HHS Public Access

Author manuscript

Med Image Anal. Author manuscript; available in PMC 2024 July 08.

Published in final edited form as:

Med Image Anal. 2023 December ; 90: 102939. doi:10.1016/j.media.2023.102939.

UNesT: Local spatial representation learning with hierarchical transformer for efficient medical segmentation

Xin Yu[a], **Qi Yang**[a], **Yinchi Zhou**[a], **Leon Y. Cai**[b], **Riqiang Gao**[a,c], **Ho Hin Lee**[a], **Thomas Li**[b], **Shunxing Bao**[d], **Zhoubing Xu**[c], **Thomas A. Lasko**[e], **Richard G. Abramson**[b,f], **Zizhao Zhang**[g], **Yuankai Huo**[a,d], **Bennett A. Landman**[a,b,d,e], **Yucheng Tang**[d,h,*]

[a]Department of Computer Science, Vanderbilt University, Nashville TN, 37212, USA

[b]Department of Biomedical Engineering, Vanderbilt University, Nashville, TN, 37212, USA

[c]Digital Technology and Innovation, Siemens Healthineers, Princeton, NJ, 08540, USA

[d]Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN, 37212, USA

[e]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, 37212, USA

[f]Annalise-AI, Pty, Ltd, USA

[g]Google Cloud AI, USA

[h]Nvidia Corporation, USA

Abstract

Transformer-based models, capable of learning better global dependencies, have recently demonstrated exceptional representation learning capabilities in computer vision and medical image analysis. Transformer reformats the image into separate patches and realizes global communication via the self-attention mechanism. However, positional information between patches is hard to preserve in such 1D sequences, and loss of it can lead to sub-optimal performance when dealing with large amounts of heterogeneous tissues of various sizes in 3D medical image segmentation. Additionally, current methods are not robust and efficient for heavy-duty medical segmentation tasks such as predicting a large number of tissue classes or modeling globally inter-connected tissue structures. To address such challenges and inspired by the nested hierarchical structures in vision transformer, we proposed a novel 3D medical image segmentation method (UNesT), employing a simplified and faster-converging transformer encoder design that achieves local communication among spatially adjacent patch sequences by aggregating them hierarchically. We extensively validate our method on multiple challenging datasets, consisting of

*Corresponding author at: Nvidia Corporation, USA. yuchengt@nvidia.com (Y. Tang).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yucheng Tang reports financial support was provided by Vanderbilt University. Yucheng Tang reports a relationship with Vanderbilt University that includes: employment.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2023.102939.

multiple modalities, anatomies, and a wide range of tissue classes, including 133 structures in the brain, 14 organs in the abdomen, 4 hierarchical components in the kidneys, inter-connected kidney tumors and brain tumors. We show that UNesT consistently achieves state-of-the-art performance and evaluate its generalizability and data efficiency. Particularly, the model achieves whole brain segmentation task complete ROI with 133 tissue classes in a single network, outperforming prior state-of-the-art method SLANT27 ensembled with 27 networks. Our model performance increases the mean DSC score of the publicly available Colin and CANDI dataset from 0.7264 to 0.7444 and from 0.6968 to 0.7025, respectively. Code, pre-trained models, and use case pipeline are available at: https://github.com/MASILab/UNesT.

Keywords

Hierarchical transformer; Whole brain segmentation; Renal substructure segmentation

1. Introduction

Medical image segmentation tasks have become increasingly challenging due to the need for modeling hundreds of tissues (Huo et al., 2019; Wasserthal et al., 2022) or hierarchically inter-connected structures (Landman et al., 2015) in 3D volumes. In the past few years, convolutional neural networks (CNNs) have dominated medical image segmentation due to their superior performance. Among all the CNNs, the U-Net (Ronneberger et al., 2015) and its variants have been the most widely used for medical image segmentation. A ‘‘U-shape’’ model generally consists of an encoder for global representation learning and a decoder to gradually decode the learned representation to a pixel-wise segmentation. However, CNN-based models’ encoding performance is limited because of their localized receptive fields (Hu et al., 2019).

Vision Transformers (ViT), on the other hand, are capable of learning long-range dependencies and have recently demonstrated exceptional representational learning capabilities and effectiveness in computer vision and medical image applications (Dosovitskiy et al., 2020; Hatamizadeh et al., 2022; Zhou et al., 2021a). Unlike CNNs, ViTs learn better long-range information by tokenizing images into 1D sequences and leveraging the self-attention blocks to facilitate global communication (Hatamizadeh et al., 2022), which makes transformers better encoders. However, by tokenizing the image into 1D patches, transformers are less able to capture local positional information compared to CNNs, due to the lack of locality inductive bias inherent to CNNs (Cordonnier et al., 2019; Dosovitskiy et al., 2020). To overcome this, ViT usually requires a large amount of training data which is expensive to acquire (Tang et al., 2022; Zhou et al., 2021a). With small datasets in the medical field, insufficient data can lead to model inefficiency, especially when dealing with a large number of tissues of various sizes. Moreover, the self-attention mechanism for modeling multi-scale features for high-resolution medical volumes is computationally expensive (Beltagy et al., 2020; Han et al., 2021; Liu et al., 2021).

To improve representation learning in transformers in small datasets, recent works envision the use of local self-attention (Liu et al., 2021; Cao et al., 2021; Han et al., 2021). To

leverage information across embedded sequences, ‘‘shifted window’’ transformers (Liu et al., 2021) have been proposed for dense predictions and modeling multi-scale features. However, these attempts aiming to adapt the self-attention mechanism by modifying patch communication often yield high computational complexity. In addition, the Swin transformer under-performs when datasets are small, or there are a large number of structures (Liu et al., 2021).

Considering the advantages of hierarchical models (Çiçek et al., 2016; Roth et al., 2018; Tang et al., 2022) on modeling heterogeneous high-resolution radiographic images and inspired by the aggregation function in the nested ViT (Zhang et al., 2022), we propose a Hierarchical hybrid 3D U-shape medical segmentation model with Nested Transformers (UNesT). Specifically, with nested transformers as the encoder, UNesT hierarchically encodes features with the 3D block aggregation function and merges with the convolutional-based decoder via skip connections at various resolutions to enable learning of local behaviors for small structures or small datasets. The 3D nested structure retains the original global self-attention mechanism and achieves information communication across patches by stacking transformer encoders hierarchically.

We perform extensive experiments to validate the performance of UNesT on the challenging whole brain segmentation task with 133 classes using T1 weighted (T1w) MRI images and a collected renal substructures 3D CT volumetric dataset with 116 patients on characterizing multiple kidney components including renal cortex, medulla and pelvicalyceal system with kidney function. We further evaluate UNesT on three widely-used public datasets Beyond The Cranial Vault (BTCV) (Landman et al., 2015), KiTS19 (Heller et al., 2021), and BraTS21 (Baid et al., 2021) to illustrate the generalizability of UNesT. We compare UNesT to recent convolutional and transformer-based 3D medical segmentations baselines and conduct scalability and data efficiency analysis in a low-data regime.

Our contributions to this work can be summarized as:

- We introduce a novel 3D hierarchical block aggregation module, and propose a new transformer-based 3D medical segmentation model, dubbed UNesT. The model provides local spatial patch communication to better capture various tissues. This method achieves hierarchical modeling of high-resolution medical images and outperforms local self-attention variants with a simplified design compared to the ‘‘shifted window’’ module leading to improved data efficiency.
- We validate UNesT on a whole brain segmentation task that contains hundreds of classes. UNesT outperforms the current convolutional- and transformer-based single model methods. Our single model also outperforms the prior top method SLANT27 (Huo et al., 2019) ensembled with 27 networks, and achieves new state-of-the-art performance.
- We collect and manually delineate the first in-house renal substructure dataset (116 CT subjects). We show that our method achieves state-of-the-art performance for accurately measuring cortical, medullary, and pelvicalyceal system volumes. We demonstrate the clinical utility of this work through accurate volumetric analysis, strong correlations, and robust reproducibility. We

also introduce MONAI Bundle, a new plug-and-use framework for deploying models. Our codes, trained models, and tutorials are released for public availability.

- We investigate model scalability and data efficiency in low-data regimes as well as the impact of the size of pre-training dataset. We show the proposed method's generalizability by validating it on public datasets: BTCV, KiTS19, and BraTS2021.

2. Related works

Medical Segmentation with Transformers.

Transformer models demonstrate the ability of modeling longer-range dependencies for high dimension and high-resolution medical images in 3D Space. The scalability, generalizability, and efficiencies of ViT and hierarchical transformers enable stronger representation learning for dense predictions (e.g., pixel-to-pixel segmentation). Medical image segmentation tasks embed learning problems with multi-scale features instead of fixed scale, such as word tokens. To employ the vanilla Transformer (Dosovitskiy et al., 2020) for medical images, recent works proposed variant architectures that use ViT as network components.

Transformer is known for its capability of capturing long-range dependencies but lacks inductive bias, which is inherent in CNNs. To reap the advantages of both CNNs and transformers, many efforts have been made to integrate the benefits of CNNs and transformers into a hybrid network. In the medical image segmentation domain, these works can be classified into three types: Transformer as main encoder, transformer as secondary encoder, and fusion model of both transformer encoder and CNNs encoder (Li et al., 2023).

When utilizing the Transformer as the primary encoder, the segmentation model usually includes a sequence of successive transformer blocks as the encoder. Various studies have used this design, such as UNETR, VT-UNet, and SwinUNETR (Hatamizadeh et al., 2022; Peiris et al., 2021; Tang et al., 2022). The advantage of sequence-to-sequence modeling as the first embedding for medical images is to directly generate tokenized patches for the feature representation. Most of these methods connect a convolutional neural network (CNN)-based decoder and form the ‘‘U-shaped’’ architecture for segmentation. This design features the long-range modeling ability for input images with a transformer encoder and better inductive bias with CNN decoder.

The second design utilizes a transformer as a secondary encoder after the CNNs encoder. The reason for this design is two-fold. Firstly, due to the lack of inductive bias in transformer models, encoding image feature with CNN networks leads to superior global feature modeling. Secondly, performing global self-attention on voxels in high-resolution medical images is computationally intensive. By using the CNN encoder first, the computational workload can be significantly reduced (Li et al., 2023). One early use of vanilla transformer blocks for medical segmentation is the TransUNet (Chen et al., 2021b), which used 12 2D transformer layers for encoding bottleneck features. TransUNet++ (Wang et al., 2022), AFTer-UNet (Yan et al., 2022), TransClaw (Chang et al., 2021), Ds-TransUNet

(Lin et al., 2021), TransAttUNet (Chen et al., 2021a) and GT-UNet (Li et al., 2021b) improved the self-attention blocks and achieved promising performance in CT segmentation. In addition, TransBTS (Wang et al., 2021a), CoTr (Xie et al., 2021b), and TrasnBridge (Deng et al., 2021) explored variant modules such as deformable transformer blocks for 3D image segmentation tasks. Later, SegTrans (Li et al., 2021a), MT-UNet (Wang et al., 2021c) introduced squeeze and expansion mechanisms and mixed structure for modeling context affinities. BAT (Wang et al., 2021b) and Poly-PVT (Dong et al., 2021) used grouping or boundary-aware designs to improve transformer robustness with cross-slice attention.

The third design utilizes both transformer and CNNs encoders in parallel, which are also called fusion models. This design aims to take the global and local information from the transformer and CNNs encoder, respectively, for better representation learning. The encoded representations by two encoders are then fused into a single decoder. TransFuse (Zhang et al., 2021a), and FusionNet (Meng et al., 2021) are pioneering works that benefit from learning global and local features. The PMtrans (Zhang et al., 2021b) and X-Net (Li et al., 2021c) introduce a multi-branch pyramid and a dual encoding network which demonstrate leading results on pathology images. MedT (Valanarasu et al., 2021) and Ds-TransUNet (Lin et al., 2021) proposed a CNN global branch and a local transformer branch with an axial self-attention module. With a fusion model, input medical images are split into both whole feature and non-overlapping patches followed by two encoder branches. With fusion designs, model complexities are commonly large due to the additional encoding branches, which is a disadvantage of these models. To the best of our knowledge, no fusion model has been proposed for volumetric medical image segmentation.

Recently, scientists have investigated the full adoption of transformer models for medical image segmentation. There are challenges in using pure transformer models, especially for 3D images, due to the limitation of inductive bias and the high complexity of transformers. Swin UNet (Cao et al., 2021) is a pure transformer model designed for 2D medical images. It adopted the ‘‘U-shape’’ architecture and used a skip connection that connected the encoded features to the transformer decoder. D-Former (Wu et al., 2022) utilized dynamic position encoding blocks and local scope modules for improving local feature representation learning. MISSformer (Huang et al., 2021) is a pure transformer network with feed-forward enhanced blocks in its transformer modules. This design leveraged long-range dependencies with local features at different scales. The nnFormer (Zhou et al., 2021a) is another promising network that used 3D transformers and combined encoder and decoder with self-attention operations. nnFormer incorporated a skip attention mechanism to replace simple skip connections, which outperformed CNN-based methods significantly. Though the use of pure transformers as the model is more intuitive and better for design consistency; Yet, there are still uncharted areas using self-attention in the decoder. High model complexity can cause unsatisfied robustness and is challenging to explore in 3D context.

