Contents lists available at ScienceDirect

Computers & Graphics

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

Special Section on SMI 2024

Multi-scale Knowledge Transfer Vision Transformer for 3D vessel shape segmentation

Michael J. Hua [a,*], Junjie Wu [b], Zichun Zhong [b]

[a] *Cranbrook Schools, Bloomfield Hills, MI 48303, USA*
[b] *Wayne State University, Detroit, MI 48202, USA*

ARTICLE INFO

Keywords:
Vision transformer
Convolutional embedding
Knowledge transfer
Multi-scale encoding

ABSTRACT

In order to facilitate the robust and precise 3D vessel shape extraction and quantification from in-vivo Magnetic Resonance Imaging (MRI), this paper presents a novel multi-scale Knowledge Transfer Vision Transformer (i.e., KT-ViT) for 3D vessel shape segmentation. First, it uniquely integrates convolutional embeddings with transformer in a U-net architecture, which simultaneously responds to local receptive fields with convolution layers and global contexts with transformer encoders in a multi-scale fashion. Therefore, it intrinsically enriches local vessel feature and simultaneously promotes global connectivity and continuity for a more accurate and reliable vessel shape segmentation. Furthermore, to enable using relatively low-resolution (LR) images to segment fine scale vessel shapes, a novel knowledge transfer network is designed to explore the inter-dependencies of data and automatically transfer the knowledge gained from high-resolution (HR) data to the low-resolution handling network at multiple levels, including the multi-scale feature levels and the decision level, through an integration of multi-level loss functions. The modeling capability of fine-scale vessel shape data distribution, possessed by the HR image transformer network, can be transferred to the LR image transformer to enhance its knowledge for fine vessel shape segmentation. Extensive experimental results on public image datasets have demonstrated that our method outperforms all other state-of-the-art deep learning methods.

1. Introduction

Extensive animal and human studies shows that small cerebrovascular abnormalities are the cause of many neurologic disorders [1–3]. Vascular diseases such as hypertension and arteriosclerosis, cerebral amyloid angiopathy, diabetes, ischemia, stroke, and many neurodegenerative and inflammatory diseases (e.g., dementia, Alzheimer's disease, Parkinson's disease, epilepsy, traumatic brain injury, etc.), usually do not have clear pathogenetic mechanisms. All of these diseases are characterized by significant small blood vessel involvement. Therefore, there is an urgent need for better detection and understanding of vascular abnormalities in vivo at the micro-level [4], where critical vascular nourishment and cellular metabolic changes occur. Being able to monitor and diagnose these changes earlier in the disease process will open the door to better understanding the etiology of the disease and to better treatment of the disease. There is an emerging trend and urgent demand for high resolution magnetic resonance imaging (MRI) images that are capable of extracting vessels and pinpointing vascular abnormalities. However, the super high-resolution (HR) MRI requires impractical longer scanning time, more expensive acquisition devices, and more labor-intensive processing work. It also suffers from lower signal-to-noise ratio (SNR) and smaller spatial coverage, which become the notorious bottlenecks and challenges in many medical and research explorations as well as applications. The motivation of this project is to extract the high-fidelity vascular structures from relatively low-resolution images, which are commonly used in current clinical practice for vessel imaging, to offer a practical solution for high-quality brain vessel analysis.

The advancement of blood vessel shape segmentation has gone through multiple stages. Early research on vessel shape segmentation mainly focused on manual feature extraction, filtering-based models, line detection and wavelet transform [5–7]. These methods aim to enhance boundary gradients, remove unwanted background information, and filter image noise, thereby reducing the segmentation problem to a mathematical optimization problem with fixed patterns. However, these methods cannot cope well with brain vessel shape segmentation, especially for geometrically sophisticated and topologically complex brain vessels. In recent years, deep learning-based methods have developed rapidly. They have been widely applied in the field of medical image analysis due to their excellent representation learning capabilities and they surpass the performance of those traditional

[*] Corresponding author.
E-mail address: mhua27@cranbrook.edu (M.J. Hua).

https://doi.org/10.1016/j.cag.2024.103976
Received 19 May 2024; Accepted 11 June 2024
Available online 19 June 2024

methods [8]. Fully convolutional networks (FCNs) provide an end-to-end solution to utilize adjacent information to accomplish the task of structured prediction [9,10]. However, the max pooling operation in FCNs sacrifices network localization accuracy and produces relatively low segmentation accuracy [11]. In order to solve this problem, a U-Net structure [12] is designed based on the structure of FCN. In addition to the context, image feature can be propagated from its downsampling parts to its upsampling parts to allow feature aggregation from different resolution levels. Following the U-Net [12], various viable variant models have emerged for vessel shape segmentation. Although these models achieve certain segmentation improvement according to the evaluation metrics, there are still significant undetected blood vessel pixels in the segmentation maps, resulting in poor vessel connectivity and accuracy, especially for smaller vessels. This is because these vessel shape segmentation algorithms only rely on convolutional neural networks (CNN) for feature extraction. However, due to the local and limited receptive field of CNNs, it is difficult for the CNN-based algorithms to extract the global features of the image. Therefore, they are insensitive to the position information of each extracted feature. This results in poor connections between segmented vessels and insufficient segmentation of fine/small blood vessels.

Recently, transformer, originally designed for sequence prediction, has emerged as an alternative architecture, which completely abandons convolutional operations and relies only on self-attention [13]. Unlike CNN-based methods, transformer, with a global sensing, shows strong capabilities in global feature extraction. Based on the architecture of a transformer, Vision Transformer (ViT) model represents an input image as a series of image patches, like the series of word embeddings when applying transformer to text, and directly predicts class labels for the image [14]. ViT has also achieved great success in various image analysis tasks [15–17]. However, since a transformer treats the input as a one-dimensional sequence and focuses on globally relational attentions at each stage, the local feature extraction ability of the transformer is relatively poor. The algorithms, which only use the transformer to encode features, often result in the insufficient capability in segmenting brain vessels. In contrast, CNNs, with strong local feature extraction capabilities, can potentially complement the transformer regarding this shortcoming. Hence, one of our aims is to develop a network model that can take advantage of the positives of both CNN and transformer to solve the aforementioned issues and achieve an improved performance on vessel shape segmentation.

On the other hand, conducting segmentation on enhanced high-resolution images is likely to improve segmentation accuracy. There are a limited number of SR-based image segmentation methods, such as DSRL [18], SegSRGAN [19], and PFSeg [20], to explore this direction. However, there exists no work investigating the correlational interaction between high-resolution (HR) and low-resolution (LR) and how to leverage the knowledge in HR-image-based embedding learning to improve the small scale vessel segmentation from LR images. This is critical in clinical practice because HR MRI are usually difficult to obtain, which requires impractical longer scanning time and more expensive acquisition devices as well as suffers from lower signal-to-noise ratio (SNR) and smaller spatial coverage. If LR images can be enhanced with knowledge regarding fine-scale vessels to facilitate near-HR vessel segmentation quality, it has potential to greatly enhance the current vessel analysis in clinical practice.

