Medical Image Analysis 97 (2024) 103280

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers

Jieneng Chen [a], Jieru Mei [a], Xianhang Li [b], Yongyi Lu [a], Qihang Yu [a], Qingyue Wei [c], Xiangde Luo [d], Yutong Xie [h], Ehsan Adeli [e], Yan Wang [g], Matthew P. Lungren [e], Shaoting Zhang [d], Lei Xing [c], Le Lu [f], Alan Yuille [a], Yuyin Zhou [b,*]

[a] Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
[b] Department of Computer Science and Engineering, University of California, Santa Cruz, CA 95064, USA
[c] Department of Radiation Oncology, Stanford University, Stanford, CA 94305, USA
[d] Shanghai AI Lab, Xuhui District, Shanghai, 200000, China
[e] The School of Medicine, Stanford University, Stanford, CA 94305, USA
[f] DAMO Academy, Alibaba Group, New York, NY 10014, USA
[g] The East China Normal University, Shanghai 200062, China
[h] The Australian Institute for Machine Learning, University of Adelaide, Australia

ARTICLE INFO

Keywords:
Medical image segmentation
Vision Transformers
U-Net

ABSTRACT

Medical image segmentation is crucial for healthcare, yet convolution-based methods like U-Net face limitations in modeling long-range dependencies. To address this, Transformers designed for sequence-to-sequence predictions have been integrated into medical image segmentation. However, a comprehensive understanding of Transformers' self-attention in U-Net components is lacking. TransUNet, first introduced in 2021, is widely recognized as one of the first models to integrate Transformer into medical image analysis. In this study, we present the versatile framework of TransUNet that encapsulates Transformers' self-attention into two key modules: (1) a Transformer encoder tokenizing image patches from a convolution neural network (CNN) feature map, facilitating global context extraction, and (2) a Transformer decoder refining candidate regions through cross-attention between proposals and U-Net features. These modules can be flexibly inserted into the U-Net backbone, resulting in three configurations: Encoder-only, Decoder-only, and Encoder+Decoder. TransUNet provides a library encompassing both 2D and 3D implementations, enabling users to easily tailor the chosen architecture. Our findings highlight the encoder's efficacy in modeling interactions among multiple abdominal organs and the decoder's strength in handling small targets like tumors. It excels in diverse medical applications, such as multi-organ segmentation, pancreatic tumor segmentation, and hepatic vessel segmentation. Notably, our TransUNet achieves a significant average Dice improvement of 1.06% and 4.30% for multi-organ segmentation and pancreatic tumor segmentation, respectively, when compared to the highly competitive nn-UNet, and surpasses the top-1 solution in the BrasTS2021 challenge. 2D/3D Code and models are available at https://github.com/Beckschen/TransUNet and https://github.com/Beckschen/TransUNet-3D, respectively.

1. Introduction

Convolutional neural networks (CNNs), particularly fully convolutional networks (FCNs) (Long et al., 2015), have risen to prominence in the domain of medical image segmentation. Among their various iterations, the U-Net model (Ronneberger et al., 2015), 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 (Yu et al., 2017), organ delineation using computed tomography (CT) scans (Zhou et al., 2017; Li et al., 2018b; Yu et al., 2018; Luo et al., 2021), and polyp segmentation in colonoscopy recordings (Zhou et al., 2019).

Despite CNNs' unparalleled representational capabilities, they often falter when modeling long-range relationships due to the inherent

* Corresponding author.
E-mail address: yzhou284@ucsc.edu (Y. Zhou).

https://doi.org/10.1016/j.media.2024.103280
Received 16 February 2024; Received in revised form 16 June 2024; Accepted 15 July 2024
Available online 22 July 2024

locality of convolution operations. This limitation becomes particularly pronounced in cases with large inter-patient texture, shape, and size variations. 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 (Vaswani et al., 2017). However, Transformers process inputs as 1D sequences and prioritize global context modeling, inadvertently producing features of low resolution. A more promising hybrid approach involves combining CNN and Transformer encoders. TransUNet (Chen et al., 2021), first introduced in 2021, is among the first models to integrate Transformer into medical image analysis. This approach capitalizes on U-Net encoders' high-resolution spatial details while leveraging Transformers' global context modeling, which is crucial in medical image segmentation. This innovation spurred a number of subsequent studies (Cao et al., 2022; Xie et al., 2021; Hatamizadeh et al., 2021). Despite this, a comprehensive understanding of Transformers' self-attention in different U-Net components remains a missing piece.

In this study, we introduce TransUNet, a flexible framework offering a comprehensive exploration of strategic Transformer integration in both encoding and decoding processes, as an extension of Chen et al. (2021). TransUNet encapsulates Transformers' self-attention into two opted modules. 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. These queries are progressively refined by synergizing cross-attention with localized multi-scale CNN features. We also introduce a coarse-to-fine attention refinement in the Transformer decoder, by constraining the ongoing cross-attention exclusively to the foreground of the preceding coarse prediction for each query. The details of the two modules as well as the TransUNet framework are outlined in Fig. 1.

