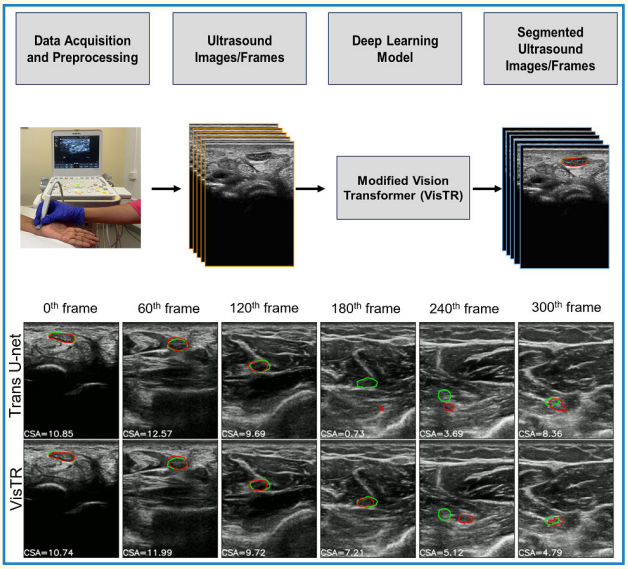
Transformer-Based Automated Segmentation of the Median Nerve in Ultrasound Videos of Wrist-to-Elbow Region

Karan R. Gujarati, Lokesh Bathala, Vaddadi Venkatesh, Raji Susan Mathew, *Member, IEEE*, and Phaneendra K. Yalavarthy, *Senior Member, IEEE*

***Abstract*—Segmenting the median nerve is essential for identifying nerve entrapment syndromes, guiding surgical planning and interventions, and furthering understanding of nerve anatomy. This study aims to develop an automated tool that can assist clinicians in localizing and segmenting the median nerve from the wrist, midforearm, and elbow in ultrasound videos. This is the first fully automated single deep learning model for accurate segmentation of the median nerve from the wrist to the elbow in ultrasound videos, along with the computation of the cross-sectional area (CSA) of the nerve. The visual transformer architecture, which was originally proposed to detect and classify 41 classes in YouTube videos, was modified to predict the median nerve in every frame of ultrasound videos. This is achieved by modifying the bounding box sequence matching block of the visual transformer. The median nerve segmentation is a binary class prediction, and the entire bipartite matching sequence is eliminated, enabling a direct comparison of the prediction with expert annotation in a frame-by-frame fashion. Model training, validation, and testing were performed on a dataset comprising ultrasound videos collected from 100 subjects, which were partitioned into 80, ten, and ten subjects, respectively. The proposed model was compared with U-Net, U-Net++, Siam U-Net, Attention U-Net, LSTM U-Net, and Trans U-Net. The proposed transformer-based model effectively leveraged the temporal and spatial information present in ultrasound video frames and efficiently segmented the median nerve with an average dice similarity coefficient (DSC) of approximately 94% at the wrist and 84% in the entire forearm region.**



***Index Terms*—Cross-sectional area (CSA), median nerve segmentation, ultrasound video, vision transformer (ViT).**

Manuscript received 17 October 2023; accepted 2 November 2023. Date of publication 6 November 2023; date of current version 11 January 2024. This work was supported in part by the General Electric (GE) Healthcare M.Tech. Fellowship, in part by C. V. Raman Postdoctoral Fellowship of IISc, and in part by the Wipro-GE Collaborative Laboratory on Artificial Intelligence in Health Care and Medical Imaging. *(Corresponding author: Phaneendra K. Yalavarthy.)*

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by Aster-CMI Hospital, Bangalore, India, under Application No. Aster/IEC/049/2020-21, Dated June 27, 2020.

Karan R. Gujarati, Vaddadi Venkatesh, Raji Susan Mathew, and Phaneendra K. Yalavarthy are with the Department of Computational and Data Sciences, Indian Institute of Science, Bangalore 560012, India (e-mail: karang@iisc.ac.in; venkateshvad@iisc.ac.in; rajisusanm@iisc.ac.in; yalavarthy@iisc.ac.in).

Lokesh Bathala is with Aster CMI Hospital, Hebbal, Bangalore 560092, India (e-mail: drlokesh.b@asterhospital.com).

Digital Object Identifier 10.1109/TUFFC.2023.3330539

I. Introduction

THE median nerve is a major peripheral nerve that serves as a critical communication pathway between the hand and central nervous system. It originates in the arm and acts as a major channel for motor and sensory transmission between the brain and the upper limb. This vital nerve supplies various muscles, allowing forearm and hand movements such as wrist and finger flexion. Furthermore, it transports sensory information from the skin of the palm and fingers to the central nervous system, thereby providing critical feedback for touch, temperature, and pain perception. Owing to its central role in coordinating intricate hand functions and sensory perception, any damage or impairment to the median nerve can result in significant motor deficits and sensory loss.