Based on the above rationals, in order to facilitate the robust and precise 3D cerebrovascular extraction and analysis from the in-vivo MR data, we propose a novel multi-scale Knowledge Transfer Vision Transformer (i.e., KT-ViT) for vessel shape segmentation. First, it uniquely integrates convolutional embeddings with a transformer in a U-net architecture, which simultaneously responds to local receptive fields with convolution layers and global contexts with transformer encoders in a multi-scale fashion. Therefore, it intrinsically enriches local vessel feature and promotes global vessel connectivity information exchange for a more accurate and robust segmentation. Furthermore, to facilitate using relatively low-resolution images to segment fine scale vessels, a novel knowledge transfer network is designed to automatically transfer the knowledge gained from high-resolution data to the low-resolution handling transformer at multiple levels, including the multi-scale feature levels and the decision level, through an integration of multi-level loss functions. The modeling capability of fine-scale vessel data distribution possessed by the high-resolution image learning network can be transferred to the low-resolution image learning network to enhance its knowledge. Therefore, when only low-resolution image data is available, the low-resolution transformer network can deliver an improved vessel shape segmentation, i.e., inferring fine vessel detail from subtle clues on low-resolution images. Experimental results are evaluated and compared with the state-of-the-art deep learning methods, using extensive public image datasets. Our main contributions can be summarized as follows:

- A novel knowledge transfer transformer is proposed to segment and analyze high-fidelity fine-scale 3D vasculature with complicated geometry and small size from the raw high resolution or low-resolution volumetric images. It provides robust and accurate 3D cerebrovascular segmentation and quantification for the in-vivo MR data.
- A unique U-shape like multi-scale vision transformer is designed to integrate both local and global structure information when processing input image patches. It takes advantage of local convolutional embeddings and global transformer encodings to allow the transformer to have both local and global receptions as well as attentive information exchange capabilities towards better vessel shape segmentation.
- This is the first time that a knowledge transfer transformer is proposed for multi-resolution learning, which can explore the inter-dependencies in vessel data distributions between HR and LR images and pass knowledge at multiple levels to LR network. The computation of vessel probabilities from LR images can be synergistically enhanced from the HR data distribution modeling, especially at the fine scales.
- Extensive experiments on publicly available dataset indicate the effectiveness of the proposed KT-ViT network.

2. Related work

In this section, related works in deep neural networks and their applications in brain vessel shape segmentation as well as knowledge distilling are reviewed.

2.1. Deep neural networks for brain vessel segmentation

With the development of deep neural networks, many types of networks have been applied to vessel shape segmentation.

2.1.1. Convolutional neural network for segmentation

Due to the powerful ability of CNNs in performing image classification and pattern recognition, many researchers have explored applying CNNs to blood vessel shape segmentation tasks, using the feature extraction capabilities of CNNs to replace traditional hand-designed features. These methods usually use 2D, 2.5D, or 3D CNN structures to process each pixel or region of the image as a classification problem and output labels or probabilities of blood vessels or non-vessels.

The first work to apply CNNs for brain vessel segmentation is proposed by Phellan et al. [21]. They utilized a basic CNN architecture that consists of two convolution layers and two fully connected layers. To segment vessels more precisely, more complex deep CNN models [22–25] have been proposed in recent years. For example, Joo et al. [24] used a 3D ResNet architecture and a pixel-wise voting technique to generate bounding boxes around the vessels. DeepMedic [25] employs a generic 11-layer 3D CNN model with a dual pathway architecture to extract features at different scales. Although these CNN-based methods can automatically learn image features and improve the segmentation accuracy compared to hand-crafted features, they ignore the spatial structure and connectivity of blood vessels, leading to segmentation results that are broken or thinned.

Therefore, some researchers resort to the fully convolutional network FCN [11] for blood vessel segmentation. In particular, U-Net [12] has become one of the most important backbones for medical segmentation tasks. It is a symmetric architecture consisting of a encoder with layers for downsampling, a decoder with layers for upsampling, and skip connections to improve local features. Thus it is able to restore the resolution and details of the image and output a complete segmentation map. Recently, many variants of the U-Net models have been proposed for blood vessel segmentation [8,26,27]. For example, Sanchesa et al. [27] inserted the Inception module into U-Net to better handle cerebrovascular regions of different scales. Lee et al. [28] proposed spider U-Net, which employed LSTM between the encoder and decoder to capture the connectivity of different slices. To adaptively combine local features of blood vessel images with global dependencies, Mou et al. [29] added a dual self-attention mechanism consisting of spatial attention and channel attention between the encoder and decoder. Wu et al. [30] proposed SCS-Net to capture contextual information and promote feature fusion at different levels to obtain more semantic representations. However, they require more computing resources and training time, as well as the size and quality of the dataset.

Even though these CNN-based methods try to process the entire image simultaneously to retain the spatial information and connectivity of blood vessels, it is still difficult for CNN-based methods to learn explicit global and long-range semantic information interactions due to the inherent locality of convolutional operations.

2.1.2. Transformers for segmentation

Transformer models [31] have recently demonstrated state-of-the-art performance on a broad range of language tasks, e.g., text classification, machine translation, etc. They are based on a self-attention mechanism that can learn long-range relationships between elements of a sequence. Since attention is permutation invariant, the position of each word needs to be encoded (positional encoding) after tokenization and embedding. To adapt transformer models for computer vision tasks, novel network designs and training schemes have been adopted. For example, Parmar et al. [32] applies self-attention only in the local neighborhood of each query pixel, rather than globally. Child et al. [33] proposed Sparse Transformer, which adopts a scalable approximation to global self-attention. Recently, Vision Transformer (ViT) [34] achieved state-of-the-art results on ImageNet classification by directly applying a transformer to sequences of image patches with global self-attention. Liu et al. proposed Swin Transformer [35], a pure transformer network based upon sliding windows and hierarchical structures, which can be used for different vision tasks e.g., classification, detection and segmentation.

Inspired by transformers in vision, Cao et al. proposed Swin-Unet [16], a pure transformer-based U-Net network based on Swin Transformer, which is used for medical image segmentation. It can effectively capture features and context information, and improve segmentation accuracy. Recently, Chen et al. [36] proposed TRSF-Net, which consists of a transformer-based encoder for feature extraction and a CNN-based decoder for cerebrovascular segmentation. However, they only concatenate convolutional features in the decoder level, which misses local feature extraction in the encoding process. VT-UNet [37] places transformer encoders and decoders in U-Net architecture to computer feature representation, and U-Net [38] Transformer inserts multi-head self-attention and multi-head cross-attention modules in the decoders to express a new representation. These networks can simultaneously utilize global context and multi-scale hierarchy to improve segmentation accuracy and generalization capabilities. However, the potential of transformer-based segmentation network with an integration of local and global feature processing is still under-explored.

2.2. Knowledge distillation

Knowledge distillation (KD) was formally proposed in [39] and was aimed to transfer the knowledge learned by teacher models (high-performance networks) to student models (lightweight networks). As a representative technique of model compression and acceleration, knowledge distillation has received increasing attention from the research community in recent years. The main idea is that the student model mimics the teacher model in order to obtain a competitive performance. Basically, a knowledge distillation system is composed of three key components: knowledge, a distillation algorithm, and a teacher–student architecture [40].

In coronary vessel shape segmentation, Guo et al. [41] proposed a lightweight segmentation framework based on similarity knowledge distillation. The framework uses a feature-level similarity distillation (FSD) module and an adversarial similarity distillation (ASD) module to transfer semantic and spatial similarity knowledge from a complex teacher network to a simple student network. Cai et al. [42] proposed an automatic segmentation method based on knowledge distillation. The method designs a teacher–student architecture that utilizes channel-level knowledge distillation techniques to transfer knowledge from a teacher model trained using accurately cropped regions of interest (ROI) to a student model trained using full-size images. Dang et al. [43] proposed LightVessel, a similarity knowledge distillation framework for lightweight coronary artery vessel segmentation. The feature similarity between the symmetric layers from the encoder and decoder is transferred as knowledge from a cumbersome teacher network to a non-trained lightweight student network. The feature similarity between the encoder and decoder is used for a teacher network to guide a non-trained lightweight student network.

However, existing knowledge distillation works have not explored knowledge transfer between analogous datasets with similar probability distribution, e.g., high-definition and low-definition of a dataset, for image analysis tasks.

3. Multi-scale knowledge transfer transformer for multi-resolution learning

Our novel multi-scale knowledge transfer vision transformer (i.e., KT-ViT) learns the knowledge from high-resolution (HR) image datasets at multiple scales, and then transfers multi-level attentive knowledge to the low-resolution (LR) processing network to enhance its capability. First, a novel multi-scale vision transformer (see Section 3.1) is designed and used for both HR net and LR net. Then, the knowledge gained by HR net is transferred to a symmetric LR net at multiple locations (see Section 3.2), including the multi-scale feature encoding levels, decoding levels, and the decision level, through an integration of multiple loss functions (see Section 3.2.1). The details of our approach are explained as follows.

