3D TransUNet: Advancing Medical Image Segmentation through Vision Transformers

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew Lungren, Lei Xing, Le Lu, Alan Yuille, Yuyin Zhou

arXiv:2310.07781v1 [cs.CV] 11 Oct 2023

Abstract—**Medical image segmentation plays a crucial role in advancing healthcare systems for disease diagnosis and treatment planning. The u-shaped architecture, popularly known as U-Net, has proven highly successful for various medical image segmentation tasks. However, U-Net's convolution-based operations inherently limit its ability to model long-range dependencies effectively.**

To address these limitations, researchers have turned to Transformers, renowned for their global self-attention mechanisms, as alternative architectures. One popular network is our previous TransUNet, which leverages Transformers' self-attention to complement U-Net's localized information with the global context. In this paper, we extend the 2D TransUNet architecture to a 3D network by building upon the state-of-the-art nnU-Net architecture, and fully exploring Transformers' potential in both the encoder and decoder design. We introduce two key components: 1) A *Transformer encoder* that tokenizes image patches from a convolution neural network (CNN) feature map, enabling the extraction of global contexts, and 2) A *Transformer decoder* that adaptively refines candidate regions by utilizing cross-attention between candidate proposals and U-Net features. Our investigations reveal that different medical tasks benefit from distinct architectural designs. The Transformer encoder excels in multi-organ segmentation, where the relationship among organs is crucial. On the other hand, the Transformer decoder proves more beneficial for dealing with small and challenging segmented targets such as tumor segmentation. Extensive experiments showcase the significant potential of integrating a Transformer-based

encoder and decoder into the u-shaped medical image segmentation architecture. TransUNet outperforms competitors in various medical applications, including multi-organ segmentation, pancreatic tumor segmentation, and hepatic vessel segmentation. It notably surpasses the top solution in the BraTS2021 challenge. Code and models are available at `https://github.com/Beckschen/3D-TransUNet`.

I. INTRODUCTION

Convolutional neural networks (CNNs), particularly fully convolutional networks (FCNs) [1], have risen to prominence in the domain of medical image segmentation. Among their various iterations, the U-Net model [2], characterized by its symmetric encoder-decoder design augmented with skip-connections for improved detail preservation, stands out as the preferred choice for many researchers. Building on this methodology, remarkable progress has been witnessed across diverse medical imaging tasks. These advancements encompass cardiac segmentation in magnetic resonance (MR) imaging [3], organ delineation using computed tomography (CT) scans [4]–[7], and polyp segmentation in colonoscopy recordings [8].

Despite CNNs' unparalleled representational capabilities, they often falter when modeling long-range relationships due to the inherent locality of convolution operations. This limitation becomes particularly pronounced in cases with large inter-patient variations in texture, shape, and size. Recognizing this limitation, the research community has been increasingly drawn to Transformers, models built entirely upon attention mechanisms due to their innate prowess in capturing global contexts [9]. In the realm of medical image segmentation, our prior work with TransUNet [10] stands as a testament to the potential of transformers. However, a pivotal observation from our research indicates that simply substituting a CNN encoder with a Transformer can lead to suboptimal outcomes. Transformers process inputs as 1D sequences and prioritize global context modeling, inadvertently producing features of low resolution. Directly upsampling such features fails to reintroduce the lost granularity. In contrast, a hybrid approach combining CNN and Transformer encoders seems more promising. It effectively harnesses the high-resolution spatial details from CNNs while also benefiting from the global context provided by Transformers.

In this study, we extend the original 2D TransUNet architecture to a 3D configuration, delving deeper into the strategic incorporation of Transformers in both encoding and decoding processes. This leap is rooted in the prowess of the nnU-Net framework, with a vision to surpass its established standards. Our 3D TransUNet unfolds through two primary mechanisms: Firstly, the Transformer Encoder tokenizes image patches from CNN feature maps, allowing a seamless fusion of global self-attentive features with high-resolution CNN features skipped from the encoding path, for enabling precise localization. Secondly, the Transformer Decoder redefines conventional per-pixel segmentation as a mask classification, framing prediction candidates as learnable queries. Specifically, these queries are progressively refined by synergizing cross-attention with localized multi-scale CNN features. In addition, we introduce coarse-to-fine attention refinement in the Transformer decoder where for each segmentation class, an initial candidate set is meticulously refined using attention mechanisms focused on the prediction's foreground, ensuring that each iterative refinement sets a new standard for the subsequent, culminating in continuously improved segmentation accuracy.

By integrating Transformers in the encoder and decoder components of the U-Net-like architectures, we show that our designs allow the framework to preserve the advantages of Transformers while enhancing medical image segmentation. Intriguingly, multi-organ segmentation, which leans heavily on global context information—such as the interplay among diverse abdominal organs—tends to gravitate towards the Transformer encoder design. Conversely, tasks like small target segmentation, such as tumor detection, generally benefit more from the Transformer decoder design. Our extensive experiments demonstrate the superior performance of our method compared to competing approaches across various medical image segmentation tasks. In summary, our contributions can be summarized as follows:

- We introduce a Transformer-centric medical image segmentation framework, incorporating self-attention within the sequence-to-sequence prediction context, applicable to both 2D and 3D medical image segmentation tasks.
- We thoroughly investigate the effects of integrating vision transformers into the encoder and the decoder of the u-shaped segmentation architectures, providing insights on tailoring designs to cater to distinct medical image segmentation challenges.
- We achieve the state-of-the-art results on various medical image segmentation tasks, and release our codebase to encourage further exploration in applying Transformers to medical applications.

II. RELATED WORKS

Combining CNNs with self-attention mechanisms. Various studies have attempted to integrate self-attention mechanisms into CNNs by modeling global interactions of all pixels based on the feature maps. For instance, Wang *et al.* designed a non-local operator, which can be plugged into multiple intermediate convolution layers [11]. Built upon the encoder-decoder u-shaped architecture, Schlemper *et al.* [12] proposed

additive attention gate modules which are integrated into the skip-connections. Different from these approaches, we employ Transformers for embedding global self-attention in our method.

Transformers. Transformers were first proposed by [9] for machine translation and established state-of-the-arts in many NLP tasks. To make Transformers also applicable for computer vision tasks, several modifications have been made. For instance, Parmar *et al.* [13] applied the self-attention only in local neighborhoods for each query pixel instead of globally. Child *et al.* [14] proposed Sparse Transformers, which employ scalable approximations to global self-attention. Recently, Vision Transformer (ViT) [15] achieved state-of-the-art on ImageNet classification by directly applying Transformers with global self-attention to full-sized images. To the best of our knowledge, our TransUNet originally proposed in 2021 is the first Transformer-based medical image segmentation framework, which builds upon the highly successful ViT. Based on TransUNet, nnformer [16] improves the methodology by interleaving convolution with self-attention. In another development, CoTR [17] offers a more efficient self-attention in the Transformer encoder. Another Transformer architecture is the Swin Transformer [18] a hierarchical vision transformer that employs shifted windows to capture local and global information for efficient and scalable visual processing. Subsequent models such as SwinUNet [19] and SwinUETR [20] have been developed for medical image segmentation.