To study the role of Transformers in the U-Net architecture, we alternately insert Transformer encoder and decoder into different parts of the U-Net backbone, yielding three configurations: Encoder-only (Transformers solely applied in the encoder), Decoder-only (Transformers solely applied in the decoder), and Encoder+Decoder (Transformers applied both in the encoder and the decoder). We also provide a library with both 2D and 3D implementations, facilitating user customization of the chosen architecture. Extensive experiments validate the superior performance of our method over competing approaches in diverse medical image segmentation tasks. Our study also offers valuable insights for optimal configuration selection—Multi-organ segmentation benefits from the Transformer encoder, while tumor segmentation benefits from the Transformer decoder. Our contributions are in four-folds:

- We introduce a Transformer-centric encoder–decoder framework, incorporating self-attention and cross-attention within the sequence-to-sequence prediction context for medical image segmentation.
- We reformulate the decoding of medical image segmentation by the introduced Transformer decoder, with learnable queries redefines conventional per-pixel segmentation as a mask classification. We further propose *coarse-to-fine attention refinement* in the Transformer decoder, boosting small target/tumor segmentation.
- We provide the first comprehensive study of strategic Transformer integration in both encoding and decoding processes of U-Net, providing insights on tailoring designs to cater to distinct medical image segmentation challenges.
- We outperform the state-of-the-art nnUNet architecture on various medical image segmentation tasks, and release our codebase to encourage further exploration in applying Transformers to medical applications.

2. Related work

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 (Wang et al., 2018). Built upon the encoder–decoder u-shaped architecture, Schlemper et al. (2019) proposed additive attention gate modules which are integrated into the skip-connections. Unlike these approaches, we employ Transformers to embed global self-attention in our method.

Transformers. Transformers were first proposed by Vaswani et al. (2017) for machine translation and established state-of-the-art methods in many NLP tasks. Several modifications have been made to make Transformers also applicable to computer vision tasks. For instance, Parmar et al. (2018) applied the self-attention only in local neighborhoods for each query pixel instead of globally. Child et al. (2019) proposed Sparse Transformers, which employ scalable approximations to global self-attention. Recently, Vision Transformer (ViT) (Dosovitskiy et al., 2021) achieved state-of-the-art on ImageNet classification by directly applying Transformers with global self-attention to full-sized images. Combining Transformers and U-Net to enable more precise medical image segmentation has also drawn increasing attention in the community. first Introduced in 2021, TransUNet (Chen et al., 2021) marks one of the first models to integrate Transformer into medical image analysis. Along this research direction, Swin-UNet (Cao et al., 2022) and SwinUETR (Hatamizadeh et al., 2021) improves the self-attention mechanisms by using the more computation-efficient Swin Transformers (Liu et al., 2021); nnFormer (Zhou et al., 2023) further improves by interleaving convolution with self-attention.

Mask classification for segmentation. DETR (Carion et al., 2020) 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 (Strudel et al., 2021; Wang et al., 2021; Cheng et al., 2021, 2022; Yu et al., 2022b,a), a novel variation known as mask Transformers has emerged. This variant introduces segmentation predictions by employing a collection of query embeddings to represent the object and its associated mask. Wang et al. (2021) first develop a mask Transformer with memory embedding, and Cheng et al. (2021) further formulate the query update in a manner of DETR (Carion et al., 2020). 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 (Cheng et al., 2021, 2022; Yu et al., 2022b,a). This process enables the model to effectively handle segmentation tasks and produce accurate results.

3. 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. We aim 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 3.1, we delve into the direct application of Transformers for encoding feature representations from segmented image patches. Following this, in Section 3.2, we elaborate on implementing the query-based Transformer, which serves as our decoder. The detailed architecture of TransUNet is then presented in Section 3.3.

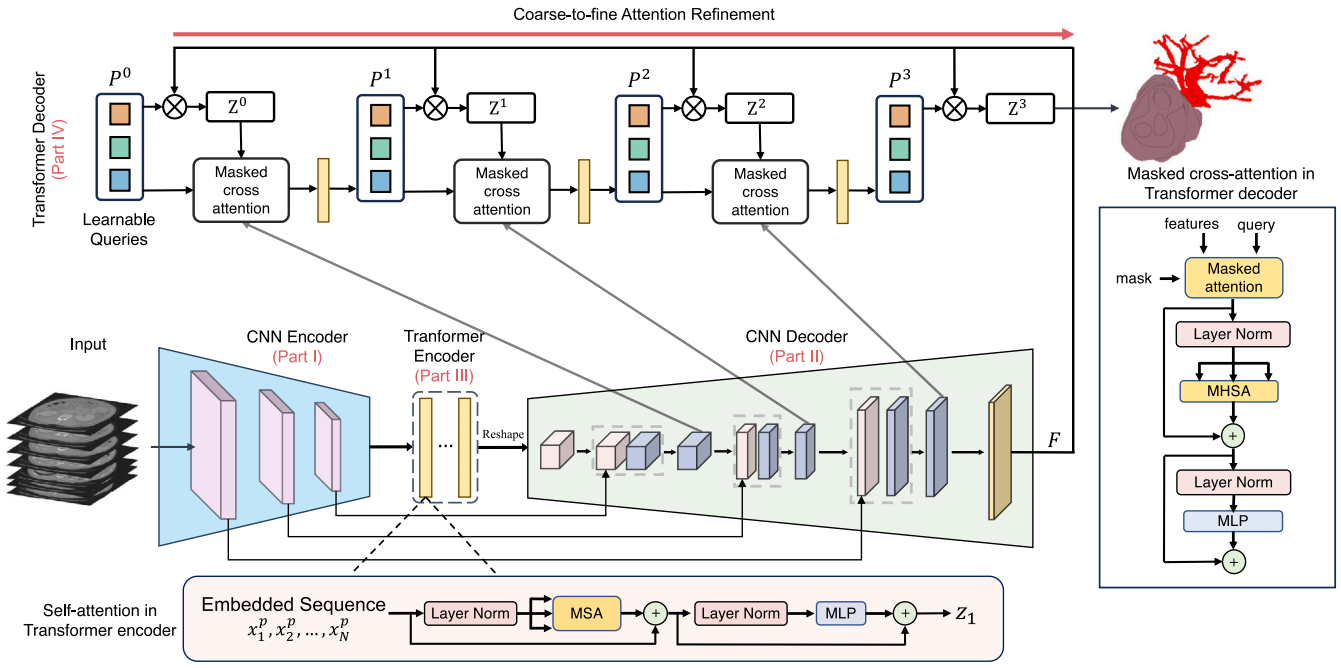

Fig. 1. Overview of TransUNet. Our proposed architecture consists of two components: (1) the Transformer encoder where a CNN encoder is firstly used for local image feature extraction, followed by a pure Transformer encoder for global information interaction; and (2) the Transformer decoder that reframes 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.

