by treating it as a mask classification task and dynamically refining organ queries through cross-attention with multi-scale CNN decoding features. Specifically, we experiment with two architectures: **Encoder-only** (CNN encoder + Transformer encoder + CNN decoder) and **Decoder-only** (CNN encoder + Transformer decoder + CNN decoder). Notably, Masked-Autoencoder (MAE) Pre-training can be used to accelerate the training of the Encoder-only model. To introduce stronger supervision, we employ deep supervision across all levels of our decoder. Our model yielded average lesion-wise Dice scores of 59.6% and 59.8%, respectively, on the validation set and test set of BraTS-METS 2023 datasets.

2 Method

In this work, we adopt 3D-TransUNet [6] to segment brain metastases. This model leverages the advantages of integrating transformers within the encoder and decoder of the U-Net architecture, as shown in Figure 1. We first studied the encoder to verify if transformer blocks can extract representative features. We also explore combining the U-Net pixel decoder with a Transformer decoder for prediction. The U-Net pixel decoder upsamples the low-resolution features generated by the image encoder. Simultaneously, the Transformer decoder enhances these features through a cross-attention mechanism, effectively refining the final prediction.

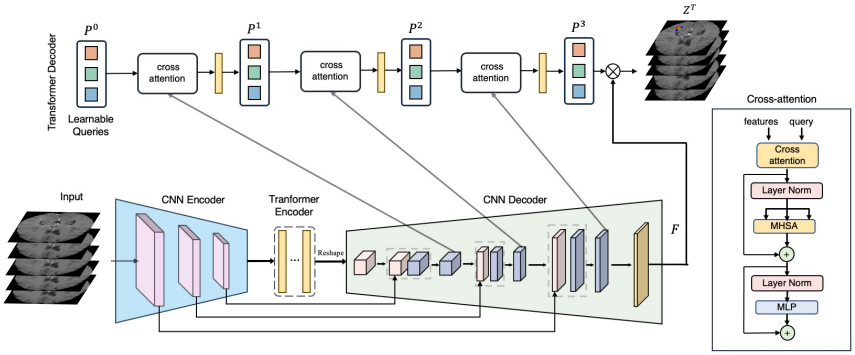

Fig. 1. Overview of our adaptation of 3D-TransUNet [6] for BraTS-METS 2023 [2].

2.1 Transformer as Encoder

Image Sequentialization An input feature map $\mathbf{x}$ are first tokenized and reshaped into a sequence of 3D patches, noted as $\{\mathbf{x}_i^p \in \mathbb{R}^{P^3 \cdot C} | i = 1, .., N\}$. The size of each 3D patch is $P \times P \times P$, and the total patch number is $N = \frac{DHW}{P^3}$.

Patch Embedding 3D patches $\mathbf{x}^p$ are linearly projected into a d_{enc}-dimensional embedding space. Learnable positional embeddings are added to retain spatial information. The final embeddings are formulated as follows:

$$\mathbf{z}_0 = [\mathbf{x}_1^p\mathbf{E}; \mathbf{x}_2^p\mathbf{E}; \cdots ; \mathbf{x}_N^p\mathbf{E}] + \mathbf{E}^{pos}, \tag{1}$$

where $\mathbf{E} \in \mathbb{R}^{(P^2C) \times d}$ and $\mathbf{E}^{pos} \in \mathbb{R}^{N \times d}$ denotes the linear projection and position embedding accordingly.

The Transformer encoder consists of L_{enc} layers of Multi-head Self-Attention (MSA) Equation. (2) and Multi-Layer Perceptron (MLP) blocks Equation. (3). Therefore, the final output $\mathbf{z}_\ell$ of the ℓ-th layer is

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \tag{2}$$
$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \tag{3}$$

where $\text{LN}(\cdot)$ is layer normalization operator and $\mathbf{z}_\ell$ is the encoded image representation.