Mask classification for segmentation. DETR [21] is the first work that uses Transformer as a decoder with learnable object queries for object detection. In the context of recent advancements in transformers [22]–[27], a novel variation known as mask Transformers has emerged. This variant introduces segmentation predictions by employing a collection of query embeddings to represent object and its associated mask. Wang *et al.* [23] first develops mask Transformer with memory embedding, and Cheng *et al.* [24] further formulate the query update in a manner of DETR [21]. At the core of mask transformers lies the decoder, which is responsible for processing object queries as input and progressively transforming them into mask embedding vectors [24]–[27]. This process enables the model to effectively handle segmentation tasks and produce accurate results.

III. METHOD

Given a 3D medical image (*e.g.*, CT/MR scan) $\mathbf{x} \in \mathbb{R}^{D \times H \times W \times C}$ with the spatial resolution of $D \times H \times W$ and C number of channels. Our goal is to predict the corresponding pixel-wise labelmap with size $D \times H \times W$. The most common way is to directly train a CNN (*e.g.*, U-Net) to first encode images into high-level feature representations, which are then decoded back to the full spatial resolution. Our approach diverges from conventional methods by thoroughly exploring the attention mechanisms utilized in both the encoder and decoder phases of standard U-shaped segmentation architectures, employing Transformers. In Section III-A, we delve into the direct application of Transformers for encoding feature representations from segmented image patches. Following this,

in Section III-B, we elaborate on the implementation of the query-based Transformer, which serves as our decoder. The detailed architecture of TransUNet is then presented in Section III-C.

A. Transformer as Encoder

Image sequentialization. Following [15], we first perform tokenization by reshaping the input $\mathbf{x}$ into a sequence of flattened 3D patches $\{\mathbf{x}_i^p \in \mathbb{R}^{P^3 \cdot C} | i = 1, .., N\}$, where each patch is of size $P \times P \times P$ and $N = \frac{DHW}{P^3}$ is the number of image patches (*i.e.*, the input sequence length).

Patch embedding. We map the vectorized patches $\mathbf{x}^p$ into a latent d_{enc}-dimensional embedding space using a trainable linear projection. To encode the patch spatial information, we learn specific position embeddings which are added to the patch embeddings to retain positional information as follows:

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

where $\mathbf{E} \in \mathbb{R}^{(P^3 \cdot C) \times d_{enc}}$ is the patch embedding projection, and $\mathbf{E}^{pos} \in \mathbb{R}^{N \times d_{enc}}$ denotes the position embedding.

Each Transformer layer consists of Multihead Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks (Eq. (2)(3)). Therefore the output of the ℓ-th layer can be written as follows:

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

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \quad (3)$$

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

B. Transformer as Decoder

1) Coarse candidate estimation: Traditional approaches, such as U-Net, predominantly view medical image segmentation as a per-pixel classification task. In this paradigm, each pixel is classified into one of the possible K categories, typically achieved by training a segmentation model with the per-pixel cross-entropy (or negative log-likelihood) loss.

Instead of considering individual pixels, our approach in this paper treats medical image segmentation as a mask classification problem. We introduce the concept of an "organ query", a d_{dec}-dimensional feature vector that represents each organ in the image. With a predefined set of N organ queries, our goal for an image comprising K segmentation classes is to segregate the image into N distinct candidate regions. Subsequently, we aim to assign the corresponding organ label to each region. Importantly, it is worth noting that the value of N does not have to align with the number of classes, as demonstrated in prior studies [22]. In fact, we intentionally set N to be significantly larger than K, to minimize the risk of false negatives. Assume the dimension of the object queries is d_{dec}, the coarse predicted segmentation map can be computed by the dot product between the initial organ queries $\mathbf{P}^0 \in \mathbb{R}^{N \times d_{dec}}$ and the embedding of the U-Net last block feature $\mathbf{F} \in \mathbb{R}^{D \times H \times W \times d_{dec}}$:

$$\mathbf{Z}^0 = g(\mathbf{P}^0 \times \mathbf{F}^\top), \quad (4)$$

where $g(\cdot)$ is sigmoid activation followed by a hard thresholding operation with a threshold set at 0.5. As shown in Eq. (4), $\mathbf{P}$ can be seen as the $1 \times 1 \times 1$ convolutional kernel in the standard U-Net segmentation head.

2) Transformer decoder: Figure 1 illustrates our customized 3D Transformer decoder designed specifically to refine organ queries, thereby enhancing the coarse prediction $\mathbf{Z}^0$. Similar to the structure seen in the Transformer encoder (detailed in Section III-A), the self-attention mechanism (*i.e.*, the MSA block) in each layer will enable the Transformer decoder to comprehensively engage with image features and capture organ query interrelations. Recognizing the rich localization in intermediate CNN features, which complements the Transformer's global image context, we refine organ queries in each decoder layer by integrating cross-attention with localized multi-scale CNN features.

Our strategy involves concurrent training of a CNN decoder and the Transformer decoder. In the t-th layer, the refined organ queries are denoted as $\mathbf{P}^t \in \mathbb{R}^{N \times d_{dec}}$. Simultaneously, an intermediate U-Net feature is mapped to a d_{dec}-dimensional feature space denoted as $\mathcal{F}$ to facilitate cross-attention computation. Notably, when the count of upsampling blocks aligns with the Transformer decoder layers, multi-scale CNN features can be 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 specify the spatial dimensions of the feature map at the t-th upsampling block. Moving to the $t+1$-th layer, organ queries are updated using cross-attention as follows:

$$\mathbf{P}^{t+1} = \mathbf{P}^t + \text{Softmax}((\mathbf{P}^t \mathbf{w}_q)(\mathcal{F}\mathbf{w}_k)^\top) \times \mathcal{F}\mathbf{w}_v, \quad (5)$$

where the t-th query features undergo linear projection to form queries for the next layer using the weight matrix $\mathbf{w}_q \in \mathbb{R}^{d_{dec} \times d_q}$. The U-Net feature, $\mathcal{F}$, is similarly transformed into keys and values using parametric weight matrices $\mathbf{w}_k \in \mathbb{R}^{d_{dec} \times d_k}$ and $\mathbf{w}_v \in \mathbb{R}^{d_{dec} \times d_v}$. Note a residual path is used for updating $\mathbf{P}$ following previous studies [25]. Next, we will introduce how to incorporate a coarse-to-fine attention refinement to further enhance the accuracy of segmentation results.