3.1. Novel multi-scale vision transformer for segmentation

We uniquely integrate convolutional embeddings and transformer encoders in a U-shaped encoder–decoder structure with skip connections, using convolutional layers to respond to local receptive fields while transformer encoders respond to global contexts in a multi-scale manner. The novel multi-scale vision transformer enriches essential local vessel features and simultaneously facilitates global vessel connectivity information exchange for a more accurate and robust segmentation. The network architecture diagram of our multi-scale vision transformer is shown in Fig. 1. When a 3D volume sample is input into the multi-scale vision transformer, it first passes through the encoder of 3D U-Net [44] to obtain the feature map with the size of $H \times W \times D \times C$, with an increasing number of channels and without resolution change. These features are then tokenized to N patches of (P, P, P) voxels with $N = (H \times W \times D)/P^3$ and fed into transformer

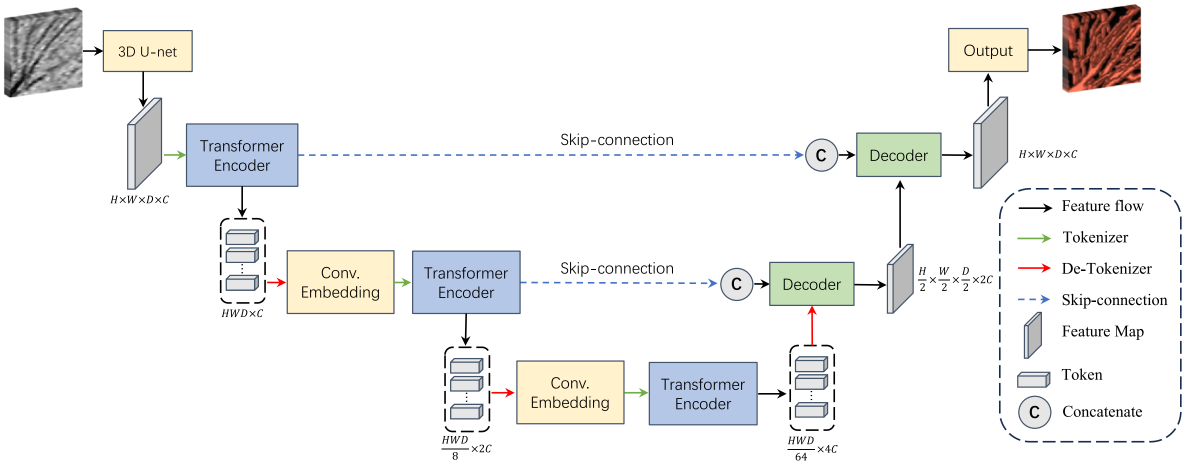

Fig. 1. The network architecture of our novel multi-scale 3D vision transformer, illustrated with a 3D volumetric image input. A 3D input volume is divided into a sequence of uniform non-overlapping patches and projected into an embedding space using a 3D U-net, and used as an input to a transformer model. The encoded representations after the first and the second encoders in the transformer are reshaped and convoluted before the next transformer encoder. These composite representations are extracted and merged with a decoder via skip connections to predict the final segmentation.

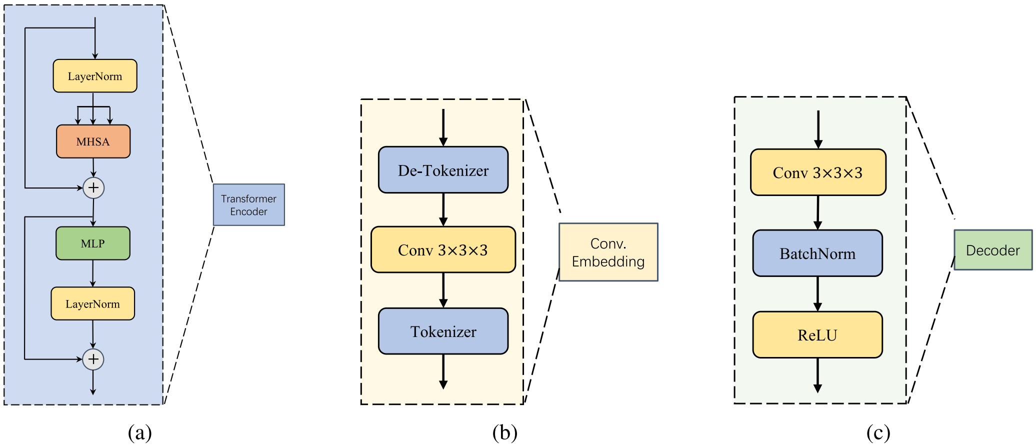

Fig. 2. The major components in the multi-scale vision transformer. (a) shows the transformer encoder with multi-head self-attention mechanism; (b) shows the joint convolutional embedding which is used to increase local receptive field response in the multi-scale vision transformer; (c) shows the light-weight transformer decoder.

encoders (as illustrated in Fig. 2(a)) to obtain encoded representations with richer global information. These encoded representations carry information about the tokens and their contextual relationships within the sequence. In order to capture multi-scale local features and compensate for the positional information during transformer encoding, the encoded representations from the first two transformer encoders are reshaped and input into the joint convolutional embedding module (as illustrated in Fig. 2(b)), respectively. Following U-net processing, a max pooling with 2 × 2 × 2 is applied to downsample the feature map size to $1/2$ at each dimension. In the decoder path, multi-scale encoded representations from the transformer encoders are fed into and merged by a lightweight decoder (as illustrated in Fig. 2(c)) via skip connections to restore its original resolution. The multi-scale processing is enabled with skip-connection in the network structure. Through skip connection, the feature aggregation is performed to include feature vectors from two different scales, so that the decoder can observe more comprehensive information at multiple scales. At the end of the decoder path, a 1 × 1 × 1 convolution is applied to obtain the final segmentation map.

Inspired by the success of Vision Transformer (ViT) [14], our transformer encoder adopts the transformer encoder part of ViT, as shown in Fig. 2(a). The input feature map is divided into multiple small blocks, called patches, and each patch is regarded as a token. Note that, in order to obtain a vessel shape with only one voxel, the smallest patch we can use is 1 × 1 × 1. Global features are then extracted through the self-attention mechanism. The self-attention mechanism is a method of calculating the correlation between each token, which can capture the important long-distance contextual dependencies of the tokens.

Each token calculates an attention score based on its own query vector and the key vectors of other tokens, indicating its level of attention to other tokens. Then, each token weighted-averages the value vectors of other tokens according to its attention score to obtain the output vector. In this way, each token can integrate information from other tokens to achieve global feature extraction, which is helpful for both small vessel and long vessel segmentation. The formula of the self-attention mechanism is as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V, \tag{1}$$

where Q, K, and V are the projections of query, key, and value, respectively. The dot product operation between Q and K is to calculate the score of each vector. d_k is the dimension of the key vector K. Divide by $\sqrt{d_k}$ to make the gradient more stable. The correlation between vectors is then standardized by the softmax function, and then multiplied by value vectors to focus on relevant information.

In order to improve the global feature extraction ability and generalization ability of the model, the transformer encoder adopts the multi-head self-attention mechanism (MHSA). The multi-head self-attention mechanism divides the self-attention mechanism into multiple parallel sub-modules, which allows the model to focus on different features and relationships at the same time, thereby improving the model's expression ability and performance. The input of MHSA is divided into M heads $(head_1, \ldots, head_M)$, and the self-attention of each head is calculated at the same time, and they are spliced together as the final output. The formula for MHSA is as follows:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V), \tag{2}$$

$$MHSA(Q,K,V) = \text{Concat}((head_1, \ldots, head_i, \ldots, head_M)W^0), \tag{3}$$

where W_i^Q, W_i^K, W_i^V $(i = 1, 2, \ldots, M)$, and W^0 are independently learnable weight matrices.