2.2 Transformer as Decoder

The segmentation task can be reformulated into a binary mask classification problem inspired by the set prediction mechanism proposed in DETR [4]. As

shown in Figure. 1, we train the CNN decoder and the Transformer decoder simultaneously, allowing for the refinement of organ queries and feature maps. Specifically, in the t-th layer of the Transformer decoder, the refined organ queries are denoted by $\mathbf{P}^t \in \mathbb{R}^{N \times d_{dec}}$. Alongside, an intermediate feature from the U-Net is transformed into a d_{dec}-dimensional feature, represented by $\mathcal{F}$. The number of upsampling blocks in the CNN decoder aligns with the Transformer decoder layers, so multi-scale CNN features are effectively projected into the feature space $\mathcal{F} \in \mathbb{R}^{(D_t H_t W_t) \times d_{dec}}$, where D_t, H_t, and W_t define the spatial dimensions of the feature map at the t-th upsampling block. Transitioning from the t-th to the $t+1$-th layer, the organ queries $\mathbf{P}^t$ are updated through the cross-attention mechanism as described by the following formula:

$$\mathbf{P}^{t+1} = \mathbf{P}^t + \text{Softmax}\left((\mathbf{P}^t \mathbf{w}_q)(\mathcal{F}^t \mathbf{w}_k)^\top\right) \mathcal{F}^t \mathbf{w}_v, \tag{4}$$

where $\mathbf{w}_q \in \mathbb{R}^{d_{dec} \times d_q}$, $\mathbf{w}_k \in \mathbb{R}^{d_{dec} \times d_k}$, and $\mathbf{w}_v \in \mathbb{R}^{d_{dec} \times d_v}$ are the weight matrices that linearly project the t-th query features, keys, and values for the subsequent layer. This process is repeated, with a residual connection updating $\mathbf{P}$ after each layer, in line with the previous method ([7]). The final prediction, $\mathbf{Z}^T$, is derived through Equation 5, which details the process of converting $\mathbf{P}^T$ into the binarized segmentation map. It involves a dot product with U-Net's last block feature, $\mathbf{F}$, resulting in $\mathbf{Z}^T$.

$$\mathbf{Z}^T = g(\mathbf{P} \times \mathbf{F}^\top), \tag{5}$$

where $g(\cdot)$ is sigmoid activation followed by a hard thresholding operation with a threshold set at 0.5, such to decode region-wise binary brain tumor masks.

Note that unlike [6], we do not use masked attention here due to the observed training instability.

2.3 3D-TransUNet Variants

Two 3D-TransUNet variants, *i.e.*, Encoder-only and Decoder-only, are involved in the experiment. Encoder-only 3D-TransUNet is used as the main architecture for all of our submissions since Decoder-only 3D-TransUNet requires longer training compared to the Encoder-only 3D-TransUNet. This is mainly due to its use of high-resolution features in the Transformer decoder. Additionally, the Hungarian matching process used to match binary masks to ground truth in the Decoder-only model is slower compared to directly computing cross-entropy and Dice loss.

Encoder-Only The Transformer encoder along with the CNN encoder compose a CNN-Transformer hybrid encoder in this variant. Feature maps are first extracted by CNN then patchified and tokenized before being fed to the Transformer encoder. A standard U-Net is used as the decoder without the Transformer decoder.

Decoder-Only This variant uses a conventional CNN encoder only for the encoding phase while both CNN and Transformer are used as the decoder. Before being processed by the Transformer decoder, they are augmented with learnable positional embeddings following Eq. (1).

2.4 Training Details

Masked-Autoencoder (MAE) Pre-training. For our **Encoder-only 3D-TransUNet**, we first pre-train the transformer encoder in an MAE style [12]. Specifically, We initially tokenized 3D input using a 2D patch embedding layer along the z-axis, then flattened it into a 1D sequence. We randomly mask out 75% of the tokens and then utilize just one lightweight decoder block to predict the masked tokens. Following [12], we calculate the reconstruction loss $L_{\text{recon}} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_{i,\text{masked}})^2$, where L_{recon} is the mean square error, x_i are the original pixel values, $\hat{x}_{i,\text{masked}}$ denotes the predicted pixel values for the masked tokens, and N represents the total count of masked tokens. Following [12], we adopt pixel normalization on each patch. Specifically, we compute the mean and standard deviation of all pixels in a patch and use them to normalize this patch. After pre-training, we discard the patch embedding layer and decoder block and solely initialize all transformer encoder blocks with the pre-trained weights. We observe that utilizing a pre-trained encoder significantly speedup model convergence. For instance, with MAE pre-training, the model attains an average Dice score across five folds of 58.2% with only 300 epochs of training, whereas a model trained from scratch needs 600 epochs to achieve comparable performance.