3) Coarse-to-fine attention refinement: The value of coarse-to-fine refinement in medical image segmentation, particularly for small target segmentation, is well-established [4], [28], [29]. This technique employs a coarse mask from an initial stage to guide subsequent refinements. Here, to integrate a seamless coarse-to-fine refinement process within the Transformer decoder, we have incorporated a mask attention module [25]. This enhancement aims to ground the cross-attention within the foreground region based on the former coarse prediction for each category, to reduce the background noise and better focus on the region of interest. This improved attention map iteratively aids subsequent, finer segmentation stages.

Concretely, we start by setting the organ queries and the coarse-level mask prediction as $\mathbf{P}^0$ and $\mathbf{Z}^0$ (based on Eq. (4)) respectively, and then begin the iterative refinement process. At the t-th iteration, using the current organ query features $\mathbf{P}^t$ and coarse prediction $\mathbf{Z}^t$, we compute the masked cross-attention, which refines $\mathbf{P}^{t+1}$ for the subsequent iteration. This

---

Algorithm 1: Iterative coarse-to-fine refinement

Input : Parametric weight matrices $\mathbf{w}_q, \mathbf{w}_k, \mathbf{w}_v$;
　　　　Organ embedding $\mathbf{P}$, U-Net last feature $\mathbf{F}$;
　　　　The U-Net t-th layer feature $\mathcal{F}$;
　　　　Max number of iterations T;

Output: Fine segmentation map $\mathbf{Z}^T$, predicted
　　　　class label $\hat{\mathbf{y}}$;

1 $t \leftarrow 0$;
2 $\mathbf{P}^0 \leftarrow \mathbf{P}$;
3 $\mathbf{Z}^0 \leftarrow g(\mathbf{P}^0 \times \mathbf{F}^\top)$;
4 **repeat**
5 　　Update $\mathbf{P}^{t+1}$ according to Eq. (6);
6 　　Update $\mathbf{Z}^{t+1} \leftarrow g(\mathbf{P}^{t+1} \times \mathbf{F}^\top)$;
7 　　$t \leftarrow t+1$;
8 **until** $t = T$;
9 Compute the class label $\hat{\mathbf{y}}$ by Eq. (8) and Eq. (9);
Return: $\mathbf{Z}^T, \hat{\mathbf{y}}$.

---

computation incorporates the existing coarse prediction $\mathbf{Z}^t$ into the affinity matrixn, as detailed in Eq.(5):

$$\mathbf{P}^{t+1} = \mathbf{P}^t + \mathrm{Softmax}((\mathbf{P}^t\mathbf{w}_q)(\mathcal{F}\mathbf{w}_k)^\top + h(\mathbf{Z}^t)) \times \mathcal{F}\mathbf{w}_v, \tag{6}$$

where

$$h\left(\mathbf{Z}^t(i,j,s)\right) = \begin{cases} 0 & \text{if } \mathbf{Z}^t(i,j,s) = 1 \\ -\infty & \text{otherwise} \end{cases} \tag{7}$$

where i, j, s are the coordinate indices. This formula restricts the cross-attention mechanism to focus solely on the foreground, nullifying it for all other regions. By iteratively updating both the organ queries and the corresponding mask predictions, our Transformer decoder systematically refines the segmentation results across multiple iterations. A detailed description of this iterative process is outlined in Algorithm 1. The refinement cycle persists until the iteration count t reaches the maximum threshold T, which is equivalent to the number of layers in the Transformer decoder.

Fine segmentation decoding. After the final iteration, the updated organ queries $\mathbf{P}^T$ can be decoded back to the finalized refined binarized segmentation map $\mathbf{Z}^T$ by the dot product with U-Net's last block feature $\mathbf{F}$, following Eq. (4). To associate each binarized mask with one semantic class, we further use a linear layer with weight matrices $\mathbf{w}_{fc} \in \mathbb{R}^{d \times K}$ that projects the refined organ embedding $\mathbf{P}^T$ to the output class logits $\mathbf{O} \in \mathbb{R}^{N \times K}$. Formally, we have:

$$\mathbf{O} = \mathbf{P}^T \mathbf{w}_{fc} \tag{8}$$

$$\hat{\mathbf{y}} = \arg\max_{k=0,1,\ldots,K-1} \mathbf{O} \tag{9}$$

where k is the label index. The final class labels associated with the refined predicted masks $\mathbf{Z}^T$ is $\hat{\mathbf{y}} \in \mathbb{R}^N$.

C. TransUNet

Our 3D TransUNet is built upon the state-of-the-art nnU-Net architecture, with the aim of surpassing its established standards. We illustrate the overall framework of the proposed TransUNet in Figure 1. We instantiate our 3D TransUNet with three distinct configurations:

1) Encoder-only: **A CNN-Transformer hybrid encoder** is employed where CNN is first used as a feature extractor to generate a feature map for the input. Patch embedding is applied to feature patches instead of from raw images. For the decoding phase, we use a standard U-Net decoder. We choose this design since 1) it allows us to leverage the intermediate high-resolution CNN feature maps in the decoding path; and 2) we find that the hybrid CNN-Transformer encoder performs better than simply using a pure Transformer as the encoder. The Encoder-only model will be trained using a hybrid segmentation loss consisting of pixel-wise cross entropy loss and dice loss.

2) Decoder-only: In this configuration, we use a conventional CNN encoder for the encoding phase. As for the decoding phase, we use a **CNN-Transformer hybrid decoder** in the segmentation model. Initially, the organ queries $\mathbf{P}$ are set to zero. Before being processed by the Transformer decoder, they are augmented with learnable positional embeddings following Eq. (1). Then as aforementioned in Section III-B, $\mathbf{P}$ will be gradually refined conditioned on the U-Net features and be decoded back into the full-resolution segmentation map. We train the network with the Hungarian matching loss following previous works [21], [23] to update the organ queries throughout the decoding layers. This loss aims to match pairs between predictions and ground-truth segments. It combines pixel-wise classification loss and binary mask loss for each segmented prediction:

$$\mathcal{L} = \lambda_0(\mathcal{L}_{ce} + \mathcal{L}_{dice)} + \lambda_1 \mathcal{L}_{cls}, \tag{10}$$

where the pixel-wise classification loss $\mathcal{L}_{ce}$ and $\mathcal{L}_{dice}$ denote binary cross-entropy loss and dice loss, respectively [30]. The classification loss $\mathcal{L}_{cls}$ is instantiated by the cross-entropy loss for each candidate region. λ_0 and λ_1 are the hyper-parameters for balancing the per-pixel segmentation loss and the mask classification loss.