3.1. Transformer as encoder

Image sequentialization. Following (Dosovitskiy et al., 2021), we first perform tokenization by reshaping the input $\mathbf{x}$ into a sequence of flattened 3D patches $\{\mathbf{x}_i^p \in \mathbb{R}^{P^3 \cdot C} | i = 1, \ldots, N\}$, 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:

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

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:

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

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

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

3.2. Transformer as decoder

3.2.1. 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 representing 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 (Strudel et al., 2021). 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 $\mathbf{F} \in \mathbb{R}^{D \times H \times W \times d_{dec}}$:

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

where $g(\cdot)$ is sigmoid activation followed by a hard thresholding operation with a threshold set at 0.5. Note that we use the sigmoid activation function here instead of the softmax function because some of our experimenting datasets (*e.g.*, BraTS) have overlapped classes.

3.2.2. Transformer decoder

Fig. 1 illustrates our Transformer decoder, designed to refine organ queries, thereby enhancing the coarse prediction $\mathbf{Z}^0$. Similar to the structure seen in the Transformer encoder (detailed in Section 3.1), 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 tth 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 tth upsampling block. Moving to the $t+1$-th layer, organ queries are updated using cross-attention as follows:

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

where the tth 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 (Cheng et al., 2022). Next, we will introduce how to incorporate a coarse-to-fine attention refinement to further enhance the accuracy of segmentation results.

3.2.3. Coarse-to-fine attention refinement

The value of coarse-to-fine refinement in medical image segmentation, particularly for small target segmentation, is well-established (Zhou et al., 2017; Zhu et al., 2018; Xie et al., 2019). 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 (Cheng et al., 2022). 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 tth 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 computation incorporates the existing coarse prediction $\mathbf{Z}^t$ into the affinity matrix, as detailed in Eq. (5):

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

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:

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

$$\hat{\mathbf{y}} = \text{argmax}_{k=0,1,\ldots,K-1}\ \mathbf{O}, \tag{9}$$

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

3.3. TransUNet

As shown in Fig. 1, there are four components in TransUNet: (1) CNN encoder (part I), (2) CNN decoder (part II), (3) Transformer encoder (part III), and (4) Transformer decoder (part IV). To conduct a thorough analysis of the Transformer encoder and Transformer decoder, and to explore their optimal integration within U-Net architectures, we instantiate our TransUNet model with three distinct configurations as outlined below.

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 $\hat{\mathbf{y}}$;

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

3.3.1. Encoder-only

A CNN-Transformer hybrid encoder (part I + part II + part III in Fig. 1) 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.

3.3.2. 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 (part I + part II + part IV in Fig. 1)** in the segmentation model. The organ queries $\mathbf{P}$ are initially 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 3.2, $\mathbf{P}$ 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 (Carion et al., 2020; Wang et al., 2021) 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 (Milletari et al., 2016). 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.

3.3.3. Encoder + decoder

Here, we integrate both the Transformer encoder and the Transformer decoder into the U-Net model **(part I + part II + part III + part IV** in Fig. 1). And then similar to the decoder-only model, here we also use the Hungarian matching loss to train the whole network.

4. Experiments and discussion

We evaluate our method on 4 different datasets, including BTCV multi-organ segmentation dataset, *i.e.*, BraTS2021 brain tumor segmentation challenge, Medical Segmentation Decathlon (MSD) HepaticVessel Dataset, and a large-scale in-house pancreatic mass dataset to confirm the effectiveness of our approach. To ensure a fair comparison across all methods, we use identical experimental settings for both TransUNet and the compared approaches. Except for the BTCV dataset where we use a hard split strictly following the setting in Fu et al. (2020), Huang et al. (2022), Zhou et al. (2023) for multi-organ segmentation evaluation, the remaining four datasets undergo a comprehensive 5-fold cross-validation, ensuring a rigorous and unbiased assessment across diverse datasets.

4.1. Dataset and evaluation

BTCV multi-organ segmentation dataset (Landman et al., 2017). We use the 30 abdominal CT scans in the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge (Landman et al., 2017), with a total of 3779 axial contrast-enhanced abdominal clinical CT images.

Each CT volume consists of 85 ∼ 198 slices of 512 × 512 pixels, with a voxel spatial resolution of ([0.54 ∼ 0.54] × [0.98 ∼ 0.98] × [2.5 ∼ 5.0]) mm^3. We report the average DSC on eight 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 Fu et al. (2020).

BraTS2021 brain tumor segmentation challenge.[1] 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 1 mm^3 resolution and skull-stripping. The image sizes of all MRI scans and associated labels are 240 × 240 × 155. In our experiments, we apply 5-fold cross-validation with the same data split used by the No. 1 solution (Luu and Park, 2021) in the BraTS2021 challenge.

Medical Segmentation Decathlon (MSD) HepaticVessel.[2] The MSD HepaticVessel, a task of Medical Segmentation Decathlon (Antonelli et al., 2021), 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 datasets. 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 Cysts, and 300 normal pancreases. The evaluation metrics include the dice score, the sensitivity and the specificity following the criterion in Zhu et al. (2019).