Highlights

- **The visual transformer architecture was modified for automated segmentation of the median nerve in ultrasound videos of wrist-to-elbow region.**
- **The proposed model used both temporal and spatial information of ultrasound video frames resulting in a dice similarity coefficient (DSC) of approximately 84% in the entire forearm region.**
- **The proposed model provides accurate segmentation of the median nerve and computes the cross-sectional area in every frame in an automated fashion leading to improved quantitative measurements.**

Peripheral nerve block using high-frequency alternating currents as a therapeutic alternative requires precise nerve diameter/cross-sectional area (CSA) to determine the minimum frequency required to produce nerve block [1]. Segmenting the median nerve is essential for diagnosing carpal tunnel syndrome (CTS) [2], [3], [4], guiding surgical planning and interventions such as ultrasound-guided regional anesthesia (UGRA) [5], [6], identifying nerve entrapment syndromes [7], and furthering our understanding of nerve anatomy [8], [9]. The accurate segmentation of the median nerve in medical imaging facilitates improved patient care, precise surgical procedures, and advancements in medical research. For example, regional anesthesia is an effective alternative to general anesthesia in many surgical procedures. The traditional approach is to blindly guide the needle to the target nerve. However, blind needle guidance can lead to nerve injury and local anesthetic toxicity in cases of erroneous needle placement [10]. In this context, the UGRA technique is becoming increasingly popular because of its ability to provide real-time visualization of nerves, needle advancement, and local anesthetic dissemination. Failure in accurately localizing the nerve can lead to nerve trauma or local anesthetic toxicity.

Several medical image analysis tools have been developed to assist radiologists with median nerve segmentation from US images. An automated median nerve localization method was proposed in [11], particularly for UGRA. This method relies on a machine learning technique that uses despeckling filtering, feature extraction, and selection, followed by pixel-wise classification based on a support vector machine with a Gaussian kernel. Similarly, a computer-aided machine learning algorithm for median nerve localization was introduced in [12]. However, feature extraction and feature selection in these methods are purely handcrafted, and the accuracy of the predicted segmented image depends on the selection of the features.

Recently, the advent of deep learning methods has shown huge potential for medical US analysis and has been applied to nerve segmentation as well. To segment the nerve region, a deep learning model using a convolutional neural network (CNN) and spatiotemporal consistency was proposed in [13]. Another deep-learning-based method that uses similarity measures to track the median nerve was introduced in [14] and was evaluated for tracking a set of predefined continuous wrist motions. An automated median nerve segmentation framework based on the U-Net-based encoder–decoder architecture [15] was proposed in [16]. Another U-Net-based approach called DeepNerve that uses the features of both MaskTrack and convolutional long short-term memory (LSTM) for median nerve segmentation was proposed in [17]. A comparative study on the performances of various pretrained CNN-based architectures including DeepLabV3+ [18], U-Net [15], feature pyramid network (FPN) [19], and Mask R-CNN [20] for median nerve localization/segmentation task was performed in [21]. A Mask R-CNN-based approach [20] with two additional transposed layers was used in [22] to segment the median nerve and predict the CSA.

The methods discussed to date are exclusively for segmenting the median nerve at the wrist, where the nerve is comparatively easy to localize. Typically, the UNet encoder–decoder architecture performs well at the wrist, and the efficiency decreases considerably as one moves from the wrist to the elbow region. When the nerve is away from the carpal tunnel, its depth is typically larger, and the shape of the nerve is different. Furthermore, UNet-based approaches do not consider the temporal relationships. To overcome this limitation, LSTMs are introduced, wherein a specific number of input frames are used to learn the temporal relationships. Although LSTMs are designed to capture long-range dependencies in sequential data, they may still struggle to capture extremely long dependencies or complex spatial relationships in nerve segmentation tasks.

Unlike computer vision tasks, where the background and receiver positions are fixed, the object of interest changes across frames in the US video, along with the background and receiver positions. This means that unlike object tracking, the field of view changes across frames for median nerve segmentation. The video mosaicing method has been shown to yield better performance by providing larger field-of-view mosaic images in medical imaging applications, such as probe-based confocal laser endomicroscopy (pCLE) [23]. However, the accurate alignment of frames in video mosaicing can be difficult, especially in cases where there is a possibility of movement of the ultrasound probe or the tissue being imaged. These artifacts can degrade the quality of the mosaic image and reduce the diagnostic accuracy of deep learning methods. Mosaicing also requires significant additional computation, limiting near-real-time clinical inferencing. Second, speckle noise is a natural part of ultrasound imaging that reduces the image resolution and contrast. The SNR of ultrasound videos is usually lower than that of general computer vision videos [24] and depends on the imaging parameters. Third, annotating US videos is a difficult task that requires an experienced clinician, thus limiting the amount of labeled data to serve as the ground truth.