We also employ deep supervision, applying the training loss to the output at each stage of the TransUNet decoder.

3) Encoder + Decoder: Here, we integrate both the Transformer encoder and the Transformer decoder into the 3D nnU-Net model. And then similar to the decoder-only model, here we also use the Hungarian matching loss to train the whole network.

Lastly, we would like to note that our method, though built upon the 3D nnU-Net, can be easily modified to fit 2D tasks by simply switching the backbone model and reducing all operations back to 2D.

IV. Experiments and Discussion

A. Dataset and Evaluation

Synapse multi-organ segmentation dataset[1]. We use the 30 abdominal CT scans in the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge, with 3779 axial contrast-enhanced abdominal clinical CT images in total.

Each CT volume consists of $85 \sim 198$ slices of 512×512 pixels, with a voxel spatial resolution of $([0.54 \sim 0.54] \times$

---

[1]https://www.synapse.org/#!Synapse:syn3193805/wiki/217789
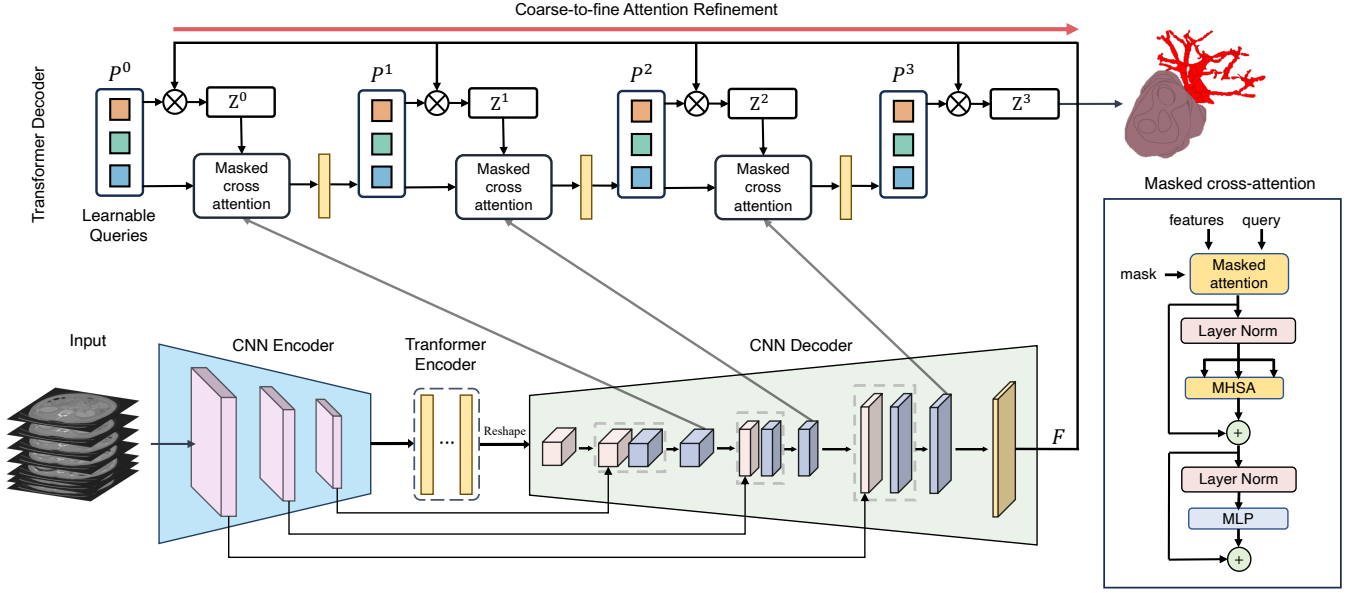
Fig. 1. Overview of 3D TransUNet. We develop two opted architectural components: 1) The Transformer encoder where a CNN encoder is firstly used for local image feature extraction, followed by a pure Transformer encoder for global information interaction; and 2) the Transformer decoder that reframes per-pixel segmentation as mask classification using learnable queries, which are refined through cross-attention with CNN features, and employs a coarse-to-fine attention refinement approach for enhanced segmentation accuracy.

$[0.98 \sim 0.98] \times [2.5 \sim 5.0])\text{mm}^3$. Following [31], we report the average DSC and average Hausdorff Distance (HD) on 8 abdominal organs (aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, spleen, stomach with a random split of 18 training cases (2212 axial slices) and 12 cases for validation, following the split setting in [10].

BraTS2021 brain tumor segmentation challenge[2]. BrasTS2021 Challenge is the most recent and largest dataset for brain tumor segmentation. 1251 multi-parametric MRI scans were provided with segmentation labels to the participants. 4 contrasts are available for the MRI scans: Native T1-weighted image, post-contrast T1-weighted (T1Gd), T2-weighted, and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR). Annotation were manually performed by one to four raters, with final approval from experienced neuro-radiologists. The labels include regions of GD-enhancing tumor (ET), the peritumoral edematous/invaded tissue (ED), and the necrotic tumor core (NCR). All MRI scans were pre-processed by co-registration to the same anatomical template, interpolation to isotropic 1mm^3 resolution and skull-stripping. The image sizes of all MRI scans and associated labels are $240 \times 240 \times 155$. In our experiments, we apply 5-fold cross-validation with the same data split used by the No.1 solution [32] in BraTS2021 challenge.

Medical Segmentation Decathlon (MSD) HepaticVessel[3]. The MSD HepaticVessel, a task of Medical Segmentation Decathlon [33], consists of 443 portal venous phase CT scans obtained from patients with a variety of primary and metastatic liver tumors. The corresponding target ROIs were the vessels and tumors within the liver. This data set was

selected due to the tubular and connected nature of hepatic vessels neighboring heterogeneous tumors. We apply 5-fold cross-validation to evaluate the methods on this dataset.

Large scale pancreatic mass dataset. Our dataset of venous phase 2930 CT scans, is collected from a high-volume US hospital. To the best of our knowledge, it is one of the largest scale pancreatic tumor CT dataset. Pancreatic ductal adenocarcinoma (PDAC) is of the highest priority among all pancreatic abnormalities with a 5-year survival rate of approximately 10% and is the most common type (about 90% of all pancreatic cancers). The labels include Pancreas, PDAC and Cyst. The dataset is randomly split into a training of 2123 CT scans and a testing dataset of 807 CT scans. The model validation is conducted on a subset of training set. The training set includes 1017 PDACs, 462 Cyst, and 644 normal pancreases. The testing set includes 506 PDACs, 271 Cyst, 300 normal pancreases. The evaluation metrics include the dice score, the sensitivity and the specificity following the criterion in [34].