Training Loss. Unless indicated otherwise, we mainly follow the training details of 3D-TransUNet [6]. In addressing the challenge posed by setting the number of coarse candidates N considerably greater than the class count K, it becomes inevitable that predictions for each class will exhibit false positives. To mitigate this, we employ a post-processing step to refine the coarse candidates, drawing on a matching process between predicted and ground truth segmentation masks. Taking cues from prior work [4,23], we utilize the Hungarian matching approach to establish the correspondence between predictions and ground-truth segments. The resulting matching loss is formulated as follows:

$$\mathcal{L} = \lambda_0(\mathcal{L}_{ce} + \mathcal{L}_{dice)} + \lambda_1\mathcal{L}_{cls}, \tag{6}$$

Here, the pixel-wise losses $\mathcal{L}_{ce}$ and $\mathcal{L}_{dice}$ denote the binary cross-entropy and dice loss [16] respectively, while $\mathcal{L}_{cls}$ represents the classification loss computed using the cross-entropy for each candidate region. The hyper-parameters λ_0 and λ_1 serve to strike a balance between per-pixel segmentation and mask classification loss.

Deep Supervision. To introduce stronger supervision, every intermediate level of the 3D-TransUNet decoder produces a prediction map on which the loss function is applied during training.

3 Experiments and results

3.1 Experimental Setting

Implementation Details. All experiments are conducted with a single NVIDIA A5000. Batch size and base learning rate are set as 2 and 2e-3 accordingly. The learning rate follows polynomial decay with a power factor of 0.9. Augmentation including random rotation, scaling, flipping, white Gaussian noise, Gaussian blurring, color jittering, low-resolution simulation, Gamma transformation. Our main experiments in Table 2 are conducted using Encoder-only 3D-TransUNet. The architecture combines a 3D nn-UNet with a pre-trained 12-layer Vision Transformer (ViT) as the Transformer encoder, utilizing Masked Autoencoder (MAE) weights. The latent dimension d is set at 768. For decoder-only, the number of layers is 3. And d_{dec} is set to 192. For training loss 6, λ_0 and λ_1 are set as 0.7 and 0.3. During the testing phase, we train our model on the entire training set for 600 epochs and submitt the predictions on the validation set. During the testing phase, we applied 10-fold cross-validation, where we trained an individual model for every fold for 1,000 epochs.

MAE pretraining settings. The input has a shape of $128 \times 128 \times 128$ after random cropping. We train all data on 8 GPUs distributedly for our MAE training, with a batch size of 2 on each GPU. We train the model in 4800 epochs, including a warm-up period of 40 epochs. The base learning rate is set to 1.5e-4, accompanied by a weight decay 0.05. AdamW optimizer is used by default.

Datasets. We report results on BraTS-MET 2023 [17] which is pivotal for crafting sophisticated algorithms to detect and segment brain metastases, aiming for easy clinical integration. This dataset encompasses a collection of untreated brain metastases mpMRI scans, sourced from multiple institutions and conducted under regular clinical protocols. It should be noted that we didn't use other officially allowed datasets, *e.g.*, NYUMets [18], UCSF-BMSR [21], BrainMetsShare [10] as these datasets don't share the same mpMRI modalities and annotation format as the BraTS-MET 2023.

It should be noted that we only use BraTS-METS 2023 for training as these datasets share the mpMRI modalities and annotation format.

Evaluation Metrics. In assessing the accuracy of a medical image segmentation model, various metrics offer insights into different aspects of performance:

1. **Lesion-wise Dice Score**: Dice score represents the similarity between two binary segmentations. It is given by:

$$\text{Dice}(\mathbf{A}, \mathbf{B}) = \frac{2|\mathbf{A} \cap \mathbf{B}|}{|\mathbf{A}| + |\mathbf{B}|} \tag{7}$$

where $\mathbf{A}$ and $\mathbf{B}$ denote the binary segmentations.
In this challenge, dice scores localized to individual lesions are adopted for lesion-specific evaluation. The Lesion-wise Dice for a designated lesion is:

$$\text{Dice}_{\text{lesion-wise}}(\mathbf{A}_i, \mathbf{B}_i) = \frac{2|\mathbf{A}_i \cap \mathbf{B}_i|}{|\mathbf{A}_i| + |\mathbf{B}_i|} \tag{8}$$