Given these challenges with automated object segmentation in ultrasound videos, a vision transformer (ViT)-based model can serve as a better solution that can work with a smaller amount of data while considering temporal long-range dependencies in an efficient manner. The success of transformer models is mainly attributed to the improved self-attention mechanism arising from their capability to model long-range dependencies [25]. Transformer-based models have also been applied in a wide range of medical imaging applications. In this regard, the Trans U-Net model [26] is designed for medical image segmentation harnesses transformer technology. In Trans U-Net, 12 transformer layers were introduced into the encoder of the UNet model. This approach effectively mitigated the limitations associated with conventional convolutional operations [26]. Furthermore, models such as detection transformers (DETRs) have shown promise in object detection [27] and have been adapted for the detection of polyps in the colon, as reported for convolution in transformer networks (COTRs) [28]. In their approach, convolutional layers were inserted into the transformer encoder for high-level image feature reconstruction and faster convergence. Another approach that combines CNN and transformer nets, called TR-Net, to detect significant stenosis was proposed by Ma et al. [29]. In this approach, a shallow 3D-CNN was used to extract the local semantic features of coronary regions, and transformer encoders were used to learn correlations between different regions of the local stenosis of a coronary artery, which aids in the accurate detection of stenosis by aggregating information from local semantic features and global semantic features. A context-aware hybrid transformer called CT-CAD [30] was proposed for the detection of chest abnormalities in X-ray images. Tao and Zheng [31] introduced a spine-transformer model designed to tackle the automatic detection and localization of vertebrae in spine CT scans with arbitrary field of view. They framed the detection task as a problem of predicting one-to-one sets. Li et al. [32] proposed the MultiIB-transformer for the segmentation of CT and ultrasound images, which consisted of a single transformer layer and multiple information bottleneck (IB) blocks. They used a deep learning structure that requires only one transformer layer, thereby significantly reducing the number of model parameters without compromising performance. The local feature information from CNNs and global context information from transformers were leveraged in the local and context-attention adaptive network (LCA-Net) for thyroid nodule segmentation in ultrasound images [33]. To simultaneously maintain sufficient global information and local details, a hybrid CNN-transformer network (HCTNet) was proposed for breast ultrasound image segmentation in [34]. Detailed reviews on the usage of transformers for medical image analysis have been presented in [35], [36], and [37].

Since their inception, ViT models have consistently advanced the forefront of various vision tasks, including image classification [38], object detection [39], semantic segmentation [40], image colorization [41], low-level vision [42], and video understanding [35], [43]. Typically, ViT has resulted in notable improvements in semantic segmentation by enabling fine-grained pixelwise interpretation of scenes and objects [40]. Compared with conventional CNNs, ViT prediction errors align more closely with human perceptual judgments, making this one of the most exciting features of ViT. Recent studies have emphasized this unique characteristic, as demonstrated by Naseer et al. [44]. ViT demonstrates a remarkable capacity to comprehend and model intricate visual relationships, allowing them to make predictions consistent with human visual reasoning. This aligns with the growing consensus that ViT, with its self-attention mechanisms and global context modeling, holds great promise in bridging the gap between artificial vision systems and human perception.

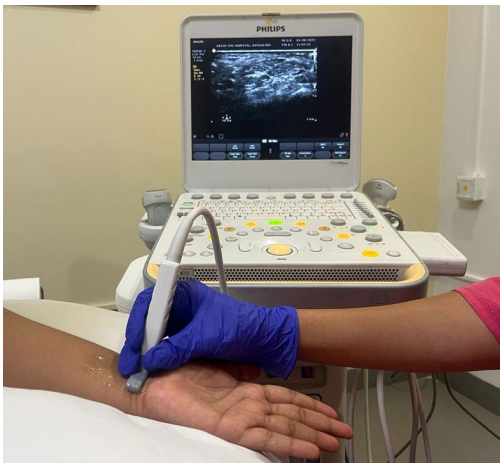


Fig. 1. Image acquisition: the probe was positioned at the wrist with the subjects seated facing the sonographer with the forearm and fingers extended in a comfortable position.

This study aims to develop an automated tool that can assist sonographers in localizing and segmenting the median nerve at the wrist, mid-forearm, and elbow in ultrasound images, which can improve the aforementioned tasks. This work proposes, for the first time, the utilization of a modified version of video instance segmentation using transformer (VisTR) architecture for median nerve segmentation from wrist-to-elbow with ultrasound video as input. Specifically, the contributions of this study are as follows.

1) A transformer-based model has been proposed for the first time for fully automated segmentation of the median nerve with ultrasound video as input providing a throughput of 76 frames in a second.
2) Modified the architecture of the visual transformer originally proposed for video instance segmentation to predict the median nerve in every frame of ultrasound videos. This is achieved by modifying the bounding box sequence matching block of the visual transformer. Since the median nerve segmentation is a binary class prediction, the entire bipartite matching sequence is eliminated, enabling a direct comparison of the prediction with expert annotations in a frame-by-frame fashion.
3) Evaluation of CSA measurement based on the median nerve (at the wrist and from wrist-to-elbow) section segmented by the proposed method in comparison to the manual tracing of nerve boundary performed by expert sonographers.