BraTS2023 brain metastases dataset. The BraTS-MET dataset [35] plays a crucial role in developing advanced algorithms for detecting and segmenting brain metastases, with the goal of ensuring seamless integration into clinical practice. This dataset comprises a variety of untreated brain metastasis mpMRI scans collected from multiple institutions [36]–[38] following standard clinical protocols. It is important to mention that our work exclusively utilizes the BraTS2023-MET version of the dataset. Currently, it consists of 238 cases, and we employ a default 5-fold cross-validation method in our analysis.

---

[2]http://braintumorsegmentation.org/
[3]http://medicaldecathlon.com/

TABLE I

IMPLEMENTATION DETAILS INCLUDING THE ARCHITECTURE HYPERPARAMETER, TRAINING SETTINGS, AND DATA AUGMENTATION.

	Synapse	MSD vessel	BraTS	Pancreas
category	multi-organ	organ&tumor	tumor	organ&tumor
crop size	$40\times224\times192$	$64\times192\times192$	$128\times128\times128$	$40\times224\times192$
batch size/gpu	2	2	2	2
lr	8e-2	3e-4	3e-4	3e-4
optimizer	sgd	adamw	adamw	adamw
lr schedule	cosine	cosine	cosine	cosine
downsample	[4, 5, 5]	[4, 5, 5]	[5, 5, 5]	[3, 5, 5]
num of query	n/a	20	20	20
C2F stage	3	3	3	3
augmentation	random rotation, scaling, flipping, white Gaussian noise, Gaussian blurring, adjusting brightness and contrast, simulation of low resolution, Gamma transformation			

B. Implementation Details

Training. In all our 3D experiments, we adhere to nnU-Net's prescribed data augmentation procedures to enhance the diversity of our training dataset. To facilitate effective training, we employ a batch size of 2 using 1 Nvidia RTX 8000 GPU. A comprehensive breakdown of our implementation details can be found in Table I, encompassing critical aspects such as architectural hyperparameters, training configurations, and data augmentation techniques customized for various datasets. We experiment with both 1-layer and 12-layer ViT for implementing the Transformer encoder. Specifically, the 12-layer ViT model is pretrained on the ImageNet21k dataset [39], with additional LayerScale [40]. The latent dimensions d_{enc} and d_{dec} are set as 768 and 192 respectively. For computing the Hungarian matching loss, λ_0 and λ_1 are set as 0.7 and 0.3.

In line with nnU-Net's pioneering adaptive architecture design, our TransUNet exhibits adaptability tailored to the characteristics of the data it processes. Similar to nnU-Net, our backbone architecture is inherently self-adaptive for different medical datasets, particularly in determining crucial factors such as the number of down-sampling layers and the allocation of channels at each stage. The details can be found in Table I. It is noteworthy that our TransUNet maintains compatibility with 2D image data, requiring minimal architectural modifications to accommodate various imaging scenarios.

Testing. Given a CT/MR scan, we do inference in a sliding-window manner. By leveraging the aggregation of all patches, we assign a probability vector to a voxel in position (i, j, s): $\sum_{n=1}^{N}(\mathbf{Z}_{n,ijs}^{T}) \in \mathbb{R}^{K}$, followed by an argmax to obtain hard prediction.

C. Analytical Study of Three Configurations

We performed a variety of ablation studies to thoroughly evaluate the proposed three configurations under the TransUNet framework, $i.e.$, Encoder-only, Decoder-only and Encoder+Decoder.

To assess how effective Transformer encoders are against CNN encoders, and likewise for decoders, we conducted comprehensive experiments of our 3D TransUNet model. This evaluation encompassed both Encoder-only and Decoder-only configurations and compared them to the baseline nnU-Net

across multi-organ segmentation and hepatic vessel tumor segmentation tasks. The results of these comparisons on 3D TransUNet are summarized in Tables II and III.

For multi-organ segmentation, while the decoder-only design demonstrates a modest performance enhancement (87.63% compared to 87.33%), the encoder-only configuration, especially when employing the 12-layer ViT encoder initialized with pre-trained weights from ImageNet, achieves a significant 0.8% improvement in Dice score, reaching 88.11%. This advantage becomes even more pronounced when adopting a 2D U-Net backbone, with performance gains of 2.80% and 3.02% at resolutions of 224×224 and 512×512, respectively, as outlined in Table VIII.

As for the vessel tumor segmentation, the encoder-only design's performance improvement, while present, remains relatively subtle. Both the 1-layer and 12-layer ViT encoders yield comparable results (66.30% and 66.35%), slightly outperforming the baseline nnU-Net's score of 66.04%. In contrast, the decoder-only configuration exhibits a substantial increment, recording a gain of 1.63% (67.67% versus 66.04%).

To summarize, our results show the encoder-only design thrives in multi-organ segmentation, while the decoder-only configuration is more adept in vessel tumor segmentation. This distinction aligns with the inherent strengths of the Transformer encoder, which captures global context information such as the intricate relationships among various targets, making it especially effective for multi-organ tasks. Conversely, the Transformer decoder, tailored to refine small and hard targets, is aptly suited for tumor segmentation. However, a combined approach of the Transformer encoder and decoder (Encoder+Decoder) does not offer further enhancements for either multi-organ or hepatic vessel segmentation. The reason may be attributed to their strengths may overlap or counteract, preventing synergistic benefits in these specific tasks. As a result, our subsequent experiments will predominantly employ the Encoder-only model for segmenting multiple objects and the Decoder-only model for segmentation tasks targeting smaller and more challenging tumors or lesions.

D. Deep Analysis for Transformer Decoder

1) Number of organ/tumor queries: For successful training, the number of learnable queries must be at least equal to the number of classes ($i.e.$, each class must have at least one query). However, when we varied the number of queries in the segmentation process, we observed that the performance of our 3D TransUNet with the Decoder-only configuration remains largely unaffected by this parameter. A detailed summary of these findings is presented in Table IV.