Pre-training transformers with a large-scale dataset are of potential value to boost transformer model performance (Dosovitskiy et al., 2020). Empirical studies (Zhai et al., 2022) show that the transformer model can have better scalability when more data are fed. In the medical domain, researchers have explored self-supervised pre-training approaches with CNNs (Zhou et al., 2021b). More recently, pre-training 3D transformers (Tang et al.,

2022) for radiological images have been presented. Furthermore, uniformed pre-training frameworks (Xie et al., 2021a,c) are shown to construct teacher–student models for medical data. However, the use of pre-training is computationally exhaustive. In this paper, we aim to simplify and evaluate the effect of the pre-training framework with empirical studies.

Hierarchical Feature Aggregation.

The aggregation of multi-level features could improve segmentation results by merging the features extracted from different layers. Modeling hierarchical features, such as U-Net (Çiçek et al., 2016) and pyramid networks (Roth et al., 2018), multi-scale representations are leveraged. The extended feature pyramids compound the spatial and semantic information through two structures, iterative deep layer aggregation which fuses multi-scale information as well as deep hierarchical aggregation which fuses representations across channels. In addition to a single network, nested UNets (Zhou et al., 2018), nnUNets (Isensee et al., 2021), coarse-to-fine (Zhu et al., 2018) and Random Patch (Tang et al., 2021a) suggest multi-stage pathways enrich the different semantic levels of features progressively with cascaded networks. Different from the above CNN-based methods, we explore the use of data-efficient transformers for modeling hierarchical 3D features by block aggregation.

3. Method

3.1. Hierarchical transformer encoder

The overall UNesT architecture is shown in Fig. 1. The input image is a sub-volume $\mathcal{x} \in \mathbb{R}^{H \times W \times D \times C}$ and the volumetric embedding token has a patch size of $S_h \times S_w \times S_d \times C$. 3D tokens are projected onto a size of $\frac{H}{S_h} \times \frac{W}{S_w} \times \frac{D}{S_d} \times C'$ in the patch projection step, where C' is the embedded dimension. Following the motivation in Zhang et al. (2022) for efficient non-local communication, all projected sequences of embeddings are partitioned to blocks (blockify) with a resolution of $\mathcal{x} \in \mathbb{R}^{b \times T \times n \times C'}$, where T is the number of blocks at the current hierarchy, b is the batch size, n is the total length of sequences, as shown in Fig. 1. The blocks are non-overlapping and n remains the same in different hierarchies. The dimensions of the embeddings follow $T \times n = \frac{H}{S_h} \times \frac{W}{S_w} \times \frac{D}{S_d}$. Each block is fed into sequential of transformer layers (Sharir et al., 2021) separately, which consist of the canonical multi-head self-attention (MSA), multi-layer perceptron (MLP) with skip connection (He et al., 2016), and layer normalization (LN) (Ba et al., 2016). We add learnable position embeddings to sequences for capturing spatial relations before the blocked transformers. The output of transformer encoder is computed as follows:

$$\begin{aligned} \hat{z}^t &= \mathrm{MSA}_{\mathrm{HRCHY}_1}\left(\mathrm{LN}\left(z^t - 1\right)\right) + z^t - 1, t = 1 \ldots L \\ z^t &= \mathrm{MLP}\left(\mathrm{LN}\left(\hat{z}^t\right)\right) + \hat{z}^t, t = 1 \ldots L \end{aligned} \tag{1}$$

where $\mathrm{MSA}_{\mathrm{HRCHY}_l}$ denotes the multi-head self-attention layer of hierarchy l, $\hat{z}^t$ and z^t are the output representations of MSA and MLP and L denotes the number of transformer layers. z^{t-1} denotes the input of the transformer encoder, as shown in Fig. 1, after undergoing layer hlnormalization and MSA, z^{t-1} is added to the output via a skip connection to produce $\hat{z}^t$. The result is then passed through another layer hlnormalization and a MLP. Finally, $\hat{z}^t$ is added to the final result to generate the output of a transformer layer, denoted by z^t, which serves as the input to the subsequent transformer layer.

In practice, $\mathrm{MSA}_{\mathrm{HRCHY}_l}$ is applied in parallel to all partitioned blocks:

$$\begin{aligned} &\mathrm{MSA}_{\mathrm{HRCHY}_l}(Q,\ K,\ V) = \mathrm{Stack}(\mathrm{BLK}_1,\ \ldots,\ \mathrm{BLK}_T) \\ &\mathrm{BLK} = \mathrm{MSA}(Q,\ K,\ V) = \mathrm{Stack}\left(\mathrm{Softmax}\left(\frac{Q_i K_i^{\mathrm{T}}}{\sqrt{\sigma}}\right)V_i\right)W^o,\ i = 1\ldots H, \end{aligned} \tag{2}$$

where Q, K, V denotes queries, keys, and value vectors in the multi-head attention, and H represents the total heads in the MSA. In each block, Q, K, V have the dimension of σ. Dot-product attention is applied between Q and K^{T} to get an attention matrix of size $\sigma \times \sigma$. To overcome the issue that when σ has a large value, the dot product between Q and K becomes magnified causing the softmax function to produce extreme value, the dot products is scaled by $\frac{1}{\sqrt{\sigma}}$ Vaswani et al. (2017). V is then multiplied with the attention matrix to get the final output with size σ. The computation of each head is performed in parallel and then combined through concatenation. The final result is then reshaped with matrix $W \in \mathbb{R}^{H \cdot \sigma \times d_{out}}$ to match the output dimension. As previously stated, each block shares a common size of $b \times n \times C'$ within the same hierarchy so that the MSA output of each block has the same size. The outputs of each block are concatenated to obtain the final results, which represent the MSA of that particular hierarchy. All blocks at each level of the hierarchy share the same parameters given the input $\mathcal{X}$, which leads to hierarchical representations without increasing complexity.

3.2. 3D block aggregation

We extend the spatial nesting operations in Zhang et al. (2022) to 3D blocks to form a local attention hierarchical design. Different to Liu et al. (2021), Tang et al. (2022), which utilizes global attention among ‘‘shift windows’’. In our design, transformer encoders are applied to each volume block separately to achieve local attention, with each block being modeled independently. Information across blocks is communicated by the aggregation module. This design leads to reduced computational complexity and improved data efficiency.

In the first hierarchy, suppose the input feature size is $H' \times W' \times D' \times C'$. The input feature is blockified into T blocks with the aforementioned size of $T \times n \times C'$. After the transformer encoder, the blocks are transformed back to the feature map with size $H' \times W' \times D' \times C'$, which serves as the input of the next hierarchy. In the following hierarchy, the input feature downsamples by a factor of 2 for each dimension and

transformed to embedded dimension before blockify with a pooling block consisting of a convolutional layer, a normalization layer and a max pooling layer to build multi-scale feature maps as in Ronneberger et al. (2015), Liu et al. (2021) for better representation learning. Applying the pooling block before blockify facilitates the information exchange between blocks and enables non-local communication because it allows convolution and pooling to be performed on the spatial area that belongs to different blocks after blockifying. The pooling block reduces the feature size by $2 \times 2 \times 2$, and the sequence length n remains the same, the number of blocks T reduces by a factor of 8 in each hierarchy as a result. Consequently, in the subsequent level of the hierarchy, the feature maps are blockified to the size of $\frac{T}{8} \times n \times C''$, where C'' represents the embedded dimension in that hierarchy. The unblocked feature maps after the transformer encoder have a size of $\frac{H'}{2} \times \frac{W'}{2} \times \frac{D'}{2} \times C''$. This process continues until the number of blocks T reach 1.

In our model design, there are three hierarchies which result in a total number of 64, 8, and 1 block in each hierarchy. In the volumetric plane, the encoded blocks are merged among adjacent block representations, as it shown in Fig. 1. The design and use of the aggregation modules in the 3D scenario leverage local attention and improved data efficiency which we demonstrate in our ablation studies.

3.3. Decoder

To better capture localized information and further reduce the effects of lacking inductive bias in transformers, we use a hybrid design with a convolution-based decoder for segmentation.

We use a patch size of $4 \times 4 \times 4$ in our encoder. The patch embedded dimension is set to 128 so that the feature size is $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 128$ after patch projection. The 3 hierarchies have a number of transformer layers (depth) of 2, 2, and 8 and embedded dimensions (width) of 128, 256, and 512, respectively. The feature size at the end of each hierarchy is $\frac{H}{4 \times 2^i} \times \frac{W}{4 \times 2^i} \times \frac{D}{4 \times 2^i} \times C$ where $i = 0, 1, 2$ and $C = 128, 256, 512$, as shown in Fig. 1 The feature map from the last hierarchy is fed into a layer normalization layer to generate the transformer encoder output.

As previously mentioned, the feature size is reduced by a factor of 2 in each dimension at each level of the hierarchy, resulting in the generation of multi-resolution feature maps. Inspired by the U-shape models (Ronneberger et al., 2015) which utilize multi-scale strategy, we merge the multi-resolution features which is the output of each hierarchy with the decoder with skip connections followed by convolutional layers.

The bottleneck is generated by feeding the output of the last hierarchy to a layer normalization followed by a $3 \times 3 \times 3$ convolutional layer. We upsample the bottleneck by applying a transpose convolutional layer. The output of the transposed convolution is concatenated with the prior hierarchical representations and fed into a residual block consisting of two $3 \times 3 \times 3$ convolutional layers, each followed by an instance

normalization (Ulyanov et al., 2016) layers (Fig. 1). In each hierarchy, excluding the last one, we have incorporated a residual block with the aforementioned layers in the skip connection between the hierarchy's output (encoder) and decoder. Since we believe these residual blocks can help minimize the semantic gap between the features from the transformer encoder and the CNN decoder. However, since the bottleneck in our architecture involves transforming the output of the last layer using a normalization layer and a convolutional layer, the semantic meaning is expected to be similar to that of the last hierarchy's output. Therefore, we choose not to include a residual block in the skip connection of the last hierarchy. The processed feature maps from the encoder are then concatenated with the feature maps from lower hierarchies or bottleneck upsampled by transposing convolutional layers. This merged feature map is then passed through another residual block with the layers mentioned earlier to merge the information from both the encoder and decoder. To enhance the semantic information, we apply the aforementioned residual block to both the input image and the features obtained after the path projection. The resulting features were then passed to the decoder through the skip connection, as illustrated in Fig. 1. The segmentation mask is acquired by $1 \times 1 \times 1$ convolutional layer with a softmax activation function. Compared to some prior related works such as TransBTS (Wang et al., 2021a) and CoTr (Xie et al., 2021b), our design employs the hierarchical transformer directly on images and extracts representations at multiple scales without convolutional layers.

4. Experiments

4.1. Dataset

Whole Brain Segmentation Dataset.—Training and testing data are MRI T1-weighted (T1w) 3D volumes from 10 different sites. The training set consists of 50 scans from the Open Access Series on Imaging Studies (OASIS) (Marcus et al., 2007) dataset which is manually traced to 133 labels based on the BrainCOLOR protocol (Klein et al., 2010) by Neuromorphometrics Inc. The size of the data is 256 × 256 × [270, 334] with 1 mm isotropic spacing. The testing cohort contains Colin27 (Colin) T1w scan (Aubert-Broche et al., 2006) and 13 T1w MRI scans from the Child and Adolescent Neuro Development Initiative (CANDI) (Kennedy et al., 2012) dataset. The Colin dataset contains one high-resolution scan averaging from 27 scans of the same subject. The label is manually traced to 130 labels based on BrainCOLOR protocol. The size of the scan is 362 × 362 × 434 with 0.5 mm isotropic spacing. The CANDI dataset is manually traced to 130 labels following the BrainCOLOR protocol. The size of the scan is 256 × 256 × 128 with spacing of 0.94 mm × 0.94 mm × 1.5 mm. A detailed class name and the 3 classes not labeled in the test sets can be found in Table A.13 in the supplementary material. The CANDI dataset contains a different age group (5–15 years old) compared to the OASIS training cohort (18–96 years old), which allowed assessment of different populations. Following the same practice in Huo et al. (2019), we use auxiliary labels comprising of 4859 T1w MRI scans from eight different sites whose labels are generated by using an existing multi-atlas segmentation pipeline (Asman and Landman, 2014) to pre-train the model and finetune the pre-trained model with the 50 manually traced data from the OASIS dataset. A detailed summary of the 4859 multi-site images is shown in Table 1.

Renal Substructure Dataset.—We construct an internal cohort of the renal substructures segmentation dataset with 116 subjects imaged under institutional review board (IRB) approval (IRB #131461). Cortex, medulla, and pelvicalyceal systems are labeled in the dataset (Fig. 2). Data with ICD codes related to kidney dysfunction are excluded since they could potentially influence kidney anatomy. The left and right renal structures are outlined manually by three interpreters under the supervision of clinical experts. The renal columns are included in the cortex label. The medulla is surrounded by the cortex, and the pelvicalyceal systems contain calyces and pelvis that drain into the ureter. All manual labels are verified and corrected independently by expert observers. For the test set of 20 subjects, we perform a second round of manual segmentation (interpreter 2) to assess the intra-rater variability and reproducibility. The image size of each scan is 512 × 512 × [90, 131] with spacing of [0.54, 0.98] mm × [0.54, 0.98] mm × 3.0 mm.