II. Methods

A. Data Acquisition and Preparation

The dataset was obtained in a clinical setting at Aster-CMI Hospital in Bangalore, India. The Philips CX50 ultrasound machine was used to acquire the dataset. A Philips L15-7io hockey stick transducer probe with a frequency range of 7–15 MHz was used. The subjects were seated facing the examiner with the forearm and the fingers extended in a comfortable position as shown in Fig. 1. The dataset consisted of acquisitions obtained at 3-cm depth from the skin. The image sequence for each subject was saved as a video of 8 s. The dimension of each acquired image was 800 × 600 pixels. In this study, a dataset consisting of 100 subjects was used, and all the images were annotated by an expert sonographer. Approval of all the ethical and experimental procedures and protocols for this research work involving human subjects was granted by Aster-CMI Hospital, Bangalore, India, under Approval No. Aster/IEC/049/2020-21, Dated June 27, 2020. Written informed consent was obtained from all the human subjects. The male-to-female ratio was 1:3. The age range of the patients was 35–65 years. The current study focused on the normal anatomy of the median nerve from the wrist-to-elbow region. As the anatomy of the tissues is altered post-surgery, subjects who underwent any major nerve surgical procedures were excluded from this study.

The sequence of frames was extracted from each video. Data de-identification was performed on all the images used in this study by removing all the personal information that could identify an individual from medical images to ensure patient privacy. Each image was cropped to a size of 448 × 336. This was done prior to any data processing carried out related to the study. For generating the training, validation, and test datasets, annotations were generated by expert sonographers using ImageJ [45], a free image analysis software made available by the National Institutes of Health (NIH), USA. The resultant annotated images were saved as binary images. The annotations included the pixelwise segmentation binary images for training the segmentation module. The dataset with expert annotation (100 videos) was then partitioned into training, validation, and test datasets, with 80, ten, and ten subjects. respectively. Apart from this, the data from 30 additional subjects were acquired for the clinical evaluation consisting of normal wrist, mild, and severe CTS, with ten subjects in each class. To increase the size of the training datasets, the available data were augmented using horizontal flips. Using this data augmentation process, the total training datasets were doubled and a total of 46 868 images were used for training. The validation and testing were performed on 3000 images each.

B. Deep Learning Model: VisTR

The architecture of VisTR [46] is a modified version of ViT to process the video sequences. The VisTR uses an architecture that consists of five components and treats the job as a parallel series decoding problem. The five blocks that constitute the VisTR architecture are the backbone, encoder, decoder, bounding box sequence matching, and sequence segmentation. Each block is explained briefly here.

1) Backbone: The CNN backbone extracts high-level feature maps for T input frames, each with a size of $H_0 \times W_0$ (i.e., $x_{\text{clip}} \in \mathbb{R}^{T\times 3\times H_0\times W_0}$). Afterward, the feature maps ($\in \mathbb{R}^{C\times(H_0/32)\times(W_0/32)}$) (with Res-Net backbone) have been concatenated to generate temporal feature maps $t_0 \in \mathbb{R}^{T\times C\times(H_0/32)\times(W_0/32)}$.

2) Encoding Using Transformer: A 1 × 1 convolution was used on the t_0 feature maps to reduce the dimensionality of the feature space by linearly transforming them into a new feature space $t_1 \in \mathbb{R}^{T\times d\times(H_0/32)\times(W_0/32)}$ with a lower dimension, $d < C$. Since the transformer model processes tokenized vectors, this 4-D feature map tensor was flattened into $N = T\cdot(H_0/32)\cdot(W_0/32)$ vectors, each with a dimension of d, and fed into the transformer encoder. The transformer encoder comprises a multihead self-attention (MHSA) and feedforward network (FFN) network. The encoder consists of the positional encoder for encoding the spatial and temporal space. The transformer architecture does not explicitly consider the order of elements in the input sequence (i.e., permutation-invariant). However, precise position information is required to accurately segment objects in an image or video in the segmentation task. To address this issue, the VisTR model introduces spatial and temporal positional encoding to provide the transformer model with information about each element's spatial and temporal positions in the input sequence. This allows the model to effectively capture the spatial and temporal relationships between objects and their surroundings, resulting in accurate instance segmentation.

3) Decoding Using Transformer: The decoder takes the T number of query embedding vectors having d dimensions ($I \in \mathbb{R}^{T\times d}$) along with the output of the transformer encoder E and generates sequence prediction O($\in \mathbb{R}^{T\times d}$) of N feature vectors that preserve the original order of the input frames.

4) Bounding Box Sequence Matching: To obtain the bounding box of median nerve, a three-layer FFN was applied to the sequence prediction O. This allows us to get the normalized center coordinates, width, and height of the box with respect to the image. A linear projection layer was with a softmax activation function to determine the class labels.

5) Segmentation Head for Sequences: This module aims to generate a series of mask predictions for median nerve. This module uses transformer decoder output (i.e., sequence prediction) (O) and encoder output (E) as input to generate similarity maps for each frame. To simplify, the module only calculates each frame's similarity maps individually, as temporal information is already available in decoder output O. The self-attention block processes the prediction features (O) and encoded features (E) to produce attention maps (A) for each frame.