2) Multi-scale CNN feature for updating queries: A defining characteristic of our Transformer decoder is its integration of multi-scale features from the CNN decoder, which are rich in localization details. These features play a pivotal role in progressively refining the learnable queries through the synergy of cross-attention with localized multi-scale CNN representations. Our experimentation, as summarized in Table V, encompasses the configuration of Decoder-only. The consistently observed performance enhancements, as compared to

TABLE II

COMPARISON ON THE SYNAPSE MULTI-ORGAN CT DATASET (AVERAGE DICE SCORE %, AND DICE SCORE % FOR EACH ORGAN). WE USE PRETRAINED MODEL FROM [15] FOR THE ENCODER WITH 12 TRANSFORMER LAYERS.

encoder		decoder	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach	Avg. Dice
1-layer	12-layer										
			93.04	78.82	84.68	88.46	97.13	81.50	91.68	83.34	87.33
✓			93.07	79.56	86.16	87.68	97.22	81.71	92.56	83.23	87.65
	✓		92.97	81.15	85.76	87.47	97.03	81.76	93.39	85.31	**88.11**
		✓	92.88	82.06	86.04	87.70	97.10	82.08	91.14	82.03	87.63
	✓	✓	92.67	81.66	85.29	87.76	97.34	82.69	91.90	85.59	**88.11**

TABLE III

ABLATION OF TRANSFORMER ARCHITECTURE ON MSD VESSEL DATASET WITH DICE SCORE METRICS (%). EXPERIMENTS ARE CONDUCTED IN FIVE-FOLD CROSS-VALIDATION.

encoder		decoder	Vessel	Tumor	Avg. Dice
1-layer	12-layer				
			63.71	68.36	66.04
✓			63.47	69.12	66.30
	✓		63.67	69.02	66.35
		✓	64.41	70.94	**67.67**
✓		✓	64.58	69.89	67.24

TABLE IV

ABLATION OF NUMBER OF QUERIES ON MSD VESSEL DATASET WITH DICE SCORE METRICS (%). EXPERIMENTS ARE CONDUCTED IN FIVE-FOLD CROSS-VALIDATION.

Number of Queries	Vessel	Tumor	Avg. Dice
5	64.75	70.32	67.53
20	64.41	70.94	67.67
40	64.32	70.41	67.37

TABLE V

ABLATION OF TRANSFORMER DECODER ON MSD VESSEL DATASET WITH DICE SCORE METRICS (%). EXPERIMENTS ARE CONDUCTED IN FIVE-FOLD CROSS-VALIDATION.

Transformer decoder	Multi-scale	masked cross attention	Vessel	Tumor	Average
✓	✓	✓	64.41	70.94	67.67
✓	✓		64.37	70.71	67.54
✓			64.19	69.89	67.04
			63.71	68.36	66.04

TABLE VI

PERFORMANCE COMPARISON ON BRATS2023 FOR BRAIN METASTASIS CHALLENGE WITH LESION-WISE DICE SCORE METRICS (%). MASKED CROSS ATTENTION IS NOT USED FOR THE DECODER-ONLY MODEL FOR TRAINING STABILITY. EXPERIMENTS ARE CONDUCTED IN FIVE-FOLD CROSS-VALIDATION.

Method	ET	TC	WT	Avg. Dice
nnU-Net	54.90	58.67	55.75	56.44
Encoder-only	54.79	58.96	56.05	56.60
Decoder-only	56.80	61.12	60.09	59.34

TABLE VII

PERFORMANCE COMPARISON ON IN-HOUSE LARGE-SCALE PANCREATIC TUMOR SEGMENTATION DATASET.

Method	DSC				Sensitivity	Specificity
	Pancreas	PDAC	Cyst	Average		
nnU-Net	83.8	56.94	56.88	65.97	89.34	91.00
Encoder-only	83.77	58.38	57.98	66.71	91.71	84.66
Decoder-only	85.35	62.66	61.04	69.69	89.94	97.33

the baseline Transformer decoder—where the segmentation mask is computed by directly employing the dot product of the learned query and the last-layer CNN feature—underscore the indispensable nature of incorporating multi-scale CNN features in the query updating process.

3) Coarse-to-fine refinement in Transformer decoder: In each Transformer decoder layer, the coarse-to-fine refinement uses the predicted mask from the current iteration to constrain the cross-attention within the foreground region, therefore refining the organ queries at the next iteration. To demonstrate the effectiveness of this strategy, we have selected vessel tumor segmentation as a representative case study. This choice is motivated by the Transformer decoder's demonstrated proficiency in segmenting small targets, such as tumors or lesions. As illustrated in Table V, the integration of coarse-to-fine refinement (masked cross-attention) consistently yielded enhanced results. For a more intuitive understanding, we provide a qualitative example in Fig. 2, elucidating how this attention refines masks for intricate targets. From the first to the third iteration, the segmentation quality of the tumor has been significantly improved.

E. Generalization to Other Tumor/Lesion Datasets

We assess the generalizability of our 3D TransUNet across different imaging modalities and tasks, including brain metastases segmentation from MRI data (BraTS2023), pancreatic tumor (PDAC) and cyst segmentation. The outcomes are detailed in Table VI and Table VII. Our findings confirm that the Encoder-only variant offers modest enhancements, whereas the Decoder-only configuration substantially boosts performance for tumor/lesion segmentation. For example, in brain metastatic lesion segmentation, the Encoder-only model provides a marginal enhancement, whereas the Decoder-only model contributes to a significant Dice improvement of

TABLE VIII

COMPARISON ON THE SYNAPSE MULTI-ORGAN CT DATASET (AVERAGE DICE SCORE %, AND DICE SCORE % FOR EACH ORGAN).

Scale	Method	Avg. Dice	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
2D (224×224)	U-Net [2]	74.68	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
	Pyramid Attn [41]	73.08	82.57	56.25	75.78	70.51	93.46	50.02	83.95	72.13
	DeepLabv3+ [42]	76.35	82.00	62.85	78.89	75.24	93.96	57.75	86.57	73.57
	UNet++ [43]	76.65	86.93	63.69	77.86	68.29	93.91	59.23	87.81	75.49
	AttnUNet [12]	75.57	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
	TransUNet	**77.48**	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
2D (512×512)	U-Net [2]	81.34	89.69	69.98	83.08	74.13	95.10	67.73	90.50	80.51
	Pyramid Attn [41]	80.08	88.59	65.91	84.45	75.15	95.30	60.06	91.84	79.33
	DeepLabv3+ [42]	82.50	88.79	72.16	88.13	79.52	95.58	65.97	90.02	79.87
	UNet++ [43]	81.6	89.65	71.68	82.92	75.15	94.92	69.06	89.42	80.01
	AttnUNet [12]	80.88	89.46	67.09	83.83	75.98	95.28	68.48	88.63	78.26
	nnU-Net [44]	82.92	91.55	73.43	82.74	73.61	96.01	71.81	94.29	79.94
	TransUNet	**84.36**	90.68	71.99	86.04	83.71	95.54	73.96	88.80	84.20
3D	V-Net [30]	68.81	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
	DARR [31]	69.77	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
	nnU-Net [44]	87.33	93.04	78.82	84.68	88.46	97.13	81.50	91.68	83.34
	CoTr [17]	85.72	92.96	71.09	85.70	85.71	96.88	81.28	90.44	81.74
	nnFormer [16]	85.32	90.72	71.67	85.60	87.02	96.28	82.28	87.30	81.69
	VT-UNet [45]	70.72	78.25	44.76	77.51	78.16	91.63	45.18	82.20	68.04
	Swin UNETR [20]	73.51	82.94	60.96	80.41	71.14	91.55	56.71	77.46	66.94
	TransUNet	**88.11**	92.97	81.15	85.76	87.47	97.03	81.76	93.39	85.31

TABLE IX

PERFORMANCE COMPARISON ON MSD VESSEL DATASET WITH DICE SCORE METRICS (%). EXPERIMENTS ARE CONDUCTED IN FIVE-FOLD CROSS-VALIDATION.