Multi-organ Segmentation (BTCV) Dataset.—We evaluate model generalizability with the Beyond The Cranial Vault (BTCV) dataset. It is comprised of 100 de-identified contrast-enhanced CT volumes with 13 labeled anatomies, including spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, inferior vena cava (IVC), portal and splenic veins (PSV), pancreas, right and left adrenal gland. The image size of each scan is 512 × 512 × [80, 255] with the spacing of [0.54, 0.98] mm × [0.54, 0.98] mm × [2.5, 7.0] mm. 50 scans are publicly available in the MICCAI 2015 Multi-atlas Labeling Challenge (Landman et al., 2015), in which 20 scans are used for public testing.

KiTS19.—To further validate the generalizability of the proposed method for characterizing renal tissues, we apply the model to the public KiTS19 dataset. The KiTS19 (Heller et al., 2021) task focuses on the whole kidney and kidney tumor segmentation. Images and labels from 210 subjects are publicly available. The image size of each scan is 512 × [512, 796] × [29, 1059] with spacing of [0.44, 1.04] mm × [0.44, 1.04] mm × [0.5, 5.0] mm.

Brats.—The BraTS 2021 dataset contains 1251 subjects, and each scan is associated with 4 MRIs: (1) native (T1) and (2) post-contrast T1-weighted (T1Gd), (3) T2-weighted (T2), and (4) T2 Fluid-attenuated Inversion Recovery (T2-FLAIR). Each subject's images are registered and resampled to 1.0 × 1.0 × 1.0 mm isotropic resolution. The input 3D volumes are of size 240 × 240 × 155.

4.2. Implementation details

To eliminate the impact of data difference on the final performance, all baseline models undergo the same data augmentation and pre-processing steps, except of the nnUNet framework-based methods. Experiments are implemented in Pytorch and MONAI[1]. All segmentation models are trained with a single Nvidia RTX 5000 16G GPU with an input volume size of $96 \times 96 \times 96$. We follow the initial learning rate and weight decay configurations used in each respective baseline, provided that the original baseline was tested on the same task. If the baseline does not test on that task, for nn-UNet based method (nn-UNet and CoTr), we keep the initial settings since the training planner in the code has

[1] https://monai.io/

the original learning rate set as default. For the other models, we adopt an identical learning rate of 1e-4 and weight decay of 1e-4. The rationale behind this decision is twofold. Firstly, these baselines' initial learning rate and weight decay settings match or closely resemble this setting. Secondly, we utilize a cosine scheduler with 500 step warm-up for all the models which will adjust the learning rate based on the models accordingly.

4.2.1. Whole brain segmentation—During pre-training with auxiliary labels, the learning rate is initialized to 0.0001 with weight decay of $1e^{-4}$ to train for 200 K iterations. During finetuning, the learning rate is set to $1e^{-5}$ to train for 50 K iterations. As shown in Fig. 3, all data are registered to the MNI space using the MNI305 (Evans et al., 1993) template and preprocessed following the method in Huo et al. (2019). All processed images have a size of 172 × 220 × 156 with isotropic spacing of 1 mm. Registered input images are randomly cropped to the size of 96 × 96 × 96 during the online augmentation. We use a five-fold cross-validation strategy during finetuning. The best-performing model in each fold is selected to test on the external testing set and ensembled to get the final prediction in the MNI space. Predictions in MNI space are inverse transformed to the original space using NiftyReg (Ourselin et al., 2001) for evaluation (Fig. 3). No data augmentation is used in all the experiments due to the negative impact on model performance observed during our experiments. Segmentation performances are evaluated using Dice similarity coefficient (DSC) and symmetric Hausdorff Distance (HD).

4.2.2. Renal substructures segmentation—Five-fold cross-validation is used for all experiments on 96 subjects, while 20 subjects are used for held-out testing. The five-fold models' ensemble is used for inference and evaluating test set performance. For experiment training, we used (1) a CT window range of [−175, 275] HU; (2) scaled intensities of [0.0, 1.0] with 1.0 mm isotropic spacing. The learning rate is initialized to 0.0001, followed by a weight decay of $1e^{-4}$ for 50 K iterations. Common data augmentation such as random flip, rotation, and change of intensity are applied with the probability of 0.1. For fair comparison and direct evaluation of the effectiveness of models, no pre-training is performed for all segmentation tasks. Segmentation results are evaluated with DSC and HD. We conduct volumetric analyses on kidney components in terms of R squared error, Pearson R, absolute deviation of volume, and the percentage difference between the proposed method and manual label.

4.2.3. Multi-organ segmentation—80 subjects are used for training/validation and 20 are used for testing. The images are resampled to 1.5 mm × 1.5 mm × 2.0 mm. We perform the same data augmentation as the renal substructure segmentation. The learning rate is initialized to 0.0001 followed by a weight decay of $1e^{-4}$ for 100 K iterations. Segmentation results are evaluated with DSC.

4.2.4. Kidney and kidney tumor segmentation—We perform five-fold cross-validation experiments on 210 subjects and show DSC results of the held-out 20%. The experiments have the same settings as the renal substructures dataset.

4.2.5. Brain tumor segmentation—Following the same data split of Swin UNETR (Tang et al., 2022), SegResNet (Myronenko, 2019), and nnUNET (Isensee et al., 2021), we train our method with five-fold cross-validation with a ratio of 0.8 and 0.2.

5. Results

We evaluate the UNesT performance against recent convolutional-(Isensee et al., 2021) and transformer-based (Wang et al., 2021a; Zhou et al., 2021a; Hatamizadeh et al., 2022; Tang et al., 2022) 3D medical segmentation baselines. UNesT presents distinguished results on the task of whole brain segmentation with 133 tissue classes. Next, we perform experiments on the first kidney substructures CT dataset. We further validate model generalizability with the publicly available BTCV, KiTS19, and BraTs2021 datasets.

5.1. Whole brain segmentation

A detailed comparison of quantitative performance is shown in Table 2 and Fig. 4. The qualitative performance is shown in Fig. 5. All the models are pre-trained with 4859 auxiliary pseudo labels and are finetuned with 50 manually traced labels from OASIS in the 5-fold ensemble setting. We first compare the proposed UNesT model with nnUNet (Isensee et al., 2021) and several transformer-based methods. Most of the methods have infinite HD on the CANDI dataset associated with 0.43 to 0.69 DSC score indicating those methods fail to predict all of the 130 classes in the external testing set. UNETR performs the best among these widely used 3D medical image segmentation methods. Compared with UNETR, UNesT improves the performance in the Colin (from 0.7320 to 0.7444) and the CANDI (from 0.6851 to 0.7025) dataset by a margin. SLANT27 (Huo et al., 2019), the prior state-of-the-art method, divides the whole brain into 27 parts and ensembles 27 tiled 3D-UNet (Çiçek et al., 2016) for the final predictions. Within the same 5-fold ensemble settings, UNesT ensembled with 5 models outperforms SLANT27 ensembled with 135 models in terms of DSC in both Colin (0.7444 vs. 0.7264) and CANDI (0.7025 vs. 0.6968) dataset and achieves the state-of-the-art performance. UNesT achieves significant improvement on the test set compared to SLANT27 with $p < 0.05$ under Wilcoxon signed-rank test and further reduces the variation of DSC score distribution with tighter quartiles (Fig. 4). In Fig. 5, we show UNesT has better captures on the boundary and correctly segments brain tissues. As the external testing set represents a high resolution and different age population cohort, we show that our method can generalize learned knowledge to different populations (see Table 3).

5.2. Characterization of renal substructures

Segmentation Results.—Compared to canonical kidney studies using shape models or random forests in Table 4, deep learning-based methods improve the performance by a large margin from 0.7233 to 0.7991. Among the nnUNet (Isensee et al., 2021) and extensive transformer models, we obtain the state-of-the-art average DSC score of 0.8564 compared to the second-best performance of 0.8411 from SwinUNETR, with a significant improvement $p < 0.05$ under Wilcoxon signed-rank test. We observe higher improvement in smaller anatomies such as the medulla and collecting systems. We compare qualitative results in Fig. 6. Our method demonstrates the distinct improvement of detailed structures for the medulla

and pelvicalyceal systems. Fig. 7 shows that the proposed automatic segmentation method achieves better agreement compared to inter-rater assessment, 0.03 against 0.29 of mean difference indicating reliable reproducibility.

Volumetric Analysis.—Table 5 lists the volume measurement with the proposed method. The UNesT achieves an R squared error of 0.9348 on the cortex. The correlation performance metric with Pearson R achieves 0.9896 for the UNesT against the manual label on the cortex. Our method obtains 2.5259 with an absolute deviation of volumes. The percent difference in the cortex is 3.8411. We observe the same trend for the Medulla and Pelvicalyceal systems. Quantitative results show that our workflow can serve as the state-of-the-art volumetric measurement compared to the prior kidney characterization state-of-the-art (Tang et al., 2021b).

5.3. Multi-organ segmentation

We present the quantitative performance and qualitative segmentation comparison on the BTCV dataset in Table 6 and Fig. 8, respectively. No pre-training or ensemble is performed in all experiments. UNesT achieves the best average performance on BTCV dataset which demonstrate the generalizability of UNesT. Compared with the other methods, UNesT achieves large improvement on organs that are small in size, such as the esophagus, pancreas, and adrenal glands, where UNesT outperforms the second best performing method by 2.5%, 1.2% and 1.9%, respectively. In Fig. 8 rows 1 and 2, UNesT successfully differentiates stomach tissues and background tissues demonstrating that UNesT has a better capability on identifying heterogeneous organs. UNesT better captures spatial information in Fig. 8 row 3, where most of the other model confuses right/left kidneys and liver/spleen tissues.

5.4. Kidney and tumor segmentation

To validate the generalizability of UNesT, we compare KiTS19 results among nnUNet (Isensee et al., 2021) and transformer-based methods. Our approach achieves moderate improvement at DSC of 0.9794 and 0.8439 for kidneys and tumors, respectively, as shown in Table 7, indicating that the designed architecture can be used as a generic 3D segmentation method. We show a qualitative comparison between our transformer-based model with the CNN-based nnUNeT in Fig. 9. Case 1 is an above average sample that shows UNesT achieves a clearer boundary between kidney and tumor, while case 2 is an under average case where the 3D DSC score of UNesT achieves 0.80 compared to 0.72.

5.5. Brain tumor segmentation

In Table 8, we compared the performance of UNesT with three top-performed methods in BraTS 2021 challenge dataset. Dice scores of the three types of brain tumors are presented in the table with 5 folds experiment design. UNesT consistently outperforms the CNN-based method SegResNet and nnUNet, and the transformer-based Swin UNETR. In particular, for ET, UNesT achieves top Dice scores of 0.898 on average and outperforms the closest competing method by 0.7%. Overall, the average Dice across 5 folds and 3 brain tumor structures is 0.917 which surpasses the state-of-the-art by 0.4%.

5.6. Ablation study

5.6.1. Model scales—To investigate the scalability of our proposed model, we designed ‘‘small’’, ‘‘base’’ and ‘‘large’’ UNesT models (UNesT-S, UNesT-B and UNesT-L) by scaling the depth, heads and width of the transformer. Detailed parameters of UNesT models with various hyperparameter settings are shown in Table 9. Experiments are performed on whole brain segmentation task with 50 T1w MRI scans from the OASIS dataset. 45 T1w scans are used for training and the other 5 for validation. No pre-training is performed for all the models. We start with 20% of the training data and add 20% each time until all data are included. All models are trained five times with 9, 18, 27, 36, and 45 samples, respectively. Fig. 10(a) and (b) shows the quantitative results of DSC in the CANDI and Colin dataset, respectively. Fig. 10(c) shows the distribution of the average DSC in each subject of the test set. Fig. 11 shows the qualitative comparison of whole brain segmentation of different model scales trained with 45 T1w scans.

We observe that larger models and additional data improve segmentation performance. Larger models are more data efficient as with the amount of training data increase, larger models perform better than smaller models. In Fig. 11, compared with UNesT-S and UNesT-B, UNesT-S evidently mis-classified a large amount of background and brain tissues pixels whereas UNesT-B has mostly clean background indicating that UNesT-B better utilizes the training data efficiently. UNesT-L further improves the segmentation results indicating that larger models are more data efficient. In terms of reducing annotation effort, both UNesT-B and UNesT-L perform better with 9 training samples than UNesT-S with all the training samples, which reduces the annotation effort by at least 80%. When adding additional data, the DSC score increases for all the models of different scales. In terms of the relationship between model size and DSC score performance, although the DSC score performance steadily increases as the model scale increases, the performance differences become smaller. In low-data regime, UNesT-B can achieve comparable DSC compared to UNesT-L, but UNesT-B marginally outperforms UNesT-S. When all the training data are included, the performance increase ratio between UNesT-B and UNesT-S is 5.41% (0.6941 versus 0.6585) compared with 1.83% between UNesT-S and UNesT-B (0.7068 versus 0.6941) on the Colin dataset and 3.93% (0.6244 versus 0.6008) versus 3.04% (0.6434 versus 0.6244) on the CANDI dataset. Although UNesT-L has 3 times more parameters than UNesT-B, the comparable performance between UNesT-L and UNesT-B indicates UNesT-B is efficient for the training data. After reaching a certain point, scaling up models may not necessarily lead to large performance improvements.

In Table 10, we summarize the training and testing/inference time with the whole brain segmentation using different model scales and sliding window overlaps. According to our benchmarks, the inference time is mostly impacted by the sliding window overlap, as the increase of overlap will result in a significant number of patches in $overlap = 0.7$. We benchmark the registered MRI volume $172 \times 220 \times 156$. In practical clinical settings, auto segmentation of a given MRI volume with a single GPU or CPU can achieve satisfactory performance (shown in Fig. 10) and inference time (e.g., 2.34 s).