Method	Lesion-wise Dice (↑)				HD95 (↓)			
	ET	TC	WT	Avg.	ET	TC	WT	Avg.
Encoder-only 3D-TransUNet [6]	54.79%	58.96%	56.05%	56.60%	108.9	107.6	109.5	108.7
Decoder-only 3D-TransUNet [6]	56.80%	61.12%	60.09%	59.34%	99.4	95.9	93.97	96.4

Table 1. Ablation Performance of 3D-TransUNet on the training set of BraTS-METS 2023 [17] under 5-fold cross-validation.

Dataset Split	Lesion-wise Dice (↑)				HD95 (↓)			
	ET	TC	WT	Avg.	ET	TC	WT	Avg.
Validation	59.2%	63.4%	56.5%	59.6%	94.8	94.8	110.9	100.1
Test	57.4%	62.0%	59.9%	59.8%	103.0	99.8	99.6	100.8

Table 2. Performance of Encoder-only 3D-TransUNet on the validation and test set of BraTS-METS 2023 [17].

This formula assesses the overlap between the i^{th} ground-truth lesion $\mathbf{B}_i$ and all the lesions that overlap with it $\mathbf{A}_i$.

2. **Hausdorff Distance (95%)**: A metric quantifying the maximum of minimum distances between two binary images at the 95-th percentile to mitigate outlier effects. Mathematically:

$$H(A,B) = \max\left\{ \sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{a\in A} d(a,b) \right\} \tag{9}$$

Model Ensemble and Test-Time Augmentation During testing phase, five models are randomly chosen from ten models trained on each fold for ensemble. Predictions from these five models were averaged to ensemble predictions. To further boost the model's performance with test-time augmentation, predictions from augmented views with flipping and rotation (90°, 180°, 270°) are averaged to produce the final predictions.

3.2 Encoder-only v.s. Decoder-only

In order to compare the effectiveness between the Encoder-only model and the Decoder-only model, we apply 5-fold cross-validation on the entire 238 training cases and report the average Dice and HD95 for all testing cases. The 5 models from the 5 folds of BraTS-METS 2023 training set are trained for 200 epochs. In Table 1, we report the comparison between Encoder-only 3D-TransUNet and Decoder-only 3D-TransUNet on BraTS-METS 2023's training set with 5-fold cross-validation are presented in Table 1. Compared to our internal validation results with the baseline nnUNet, which yielded average Dice scores of 54.90%, 58.67%, and 55.75% for segmenting ET, TC, and WT, respectively, resulting in an overall average Dice score of 56.44%, it becomes evident

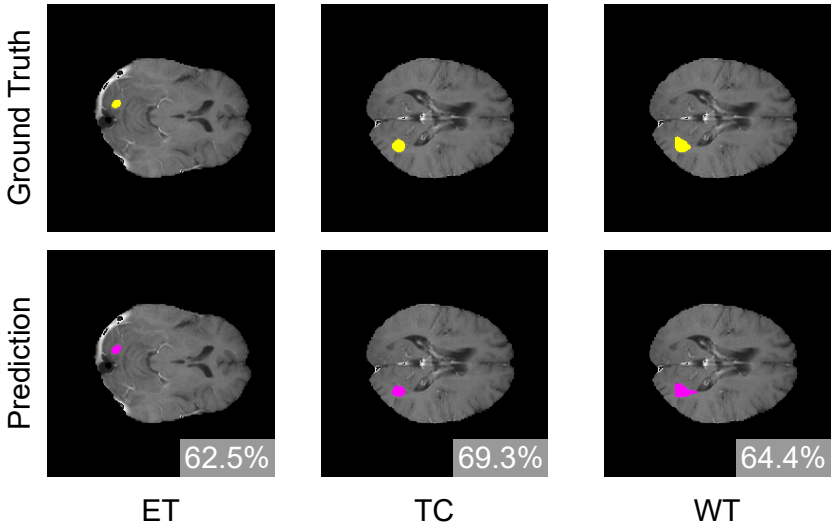

Fig. 2. Visual comparison between ground-truths and predictions from Encoder-only 3D-TransUNet on BraTS-METS 2023. The lesion-wise Dice score of each category on this patient is also listed (best viewed in color).

that while the Encoder-only model offers only marginal improvement in segmentation, the Decoder-only model demonstrates a substantial increase in Dice score by 2.9%. It is important to note that MAE pretraining was not applied to the Encoder-only model in this evaluation. However, with MAE pretraining, the advantages of the Encoder-only model are expected to be more pronounced, albeit still inferior to the Decoder-only model.