Method	Vessel	Tumor	Avg. Dice
nnU-Net	63.71	68.36	66.04
nnFormer [16]	63.21	69.37	66.29
VT-UNet [45]	60.88	59.82	60.35
Swin UNETR [20]	57.65	58.31	57.98
TransUNet	64.10	70.60	**67.67**

2.9%. This trend is even more pronounced for PDAC and cystic lesion segmentation. Specifically, Encoder-only model enhances the PDAC segmentation from 56.94% to 58.38%, while Decoder-only model pushes it further to 62.66%. In the case of cystic lesions, Encoder-only model lifts nnU-Net's performance from 56.88% to 57.98%, with Decoder-only model further elevating it to 61.04%. Consequently, the aggregate metrics of average Dice, Sensitivity, and Specificity all witness marked improvements with Decoder-only model.

F. Comparison with State-of-the-arts

We compare our TransUNet to previous 2D and 3D state-of-the-art methods on multi-organ segmentation and hepatic vessel tumor segmentation in Table VIII and Table IX. With the 2D version built on the U-Net architecture and the 3D variant grounded in the 3D nnU-Net framework, our TransUNet consistently outperforms other state-of-the-art methods, underscoring its efficacy across diverse U-Net frameworks. As discussed above, leveraging the Transformer Encoder's ability to capture global organ relationships, we use the Encoder-only design for multi-organ segmentation. Conversely, given

the Transformer Decoder's prowess in refining small targets, we opt for the Decoder-only setup for tumor segmentation. Specifically, we compare TransUNet against a spectrum of methodologies, including: 1) 2D techniques such as U-Net [2], DeepLabv3+ [42], and UNet++ [43], complemented by attention-augmented CNN methods like AttnUNet [12] and Pyramid Attn [41], evaluated across resolutions of 224×224 and 512×512; 2) 3D approaches like V-Net [30], DARR [31], and 3D nnU-Net [44], accompanied by cutting-edge Transformer-centric strategies including CoTR [17], nn-former [16], VT-UNet [45], and Swin UNETR [20]. As corroborated by the results in Table VIII, TransUNet not only surpasses traditional CNN-based self-attention models but also outperforms numerous state-of-the-art Transformer-oriented techniques. For example, when benchmarked against recent state-of-the-art Transformer-based methods like CoTr and nnformer, our TransUNet achieves approximately a 10% improvement in Dice scores for the challenging task of gall-bladder segmentation and about a 3% enhancement in overall segmentation

Notably, as evidenced in Table X, our 3D TransUNet surpasses the top-ranked solution, nnU-Net-Large [32], from the BraTS2021 challenge, underscoring the robustness and efficacy of our proposed approach.

V. CONCLUSION

While U-Net has been successful, its limitations in handling long-range dependencies have prompted the exploration of Transformer as an alternative architecture, exemplified by our previously developed TransUNet, harnessing the combined strengths of U-Net and Transformers. Our study extends the TransUNet architecture to a 3D network, building upon nnU-Net's foundation. Leveraging Transformers' capabilities in both encoder and decoder design, we introduce 1) A
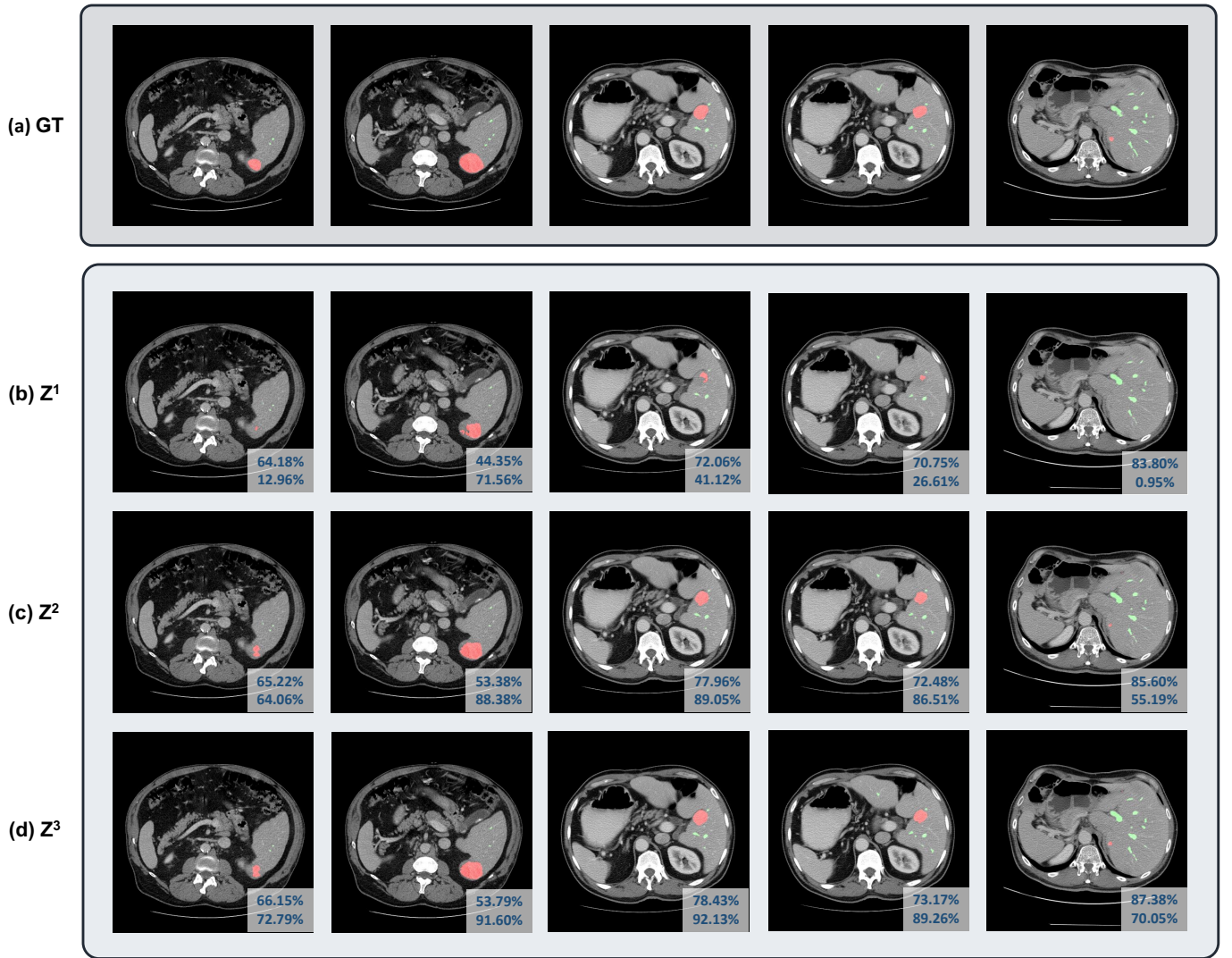
Fig. 2. Visualizations of outputs from different iterations during coarse-to-fine refinement: (a) Groundtruth. (b-d) the segmentation mask at the first to the third iteration. Different columns represent different samples from MSD Vessel Dataset. The dice coefficients of vessels and tumors are indicated in the first and the second row of the lower right corner of each image respectively.