[1] http://braintumorsegmentation.org/

[2] http://medicaldecathlon.com/

Table 1
Implementation details including the architecture hyperparameters, training settings, and data augmentation. Note that the customized hyperparameters (row3-6) are directly borrowed from nnUNet configuration.

category	Synapse	MSD vessel	BraTS	Pancreas
	multi-organ	organ&tumor	tumor	organ&tumor
crop size	40 × 224 × 192	64 × 192 × 192	128 × 128 × 128	40 × 224 × 192
batch size/gpu	2	2	2	2
downsample	[4, 5, 5]	[4, 5, 5]	[5, 5, 5]	[3, 5, 5]
augmentation	random rotation, scaling, flipping, white Gaussian noise, Gaussian blurring, adjusting brightness and contrast, simulation of low resolution, Gamma transformation			
lr	8e−2	3e−4	3e−4	3e−4
optimizer	sgd	adamw	adamw	adamw
lr schedule	cosine	cosine	cosine	cosine
num of query	n/a	20	20	20
C2F stage	3	3	3	3

4.2. Implementation details

Training. We use 3D nn-UNet as our backbone architecture and adhere to nn-UNet's prescribed data augmentation procedures to enhance the diversity of our training dataset. We employ a batch size of 2 using 1 Nvidia RTX 8000 GPU to facilitate effective training. A comprehensive breakdown of our implementation details can be found in Table 1, 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 (Russakovsky et al., 2015), with additional LayerScale (Touvron et al., 2021). 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. Following nn-UNet's framework, our TransUNet exhibits adaptability tailored to the characteristics of the data it processes. In Table 1, we show details for each segmentation dataset, including the number of down-sampling layers and the allocation of channels at each stage.

Note that our primary analysis is on 3D experiments due to the substantially superior performance exhibited by the 3D baseline compared to the 2D baseline, as illustrated in Table 2.

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): $\sum_{n=1}^{N}(\mathbf{Z}^{T}_{n,ijs}) \in \mathbb{R}^{K}$, followed by an argmax to obtain a hard prediction.

4.3. Analytical study

Our hypothesis is that the Transformer encoder should excel at capturing global context information as it encodes the high-level CNN features before transmitting them to the decoder. Therefore it should be mostly effective for multi-organ segmentation. Conversely, the Transformer decoder, employing a coarse-to-fine attention mechanism to refine small and challenging targets, should be more suitable for tumor segmentation. To test this hypothesis, we performed various ablation studies to thoroughly evaluate the three configurations of TransUNet, *i.e.*, Encoder-only, Decoder-only and Encoder+Decoder, as mentioned in Section 3.3.

4.3.1. Comparison of the three configurations

To assess how effective Transformer encoders are against CNN encoders, and likewise for decoders, we conducted comprehensive

Table 2
Comparison of different configurations of TransUNet on the BTCV multi-organ CT dataset (average dice score %, and dice score % for each organ).

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
✓		✓	93.04	82.04	85.67	88.87	97.18	82.92	92.36	85.06	**88.39**

Table 3
Comparison of different configurations of TransUNet 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**
	✓	✓	63.91	70.45	67.18
✓		✓	**64.58**	69.89	67.24

Table 4
Generalization of the Transformer decoder to different pancreatic tumors on our in-house large-scale pancreatic tumor segmentation dataset.

Method	Pancreas	PDAC	Cyst	Avg. Dice (%)
nnU-Net	83.8	56.94	56.88	65.97
Encoder-only	83.77	58.38	57.98	66.71
Decoder-only	85.35	62.66	61.04	69.69
Encoder+Decoder	85.37	61.82	60.60	69.26

comparison of Encoder-only, Decoder-only and Encoder+Decoder as summarized in Tables 2 and 3. 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%. Moreover, comparative analysis against a counterpart model trained from scratch (87.53%) reveals a 0.58% enhancement when leveraging pre-trained weights.

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%). Specially, we note that all configurations employing a Transformer decoder for vessel tumor segmentation (last three rows of Table 3) outperform those without a Transformer decoder (first three rows of Table 3). This demonstrates the effectiveness of the Transformer decoder for small target segmentation.

The combined approach of the Transformer encoder and decoder (Encoder+Decoder) yields results comparable to either the encoder-only or decoder-only configurations, offering no significant additional enhancements for multi-organ or hepatic vessel segmentation. For instance, in multi-organ segmentation, the combination of 1-layer Transformer encoder and the Transformer decoder achieves an average Dice score of 88.39% while the 12-layer Transformer encoder-only configuration achieves 88.11% (Table 2). For the MSD HepaticVessel task, the best configuration, decoder-only, achieves a 67.24% average Dice, while the combination of 1-layer/12-layer Transformer encoder and the Transformer decoder achieves 67.24%/67.18%.

Note that Tables 2 and 3 guide our selection of default configurations. The 12-layer Transformer encoder is chosen for the encoder-only architecture due to the consistent superior performance on both the BTCV multi-organ and MSD vessel datasets. Despite the 12-layer encoder's individual superiority, the 1-layer encoder, when integrated with the decoder, achieves similar results to its 12-layer counterpart with much a much less computation budget (Table 5). Therefore the Encoder+Decoder architecture defaults to the 1-layer Transformer encoder.

4.3.2. Generalization of the transformer decoder to pancreatic tumor segmentation

We have verified that the Transformer decoder demonstrates greater efficacy in tumor segmentation than the Transformer encoder. To demonstrate the generalizability of this conclusion beyond vessel tumor segmentation, we also compare the results for different pancreatic tumors in Table 4. And we find that for different pancreatic tumors (*i.e.*, PDAC and Cyst), the decoder-only architecture consistently achieves better results. This, again, verifies our hypothesis that the Transformer decoder, employing a coarse-to-fine attention, is well-suited for handling small targets like tumors.