The convolutional embedding shown in Fig. 2(b) applies 3D convolution operations on the de-tokenized representation from a transformer encoder. The de-tokenization is conducted through reshaping the encoded representation to match the expected input shape for convolutional operations. Then, the convolutional layer operates filters over the representation to extract local features. After that, the feature map is reshaped and tokenized for the subsequent transformer layer. The decoder, as shown in Fig. 2(c), is light-weight, consisting of a 3 × 3 × 3 convolution layer, followed by a batch normalization and a rectified linear unit.

To supervise the learning of the multi-scale vision transformer, the Binary Cross Entropy Loss function, also known as BCE loss, is used for measuring the binary classification of vessels to the ground truth. Its formula is:

$$\mathcal{L}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)\right], \tag{4}$$

where y and $\hat{y}$ represent the true value of vessel map and predicted probability of the ith image, respectively; N represents the batch size. The objective of BCE loss is to measure the difference between two vessel probability distributions. When y and $\hat{y}$ are closer, the BCE loss is smaller, and vice versa.

Note that, the multi-scale vision transformer can handle 2D image or 3D volume input in the same fashion. However, the number of dimensions becomes lower for 2D image handling. Suppose that the 2D output after 2D U-net [44] is a volume of $H \times W$ with C channels, i.e., $H \times W \times C$, as compared to $H \times W \times D \times C$ in 3D in Fig. 1. It is divided into N patches of (P, P) voxels with $N = (H \times W)/P^2$. The convolution kernel sizes (as shown in Fig. 2(b) and (c)) are 3 × 3 instead. The computations of transformer for 2D and 3D have the same fashion.

3.2. Novel multi-scale knowledge transfer network design

The majority of low-resolution (LR) and high-resolution (HR) MR images of a human subject contain similar information about the vascular structure. Therefore, they have similar data distributions, while HR MR images retain more fine-scale vessel structure information than LR images. If the knowledge regarding the fine details from high-definition images can be effectively transferred to a LR handling network, it can be enhanced with a better vessel expression and analysis capabilities, even based on low definition. As seen in Fig. 3, to facilitate using relatively low-resolution images for fine scale vessel shape segmentation, a novel knowledge transfer architecture is needed to transfer the knowledge gained from high-resolution data (i.e., H-net) to the low-resolution handling network (i.e., L-net). The modeling capability of fine-scale vessel data distribution possessed by the HR image learning network can be transferred to the LR image learning network to enhance its knowledge, which in turn improves its accuracy. When only LR image data is available (very common in current clinical practice), the low-resolution transformer network has the capability to outperform with respect to vessel shape segmentation.

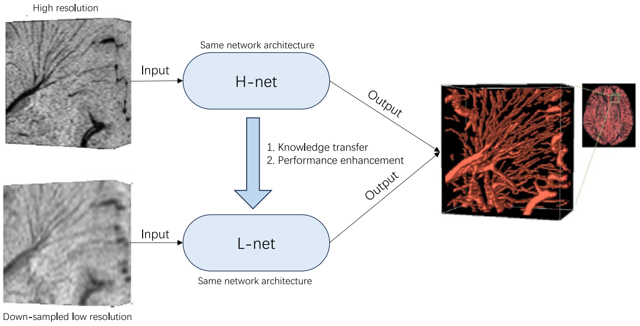
High resolution
Input
Same network architecture
H-net
Output
1. Knowledge transfer
2. Performance enhancement
Output
Input
L-net
Same network architecture
Down-sampled low resolution

Fig. 3. Our proposed multi-resolution learning architecture with knowledge transfer.

In existing student-teacher distillation networks, the inputs to both student and teacher networks are the same, with the exact data distribution. The purpose of traditional student-teacher networks are generally used to train a light-weight student network through knowledge distilling from a complex teacher network. Unlike previous distillation learning that inputs the same data into two branches, in our knowledge transfer network, high and low resolution images are input to the H-net and L-net networks, respectively. To enable a more effective transfer, we propose to use the same network architecture for both H-net and L-net. A down-sampling and up-sampling are conducted to maintain consistent dimension sizes at each layer between the two networks. This symmetric architecture design allows the knowledge transfer to happen at multiple processing steps, including the feature levels and decision level.

In order to facilitate the use of relatively low-resolution images to segment fine-scale blood vessels, a new knowledge transfer architecture is designed based on the multi-scale vision transformer, as described in Section 3.1, to automatically transfer the knowledge acquired from H-net to L-net. Thus, the fine-scale blood vessel data distribution modeling ability of the high-resolution image learning network can be transferred to the low-resolution image learning network to enhance its knowledge of fine-scale vessel segmentation. As seen in Fig. 4, H-net and L-net use the same multi-scale vision transformer, described in Section 3.1, forming a symmetric architecture. H-net takes the input images with the original resolution. The original resolution images are downsampled to one half resolution to establish low-resolution data and then resized back to their original size. That is to say, the low-resolution images have the same dimension sizes as the original high-resolution images, but with lower-resolution quality content, which are input to L-net. Since the input data dimensions are the same for both H-net and L-net, and their network processing is the same as well, the corresponding representation spaces of H-net and L-net are aligned. Therefore, the knowledge gained from high-resolution data by H-net can be transferred to the L-net at multiple levels, including the multi-scale feature encoding levels, decoding levels, and the decision level, through an integration of multi-level loss functions.

3.2.1. Multi-level loss function

Multi-level loss functions need to be designed and enforced in order to pass knowledge effectively from H-net to L-net. One type of loss function of the KT-ViT is the Decision-phase Knowledge Transfer Loss, which is used to measure the difference in final decision probability distribution between the H-net and the L-net. This knowledge transfer loss is defined using Kullback–Leibler Divergence loss, which can reflect the difference between two distributions output by the H-net and the L-net. The equation is as follows:

$$\text{KL}(P_{\text{L-net}} \parallel P_{\text{H-net}}) = \sum_i P_{\text{L-net}}(i) \cdot \log\left(\frac{P_{\text{L-net}}(i)}{P_{\text{H-net}}(i)}\right), \tag{5}$$