TABLE X
PERFORMANCE COMPARISON ON BRATS2021 CHALLENGE FOR BRAIN TUMOR SEGMENTATION WITH DICE SCORE METRICS (%). EXPERIMENTS ARE CONDUCTED IN FIVE-FOLD CROSS-VALIDATION.

Method	ET	TC	WT	Avg. Dice
nnU-Net	88.05	91.92	93.79	91.25
AxialAttn [32]	87.23	91.88	93.21	90.77
nnUNet-Large [32]	88.23	92.35	93.83	91.47
TransUNet	88.85	92.48	93.90	**91.74**

Transformer encoder that tokenizes CNN feature map patches, facilitating a richer extraction of global contexts; and 2) A Transformer decoder designed to adaptively refine segmentation regions, capitalizing on cross-attention mechanisms between candidate proposals and U-Net features. Our investigations further illuminate that medical segmentation tasks have varying architectural preferences and offer insights into tailoring design configurations to suit these specific requirements. Empirical results showcase our 3D TransUNet's superior performance in diverse medical segmentation tasks. Notably, our method excels in multi-organ, pancreatic tumor, hepatic vessel, and brain metastases segmentation. Additionally, we achieved notable results in the BraTS2021 challenge.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[3] L. Yu, J.-Z. Cheng, Q. Dou, X. Yang, H. Chen, J. Qin, and P.-A. Heng, "Automatic 3d cardiovascular mr segmentation with densely-connected volumetric convnets," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 287–295.

[4] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille, "A fixed-point model for pancreas segmentation in abdominal ct scans," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 693–701.

[5] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[6] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman, and A. L. Yuille, "Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8280–8289.

[7] X. Luo, J. Chen, T. Song, Y. Chen, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation through dual-task consistency," *AAAI Conference on Artificial Intelligence*, 2021.

[8] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[10] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[11] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[12] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[13] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4055–4064.

[14] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," *arXiv preprint arXiv:1904.10509*, 2019.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[16] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, and Y. Yu, "nnformer: Volumetric medical image segmentation via a 3d transformer," *IEEE Transactions on Image Processing*, 2023.

[17] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer, 2021, pp. 171–180.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[19] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[20] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI Brainlesion Workshop*. Springer, 2021, pp. 272–284.

[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[22] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7262–7272.

[23] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "Max-deeplab: End-to-end panoptic segmentation with mask transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5463–5474.

[24] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[25] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[26] Q. Yu, H. Wang, S. Qiao, M. Collins, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "k-means mask transformer," in *European Conference on Computer Vision*. Springer, 2022, pp. 288–307.

[27] Q. Yu, H. Wang, D. Kim, S. Qiao, M. Collins, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "Cmt-deeplab: Clustering mask transformers for panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2560–2570.

[28] Z. Zhu, Y. Xia, W. Shen, E. Fishman, and A. Yuille, "A 3d coarse-to-fine framework for volumetric medical image segmentation," in *2018 International conference on 3D vision (3DV)*. IEEE, 2018, pp. 682–690.

[29] L. Xie, Q. Yu, Y. Zhou, Y. Wang, E. K. Fishman, and A. L. Yuille, "Recurrent saliency transformation network for tiny target segmentation in abdominal ct scans," *IEEE transactions on medical imaging*, vol. 39, no. 2, pp. 514–525, 2019.

[30] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3DV*, 2016.

[31] S. Fu, Y. Lu, Y. Wang, Y. Zhou, W. Shen, E. Fishman, and A. Yuille, "Domain adaptive relational reasoning for 3d multi-organ segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 656–666.

[32] H. M. Luu and S.-H. Park, "Extending nn-unet for brain tumor segmentation," *arXiv preprint arXiv:2112.04653*, 2021.

[33] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken *et al.*, "The medical segmentation decathlon," *arXiv preprint arXiv:2106.05735*, 2021.

[34] Z. Zhu, Y. Xia, L. Xie, E. K. Fishman, and A. L. Yuille, "Multi-scale coarse-to-fine segmentation for screening pancreatic ductal adenocarcinoma," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer, 2019, pp. 3–12.

[35] A. W. Moawad, A. Janas, U. Baid, D. Ramakrishnan, L. Jekel, K. Krantchev, H. Moy, R. Saluja, K. Osenberg, K. Wilms *et al.*, "The brain tumor segmentation (brats-mets) challenge 2023: Brain metastasis segmentation on pre-treatment mri," *arXiv preprint arXiv:2306.00838*, 2023.

[36] E. Oermann, K. Link, Z. Schnurman, C. Liu, Y. J. F. Kwon, L. Y. Jiang, M. Nasir-Moin, S. Neifert, J. Alzate, K. Bernstein *et al.*, "Longitudinal deep neural networks for assessing metastatic brain cancer on a massive open benchmark." *preprint*, 2023.

[37] J. D. Rudie, R. S. D. A. Weiss, P. Nedelec, E. Calabrese, J. B. Colby, B. Laguna, J. Mongan, S. Braunstein, C. P. Hess, A. M. Rauschecker *et al.*, "The university of california san francisco, brain metastases stereotactic radiosurgery (ucsf-bmsr) mri dataset," *arXiv preprint arXiv:2304.07248*, 2023.

[38] E. Grøvik, D. Yi, M. Iv, E. Tong, D. Rubin, and G. Zaharchuk, "Deep learning enables automatic detection and segmentation of brain metastases on multisequence mri," *Journal of Magnetic Resonance Imaging*, vol. 51, no. 1, pp. 175–182, 2020.

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[40] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 32–42.

[41] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[42] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[43] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[44] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[45] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi, "A robust volumetric transformer for accurate 3d tumor segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 162–172.