4.3.3. Computation efficiency

We conduct a comprehensive comparison of results and computational efficiency across three configurations. Table 5 presents the average Dice scores and network parameters comparison on the MSD HepaticVessel dataset, our large-scale in-house pancreatic mass dataset, and the BTCV multi-organ segmentation dataset. Notably, we advocate Encoder+Decoder and Decoder-only as the superior options, as they achieves the best result on different tasks with a considerably small number of network parameters. Note that Encoder+Decoder with 41.4M parameters is also more parameter-efficient than compared methods (*e.g.*, SwinUNet's 62.0M, 3D UX-Net's 53.0M, SwinUNETR-v2's 72.8M). Next, we will mainly report results of Encoder+Decoder and Decoder-only for the remaining experiments.

4.4. Deep analysis for transformer decoder

To ablate the role of organ/tumor queries, multi-scale CNN features and coarse-to-fine refinement, we conduct experiments based on the Decoder-only configuration as below. Additionally, we examine the influence of various attention mechanisms and positional embeddings to comprehensively analyze their impact.

4.4.1. 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 TransUNet with the Decoder-only configuration remains largely unaffected by this parameter. A detailed summary of these findings is presented in Table 6.

Table 5
Performance (average Dice (%)) v.s. network parameters comparison on MSD HepaticVessel dataset, our in-house large-scale pancreatic mass dataset, and Synapse multi-organ segmentation dataset. The pancreatic tumor Dice scores are averaged from the performance of both pancreatic cyst and PDAC segmentation. **bold** denotes the best results and underline denotes the second best results.

Method	Pancreas	Tumor	Vessel	Tumor	Multi-organ	#Params
nnU-Net	83.8	56.91	63.71	68.36	87.33	30.8M
Encoder-only	83.77	58.18	63.67	69.02	88.11	116.5M
Decoder-only	85.35	**61.85**	64.41	**70.94**	87.63	33.6M
Encoder+Decoder	**85.37**	61.21	**64.58**	69.89	**88.39**	41.4M

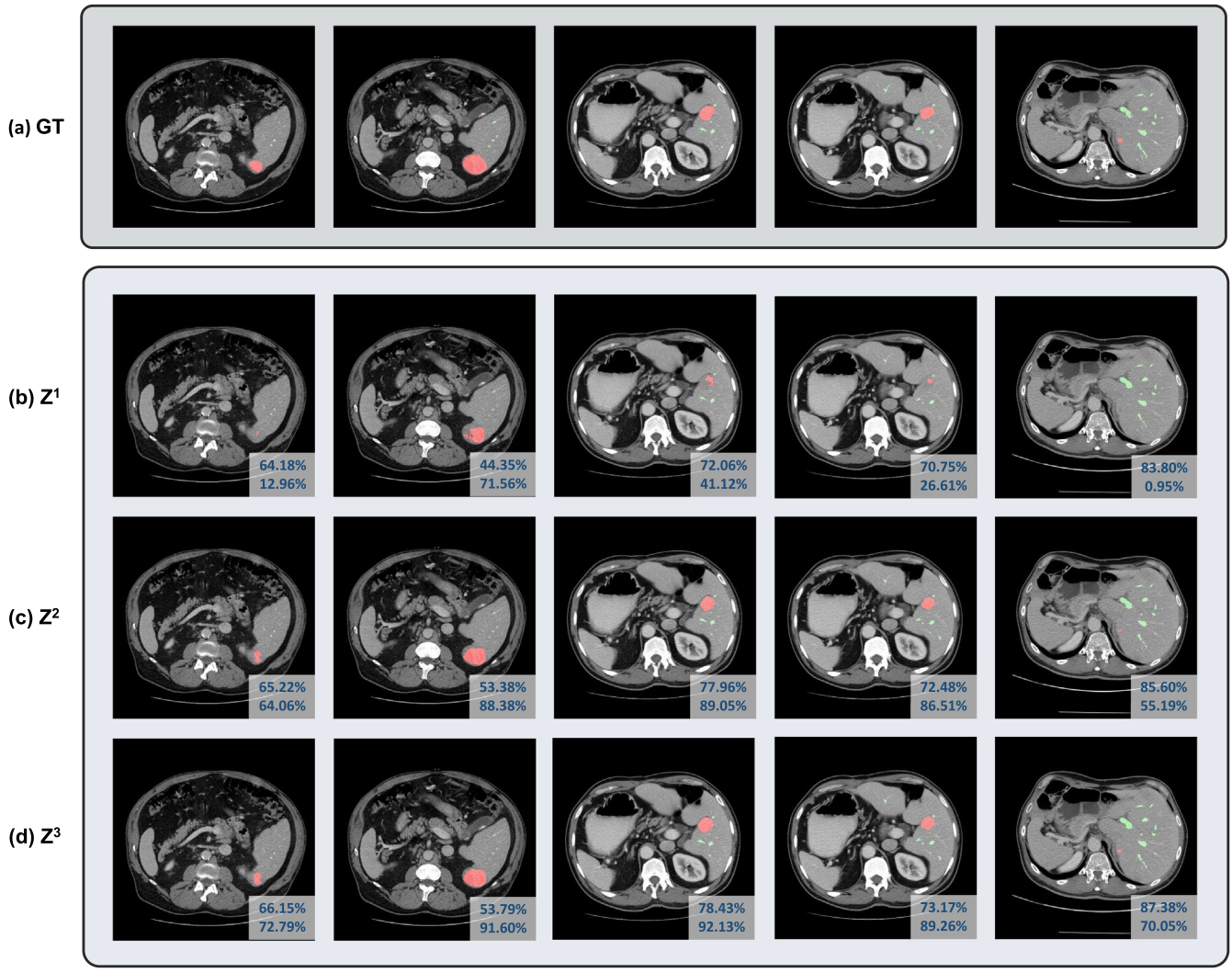

Fig. 2. Visualizations of outputs from different iterations during coarse-to-fine refinement: (a) Groundtruth. (b-d) the segmentation mask at the first to the third iteration. Different columns represent different samples from MSD Vessel Dataset. The dice coefficients of vessels and tumors are indicated in each image's first and second row of the lower right corner, respectively.