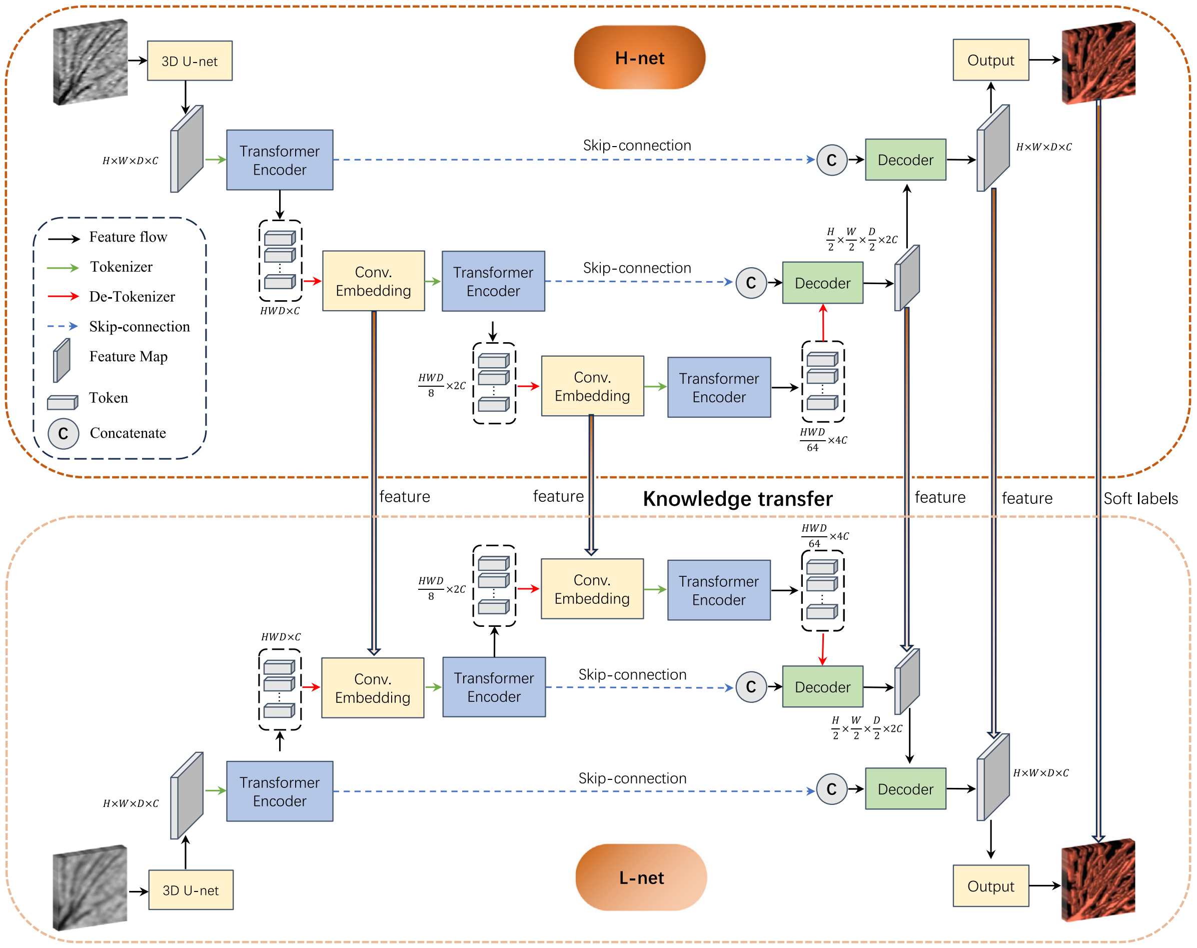

Fig. 4. The architecture of the multi-scale knowledge transfer transformer, consisting of H-net and L-net. The thick arrows indicate the knowledge transfer from H-net to L-net during training at the convolutional embeddings, transformer decoders, decision phase, and supervision phase.

where $P_{\text{H-net}}$ is the probability distribution of the H-net output; $P_{\text{L-net}}$ is the probability distribution of the L-net output; i represents a vessel map in the ith image.

The advantage of the decision-phase knowledge transfer loss is that it allows L-net to learn the soft label (Soft-Label) of H-net network, which is the probability value of each binary category, instead of just the hard label (Hard-Label), which is the maximum probability corresponding to category. This allows L-net to obtain more information and details online and improve its decision-making ability.

Another type of knowledge transfer loss function used in KT-ViT is Feature-phase Knowledge Transfer Loss, which measures the difference between the feature layers of the L-net and the H-net. The minimization of feature-level differences between L-net and H-net allows L-net to imitate the feature extraction capabilities of H-net. In our KT-ViT, feature-phase knowledge transfer loss is calculated using Mean Squared Error (i.e., MSE), which can reflect the distance or similarity between two feature vectors. The advantage of this loss is that it allows L-net to learn the feature representation capabilities of H-net, thereby improving its ability to understand and characterize the input data.

Mean Squared Error (MSE) loss calculates the mean of the squared differences between the predicted and true values. The equation for MSE is:

$$\text{MSE}_{\text{feature}} = \frac{1}{N}\sum_{i=1}^{N}(F_{\text{H-net}}(i) - F_{\text{L-net}}(i))^2. \tag{6}$$

Among them, $F_{\text{H-net}}(i)$ is the feature map of the H-net feature layer output; $F_{\text{L-net}}(i)$ is the feature map of the L-net feature layer output; N is the number of samples. By minimizing this mean square error, the L-net can better learn the feature extraction capability of the H-net and improve its model performance.

Decision-phase and feature-phase knowledge transfer losses focus on different levels of information transfer and complement to each other. When integrating them to form a comprehensive objective function and balancing the importance between them through weight coefficients, it allows the L-net to obtain leveraged vessel shape segmentation capability, even with low-resolution input. Note that the H-net is trained individually using high-resolution data and then its parameters are kept frozen when training KT-ViT. The overall loss function is as follows,

$$\begin{aligned}\text{L}_{\text{total}} = L_{L-net}(y,\hat{y}) + \sum_{i=1}^{4} w_i \text{MSE}(feature_i)\\ + w\text{KL}(P_{L-net} \parallel P_{H-net}),\end{aligned} \tag{7}$$

where $L_{L-net}(y,\hat{y})$ is the binary cross entropy loss between the L-net prediction and the groundtruth upon low-resolution input, w_i are the weights of the four feature-phase knowledge transfer losses as indicated in Fig. 4, MSE measures the feature differences, w is the weight of the KL loss between the prediction probabilities of L-net and H-net.

4. Experiments

We conduct extensive experiments including training and testing on public MRA dataset using its standard protocols. The source code will be publicly available soon.

Table 1
Quantitative performance evaluation of different methods on TubeTK dataset. '–' means 'not applicable' due to lack of their implementations or results. The best results are shown bolded.

Methods/Metrics	Sensitivity (%) ↑	Precision (%) ↑	Dice (%) ↑
Multi-scale Vision Transformer (ours)	**83.21**	**79.38**	**74.83**
3D TransUNet	80.81	77.41	72.52
VC-Net	79.33	76.66	71.81
3D U-Net	75.99	74.00	71.01
Uception	66.02	–	67.01
DeepVesselNet	63.20	63.75	64.12
2D U-Net	73.93	70.05	65.10

4.1. Datasets and settings

Public MRA Dataset. To compare KT-ViT with state-of-the-art (SOTA) methods, we use a public TubeTK Toolkit MRA dataset from the University of North Carolina at Chapel Hill, acquired by a Siemens Allegra head-only 3T MR system. The dataset contains 42 patient cases with the manual-labeled vessel shape segmentation masks. The voxel spacing of the MRA images is $0.5 \times 0.5 \times 0.8$ mm^3 with a volume size of 448 × 448 × 128 voxels.

When processing the MRA TubeTK dataset, we first strip out the voxels of skull to extract the brain for each image [45]. Due to the size of the whole brain volume, the training samples of our KT-ViT network are those regions cropped from the volume images. The 3D training samples are randomly-cropped regions with overlapping from the whole 3D brain MRA. In the experiments, 80 regions are cropped from each TubeTK case. The random split of training/validation/testing cases is 33/3/6. Therefore, the training/validation/testing samples is 2640/240/480. All the numerical evaluations are reported in terms of whole brain region instead without overlapping.

The performance of our KT-ViT network and SOTA methods are evaluated with the following three quantitative metrics, which are defined from the classifier confusion matrix from different aspects:

Sensitivity, computed as $TP/(TP + FN)$, measures a model's capability of extracting real vessel voxels as many as possible, which are very sparsely distributed in a MRI volume.

Precision, computed as $TP/(TP + FP)$, measures a model's capability of ruling out the various noises and obtaining the correct vessel voxels.

Dice Similarity (Dice), computed as $2TP/(2TP + FP + FN)$, measures the intersection over union between prediction and ground truth, which examines sparse vessel shape segmentation quality with respect to background.

4.2. Implementation details

Our KT-ViT is implemented using PyTorch and trained/tested on a single NVIDIA GeForce RTX 3090 GPU with 24 GB GDDR6X. Model parameters are optimized using the Adam optimization method [46] with weight decay of 0.00001 and an initial learning rate of 0.0001. The cosine annealing algorithm gradually reduces the learning rate over 20 epochs.

During the training phase, the images are preprocessed as follows: (1) We downsample the original high-resolution version of the input data set to $1/2$, and then upsample it back to the original size to generate low-resolution image data, but its content is at low-resolution quality because of the downsampling process; (2) Randomly crop out a region with a fixed size; (3) Randomly conduct horizontal flipping, vertical flipping and [90, 180, 270] degree rotation for data enhancement during the training phase to increase the diversity of data and reduce overfitting; (4) Train the H-net using high-definition datasets (i.e., the original resolution of TubeTK data) and save the trained H-net model weights; (5) KT-ViT uses the aforementioned downsampled and then upsampled data to learn and lets the frozen H-net serve as the guide of the L-net; (6) When KT-ViT converges, save the L-net which is the knowledge enhanced low-resolution handling network.