The encoded features were reshaped into $\mathbb{R}^{T\times d\times(H_0/32)\times(W_0/32)}$. These encoded features (E) and decoded features (O) are used to generate the mask attention map A ($\in \mathbb{R}^{T\times n\times(H_0/32)\times(W_0/32)}$) of n feature maps. To generate the feature maps of the mask, the module actively combines the attention maps (A) with the encoded features (E) and backbone features (B) (i.e., with ResNet-101 or 50 as backbone ($\mathbb{R}^{T\times 256\times(H_0/4)\times(W_0/4)}$, $\mathbb{R}^{T\times 512\times(H_0/8)\times(W_0/8)}$, $\mathbb{R}^{T\times 1024\times(H_0/16)\times(W_0/16)}$, $\mathbb{R}^{T\times 2048\times(H_0/32)\times(W_0/32)}$) of the corresponding frames. A deformable convolution [47]

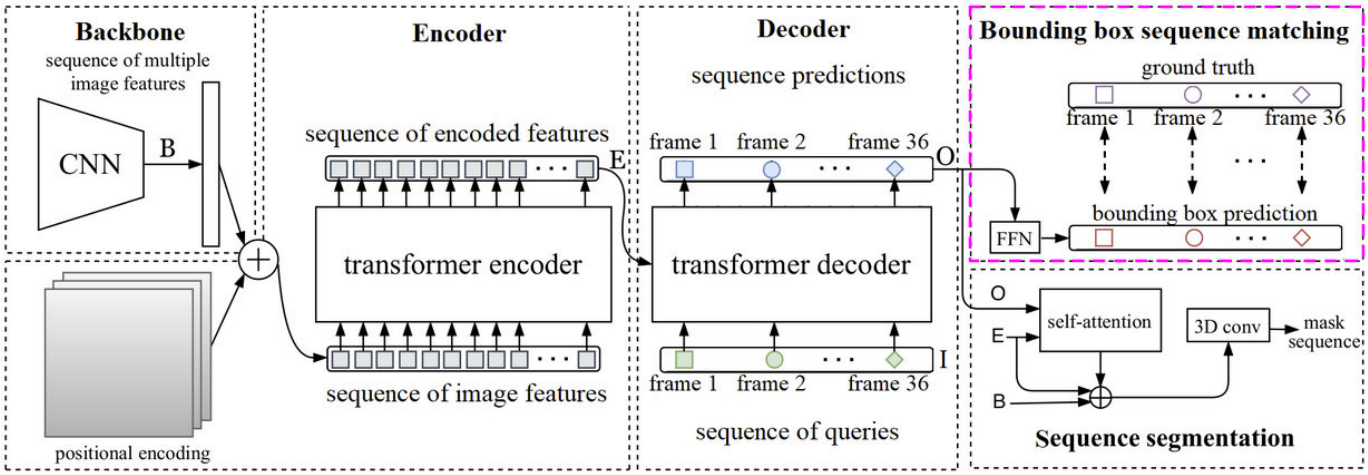


Fig. 2. VisTR architecture: 1) CNN backbone (here, ResNet-101) to extract the spatial feature maps of clip frame by frame. 2) Transformer encoder decoder blocks to extract the high-level spatial + temporal feature maps by combining all individual feature maps of the backbone. 3) Bounding box coordinates' extraction and matching to detect the median nerve in the frame. 4) Segmentation head to precisely detect the boundary of the median nerve in the frame. The magenta box indicates the modified block. All other blocks were the same as in the original implementation of VisTR [46].

operation was applied to combine features and generate the sequence of mask ($\in \mathbb{R}^{T\times k\times(H_0/4)\times(W_0/4)}$) of k feature maps. To propagate information about mask features across different frames, a 3-D convolution operation was used. Three 3-D convolution layers, group normalization layers [48], and a rectified linear unit (ReLU) activation function were used to generate the output series of mask $\in \mathbb{R}^{T\times 1\times(H_0/4)\times(W_0/4)}$. Finally, a bilinear interpolation was used to get the mask with size the same as the input image [i.e., (H0, W0)], mask $\in \mathbb{R}^{T\times 1\times H_0\times W_0}$.

6) Loss Functions for Models: The ground truth in T frames can be given by

$$y = \{(k, k\ldots, k), (\text{box}_0, \text{box}_1\ldots, \text{box}_T)\} \tag{1}$$

where k is the class label of the target for this prediction, and $\text{box}_t = (x_{\min}, y_{\min}, \text{width}, \text{height})$ is normalized box coordinates and its width and height with respect to the tth frame. So, the probability of prediction being class k is given by

$$\hat{p}(k) = \{\hat{p}_0(k), \hat{p}_1(k)\ldots, \hat{p}_T(k)\} \tag{2}$$

and bounding box prediction sequence is given by

$$\widehat{\text{box}} = \{\widehat{\text{box}}_0, \widehat{\text{box}}_1\ldots, \widehat{\text{box}}_T\}. \tag{3}$$