5.6.2. Data efficiency—We investigate the data efficiency of our proposed method using the whole brain and renal substructures dataset. Fig. 11 shows the performance comparison between different UNesT variants, base and larger model are of better data-efficient when training with less data (e.g., 9 or 18). We show the UNesT-B model achieves 133 classes segmentation of DSC 0.6131 with only 9 training samples. Fig. 12 shows the data efficiency evaluated and compared on the renal substructure dataset. UNesT achieves DSC of 0.7903 compared to the second-best SwinUNETR 0.7681 when training with 20% samples. With the increase of training data, our method performs consistently higher DSC compare to baseline methods. We observe the UNesT model trained with 20% data is comparable to nnUNet or TransBTS using full training data, which shows superior data efficiency.

5.6.3. Effects of block aggregation—We show the hierarchical architecture design (with 3D block aggregation) provides significant improvement for medical image segmentation (as shown in Fig. 12). The result shows that the hierarchy mechanism achieves superior performance at 20% to 100% of training data. Under a low-data regime, block aggregation achieves a higher improvement (> 3% of DSC) compared to the second-best method. We notice that the model without block aggregation (canonical transformer layers) obtains lower performance. In addition, UNesT with block aggregation demonstrates a faster convergence rate (15% and 4% difference at 2 K/30 K iterations) compared to the backbone model without hierarchies. The results show block aggregation is an effective component for representation learning for transformer-based models. In addition, compared with the Swintransformer-based method, our UNesT shows consistently superior performance, especially in whole brain segmentation, which indicates the 3D aggregation modules perform better than shifted window module for local patch communication.

5.6.4. Size of pre-training dataset in whole brain segmentation—Acquiring human annotation is labor intensive, thus many studies (Yang et al., 2022a; Roy et al., 2017; Huo et al., 2019; Yang et al., 2022b) adopt the strategy of pre-training with pseudo labels and then finetuning with human annotations to get around this limitation and increase model performance. Herein, we perform experiments using different amounts of pre-training data in the whole brain segmentation to investigate the impact of pre-training data quantity on the final results. We repeat the experiments 6 times with 5, 25, 125, 625, 3125, and all available pre-training data, respectively. All the pre-trained models are finetuned using the OASIS dataset in the same 5-fold cross-validation setting. We include the results of the training from scratch using the OASIS dataset for comparison. The results are shown in Fig. 13. We observe that increasing the number of pre-trained examples up to 125 resulted in a rapid improvement in the DSC score. Pre-training sizes greater than 125 do not further advance performance and the results fluctuate in a small range. This observation demonstrates that UNesT can benefit from pre-training using pseudo data, but a large pre-training dataset is not a necessity. When the amount of pre-training data reaches a certain limit, the performance gains are reduced. Instead, adding more pseudo data could possibly confuse the network.

6. Discussion

6.1. Why do we need an efficient hierarchical transformer-based medical segmentation model?

In this paper, we target the critical problem that transformed-based models commonly lack of local positional information resulting in sub-optimal performance when handling considerable tissue classes in 3D medical image segmentation. Specifically, medical segmentation datasets are small where images are of spatially high-resolution and high dimensionality which can lead to data inefficiency. Our proposed UNesT addresses the above problem by hierarchically aggregating the spatially adjacent patches and leveraging the global self-attention mechanism to combine global and local information efficiently. SwinUNETR (Tang et al., 2022), which uses ‘‘shifted window’’ for local patch communication, observed good but inconsistent performance. Specifically, it achieves second-best performance in the renal substructures segmentation, but in the whole brain segmentation, its DSC scores in test datasets under-perform the second-best performing SLANT27 model by a large margin. Our method consistently achieves superior results on the four evaluated heterogeneous tasks.

We highlight our method on dealing with multiple tissues and inter-connected structures. Compared to the prior state-of-the-art method SLANT27 (Huo et al., 2019), which used 27 ensembled networks, UNesT successfully achieves better performance with a single model. Among current 3D medical image segmentation methods, we address the challenging tasks, including more than one hundred structures in T1w MRI, three inter-connected components in kidneys, thirteen major organs in the abdomen, kidney-tumor, and brain-tumor connected tissue.

When predicting multiple tissues that have various sizes and shapes simultaneously, the model performance is often susceptible to the tissues that are small in size. And when dealing with inter-connected tissues, model performance is particularly sensitive to boundary prediction, where missing boundary prediction will jeopardize the results of adjacent classes. However, boundary prediction is not a trivial task in medical image segmentation since the images usually have blurry boundaries and similar intensity/appearance which make it hard to characterize one tissue from another. In our UNesT model, we adopt hierarchical design which utilizes multi-scale strategy to handle the difference in tissue size and blurry boundary problems. At a coarse scale, the model can focus more on the overall structure of the image, and at a finer scale, the model can focus more on the detail of the tissues. Additionally, our design of 3D block aggregation provides additional adjacent positional information to the model, which gives it better tissue distinguish capability. Therefore, UNesT achieves better performance on handling multiple tissues and inter-connected tissue problems.

UNesT shows consistent competitive performance for the brain tumor segmentation task, which is a difficult problem. UNesT contains several hierarchical blocks as its encoder, and it can efficiently encode the multi-scale features of the 3D multi-modal inputs. And multi-scale embeddings are of significant value to medical image segmentation. We also observe that Swin UNETR, SegResNet, and nnU-Net achieve close competitive performance in this dataset. The three baselines contain feature downsample and upsample modules,

where multi-scale feature maps are utilized and output to the decoder. The quantitative results show that our method can be effective at modeling tumor tissues and efficiently learning multi-scale features.

The superior performance of UNesT in these inferences 5 different tasks demonstrates the effectiveness and efficiency of UNesT in segmenting multiple structures/tissues with small medical datasets. Specifically, we validate that our model is data efficient in low-data regimes. Moreover, our experiments show that larger models are more data efficient, suggesting the proposed network is easily scalable if necessary. Furthermore, we study the impact of the number of pseudo labels used for pre-training. We observed that pre-training sizes exceeding a certain number do not further advance model performance. On the contrary, adding more pseudo labels may confuse the network and decrease performance (Fig. 13).

6.2. The combination of CNN and transformers

The proposed UNesT follows the first class of the CNNs and transformer combination design where transformer is the main encoder and CNNs served as the decoder. To evaluate the effectiveness of our proposed UNesT in comparison to other CNN and transformer hybrid models, we compare two models falling under the first category - UNETR and SwinUNETR. Additionally, we include TransBTS and CoTr from the second category, where transformers serve as secondary encoders and CNNs serve as the main encoder and decoder. All these baseline methods are specifically designed for 3D volumetric medical image segmentation. As there are currently no existing techniques in the third category that are optimized for this task, we do not include them in our comparative analysis.

CoTr achieves good performance in whole brain segmentation and multi-organ segmentation task in terms of DSC. However, in whole brain segmentation, both TransBTS and CoTr have inf value in the HD of the CANDI dataset. This may indicate that this type of model is susceptible to outliers. On the other hand, UNETR, SwinUNETR, and UNesT achieve relatively stable performance among four datasets, especially when dealing with outlier cases in the whole brain segmentation task.

6.3. Single models performance vs. Ensembles

According to the single instance benchmarks (Tables 11 and 12), we observe a similar performance trend to what is observed in the ensemble results (Table 2, Table 4) and Table 7. Some models have improved performance without using an ensemble, while others experience a minor decrease in performance. In general, the model’s performance is comparable whether or not an ensemble is utilized, but the use of an ensemble could potentially enhance the stability of the model’s performance. Ensemble five-fold models can effectively remove outlier predictions and improve overall performance. With single model testing, our method consistently achieves the best performance across different datasets. Compared to the five-fold ensemble strategy, a single instance is more practical and used in clinical workflow with faster inference time and less computational effort.

6.4. Reproducibility against clinical radiologists

In this work, we develop the first in-house renal sub-structures CT cohort for segmentation, including the renal cortex, medulla, and pelvicalyceal system which are manually annotated by radiologists. We show that the proposed method is data-efficient for accurately quantifying kidney components and can be used for volumetric analysis such as in the medullary pyramids. Fig. 7 shows the proposed automatic segmentation method achieves better agreement compared to inter-rater assessment, with 0.03 versus 0.29 mean difference, respectively, indicating robust reproducibility. Visual quantitative analysis of renal structures remains a complex task for radiologists. Some of the histomorphometric features in regions of the kidney (e.g., textural or graph features) are poorly adapted for manual identification. In this study, we show that UNesT achieves consistently reliable performance. Compared with previous studies on cortex segmentation, the proposed approach significantly facilitates the derivation of the visual and quantitative results.

6.5. Limitation and sensitivity study

For whole brain segmentation, we observe current performance is limited by registration. Specifically, the DSC score in the MNI space is around 0.90 and around 0.87 in the Colin and CANDI dataset, respectively. However, the performance drops around 0.17 DSC score after inverse transformation to the original space. Investigation of registration performance should be considered in the future.

We study outlier cases of renal structure segmentation to demonstrate potential limitations. In reviewing most computer-automated segmentation methods, we found about 90% of the segmentation is promising, but about 10% are also found to be outliers. As shown in Fig. 14, typical outliers under-segment and fail to capture parts of tissue labels (left two images). The missing parts result in a lower DSC score of about 0.80 (cortex) and 0.62 (medulla). The right two images show the other type of failure: over-segmentation, where we observe a complete renal segmentation but mis-labeling of nearby tissues. This issue can potentially be resolved by component analysis in a post-processing step. These two types of outlier segmentation are easily spotted with a rudimentary visual quality check.

7. Conclusions

In this paper, we propose a novel hierarchical transformer-based 3D medical image segmentation approach (UNesT) with a 3D block aggregation module to achieve local communication. We validate the effectiveness of UNesT on 5 different tasks in both CT and MRI modalities including a whole brain segmentation task with 133 classes, a renal substructure segmentation task, a multi-organ abdominal segmentation task, and a kidney/tumor segmentation task as well as a brain tumor segmentation task. We consistently achieve state-of-the-art performance on the four datasets. Our single model outperforms 27 ensembled models in the prior state-of-the-art method, SLANT27, for whole brain segmentation. In addition, we develop the first in-house renal sub-structures CT dataset with radiologists. UNesT achieves the best performance among recent popular convolutional- and transformer-based volumetric medical segmentation methods. We show the major

contribution of the proposed method on successfully modeling hundreds of tissues (e.g., 133 classes) and hierarchically inter-connected structures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research is supported by NIH Common Fund and National Institute of Diabetes, Digestive and Kidney Diseases U54DK120058, NSF CAREER 1452485, NIH grants, 2R01EB006136, 1R01EB017230 (Landman), and R01NS09529. The identified datasets used for the analysis described were obtained from the Research Derivative (RD), database of clinical and related data. The imaging dataset(s) used for the analysis described were obtained from ImageVU, a research repository of medical imaging data and image-related metadata. ImageVU and RD are supported by the VICTR CTSA award (ULTR000445 from NCATS/NIH) and Vanderbilt University Medical Center institutional funding. ImageVU pilot work was also funded by PCORI (contract CDRN-1306-04869).

Data availability

The author do not have permission to share in-house dataset. The code is publicly available at the link provided in the abstract.

References

Asman AJ, Landman BA, 2014. Hierarchical performance estimation in the statistical label fusion framework. Med. Image Anal. 18 (7), 1070–1081. [PubMed: 25033470]

Aubert-Broche B, Evans AC, Collins L, 2006. A new improved version of the realistic digital brain phantom. NeuroImage 32 (1), 138–145. [PubMed: 16750398]

Ba JL, Kiros JR, Hinton GE, 2016. Layer normalization.

Baid U, Ghodasara S, Mohan S, Bilello M, Calabrese E, Colak E, Farahani K, Kalpathy-Cramer J, Kitamura FC, Pati S, et al. 2021. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314.

Beltagy I, Peters ME, Cohan A, 2020. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M, 2021. SwinUnet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537.

Chang Y, Menghan H, Guangtao Z, Xiao-Ping Z, 2021. Transclaw u-net: Claw u-net with transformers for medical image segmentation. arXiv preprint arXiv:2107.05188.

Chen B, Liu Y, Zhang Z, Lu G, Zhang D, 2021a. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. arXiv preprint arXiv:2107.05274.

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

Chen X, Summers RM, Cho M, Bagci U, Yao J, 2012. An automatic method for renal cortex segmentation on CT images: evaluation on kidney donors. Academic Radiol. 19 (5), 562–570.

Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O, 2016. 3D U-net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 424–432.

Cordonnier JB, Loukas A, Jaggi M, 2019. On the relationship between self-attention and convolutional layers. arXiv preprint arXiv:1911.03584.

Deng K, Meng Y, Gao D, Bridge J, Shen Y, Lip G, Zhao Y, Zheng Y, 2021. TransBridge: A lightweight transformer for left ventricle segmentation in echocardiography. In: International Workshop on Advances in Simplifying Medical Ultrasound. Springer, pp. 63–72.

Dong B, Wang W, Fan DP, Li J, Fu H, Shao L, 2021. Polyp-PVT: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. 2020. An image is worth 16×16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations.

Evans AC, Collins DL, Mills S, Brown ED, Kelly RL, Peters TM, 1993. 3D statistical neuroanatomical models from 305 MRI volumes. In: 1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference. IEEE.