4.3. Comparison with the state-of-the-art methods

We conduct extensive experiments and compare the performance of our KT-ViT on 3D brain vessel shape segmentation with a number of SOTA deep learning methods.

4.3.1. Performance of the multi-scale vision transformer on the original TubeTK

First, we conduct experiments with our multi-scale vision transformer (i.e., single H-net without knowledge transfer) and other SOTA deep learning methods, including 3D TransUNet [47], VC-net [48], 3D U-Net [44], 2D U-Net [12], DeepVesselNet [23], and Uception [27], on original TubeTK data resolution. All deep learning methods in comparison are trained until convergence with the same dataset split or using the results reported in their original publication (i.e., Uception). The numerical evaluations are shown in Table 1, where '–' means 'not applicable' due to a lack of their implementations or results. The best results are shown bolded. It is clear that the multi-scale vision transformer network significantly outperforms all other recent approaches. The performance gain is mainly because that our convolutional embedding is interleaved with the transformer encoders in a U-shape like network and it captures rich multi-scale local feature well in advance for transformer's contextual analysis. Collaboratively, it captures distinctive local and global characteristics of vessels for segmentation, especially for small vessels and complex connectivities.

The final vessel shape segmentation from our method shows better connectivity and better small vessel extraction. Fig. 5 shows qualitative analysis and visual comparison between our method and the runner-up method, 3D TransUNet [47], on the TubeTK case. The dotted circles highlight the differences in segmentation results between different methods and the ground truth. The yellow circles represent the blood vessels that are not detected by 3D TransUNet but correctly segmented out by our algorithm. Green circles indicate blood vessels that are incorrectly detected by 3D TransUNet but not by our algorithm, which shows that multi-scale convolutional transformer has better segmentation in certain error-prone areas. The blue box indicates a situation where the multi-scale convolutional transformer segments a more correct portion than 3D TransUNet in this region as compared to the ground truth. Visually, our multi-scale vision transformer can segment smaller blood vessels and more connectivities than 3D TransUNet. The qualitative analysis and visual comparison to VC-Net and 3D U-net can be found in Fig. 6.

Note that the number of our network parameters is 25.5 million and the training of our model takes about 110 epochs with 12 h, and the inference time is 175 ms for each patch of 64 × 64 × 32 and 35 s for a whole brain.

4.3.2. Performance of the KT-ViT on the low-resolution TubeTK

In this section, we conduct experiments with our KT-ViT and other SOTA methods, including VC-net [48] and 3D U-Net [44], on the low-resolution TubeTK data. The low-resolution input volume is downsampled from 448 × 448 × 128 to 224 × 224 × 64 and then upsampled back to 448 × 448 × 128 to maintain its size. The image quality is at low-resolution. Using low-resolution data as the input, the numerical evaluations are shown in Table 2, where the best results are shown

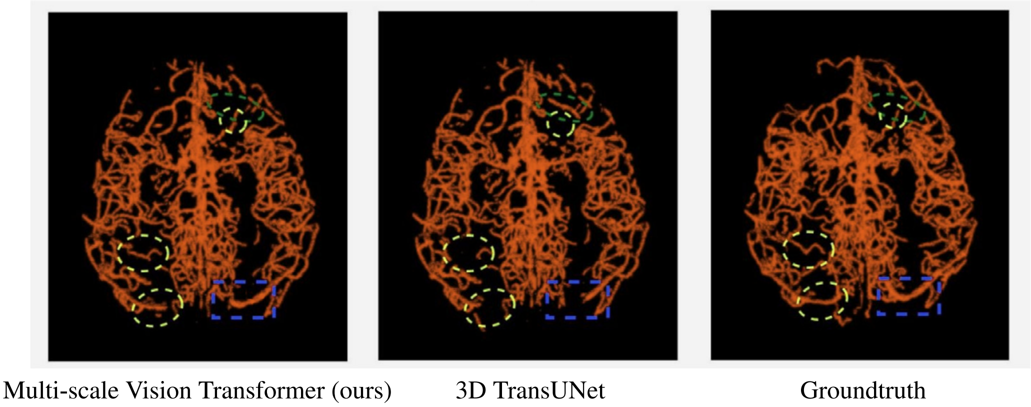

Fig. 5. Qualitative comparison results on TubeTK dataset: The 3D global vessel segmentations from multi-scale convolutional transformer, 3D TransUNet, and the ground truth are shown. The highlighted comparison areas are marked in circles. The yellow circles represent the blood vessels that are not detected by 3D TransUNet but correctly segmented out by our algorithm. Green circles indicate blood vessels that are incorrectly detected by 3D TransUNet but not by our algorithm. The blue box indicates a region where multi-scale convolutional transformer segments more correct portion than 3D TransUNet.

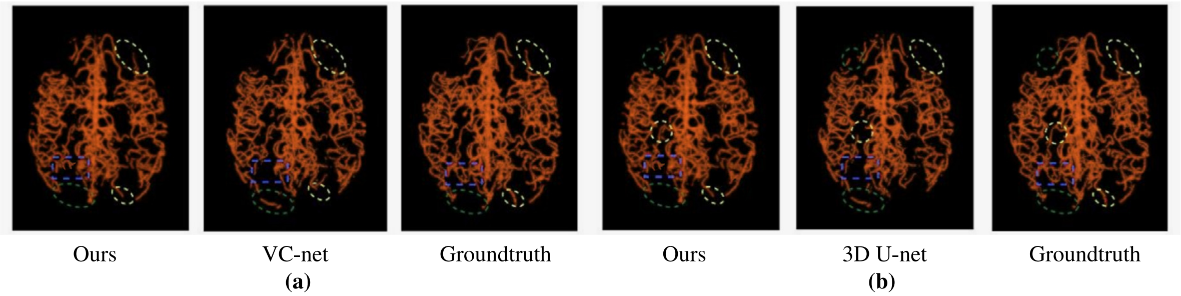

Fig. 6. Qualitative comparison results on TubeTK dataset: (a) The 3D global vessel segmentations from multi-scale vision transformer, VC-net, and the ground truth are shown. The yellow circles represent the blood vessels that are not detected by VC-net but correctly segmented out by our algorithm. Green circles indicate blood vessels that are incorrectly detected by VC-net but not by our algorithm. The blue box indicates a region where multi-scale vision transformer segments more correct portion than VC-net. (b) The 3D global vessel segmentations from multi-scale vision transformer, 3D U-net, and the ground truth are shown. The yellow circles represent the blood vessels that are not detected by 3D U-net but correctly segmented out by our algorithm. Green circles indicate blood vessels that are incorrectly detected by 3D U-net but not by our algorithm. The blue box indicates a region where multi-scale vision transformer segments a more correct portion than 3D U-net.

Table 2
Quantitative evaluation of different methods on low-resolution TubeTK dataset. The input volumes to all networks are at low-resolution.

Methods/Metrics	Sensitivity (%) ↑	Precision (%) ↑	Dice (%) ↑
L-net from full KT-ViT (ours)	**81.73**	**78.06**	**72.93**
L-net only without knowledge transfer (ours)	80.03	76.45	72.01
VC-Net	76.92	73.89	69.93
3D U-Net	72.85	71.11	69.24

bolded. The L-net, even without knowledge transfer from H-net, outperforms other SOTA methods which lack knowledge transfer capabilities. The performance difference between models with and without knowledge transfer from H-net shows that the knowledge transfer from the frozen H-net is beneficial to L-net, which improves its performance with 1.7% on sensitivity, 1.61% on precision, and 0.92% on dice. The knowledge transfer improves the L-net's performance to a near-HR vessel segmentation quality. Thus, L-net has the potential to use current clinical MR scans to provide more meaningful vessel analysis for diagnosis.