The loss function is a combination of a negative log-likelihood for predicting the class, as well as separate losses for the bounding box and mask prediction for the sequence prediction. The overall loss is a linear combination of these losses

$$\mathcal{L}(y, \hat{y}) = -\log \hat{p}(k) + \mathcal{L}_{\text{box}}(\text{box}, \widehat{\text{box}}) + \mathcal{L}_{\text{mask}}(m, \hat{m}) \tag{4}$$

where $\mathcal{L}_{\text{box}}$ is a linear combination of generalized intersection over union (GIoU) loss and L1 loss given as

$$\mathcal{L}_{\text{box}}(\text{box}, \widehat{\text{box}}) = \frac{1}{T}\sum_{t=1}^{T}\Big[\eta_{\text{giou}}\cdot\mathcal{L}_{\text{giou}}(\text{box}_t, \widehat{\text{box}}_t) + \eta_{L1}\Big|\Big|\text{box}_t - \widehat{\text{box}}_t\Big|\Big|_1\Big] \tag{5}$$

and $\mathcal{L}_{\text{mask}}$ is a linear combination of focal loss and dice loss given as

$$\mathcal{L}_{\text{mask}}(m, \hat{m}) = \eta_{\text{mask}}\frac{1}{T}\sum_{t=1}^{T}\Big[\mathcal{L}_{\text{Dice}}(m_t, \hat{m}_t) + \mathcal{L}_{\text{Focal}}(m_t, \hat{m}_t)\Big]. \tag{6}$$

7) Architectural Modifications to VisTR: In this study, the visual transformer architecture, which was originally designed to detect and classify 41 different classes in the YouTube VIS dataset, was modified for median nerve segmentation. The modification aims to adapt the visual transformer to a more specialized task, that is, segmentation of the median nerve presence in each frame of an ultrasound video. To achieve this, a critical component of the visual transformer architecture, known as the bounding box sequence matching block, was modified. This block, in its original form, was designed for tasks involving object detection and classification, relying on bipartite matching to establish correspondence between the predicted bounding boxes and ground-truth objects. As median nerve segmentation corresponds to binary classification (presence or absence of the median nerve in a given frame), the process is streamlined by eliminating the entire bipartite matching sequence, which was originally designed for handling multiple classes and complex object interactions. This simplification enabled direct frame-by-frame comparisons between the model's predictions and expert annotations, making the process more efficient and tailored to median nerve segmentation in ultrasound videos. These strategic adjustments to the visual transformer architecture enabled it to leverage its capabilities to focus exclusively on the binary classification problem of median nerve segmentation in each frame of the ultrasound video sequence, ultimately enhancing its performance and applicability. A schematic of the modified VisTR architecture is presented in Fig. 2.

C. Testing

Following end-to-end training, the architecture was subjected to classify each pixel of the test sample into two