3.3 Main Results

Since the Decoder-only 3D-TransUNet requires longer training, due to the time and computation limit, we were only able to submit results from the Encoder-only 3D-TransUNet during the validation phase and testing phase, where the official evaluation results are presented in Table 2. Specifically, we achieve average lesion-wise Dice scores of 59.6% and 59.8% on the validation and test set, securing the second place in the BraTS 2023 challenge. We also display a qualitative example to further demonstrate our method's effectiveness, as shown in Fig. 2.

4 Conclusion

Brain tumors, especially brain metastases, present challenges in segmentation due to their diverse appearances and sizes. The TransUNet model, combining

Transformer self-attention and U-Net's features, shows promise for this task. We trained the 3D-TransUNet model on the BraTS-METS 2023 dataset for brain metastases segmentation, exploring Encoder-only and Decoder-only configurations. Pre-training the Encoder-only model with Masked-Autoencoder improves initialization, facilitating faster training. Although the Decoder-only model is expected to perform better, the Encoder-only model achieved notable results in a shorter timeframe, securing second place in the BraTS-METS 2023 challenge with a 59.8% average lesion-wise Dice score on the test set.

References

1. Adewole, M., Rudie, J.D., Gbadamosi, A., Toyobo, O., Raymond, C., Zhang, D., Omidiji, O., Akinola, R., Suwaid, M.A., Emegoakor, A., Ojo, N., Aguh, K., Kalaiwo, C., Babatunde, G., Ogunleye, A., Gbadamosi, Y., Iorpagher, K., Calabrese, E., Aboian, M., Linguraru, M., Albrecht, J., Wiestler, B., Kofler, F., Janas, A., LaBella, D., Kzerooni, A.F., Li, H.B., Iglesias, J.E., Farahani, K., Eddy, J., Bergquist, T., Chung, V., Shinohara, R.T., Wiggins, W., Reitman, Z., Wang, C., Liu, X., Jiang, Z., Familiar, A., Leemput, K.V., Bukas, C., Piraud, M., Conte, G.M., Johansson, E., Meier, Z., Menze, B.H., Baid, U., Bakas, S., Dako, F., Fatade, A., Anazodo, U.C.: The brain tumor segmentation (brats) challenge 2023: Glioma segmentation in sub-saharan africa patient population (brats-africa) (2023) 1
2. Adewole, M., Rudie, J.D., Gbadamosi, A., Toyobo, O., Raymond, C., Zhang, D., Omidiji, O., Akinola, R., Suwaid, M.A., Emegoakor, A., et al.: The brain tumor segmentation (brats) challenge 2023: Glioma segmentation in sub-saharan africa patient population (brats-africa). arXiv preprint arXiv:2305.19369 (2023) 3
3. Boire, A., Brastianos, P.K., Garzia, L., Valiente, M.: Brain metastasis. Nature Reviews Cancer **20**(1), 4–11 (2020) 2
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020) 3, 5
5. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021) 2
6. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M., Xing, L., Lu, L., Yuille, A.L., Zhou, Y.: 3d transunet: Advancing medical image segmentation through vision transformers. arXiv preprint arXiv:2310.07781 (2023) 1, 2, 3, 4, 5, 7
7. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022) 4
8. Fathi Kazerooni, A., Arif, S., Madhogarhia, R., Khalili, N., Haldar, D., Bagheri, S., Familiar, A.M., Anderson, H., Haldar, S., Tu, W., et al.: Automated tumor segmentation and brain tissue extraction from multiparametric mri of pediatric brain tumors: A multi-institutional study. Neuro-Oncology Advances **5**(1), vdad027 (2023) 1