Han K, Xiao A, Wu E, Guo J, Xu C, Wang Y, 2021. Transformer in transformer. Adv. Neural Inf. Process. Syst. 34, 15908–15919.

Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth HR, Xu D, 2022. Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584.

He K, Zhang X, Ren S, Sun J, 2016. Identity mappings in deep residual networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer, pp. 630–645.

Heller N, Isensee F, Maier-Hein KH, Hou X, Xie C, Li F, Nan Y, Mu G, Lin Z, Han M, et al. 2021. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. Med. Image Anal. 67, 101821. [PubMed: 33049579]

Hu H, Zhang Z, Xie Z, Lin S, 2019. Local relation networks for image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3464–3473.

Huang X, Deng Z, Li D, Yuan X, 2021. MISSFormer: An effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162.

Huo Y, Xu Z, Xiong Y, Aboud K, Parvathaneni P, Bao S, Bermudez C, Resnick SM, Cutting LE, Landman BA, 2019. 3D whole brain segmentation using spatially localized atlas network tiles. NeuroImage 194, 105–119. [PubMed: 30910724]

Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH, 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211. [PubMed: 33288961]

Jin C, Shi F, Xiang D, Jiang X, Zhang B, Wang X, Zhu W, Gao E, Chen X, 2016. 3D fast automatic segmentation of kidney based on modified AAM and random forest. IEEE Trans. Med. Imaging 35 (6), 1395–1407. [PubMed: 26742124]

Kennedy DN, Haselgrove C, Hodge SM, Rane PS, Makris N, Frazier JA, 2012. CANDIShare: a resource for pediatric neuroimaging data. Neuroinformatics 10, 319–322. [PubMed: 22006352]

Klein A, Dal Canton T, Ghosh SS, Landman B, Lee J, Worth A, 2010. Open labels: online feedback for a public resource of manually labeled brain images. In: 16th Annual Meeting for the Organization of Human Brain Mapping, Vol. 84358.

Landman B, Xu Z, Igelsias J, Styner M, Langerak T, Klein A, 2015. MICCAI multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling beyond Cranial Vault—Workshop Challenge.

Li J, Chen J, Tang Y, Wang C, Landman BA, Zhou SK, 2023. Transforming medical imaging with transformers? A comparative review of key properties, current progresses, and future perspectives. Med. Image Anal. 102762. [PubMed: 36738650]

Li S, Sui X, Luo X, Xu X, Liu Y, Goh R, 2021a. Medical image segmentation using squeeze-and-expansion transformers. arXiv preprint arXiv:2105.09511.

Li Y, Wang S, Wang J, Zeng G, Liu W, Zhang Q, Jin Q, Wang Y, 2021b. GT U-net: A U-net like group transformer network for tooth root segmentation. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 386–395.

Li Y, Wang Z, Yin L, Zhu Z, Qi G, Liu Y, 2021c. X-Net: a dual encoding–decoding method in medical image segmentation. Vis. Comput. 1–11.

Lin A, Chen B, Xu J, Zhang Z, Lu G, 2021. DS-TransUNet: Dual swin transformer U-net for medical image segmentation. arXiv preprint arXiv:2106.06716.

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.

Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL, 2007. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19 (9), 1498–1507. [PubMed: 17714011]

Meng X, Zhang X, Wang G, Zhang Y, Shi X, Dai H, Wang Z, Wang X, 2021. Exploiting full resolution feature context for liver tumor and vessel segmentation via fusion encoder: Application to liver tumor and vessel 3D reconstruction. arXiv preprint arXiv:2111.13299.

Myronenko A, 2019. 3D MRI brain tumor segmentation using autoencoder regularization. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4. Springer, pp. 311–320.

Ourselin S, Roche A, Subsol G, Pennec X, Ayache N, 2001. Reconstructing a 3D structure from serial histological sections. Image Vis. Comput. 19 (1–2), 25–31.

Peiris H, Hayat M, Chen Z, Egan G, Harandi M, 2021. A volumetric transformer for accurate 3D tumor segmentation. arXiv preprint arXiv:2111.13300.

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.

Roth HR, Shen C, Oda H, Sugino T, Oda M, Hayashi Y, Misawa K, Mori K, 2018. A multi-scale pyramid of 3D fully convolutional networks for abdominal multi-organ segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part IV 11. Springer, pp. 417–425.

Roy AG, Conjeti S, Sheet D, Katouzian A, Navab N, Wachinger C, 2017. Error corrective boosting for learning fully convolutional networks with limited data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 231–239.

Sharir G, Noy A, Zelnik-Manor L, 2021. An image is worth 16×16 words, what is a video worth? arXiv preprint arXiv:2103.13915.

Tang Y, Gao R, Lee HH, Han S, Chen Y, Gao D, Nath V, Bermudez C, Savona MR, Abramson RG, et al. 2021a. High-resolution 3D abdominal segmentation with random patch network fusion. Med. Image Anal. 69, 101894. [PubMed: 33421919]

Tang Y, Gao R, Lee HH, Xu Z, Savoie BV, Bao S, Huo Y, Fogo AB, Harris R, de Caestecker MP, et al. 2021b. Renal cortex, medulla and pelvicaliceal system segmentation on arterial phase CT images with random patch-based networks. In: Medical Imaging 2021: Image Processing, Vol. 11596. SPIE, pp. 379–386.

Tang Y, Yang D, Li W, Roth HR, Landman B, Xu D, Nath V, Hatamizadeh A, 2022. Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740.

Ulyanov D, Vedaldi A, Lempitsky V, 2016. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022.

Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM, 2021. Medical transformer: Gated axial-attention for medical image segmentation. arXiv preprint arXiv:2102. 10662.

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

Wang W, Chen C, Ding M, Li J, Yu H, Zha S, 2021a. TransBTS: Multimodal brain tumor segmentation using transformer. arXiv preprint arXiv:2103.04430.

Wang B, Dong P, et al. 2022. Multiscale transunet++: dense hybrid U-net with transformer for medical image segmentation. Signal Image Video Process. 1–8.

Wang J, Wei L, Wang L, Zhou Q, Zhu L, Qin J, 2021b. Boundary-aware transformers for skin lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 206–216.

Wang H, Xie S, Lin L, Iwamoto Y, Han XH, Chen YW, Tong R, 2021c. Mixed transformer U-net for medical image segmentation. arXiv preprint arXiv: 2111.04734.

Wasserthal J, Meyer M, Breit HC, Cyriac J, Yang S, Segeroth M, 2022. TotalSegmentator: robust segmentation of 104 anatomical structures in CT images. arXiv preprint arXiv:2208.05868.

Wu Y, Liao K, Chen J, Chen DZ, Wang J, Gao H, Wu J, 2022. D-Former: A U-shaped dilated transformer for 3D medical image segmentation. arXiv preprint arXiv:2201.00462.

Xiang D, Bagci U, Jin C, Shi F, Zhu W, Yao J, Sonka M, Chen X, 2017. CorteXpert: A model-based method for automatic renal cortex segmentation. Med. Image Anal. 42, 257–273. [PubMed: 28888170]

Xie Z, Lin Y, Yao Z, Zhang Z, Dai Q, Cao Y, Hu H, 2021a. Self-supervised learning with swin transformers. arXiv preprint arXiv:2105.04553.

Xie Y, Zhang J, Shen C, Xia Y, 2021b. 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.

Xie Y, Zhang J, Xia Y, Wu Q, 2021c. Unified 2D and 3D pre-training for medical image classification and segmentation. arXiv preprint arXiv:2112.09356.

Yan X, Tang H, Sun S, Ma H, Kong D, Xie X, 2022. After-unet: Axial fusion transformer unet for medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3971–3981.

Yang Q, Yu X, Lee HH, Tang Y, Bao S, Gravenstein KS, Moore AZ, Makrogiannis S, Ferrucci L, Landman BA, 2022a. Label efficient segmentation of single slice thigh CT with two-stage pseudo labels. J. Med. Imaging 9 (5), 052405.

Yang Q, Yu X, Lee HH, Tang Y, Bao S, Gravenstein KS, Moore AZ, Makrogiannis S, Ferrucci L, Landman BA, 2022b. Quantification of muscle, bones, and fat on single slice thigh CT. In: Medical Imaging 2022: Image Processing, Vol. 12032. SPIE, pp. 422–429.

Zhai X, Kolesnikov A, Houlsby N, Beyer L, 2022. Scaling vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12104–12113.

Zhang Y, Liu H, Hu Q, 2021a. TransFuse: Fusing transformers and CNNs for medical image segmentation. arXiv preprint arXiv:2102.08005.

Zhang Z, Sun B, Zhang W, 2021b. Pyramid medical transformer for medical image segmentation. arXiv preprint arXiv:2104.14702.

Zhang Z, Zhang H, Zhao L, Chen T, Arik SÖ, Pfister T, 2022. Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 3. pp. 3417–3425.

Zhou HY, Guo J, Zhang Y, Yu L, Wang L, Yu Y, 2021a. nnFormer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201.

Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J, 2018. 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, pp. 3–11.

Zhou Z, Sodha V, Pang J, Gotway MB, Liang J, 2021b. Models genesis. Med. Image Anal. 67, 101840. [PubMed: 33188996]

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.

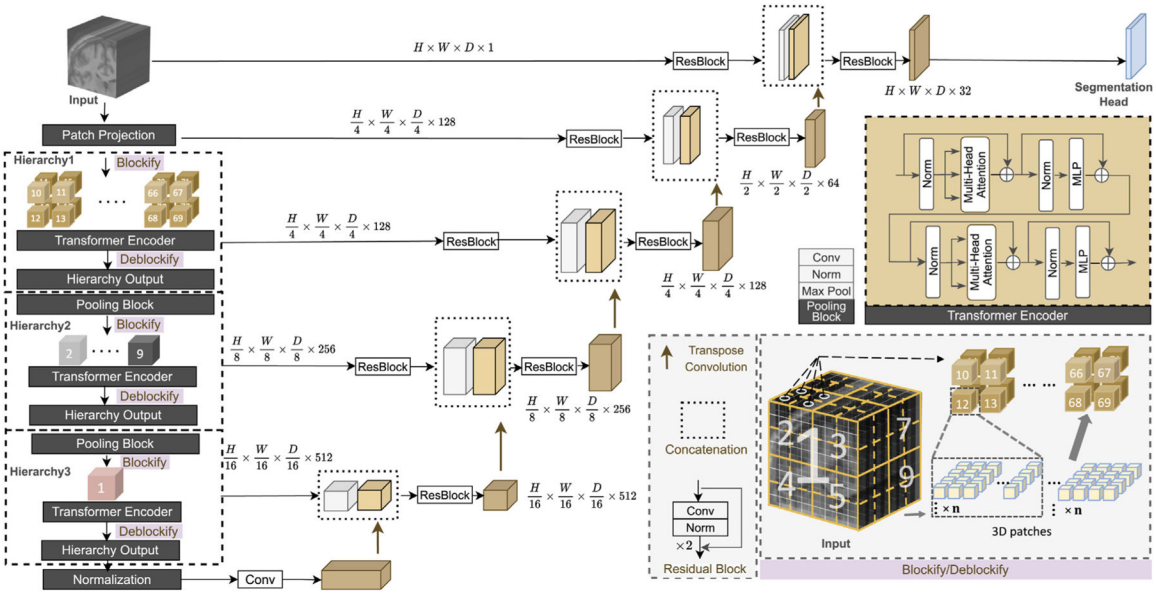

Fig. 1.

Overview of the proposed UNesT with the hierarchical transformer encoder. Input image volumes are embedded into patches. In each hierarchy, patch embeddings are downsampled and blockified before being fed into the transformer encoder. Outputs are deblockified back to the volume plane. Each hierarchy output is connected with the decoder through skip connection.

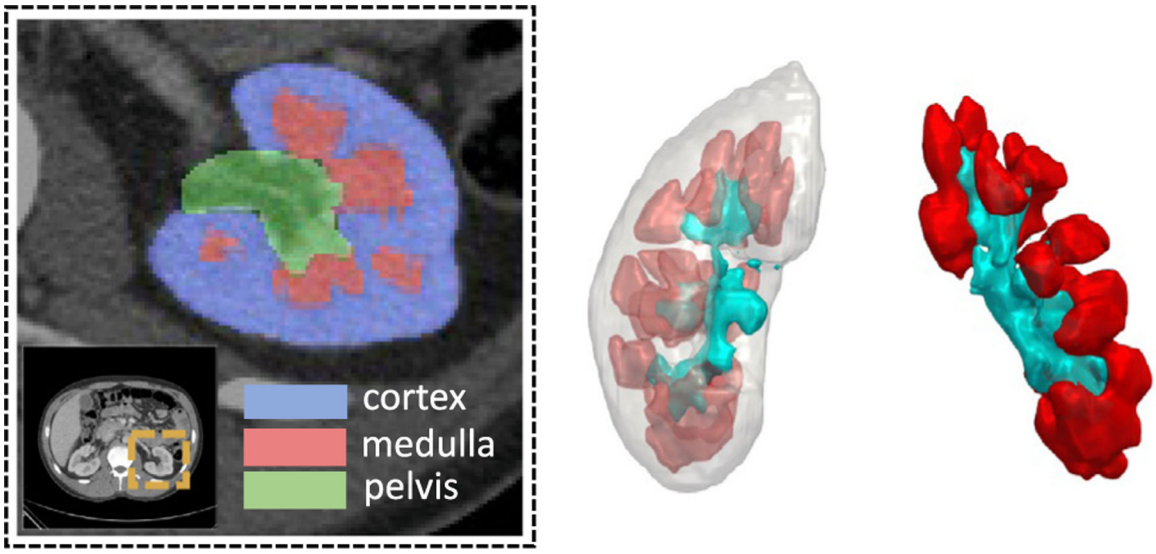

Fig. 2.
Visual and 3D illustration of the kidney components.