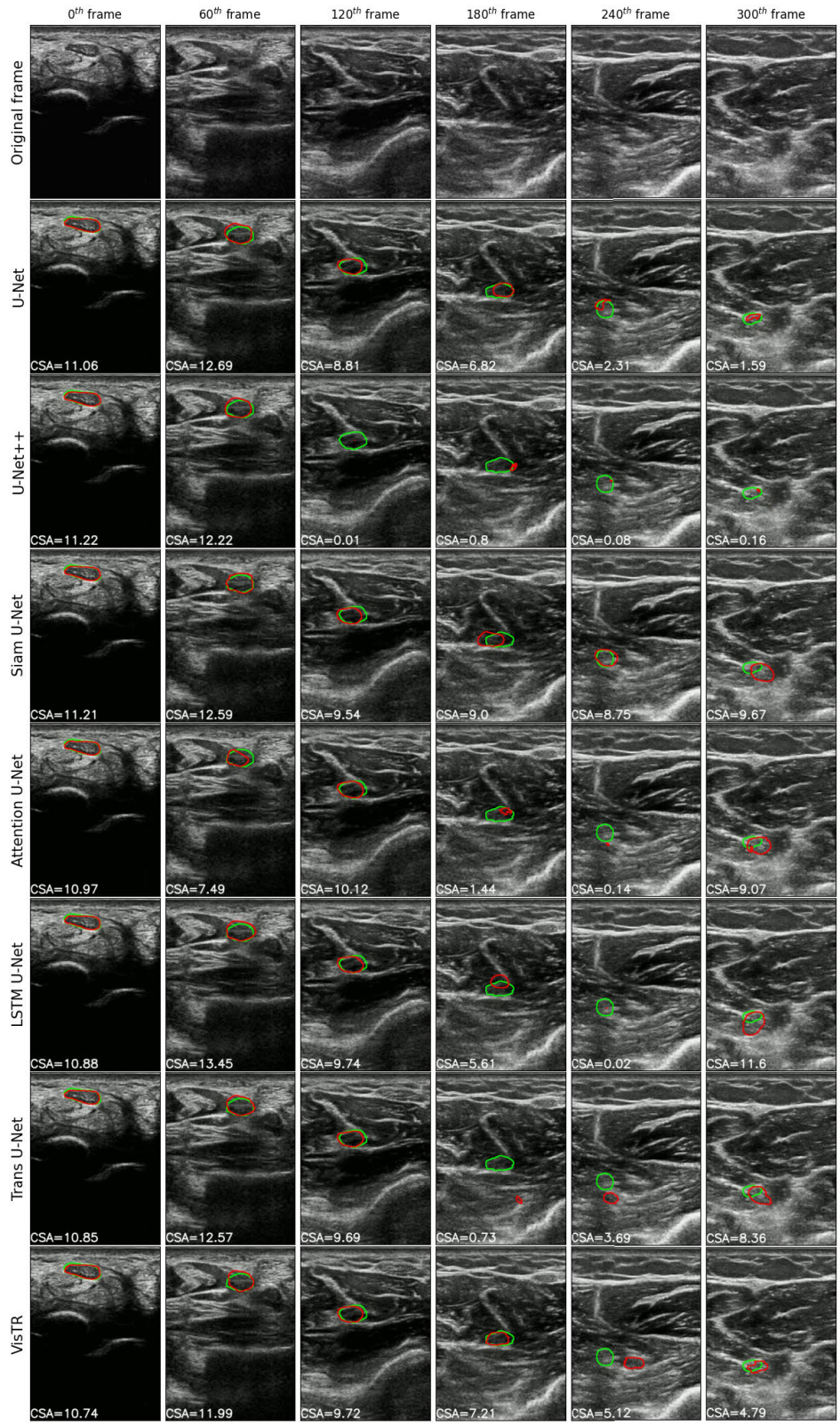


Fig. 3. Example segmentation of the median nerve using the methods discussed in this work for subject-1. The green contour indicates the expert annotation and the red contour indicates the result obtained for the corresponding method, as indicated in each row. The associated frame number is given on the top of every image (0 corresponds to the start of the wrist region and 300 corresponds to the elbow region), and the bottom of each frame has the corresponding computed CSA. This example corresponds to the lowest (minimum across test subjects) figure of merit (DSC) for the proposed VisTR (ResNet-101), which is 0.720.

categories: background and median nerve. A probability map was produced as output by the model with the same spatial dimension as the input. Each pixel was given a label based on its maximum probabilistic score across the two categories.