9. Greenwald, N.F., Miller, G., Moen, E., Kong, A., Kagel, A., Dougherty, T., Fullaway, C.C., McIntosh, B.J., Leow, K.X., Schwartz, M.S., et al.: Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. Nature biotechnology **40**(4), 555–565 (2022) 1
10. Grøvik, E., Yi, D., Iv, M., Tong, E., Rubin, D., Zaharchuk, G.: Deep learning enables automatic detection and segmentation of brain metastases on multisequence mri. Journal of Magnetic Resonance Imaging **51**(1), 175–182 (2020) 6
11. Habbous, S., Forster, K., Darling, G., Jerzak, K., Holloway, C.M., Sahgal, A., Das, S.: Incidence and real-world burden of brain metastases from solid tumors and hematologic malignancies in ontario: a population-based study. Neuro-oncology advances **3**(1), vdaa178 (2021) 2
12. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022) 5
13. Kazerooni, A.F., Khalili, N., Liu, X., Haldar, D., Jiang, Z., Anwar, S.M., Albrecht, J., Adewole, M., Anazodo, U., Anderson, H., Bagheri, S., Baid, U., Bergquist, T., Borja, A.J., Calabrese, E., Chung, V., Conte, G.M., Dako, F., Eddy, J., Ezhov, I., Familiar, A., Farahani, K., Haldar, S., Iglesias, J.E., Janas, A., Johansen, E., Jones, B.V., Kofler, F., LaBella, D., Lai, H.A., Leemput, K.V., Li, H.B., Maleki, N., McAllister, A.S., Meier, Z., Menze, B., Moawad, A.W., Nandolia, K.K., Pavaine, J., Piraud, M., Poussaint, T., Prabhu, S.P., Reitman, Z., Rodriguez, A., Rudie, J.D., Sanchez-Montano, M., Shaikh, I.S., Shah, L.M., Sheth, N., Shinohara, R.T., Tu, W., Viswanathan, K., Wang, C., Ware, J.B., Wiestler, B., Wiggins, W., Zapaishchykova, A., Aboian, M., Bornhorst, M., de Blank, P., Deutsch, M., Fouladi, M., Hoffman, L., Kann, B., Lazow, M., Mikael, L., Nabavizadeh, A., Packer, R., Resnick, A., Rood, B., Vossough, A., Bakas, S., Linguraru, M.G.: The brain tumor segmentation (brats) challenge 2023: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds) (2024) 1
14. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015) 2
15. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024) 1
16. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV (2016) 5
17. Moawad, A.W., Janas, A., Baid, U., Ramakrishnan, D., Jekel, L., Krantchev, K., Moy, H., Saluja, R., Osenberg, K., Wilms, K., et al.: The brain tumor segmentation (brats-mets) challenge 2023: Brain metastasis segmentation on pre-treatment mri. arXiv preprint arXiv:2306.00838 (2023) 2, 6, 7
18. Oermann, E., Link, K., Schnurman, Z., Liu, C., Kwon, Y.J.F., Jiang, L.Y., Nasir-Moin, M., Neifert, S., Alzate, J., Bernstein, K., et al.: Longitudinal deep neural networks for assessing metastatic brain cancer on a massive open benchmark. (2023) 6
19. Renard, F., Guedria, S., Palma, N.D., Vuillerme, N.: Variability and reproducibility in deep learning for medical image segmentation. Scientific Reports **10**(1), 13724 (2020) 1
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) 2

21. Rudie, J.D., Weiss, R.S.D.A., Nedelec, P., Calabrese, E., Colby, J.B., Laguna, B., Mongan, J., Braunstein, S., Hess, C.P., Rauschecker, A.M., et al.: The university of california san francisco, brain metastases stereotactic radiosurgery (ucsf-bmsr) mri dataset. arXiv preprint arXiv:2304.07248 (2023) 6
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017) 2
23. Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Max-deeplab: End-to-end panoptic segmentation with mask transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5463–5474 (2021) 5
24. Wang, S., Li, C., Wang, R., Liu, Z., Wang, M., Tan, H., Wu, Y., Liu, X., Sun, H., Yang, R., et al.: Annotation-efficient deep learning for automatic medical image segmentation. Nature communications **12**(1), 5915 (2021) 1