4.4. Model analysis and ablation study

In the multi-scale vision transformer, the convolutional embedding increases the response to local receptive field so that small vessel feature information can be extracted at multiple scales. In addition, it implicitly augments a position embedding information to the local reception, i.e., the weights of the convolution embedding also embeds the positional response. This learnable embedding allows transformer to compute a global feature representation suitable for vessel structure and connectivity based on the local representations. We have conducted an ablation study on the impact of the convolutional embedding in the multi-scale vision transformer. To remove the convolutional embedding, we replace it with 2 × 2 × 2 max pooling, which simply reduces the scale but does not learn local embedding. In Table 3, it shows that when we use the proposed convolutional embeddings in our transformer network, the performance values are significantly higher. Compared to the multi-scale vision transformer without convolutional embedding, its performance is increased with 7.88% on sensitivity, 5.72% on precision and 5.02% on dice metric. This indicates that the convolutional embedding captures high-quality local features and embeds positional information for the multi-scale vision transformer. As a result, it captures better vessel structure and connectivity.

The experiment shown in Table 2 already indicates that the knowledge transfer is important for L-net to learn high-quality information from H-net. Note that, the difference between the input high-resolution and low-resolution images are measured with the peak signal-to-noise ratio (PSNR) to provide insights of how the low-resolution volumes

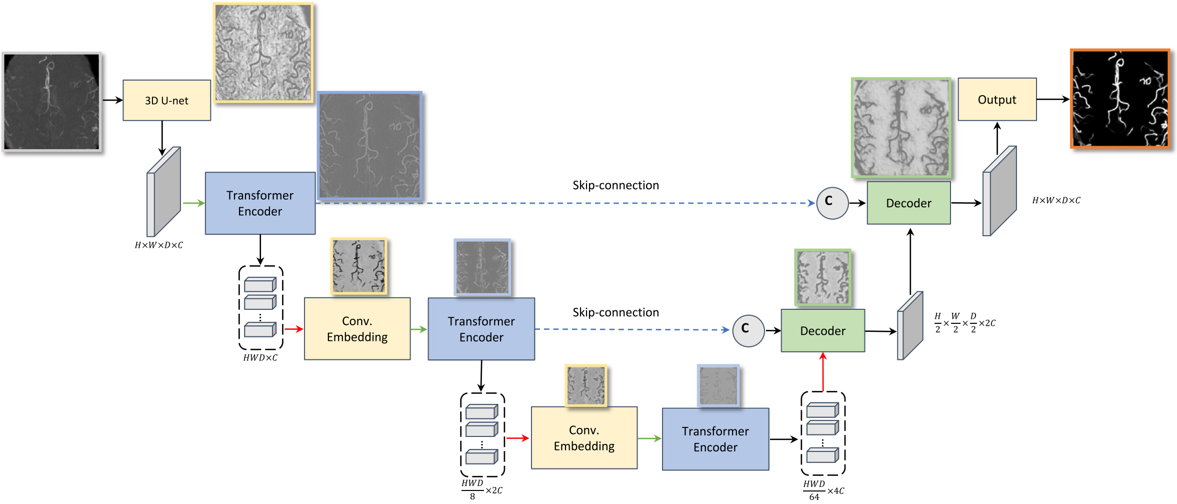

Fig. 7. Visualization of network processing with produced feature map (only showing one channel). Convolutional embeddings and transformer encoders collaboratively enrich essential local and global vessel features in a multi-scale hierarchy and simultaneously facilitate more accurate and robust segmentation of small and long vessels. The sizes of feature maps indicate the downsampling and upsampling processing.

Table 3
Quantitative performance evaluation of the impact of convolutional embedding on original resolution TubeTK dataset.

Methods/Metrics	Sensitivity (%) ↑	Precision (%) ↑	Dice (%) ↑
Multi-scale Vision Transformer (ours)	**83.21**	**79.38**	**74.83**
Our Transformer w/o convolutional embedding	75.33	73.66	69.81

Table 4
Model analysis on knowledge transfer on low-resolution TubeTK data.

Methods/Metrics	Sensitivity (%) ↑	Precision (%) ↑	Dice (%) ↑
L-net with full KT-ViT	**81.73**	**78.06**	**72.93**
L-net with only feature-level knowledge transfer	81.15	77.72	72.64
L-net with only decision-level knowledge transfer	80.53	76.89	72.30
L-net without any knowledge transfer	80.03	76.45	72.01

compare to the high-resolution volumes. In our experiment, the average PSNR is 26.4 dB. To further analyze the importance of different knowledge transfer levels, we conduct an ablation study by excluding certain knowledge transfer. From Table 4, it can be observed that the feature level knowledge transfer is more important than the decision level, because the ground truth data supervision for L-net already provides sufficient information at the decision level. Thus, the additional decision information from H-net does not help L-net as much as the feature level guidance.

To illustrate the network's processing, we take one brain subregion sample as input and visualize one of the feature channels for each network processing steps. Fig. 7 shows certain feature channel at each step. The sizes of feature maps indicate the downsampling and upsampling processing. As seen from the figure, convolutional embeddings and transformer encoders collaboratively enrich essential local and global vessel features in a multi-scale hierarchy and simultaneously facilitate more accurate and robust segmentation of small and long vessels.

5. Conclusion

In this work, we have presented a novel Multi-scale Knowledge Transfer Vision Transformer for vessel shape segmentation. KT-ViT uniquely integrates convolutional embeddings with transformer in a U-net architecture, which simultaneously responds to local receptive fields with convolution layers and global contexts with transformer encoders in a multi-scale fashion. Therefore, it intrinsically enriches multi-scale local vessel feature and promotes global vessel connectivity related contextual information exchange for a more accurate and robust segmentation. Furthermore, to facilitate using relatively low-resolution images to segment fine scale vessels, a novel knowledge transfer network is designed to automatically transfer the knowledge gained from high-resolution data to the low-resolution handling transformer at multiple levels, including the multi-scale feature levels and the decision supervision level, through an integration of multi-level loss functions. The modeling capability of fine-scale vessel data distribution possessed by the high-resolution image learning network can be transferred to the low-resolution image learning network to enhance its knowledge for fine vessel segmentation. The extensive experimental results on public image datasets have demonstrated its superiority when compared to the current state-of-the-art deep learning methods.

In the future, we plan to investigate deep learning-based approaches for MRI signal processing and image reconstruction, and hope that finer-scale vessels can be enhanced in this early signal handling stage. I also plan to explore more applications on quantitative disease diagnosis with KT-ViT.

CRediT authorship contribution statement

Michael J. Hua: Conceptualization, Methodology, Software, Validation, Visualization, Writing – original draft. **Junjie Wu:** Software, Validation, Writing – review & editing. **Zichun Zhong:** Conceptualization, Investigation, Project administration, Writing – review & editing.

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 used data is public Denmark data.

References

[1] Brown W, Thore C. Cerebral microvascular pathology in ageing and neurodegeneration. Neuropathol Appl Neurobiol 2011;37(1):56–74.

[2] Dorr A, Sahota B, Chinta LV, Brown ME, Lai AY, Ma K, et al. Amyloid-β-dependent compromise of microvascular structure and function in a model of Alzheimer's disease. Brain 2012;135(10):3039–50.

[3] Gouw A, Seewann A, Van Der Flier W, Barkhof F, Rozemuller A, Scheltens P, et al. Heterogeneity of small vessel disease: a systematic review of MRI and histopathology correlations. J Neurol Neurosurg Psychiatry 2011;82(2):126–35.

[4] Mott M, Pahigiannis K, Koroshetz W. Small blood vessels: big health problems: National institute of neurological disorders and stroke update. Stroke 2014;45(12):e257–8.

[5] Moccia S, De Momi E, El Hadji S, Mattos LS. Blood vessel segmentation algorithms—review of methods, datasets and evaluation metrics. Comput Methods Programs Biomed 2018;158:71–91.