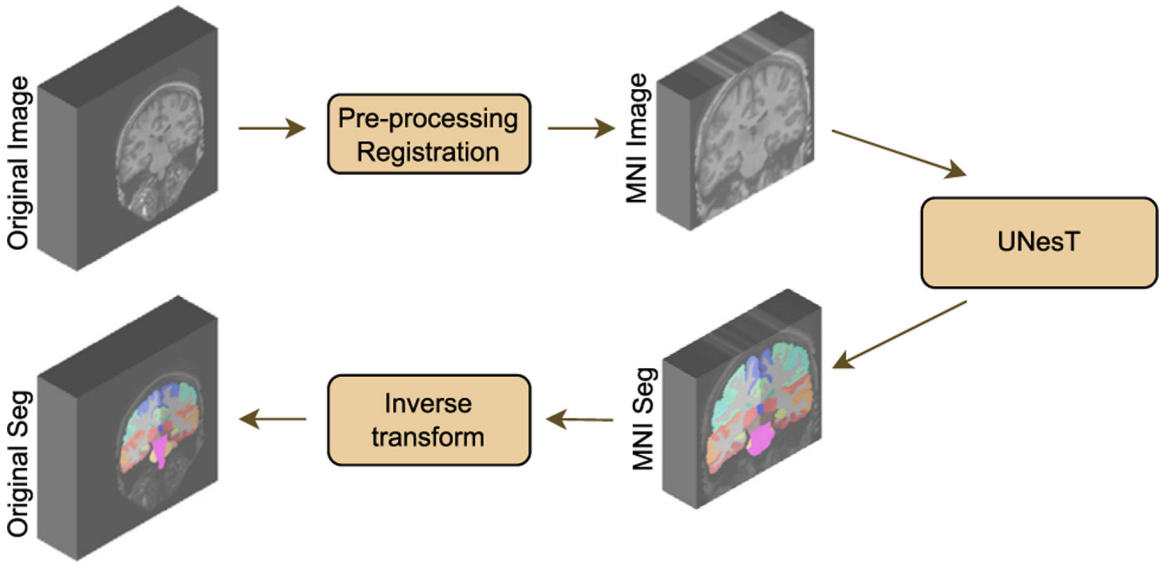

Fig. 3.

Overview of the workflow for the whole brain segmentation task. Original images are pre-processed and registered to the MNI space before feeding into the networks. Model outputs that are in MNI space are transformed back to the original space to get the final predictions.

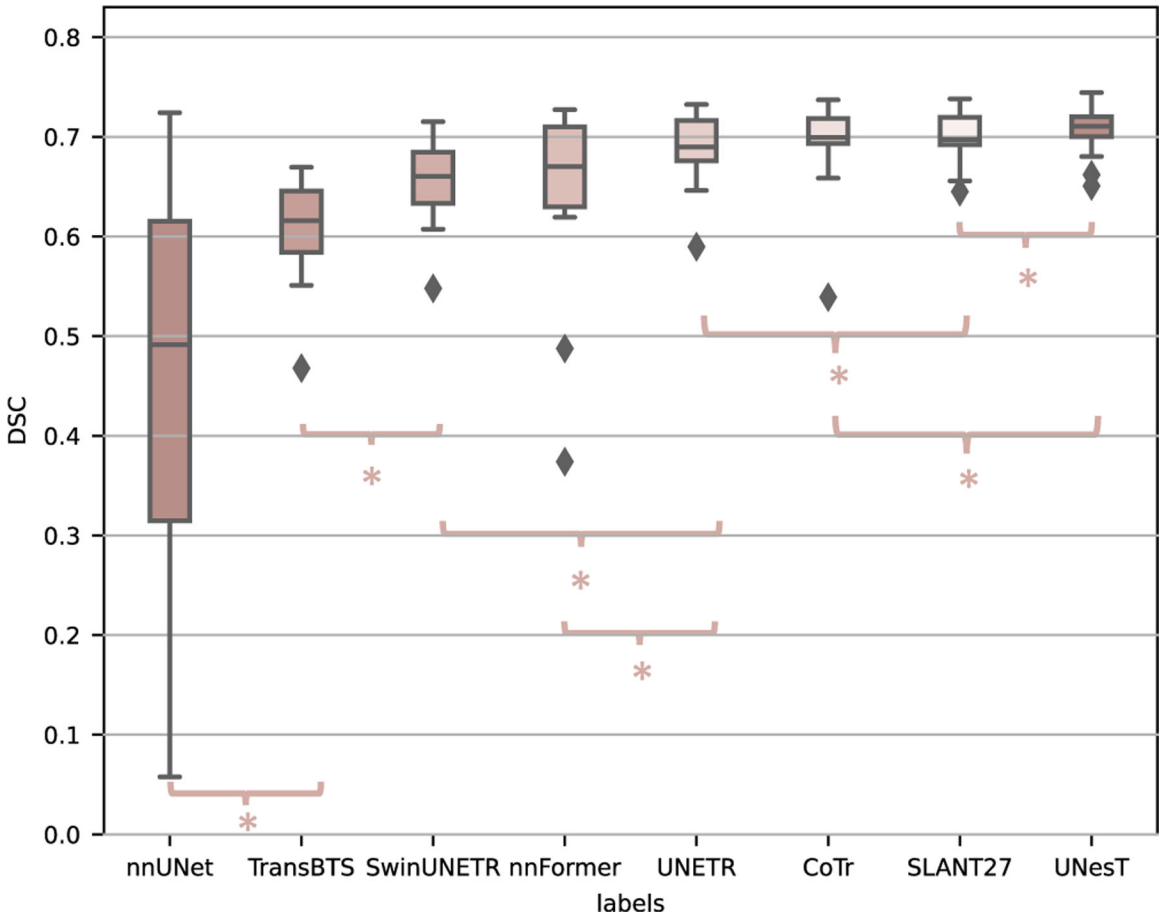

Fig. 4.

Quantitative results of the whole brain segmentation on the testing data. SLANT27 shows the smallest variation among the other baselines. UNesT achieves the overall best performance. Compared with SLANT27, UNesT further reduces the variation with improved median and quartiles of the DSC. * indicates statistically significant ($p < 0.05$) by Wilcoxon signed-rank test. Detailed quantitative performance comparison of 130 classes is shown in Fig A.17 in the supplementary material.

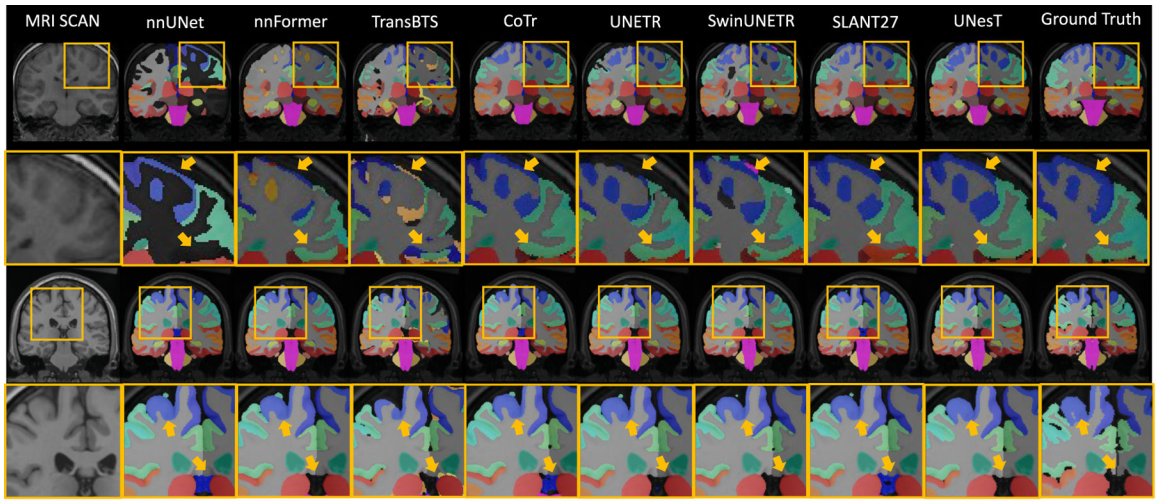

Fig. 5.

Qualitative results of whole brain segmentation on the CANDI dataset (top 2 rows) and Colin dataset (bottom 2 rows). Boxed areas are enlarged in the lower row. Differences are emphasized with the orange arrow. UNesT shows better captures the boundary and correctly segments the tissues.

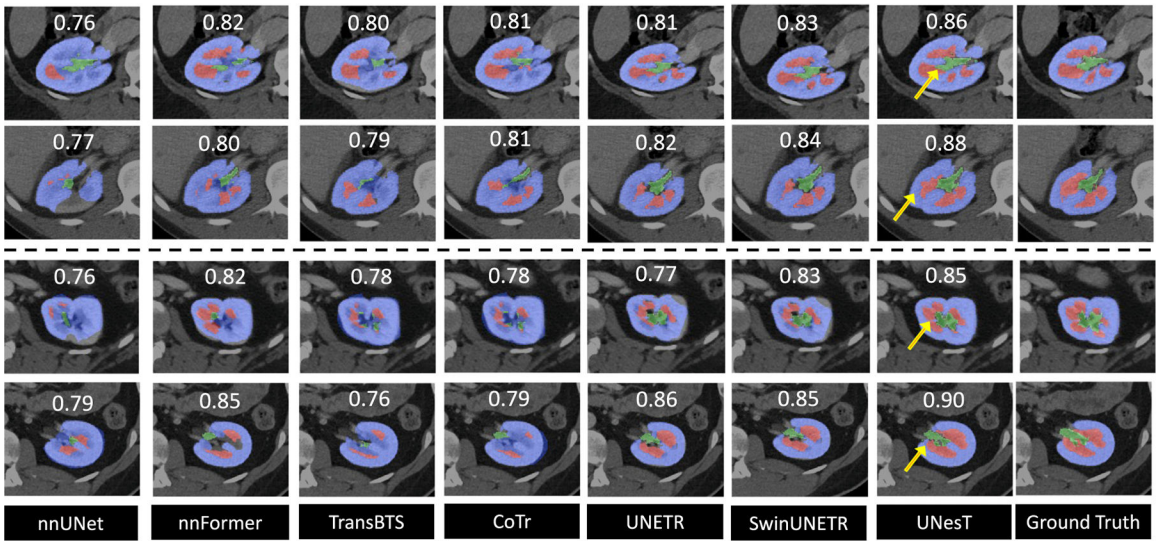

Fig. 6.

Qualitative comparisons of representative renal sub-structures segmentation on two right (top) and two left (bottom) kidneys. The average DSC is marked on each image. UNesT shows distinct improvement on the medulla (red) and pelvicalyceal system (green) against baselines. Comparisons with different baselines including the ViT and CNN hybrid approaches.

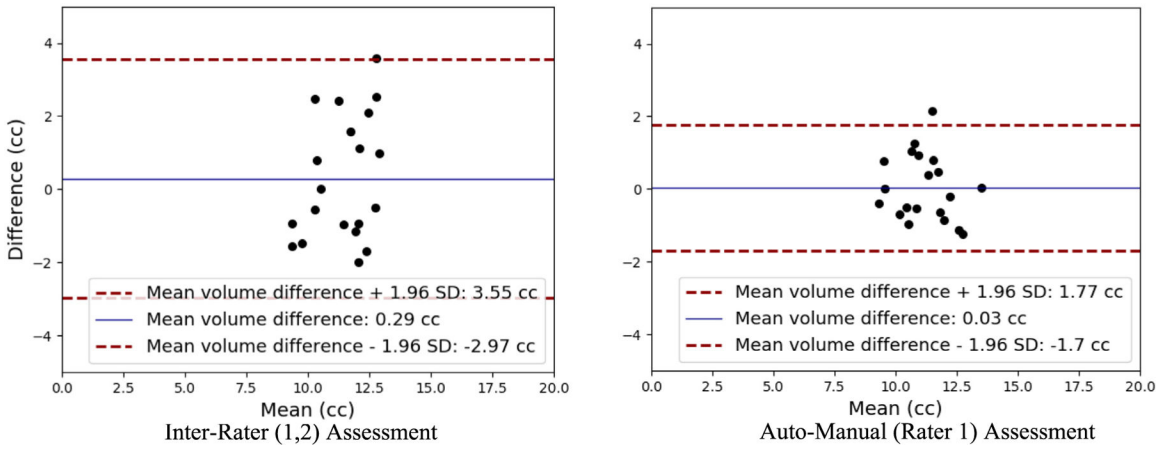

Fig. 7.

The Bland-Atman plots compare the medulla volume agreement of inter-rater and auto-manual assessment. We show cross-validation on interpreter 1 and interpreter 2 manual segmentation on the same test set. Interpreters present independent observation without communication. The auto-manual assessment shows the agreement between UNesT and interpreter 1 annotation.