Table 6
Ablation of number of queries under Transformer decoder setting 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

4.4.2. *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 7, encompasses the configuration of Decoder-only. The consistently observed performance enhancements, as compared to 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.

Table 7
Ablation of different types of attention mechanisms for the Transformer decoder on MSD vessel dataset with dice score metrics (%). Experiments are conducted in five-fold cross-validation.

cross cross-attention	Multi-scale	masked cross attention	Vessel	Tumor	Avg. Dice (%)
✓	✓	✓	64.41	70.94	67.67
✓	✓		64.37	70.71	67.54
		✓	64.10	70.60	67.35
✓			64.19	69.89	67.04
			63.71	68.36	66.04

4.4.3. 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 7, 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.

4.4.4. Ablation study

Table 7 presents the ablation results for different attention mechanisms: (1) no attention versus cross-attention, (2) multi-scale attention by leveraging multi-scale convolutional features versus single-scale attention, and (3) masked attention versus cross-attention. The results indicate that using cross-attention by the Transformer decoder improves segmentation performance by 1%, from 66.04% to 67.04%. Incorporating multi-scale features in the attention mechanism further enhances performance to 67.54%. Additionally, compared to standard cross-attention, masked cross-attention consistently achieves better performance with both single-scale and multi-scale features. Our decoder-only model, which integrates both multi-scale attention and masked cross-attention, achieves the highest performance, with a result of 67.67%.

We further investigated the impact of positional encodings on model performance. The removal of positional encoding resulted in only a marginal performance decrease of 0.1% compared to our decoder-only model. This suggests that the convolutional layers in our architecture may inherently capture rich positional information, reducing the necessity for explicit positional encodings.

4.5. 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 8 and Table 9. 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 (Ronneberger et al., 2015), DeepLabv3+ (Chen et al., 2018), and UNet++ (Zhou et al., 2019), complemented by attention-augmented CNN methods like AttnUNet (Schlemper et al., 2019) and Pyramid Attn (Li et al., 2018a), evaluated across resolutions of 224 × 224 and 512 × 512; (2) 3D approaches like V-Net (Milletari et al., 2016), DARR (Fu et al., 2020), 3D UX-Net (Lee et al., 2023), and 3D nnU-Net (Isensee et al., 2021), accompanied by cutting-edge Transformer-centric strategies including CoTR (Xie et al., 2021), nnFormer (Zhou et al., 2023), VT-UNet (Peiris et al., 2022), Swin UNETR (Hatamizadeh et al., 2021), and SwinUNETR-V2 (He et al., 2023). As corroborated by the results in Table 8, 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 including CoTr, nnFormer and Swin UNETR V2, our TransUNet at least achieves approximately a 10% improvement in Dice scores for the challenging task of gallbladder segmentation and about a 3% enhancement in overall segmentation. For all compared approaches, we download the public repositories and strictly follow their hyper-parameter settings to reproduce the results under the same split.

Notably, as evidenced in Table 11, our TransUNet surpasses the top-ranked solution, nnUNet-Large (Luu and Park, 2021), from the BraTS2021 challenge, underscoring the robustness and efficacy of our proposed approach.

4.6. Analysis of small tumor detection

We calculate the tumor segmentation performance (measured as % DSC) across various tumor sizes from the pancreatic tumor dataset. As shown in Table 10, for small PDAC tumors with a diameter less than 20 mm, our method exceeds the nnUNet baseline by 9.7% DSC. For small cysts with a diameter less than 10 mm, our method outperforms the nnUNet baseline by 4.3% DSC. Even for big tumors, our method also consistently outperforms nnUNet, despite achieving slightly less pronounced gains (5.0% improvement in PDAC segmentation and 3.0% improvement in cyst segmentation, respectively).

4.7. Efficiency analysis

We analyze the efficiency of major 3D models compared in Table 8, using five comprehensive metrics: inference speed (seconds per volume), training time (seconds per epoch), floating-point operations per second (FLOPs), and GPU memory footprint. All measurements were conducted on an NVidia A6000 GPU machine using identical profiling scripts. We maintained consistency by employing the same critical hyperparameters (*e.g.*, batch size set to 2, input size to (96, 96, 96)) to ensure a fair comparison. Based on the results in Table 12, we can see that although TransUNet has higher inference/training times compared to nnUNet due to the presence of self-attention, it is more efficient across all five metrics compared to recent models like 3D UX-Net and SwinUNETR-v2, while demonstrating superior segmentation performance (Tables 8 and 9). Notably, our model's GPU memory footprint is under 12 GB, enabling cost-effective training on resources such as the Titan-XP. Future work will focus on further reducing training costs and enhancing overall efficiency.