[6] Deshpande A, Jamilpour N, Jiang B, Michel P, Eskandari A, Kidwell C, et al. Automatic segmentation, feature extraction and comparison of healthy and stroke cerebral vasculature. NeuroImage 2021;30:102573.

[7] Cervantes J, Cervantes J, García-Lamont F, Yee-Rendon A, Cabrera JE, Jalili LD. A comprehensive survey on segmentation techniques for retinal vessel segmentation. Neurocomputing 2023;556:126626.

[8] Hilbert A, Madai VI, Akay EM, Aydin OU, Behland J, Sobesky J, et al. BRAVE-NET: fully automated arterial brain vessel segmentation in patients with cerebrovascular disease. Front Artif Intell 2020;3:78.

[9] Li Y, Qi H, Dai J, Ji X, Wei Y. Fully convolutional instance-aware semantic segmentation. In: Proc. IEEE conf. comput. vis. pattern recognit. 2017, p. 2359–67.

[10] Zhao Y, Chen Y, Chen Y, Zhang L, Wang X, He X. A fully convolutional network (FCN) based automated ischemic stroke segment method using chemical exchange saturation transfer imaging. Med Phys 2022;49(3):1635–47.

[11] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proc. IEEE conf. comput. vis. pattern recognit. 2015, p. 3431–40.

[12] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. 18th int. conf. med. image comput. comput.-assist. interv.. 2015, p. 234–41.

[13] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. In: Advances in neural information processing systems. 2017, p. 5998–6008.

[14] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In: International conference on learning representations. 2021.

[15] Strudel R, Garcia R, Laptev I, Schmid C. Segmenter: Transformer for semantic segmentation. 2021, arXiv preprint arXiv:2105.05633.

[16] Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. Springer; 2022, p. 205–18.

[17] Yu X, Wang J, Zhao Y, Gao Y. Mix-ViT: Mixing attentive vision transformer for ultra-fine-grained visual categorization. Pattern Recognit 2023;135:109131.

[18] Wang L, Li D, Zhu Y, Tian L, Shan Y. Dual super-resolution learning for semantic segmentation. In: Proc. IEEE conf. comput. vis. pattern recognit. 2020, p. 3774–83.

[19] Delannoy Q, Pham CH, Cazorla C, Tor-Díez C, Dollé G, Meunier H, et al. SegSRGAN: Super-resolution and segmentation using generative adversarial networks—Application to neonatal brain MRI. Comput Biol Med 2020;120:103755.

[20] Wang H, Lin L, Hu H, Chen Q, Li Y, Iwamoto Y, et al. Patch-free 3D medical image segmentation driven by super-resolution technique and self-supervised guidance. In: Proc. 24th int. conf. med. image comput. comput. assist. interv. Springer; 2021, p. 131–41.

[21] Phellan R, Peixinho A, Falcão A, Forkert ND. Vascular segmentation in TOF MRA images of the brain using a deep convolutional neural network. In: Intravascular imaging and computer assisted stenting, and large-scale annotation of biomedical data and expert latl synthesis: 6th joint international workshops, CVII-STENT 2017 and second international workshop, LABELS 2017, held in conjunction with MICCAI 2017, Québec City, QC, Canada, September 10–14, 2017, proceedings 2. Springer; 2017, p. 39–46.

[22] Nakao T, Hanaoka S, Nomura Y, Sato I, Nemoto M, Miki S, et al. Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography. J Magn Reson Imaging 2018;47(4):948–53.

[23] Tetteh G, Efremov V, Forkert ND, Schneider M, Kirschke J, Weber B, et al. Deepvesselnet: Vessel segmentation, centerline prediction, and bifurcation detection in 3-d angiographic volumes. Front Neurosci 2020;14:1285.

[24] Joo B, Ahn SS, Yoon PH, Bae S, Sohn B, Lee YE, et al. A deep learning algorithm may automate intracranial aneurysm detection on MR angiography with high diagnostic performance. Eur Radiol 2020;30:5785–93.

[25] Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal 2017;36:61–78.

[26] Liu Y, Kwak H-S, Oh I-S. Cerebrovascular segmentation model based on spatial attention-guided 3D inception U-Net with multi-directional MIPs. Appl Sci 2022;12(5):2288.

[27] Sanchesa P, Meyer C, Vigon V, Naegel B. Cerebrovascular network segmentation of MRA images with deep learning. In: 2019 IEEE 16th international symposium on biomedical imaging. ISBI 2019, IEEE; 2019, p. 768–71.

[28] Lee K, Sunwoo L, Kim T, Lee KJ. Spider U-Net: Incorporating inter-slice connectivity using LSTM for 3D blood vessel segmentation. Appl Sci 2021;11(5):2014.

[29] Mou L, Zhao Y, Chen L, Cheng J, Gu Z, Hao H, et al. Cs-Net: Channel and spatial attention network for curvilinear structure segmentation. In: Proc. 22nd int. conf. med. image comput. comput. assist. interv. 2019, p. 721–30.

[30] Wu H, Wang W, Zhong J, Lei B, Wen Z, Qin J. Scs-Net: A scale and context sensitive network for retinal vessel segmentation. Med Image Anal 2021;70:102025.

[31] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst 2017;30.

[32] Parmar N, Vaswani A, Uszkoreit J, Kaiser L, Shazeer N, Ku A, et al. Image transformer. In: International conference on machine learning. 2018.

[33] Child R, Gray S, Radford A, Sutskever I. Generating long sequences with sparse transformers. 2019, arXiv preprint arXiv.

[34] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In: International conference on learning representations. 2021.

[35] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[36] Chen C, Zhou K, Wang Z, Xiao R. Generative consistency for semi-supervised cerebrovascular segmentation from TOF-MRA. IEEE Trans Med Imaging 2022;42(2):346–53.

[37] Peiris H, Hayat M, Chen Z, Egan G, Harandi M. A robust volumetric transformer for accurate 3D tumor segmentation. In: Proc. 25th int. conf. med. image comput. comput. assist. interv. Springer; 2022, p. 162–72.

[38] Petit O, Thome N, Rambour C, Soler L. U-Net transformer: Self and cross attention for medical image segmentation. 2021, arXiv:2103.06104.

[39] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. 2015, arXiv preprint arXiv:1503.02531.

[40] Gou J, Yu B, Maybank SJ, Tao D. Knowledge distillation: A survey. Int J Comput Vis 2021;129:1789–819.

[41] Guo H, Wang S, Dang H, Xiao K, Yang Y, Liu W, et al. LightBTSeg: A lightweight breast tumor segmentation model using ultrasound images via dual-path joint knowledge distillation. 2023, arXiv preprint arXiv.

[42] Cai Q, Chen R, Li L, Huang C, Pang H, Tian Y, et al. The application of knowledge distillation toward fine-grained segmentation for three-vessel view of fetal heart ultrasound images. 2022, 2022,

[43] Dang H, Zhang Y, Qi X, Zhou W, Sun M. Lightvessel: Exploring lightweight coronary artery vessel segmentation via similarity knowledge distillation. In: IEEE international conference on acoustics, speech and signal processing. IEEE; 2023, p. 1–5.

[44] Çiçek Ö, Abdulkadir A, Lienkamp S, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Proc. 19th int. conf. med. image comput. comput. assist. interv. 2016, p. 424–32.

[45] Jenkinson M, Pechaud M, Smith S. BET2: MR-based estimation of brain, skull and scalp surfaces. In: Eleventh annual meeting of the organization for human brain mapping. 2005.

[46] Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014, arXiv preprint arXiv:1412.6980.

[47] Chen J, Mei J, Li X, Lu Y, Yu Q, Wei Q, et al. 3D TransUNet: Advancing medical image segmentation through vision transformers. 2023, arXiv:2310.07781.

[48] Wang Y, Yan G, Zhu H, Buch S, Wang Y, Haacke EM, et al. VC-net: Deep volume-composition networks for segmentation and visualization of highly sparse and noisy image data. IEEE Trans Visual Comput Graph 2020;27(2):1301–11.