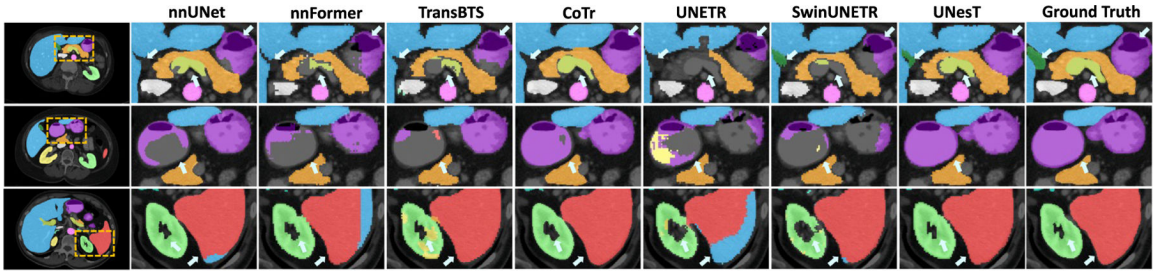

Fig. 8.

Qualitative comparison between UNesT and baseline methods on the BTCV data. Three representative cases are shown. The region with visual improvement is boxed and enlarged. White arrows emphasized the segmentation improvement on portal vein (yellow), stomach (purple), gallbladder (dark green), left kidney (light green), and spleen (red).

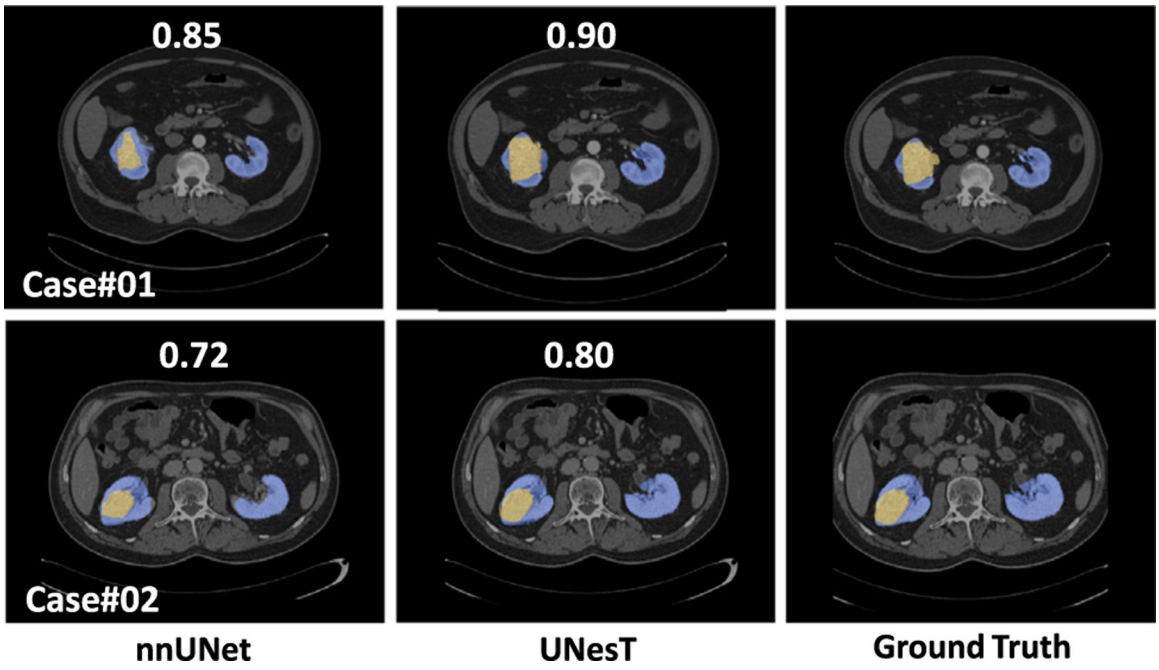

Fig. 9.

Qualitative comparison between our transformer-based segmentation method and the CNN-based model. UNesT shows better tumor segmentation, and we observe the model can better distinguish the kidney-tumor boundary.

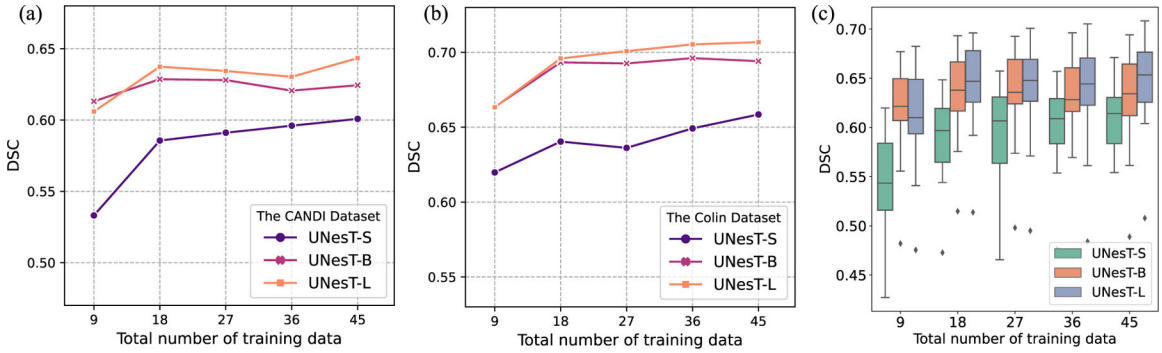

Fig. 10.

Comparison of segmentation results of models with different scales trained with different percentages of training data. (a) and (b) shows the test results of the CANDI and Colin dataset, respectively. (c) shows the results in both the CANDI and Colin dataset.

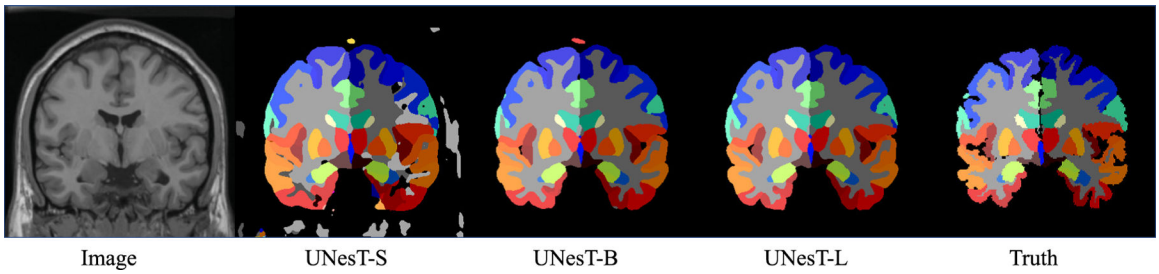

Fig. 11.

Visualization of segmentation results for each model scale trained with the same number of data. Comparing UNesT-S and UNesT-B, UNesT-S evidently mis-classified a large amount of background and brain tissues pixels whereas UNesT-B has mostly clean background indicating that UNesT-B has better capability on utilizing the training data efficiently. UNesT-L further improves the segmentation results indicating that larger models are more data efficient.

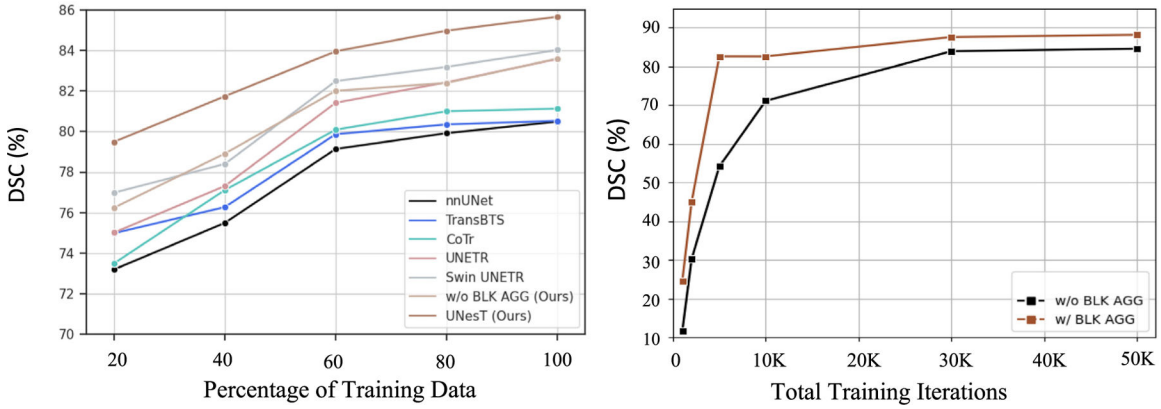

Fig. 12.

Left: DSC comparison on the test set at different percentages of training samples. Right: Comparison of the convergence rate for the proposed method with and without hierarchical modules, and validation DSC along training iterations are demonstrated. Different ViT-, CNN-based and hybrid baselines are compared.

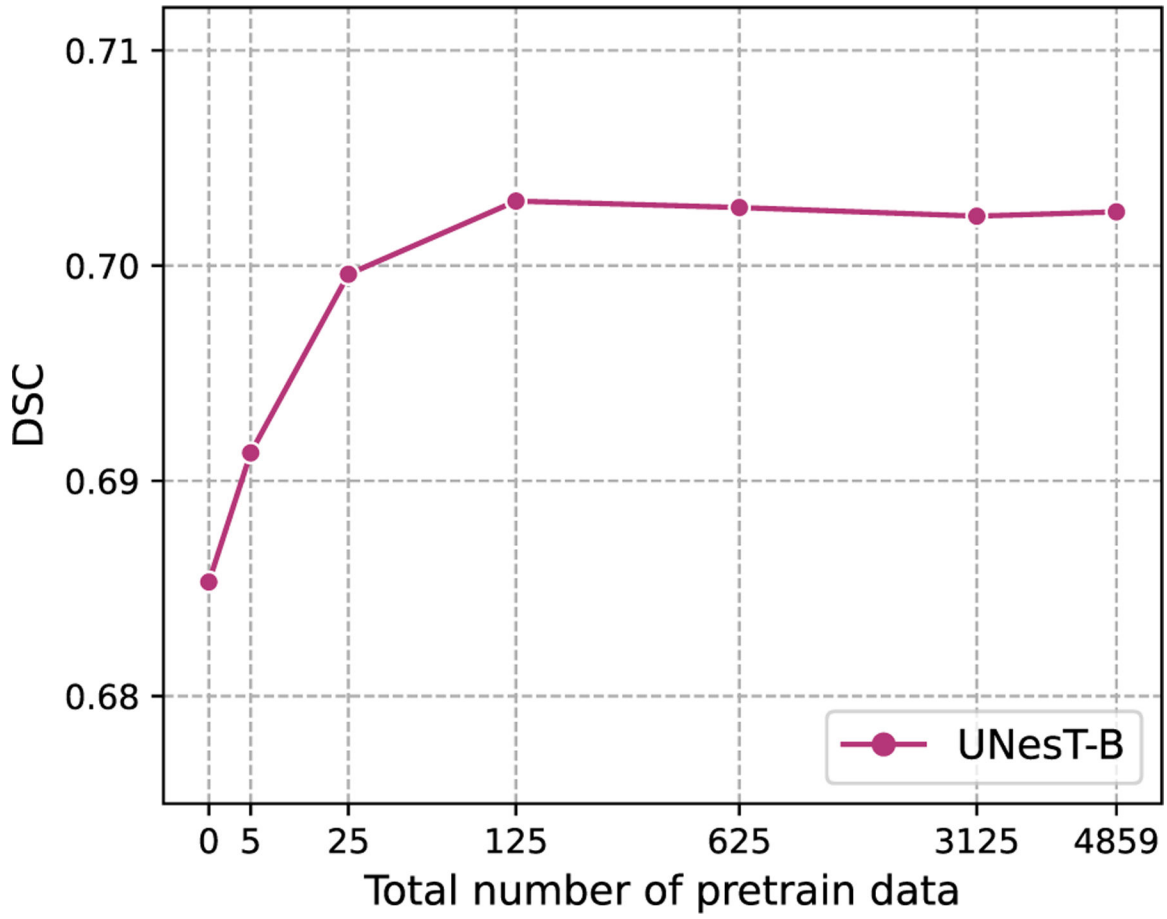

Fig. 13.
DSC comparison on whole brain segmentation CANDI dataset with different amount of data with pseudo labels for pre-training.

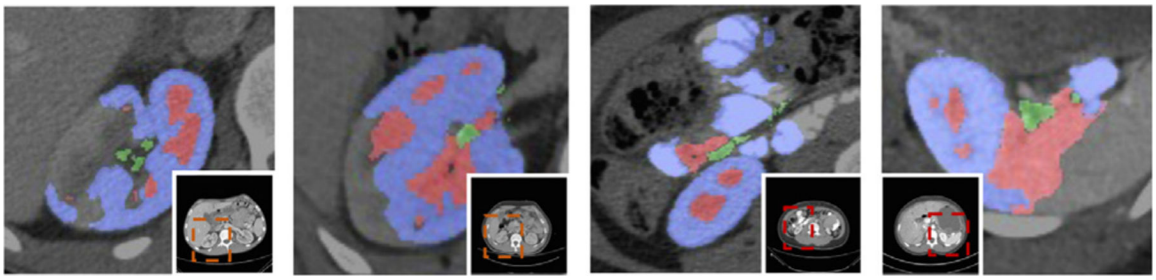

Fig. 14.