5. Conclusion

While U-Net has been successful, its limitations in handling long-range dependencies have prompted the exploration of Transformer as an alternative architecture. In this work, we introduce a Transformer-centric encoder–decoder framework, named TransUNet. Specifically, we introduce (1) A 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. We also propose a coarse-to-fine attention refinement to enhance the segmentation of small targets and tumors in the Transformer decoder. Through extensive experimentation, we

Table 8
Comparison on the BTCV multi-organ CT dataset (average dice score %, and dice score % for each organ).

Scale	Method	Param.	Memory[a]	Avg. Dice (%)	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
2D (224 × 224)	U-Net (Ronneberger et al., 2015)	32.5M	0.69G	74.68	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
	Pyramid Attn (Li et al., 2018a)	24.5M	0.67G	73.08	82.57	56.25	75.78	70.51	93.46	50.02	83.95	72.13
	DeepLabv3+ (Chen et al., 2018)	24.3M	0.66G	76.35	82.00	62.85	78.89	75.24	93.96	57.75	86.57	73.57
	UNet++ (Zhou et al., 2019)	49.0M	1.25G	76.65	86.93	63.69	77.86	68.29	93.91	59.23	87.81	75.49
	AttnUNet (Schlemper et al., 2019)	35.6M	1.49G	75.57	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
2D (512 × 512)	U-Net (Ronneberger et al., 2015)	32.5M	2.07G	81.34	89.69	69.98	83.08	74.13	95.10	67.73	90.50	80.51
	Pyramid Attn (Li et al., 2018a)	24.5M	1.57G	80.08	88.59	65.91	84.45	75.15	95.30	60.06	91.84	79.33
	DeepLabv3+ (Chen et al., 2018)	24.3M	1.84G	82.50	88.79	72.16	88.13	79.52	95.58	65.97	90.02	79.87
	UNet++ (Zhou et al., 2019)	49.0M	4.86G	81.6	89.65	71.68	82.92	75.15	94.92	69.06	89.42	80.01
	AttnUNet (Schlemper et al., 2019)	35.6M	7.24G	80.88	89.46	67.09	83.83	75.98	95.28	68.48	88.63	78.26
	nnU-Net (Isensee et al., 2021)	30.6M	2.51G	82.92	91.55	73.43	82.74	73.61	96.01	71.81	94.29	79.94
3D	V-Net (Milletari et al., 2016)	45.7M	3.81G	68.81	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
	DARR (Fu et al., 2020)	76.9M	6.17G	69.77	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
	nnU-Net (Isensee et al., 2021)	30.8M	5.81G	87.33	**93.04**	78.82	84.68	88.46	97.13	81.50	91.68	83.34
	CoTr (Xie et al., 2021)	41.9M	5.47G	85.72	92.96	71.09	85.70	85.71	96.88	81.28	90.44	81.74
	nnFormer (Zhou et al., 2023)	39.7M	7.29G	85.32	90.72	71.67	85.60	87.02	96.28	82.28	87.30	81.69
	VT-UNet (Peiris et al., 2022)	20.8M	6.55G	70.72	78.25	44.76	77.51	78.16	91.63	45.18	82.20	68.04
	Swin UNETR (Hatamizadeh et al., 2021)	62.0M	13.71G	82.33	90.17	70.83	84.76	83.89	95.50	69.39	91.37	72.76
	SwinUNETR-V2 (He et al., 2023)	72.8M	13.91G	83.23	90.97	69.33	**86.84**	86.74	94.93	69.72	89.65	77.71
	3D UX-Net (Lee et al., 2023)	53.0M	11.70G	83.82	89.94	71.42	86.12	85.86	93.25	72.63	92.03	79.32
	TransUNet (Decoder-only)	33.6M	11.16G	87.63	92.88	**82.06**	86.04	87.70	97.10	82.08	91.14	82.03
	TransUNet (Encoder+Decoder)	41.4M	11.26G	**88.39**	**93.04**	82.04	85.67	**88.87**	**97.18**	**82.92**	**92.36**	**85.06**

[a] The GPU memory is measured when batch size is set to 2 for all networks and the crop-size is set to (96,96,96) for 3D networks.

Table 9
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 (Zhou et al., 2023)	63.21	69.37	66.29
VT-UNet (Peiris et al., 2022)	60.88	59.82	60.35
Swin UNETR (Hatamizadeh et al., 2021)	57.65	58.31	57.98
TransUNet (Decoder-only)	64.41	**70.94**	**67.67**
TransUNet (Encoder+Decoder)	**64.58**	69.89	67.24

Table 10
Tumor segmentation performance (average Dice) under different tumor sizes reported in the pancreatic tumor dataset.

		Tiny <10 mm	Small [10 mm, 20 mm)	Big >=20 mm
PDAC	nnUNet	0	26.8%	63.7%
	Ours	0	36.5%	68.7%
	Gain	0	**9.7%** ↑	**5.0%** ↑
Cyst	nnUNet	55.7%	56.1%	66.8%
	Ours	60.0%	58.8%	69.8%
	Gain	**4.3%** ↑	**2.7%** ↑	**3.0%** ↑

Table 11
Performance comparison on the 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 (Luu and Park, 2021)	87.23	91.88	93.21	90.77
nnUNet-Large (Luu and Park, 2021)	88.23	92.35	93.83	91.47
TransUNet (Decoder-only)	**88.85**	**92.48**	93.90	**91.74**
TransUNet (Encoder+Decoder)	**88.85**	92.43	**93.91**	91.73

provide the first thorough investigation on the impact of integrating the Transformer encoder and decoder into U-Net architectures, providing insights for addressing diverse challenges in medical image segmentation. Empirical results showcase TransUNet's superior performance in multi-organ, pancreatic tumor, hepatic vessel and tumor segmentation. Additionally, we surpassed top-1 solution in the BraTS2021 challenge. Additionally, we have released our codebase to facilitate further exploration and encourage the adoption of Transformers in medical applications, offering both 2D and 3D implementations for user convenience.