Demonstration of potential outlier cases. The left two images show representative under-labeling of tissues. The right two images show the over-labeling of tissues. These segmentations are computed on additional contrast-enhanced CT scans without ground truth labels.

Table 1

Data summary of the 4859 multi-site images.

Study Name	Website	Images
Baltimore Longitudinal Study of Aging (BLSA)	www.blsa.nih.gov	614
Cutting Pediatrics	vkc.mc.vanderbilt.edu/ebrl	586
Autism Brain Imaging Data Exchange (ABIDE)	fcon_1000.projects.nitrc.org/indi/abide	563
Information Extraction from Images (IXI)	www.nitrc.org/projects/ixi_dataset	541
Attention Deficit Hyperactivity Disorder (ADHD200)	fcon_1000.projects.nitrc.org/indi/adhd200	950
Open Access Series on Imaging Study (OASIS)	www.oasis-brains.org	312
1000 Functional Connectome (fcon_1000)	fcon_1000.projects.nitrc.org	1102
Nathan Kline Institute Rockland (NKI_rockland)	fcon_1000.projects.nitrc.org/indi/enhanced	141

Table 2

Performance comparison for the whole brain segmentation task. Overall UNesT achieved state-of-the-art performance on the whole brain segmentation task. The number of parameters and GFLOPs (with a single input volume of 96 × 96 × 96 for the transformer-based models) are shown. ‘‘×27’’ in SLANT27 represents that 27 of the same models are used. inf denotes part of data in the testing datasets have infinite HD. Notes: the FLOPs for SLANT27 are calculated based on input size of 96 × 128 × 88 as designed in the paper (Huo et al., 2019).

Method	#Param	FLOPs(G)	Colin		CANDI	
			DSC	HD	DSC	HD
nnUNet (Isensee et al., 2021)	30.7M	358.6	0.7168	10.7321	0.4337	inf
TransBTS (Wang et al., 2021a)	33.0M	111.9	0.6537	inf	0.6043	inf
nnFormer (Zhou et al., 2021a)	158.9M	920.1	0.7113	10.2755	0.6393	inf
CoTr (Xie et al., 2021b)	42.0M	328.0	0.7209	10.3194	0.6908	inf
UNETR (Hatamizadeh et al., 2022)	92.6M	268.0	0.7320	10.3834	0.6851	11.1972
SwinUNETR (Tang et al., 2022)	62.2M	334.9	0.6854	22.0389	0.6537	34.3980
SLANT27 (Huo et al., 2019)	19.9M × 27	2051.0 × 27	0.7264	**9.9061**	0.6968	8.8851
UNesT	87.3M	261.7	**0.7444**	11.0081	**0.7025**	**8.8417**

Table 3

Single model performance for the whole brain segmentation task. Models are trained with the same training/validation data.

Method	Colin		CANDI	
	DSC	HD	DSC	HD
nnUNet (Isensee et al., 2021)	0.7062	14.0101	0.3930	inf
TransBTS (Wang et al., 2021a)	0.6542	inf	0.5991	inf
nnFormer (Zhou et al., 2021a)	0.7007	10.423	0.6420	inf
CoTr (Xie et al., 2021b)	0.7268	10.2561	0.6923	inf
UNETR (Hatamizadeh et al., 2022)	0.7328	10.216	0.6810	13.3172
SwinUNETR (Tang et al., 2022)	0.6853	21.4812	0.6536	34.5212
SLANT (Huo et al., 2019)	0.7301	**9.9470**	0.6977	9.5000
UNesT	**0.7467**	11.0358	**0.7022**	**8.8902**

Table 4

Segmentation results of the renal substructure on testing cases. The UNesT achieves state-of-the-art performance compared to prior kidney component studies and 3D medical segmentation baselines.

Method	Cortex		Medulla		Pelvicalyceal System		Avg.	
	DSC	HD	DSC	HD	DSC	HD	DSC	HD
Chen et al. Chen et al. (2012)	0.7512	40.1947	N/A	N/A	N/A	N/A	N/A	N/A
Xiang et al. Xiang et al. (2017)	0.8196	27.1455	N/A	N/A	N/A	N/A	N/A	N/A
Jin et al. Jin et al. (2016)	0.8041	34.5170	0.7186	32.1059	0.6473	39.9125	0.7233	35.5118
Tang et al. Tang et al. (2021b)	0.8601	19.7508	0.7884	18.6030	0.7490	34.1723	0.7991	24.1754
nnUNet (Isensee et al., 2021)	0.8915	17.3764	0.8002	18.3132	0.7309	31.3501	0.8075	22.3466
TransBTS (Wang et al., 2021a)	0.8901	17.0213	0.8013	17.3084	0.7305	30.8745	0.8073	21.7347
CoTr (Xie et al., 2021b)	0.8958	16.4904	0.8019	16.5934	0.7393	30.1282	0.8123	21.0707
nnFormer (Zhou et al., 2021a)	0.9094	15.5839	0.8104	15.9412	0.7418	29.4407	0.8205	20.3219
UNETR (Hatamizadeh et al., 2022)	0.9072	15.9829	0.8221	14.9555	0.7632	27.4703	0.8308	19.4696
SwinUNETR (Tang et al., 2022)	0.9182	**14.0585**	0.8344	11.9582	0.7707	14.6027	0.8411	13.5398
UNesT	**0.9262**	14.4628	**0.8471**	**8.3677**	**0.7958**	**9.735**	**0.8564** *	**10.1885**

* indicates statistically significant ($p < 0.05$) compared with the underlined performance by Wilcoxon signed-rank test.

Table 5

Comparison of volumetric analysis metrics between the proposed method and the state-of-the-art clinical study on kidney components.

Metrics	Cortex		Medulla		Pelvicalyceal System	
	Tang et al. (2021b)	UNesT	Tang et al. (2021b)	UNesT	Tang et al. (2021b)	UNesT
R Squared	0.9200	0.9348	0.6652	0.6850	0.4586	0.6126
Pearson R	0.9838	0.9896	0.8156	0.8428	0.6772	0.7454
Absolute Deviation of Volume	3.0233	2.5259	3.5496	3.0293	0.9443	0.7410
Percentage Difference	4.8280	3.8411	7.4750	6.894	19.0716	12.0171

Table 6

Quantitative comparison of the segmentation results on the BTCV testing set. All results shown are single model performances (without ensemble). Our model achieves the overall best performance. Note: RKid: right kidney, LKid: left kidney, Gall: gallbladder, Eso: esophagus, Stom: Stomach, Panc: pancreas, RAG: right adrenal gland, LAG: left adrenal gland.

Methods	Spleen	RKid	LKid	Gall	Eso	Liver	Stom	Aorta	IVC	PSV	Panc	RAG	LAG	Avg
nnUNet	**0.9595**	0.8835	0.9302	**0.7013**	0.7672	0.9651	0.8679	0.8893	0.8289	0.7851	0.7960	0.7326	0.6835	0.8316
TransBTS	0.9455	0.8920	0.9097	0.6838	0.7561	0.9644	0.8352	0.8855	0.8248	0.7421	0.7602	0.6723	0.6703	0.8131
nnFormer	0.9458	0.8862	0.9368	0.6529	0.7622	0.9617	0.8359	0.8909	0.8080	0.7597	0.7787	0.7020	0.6605	0.8162
CoTr	0.9536	0.8940	0.9330	0.6954	0.7749	0.9617	0.8801	**0.9047**	0.8376	**0.7891**	0.7964	0.7350	0.6831	0.8356
UNETR	0.9048	0.8251	0.8605	0.5823	0.7121	0.9464	0.7206	0.8657	0.7651	0.7037	0.6606	0.6625	0.6304	0.7600
SwinUNETR	0.9459	0.8897	0.9239	0.6537	0.7543	0.9561	0.7557	0.8828	0.8161	0.7630	0.7452	0.6823	0.6602	0.8044
UNesT	0.9580	**0.9249**	**0.9396**	0.7002	**0.7940**	**0.9657**	**0.8861**	0.8899	**0.8412**	0.7856	**0.8058**	**0.7372**	**0.7083**	**0.8433**

Table 7

KiTS19 DSC performance comparison with baseline methods. The UNesT achieves the highest DSC on the with-held test set.

Model	Kidney	Tumor	Avg
nnUNeT (Isensee et al., 2021)	0.9643	0.8287	0.8965
nnFormer (Zhou et al., 2021a)	0.9723	0.8348	0.9036
CoTr (Xie et al., 2021b)	0.9735	0.8341	0.9038
TransBTS (Wang et al., 2021a)	0.9740	0.8374	0.9057
UNETR (Hatamizadeh et al., 2022)	0.9746	0.8382	0.9064
Swin UNETR (Tang et al., 2022)	0.9751	0.8397	0.9074
UNEST (Ours)	**0.9794**	**0.8439**	**0.9117**

Table 8

Five-fold cross-validation performance for BraTS 2021 challenge dataset, metrics are Dice scores. Baseline methods benchmarks are directly from respective models trained by challenge participants. Brain tumor regions: ET, WT, and TC denote Enhancing Tumor, Whole Tumor, and Tumor Core.

Method	BraTS Challenge 2021															
	ET					WT					TC					Avg.
	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	
SwinUNETR (Tang et al., 2022)	0.876	0.908	0.891	0.890	0.891	0.929	**0.938**	0.931	**0.937**	0.934	0.914	0.919	**0.919**	0.920	0.917	0.913
SegResNet (Myronenko, 2019)	0.867	0.900	0.884	0.888	0.878	0.924	0.933	0.927	0.921	0.930	0.907	0.915	0.917	0.916	0.912	0.907
nnUNET (Isensee et al., 2021)	0.866	0.899	0.886	0.886	0.880	0.921	0.933	0.929	0.927	0.929	0.902	0.919	0.914	0.914	0.917	0.908
UNesT	**0.887**	**0.915**	**0.896**	**0.894**	**0.896**	**0.930**	0.936	**0.939**	0.924	**0.937**	**0.915**	**0.923**	**0.919**	**0.928**	**0.923**	**0.917**

Table 9

Model parameters of different scales. Depth: number of transformer layers, Heads: number of heads in the multi-head attention, Width: embedded dimension. Each number in the bracket represents the corresponding hyperparameter in that hierarchy.

Model	#Param	Depth	Heads	Width
UNesT-S	22.4M	(2, 2, 8)	(2, 4, 8)	(64, 128, 256)
UNesT-B	87.3M	(2, 2, 8)	(4, 8, 16)	(128, 256, 512)
UNesT-L	279.6M	(2, 2, 20)	(6, 12, 24)	(192, 384, 768)

Table 10

Training and inference time of different model scales. Training iteration time reports the time for processing a single 96×96×96 patch. For the testing cases, the latencies are reported on MNI space size of 172 × 220 × 156 using GPU RTX 3090Ti, regular inference pipeline without mixed precision, and no torchscript and TensorRT conversions. Note the inference patch size is 96 × 96 × 96, sliding window overlap has a significant impact on the inference time because the increase of overlap percentage will result in exponentially increased patches.

Model	**UNesT-S**	**UNesT-B**	**UNesT-L**
Training			
Iteration (s)	0.29	0.46	0.82
Testing			
overlap = 0.3 (s)	0.84	0.98	1.23
overlap = 0.5 (s)	2.30	2.34	2.97
overlap = 0.7 (s)	6.76	6.99	7.85

Table 11

Single model performance for the whole brain segmentation task. Models are trained with the same training/validation data.

Method	Colin		CANDI	
	DSC	HD	DSC	HD
nnUNet (Isensee et al., 2021)	0.7062	14.0101	0.3930	inf
TransBTS (Wang et al., 2021a)	0.6542	inf	0.5991	inf
nnFormer (Zhou et al., 2021a)	0.7007	10.423	0.6420	inf
CoTr (Xie et al., 2021b)	0.7268	10.2561	0.6923	inf
UNETR (Hatamizadeh et al., 2022)	0.7328	10.216	0.6810	13.3172
SwinUNETR (Tang et al., 2022)	0.6853	21.4812	0.6536	34.5212
SLANT (Huo et al., 2019)	0.7301	**9.9470**	0.6977	9.5000
UNesT	**0.7467**	11.0358	**0.7022**	**8.8902**

Table 12

Single model performance for the KiTS19 task and renal substructure segmentation. Models are trained with the same training/validation data. Pel. Sys. refers to Pelvicalyceal System.

Method	KiTS19			Renal Substructures			
	Kidney	Tumor	Avg.	Cortex	Medulla	Pel. Sys.	Avg.
nnUNet (Isensee et al., 2021)	0.9640	0.8198	0.8919	0.8881	0.7974	0.7285	0.8047
nnFormer (Zhou et al., 2021a)	0.9714	0.8321	0.9018	0.9082	0.8076	0.7370	0.8176
CoTr (Xie et al., 2021b)	0.9706	0.8336	0.9021	0.8950	0.7987	0.7304	0.8080
TransBTS (Wang et al., 2021a)	0.9710	0.8358	0.9034	0.8884	0.8009	0.7271	0.8055
UNETR (Hatamizadeh et al., 2022)	0.9737	0.8362	0.9050	0.9015	0.8143	0.7577	0.8245
SwinUNETR (Tang et al., 2022)	0.9739	0.8368	0.9054	0.9040	0.8302	0.7603	0.8315
UNesT	**0.9790**	**0.8434**	**0.9112**	**0.9211**	**0.8399**	**0.7906**	**0.8505**