Table 12
Comparison of different 3D models based on inference time (seconds per volume), training time (seconds per epoch), FLOPs, and GPU memory usage.

Model	Param.	Infer-Time	Train-Time	FLOPs	Memory
nnUNet	30.8M	0.016 s	101 s	315.7G	5.81G
CoTr	41.9M	0.026 s	141 s	295.4G	5.47G
nnFormer	39.7M	0.024 s	135 s	188.6G	7.29G
SwinUNETR	62.0M	0.057 s	198 s	336.8G	13.71G
SwinUNETR-v2	72.8M	0.062 s	215 s	362.6G	13.91G
3D UX-Net	53.0M	0.068 s	231 s	362.6G	11.70G
TransUNet	41.4M	0.058 s	188 s	362.3G	11.26G

CRediT authorship contribution statement

Jieneng Chen: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Conceptualization. **Jieru Mei:** Methodology, Investigation. **Xianhang Li:** Writing – review & editing. **Yongyi Lu:** Writing – review & editing. **Qihang Yu:** Writing – review & editing. **Qingyue Wei:** Writing – review & editing. **Xiangde Luo:** Writing – review & editing. **Yutong Xie:** Writing – review & editing. **Ehsan Adeli:** Writing – review & editing. **Yan Wang:** Writing – review & editing. **Matthew P. Lungren:** Writing – review & editing. **Shaoting Zhang:** Writing – review & editing. **Lei Xing:** Writing – review & editing. **Le Lu:** Writing – review & editing. **Alan Yuille:** Writing – review & editing, Funding acquisition. **Yuyin Zhou:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work is partially supported by the TPU Research Cloud (TRC) Program, USA, Google Cloud Research Credits Program, USA, the AWS Public Sector Cloud Credit for Research Program, USA, and a 2023 Patrick J. McGovern Foundation Award, USA.

References

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

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision. Springer, pp. 205–218.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp. 213–229.

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

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818.

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299.

Cheng, B., Schwing, A., Kirillov, A., 2021. Per-pixel classification is not all you need for semantic segmentation. Adv. Neural Inf. Process. Syst. 34.

Child, R., Gray, S., Radford, A., Sutskever, I., 2019. Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509.

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

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

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

He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D., 2023. Swinunetr-V2: Stronger swin transformers with stagewise convolutions for 3D medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 416–426.

Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y., 2022. Missformer: An effective transformer for 2d medical image segmentation. IEEE Trans. Med. Imaging.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18 (2), 203–211.

Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A., 2017. Multi-atlas labeling beyond the cranial vault-workshop and challenge.

Lee, H.H., Bao, S., Huo, Y., Landman, B.A., 2023. 3D ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. In: International Conference on Learning Representations.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A., 2018b. H-DenseUNet: hybrid densely connected unet for liver and tumor segmentation from CT volumes. IEEE Trans. Med. Imaging 37 (12), 2663–2674.

Li, H., Xiong, P., An, J., Wang, L., 2018a. Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Luo, X., Chen, J., Song, T., Chen, Y., Wang, G., Zhang, S., 2021. Semi-supervised medical image segmentation through dual-task consistency. AAAI Conf. Artif Intell.

Luu, H.M., Park, S.H., 2021. Extending nn-UNet for brain tumor segmentation. arXiv preprint arXiv:2112.04653.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D., 2018. Image transformer. In: International Conference on Machine Learning. PMLR, pp. 4055–4064.

Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M., 2022. A robust volumetric transformer for accurate 3D tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 162–172.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention gated networks: Learning to leverage salient regions in medical images. Med. Image Anal. 53, 197–207.

Strudel, R., Garcia, R., Laptev, I., Schmid, C., 2021. Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7262–7272.

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H., 2021. Going deeper with image transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 32–42.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008.

Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803.

Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.C., 2021. 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.

Xie, L., Yu, Q., Zhou, Y., Wang, Y., Fishman, E.K., Yuille, A.L., 2019. Recurrent saliency transformation network for tiny target segmentation in abdominal CT scans. IEEE Trans. Med. Imaging 39 (2), 514–525.

Xie, Y., Zhang, J., Shen, C., Xia, Y., 2021. 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, pp. 171–180.

Yu, L., Cheng, J.Z., Dou, Q., Yang, X., Chen, H., Qin, J., Heng, P.A., 2017. Automatic 3D cardiovascular MR segmentation with densely-connected volumetric convnets. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 287–295.

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

Yu, Q., Wang, H., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.C., 2022b. K-means mask transformer. In: European Conference on Computer Vision. Springer, pp. 288–307.

Yu, Q., Xie, L., Wang, Y., Zhou, Y., Fishman, E.K., Yuille, A.L., 2018. 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. pp. 8280–8289.

Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y., 2023. nnFormer: Volumetric medical image segmentation via a 3D transformer. IEEE Trans. Image Process..

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2019. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging 39 (6), 1856–1867.

Zhou, Y., Xie, L., Shen, W., Wang, Y., Fishman, E.K., Yuille, A.L., 2017. A fixed-point model for pancreas segmentation in abdominal CT scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 693–701.

Zhu, Z., Xia, Y., Shen, W., Fishman, E., Yuille, A., 2018. A 3D coarse-to-fine framework for volumetric medical image segmentation. In: 2018 International Conference on 3D Vision (3DV). IEEE, pp. 682–690.

Zhu, Z., Xia, Y., Xie, L., Fishman, E.K., Yuille, A.L., 2019. 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, pp. 3–12.