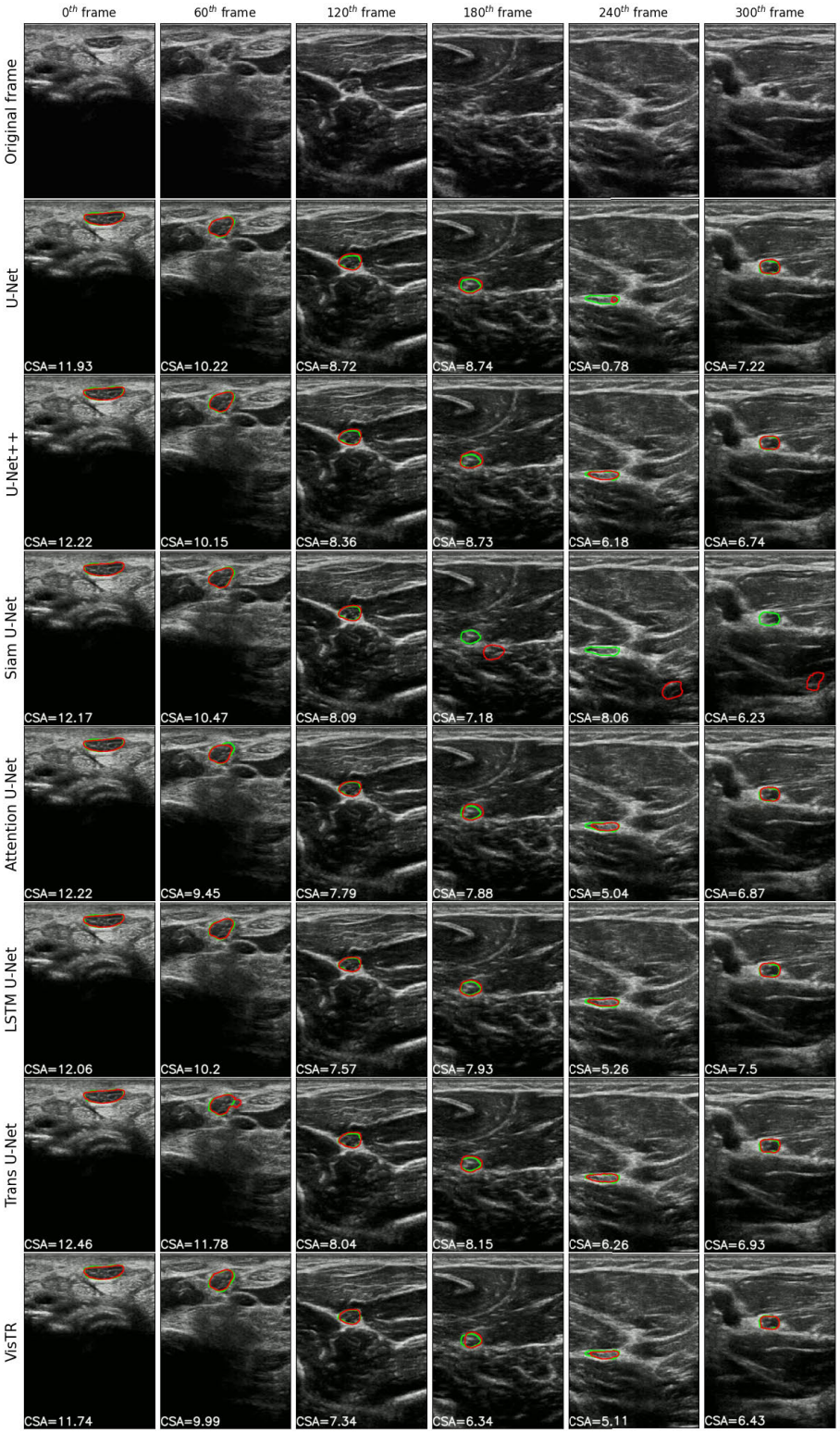


Fig. 4. Example segmentation of the median nerve using the methods discussed in this work for subject-8. The green contour indicates the expert annotation and the red contour indicates the result obtained for the corresponding method, as indicated in each row. The associated frame number is given on the top of every image (0 corresponds to the start of the wrist region and 300 corresponds to the elbow region), and the bottom of each frame has the corresponding computed CSA. This example corresponds to the best (maximum across test subjects) figure of merit (DSC) for the proposed VisTR (ResNet-101), which is 0.910.

Furthermore, the CSA of the median nerve was automatically computed by the proposed method based on the calibration of the dimensions of a single pixel, which is equal to 0.0043 mm^2 for the acquisition. Then the CSA can be computed as

$$\text{CSA} = \text{Single pixel area} \times \text{number of median nerve pixels}. \tag{7}$$

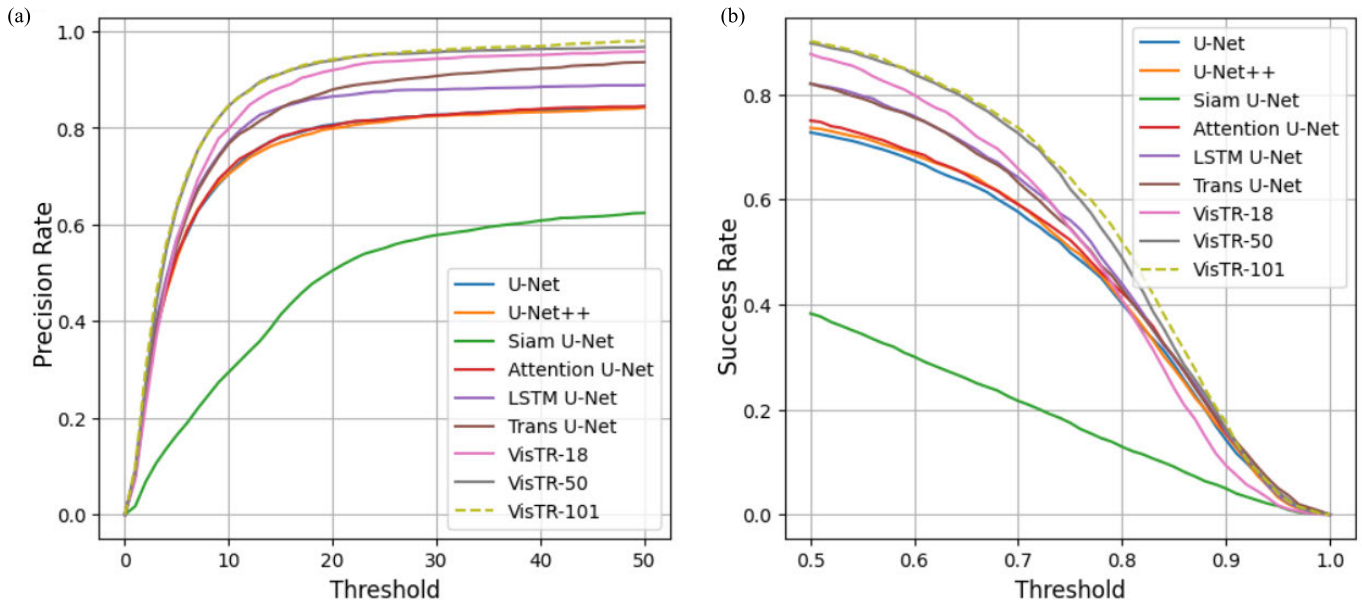


Fig. 5. Figures of merit for median nerve segmentation, (a) precision and (b) success rates, averaged over all the test subjects in the wrist-to-elbow region.

D. Evaluation

Given the expert annotations, the efficiency of the proposed deep learning model was assessed by several figures of merit including recall, precision, DSC, and Hausdorff distance (HD). Recall denotes the rate of positive samples correctly classified. The precision denotes the proportion of the classified pixels which are relevant. DSC was used to evaluate the similarity between expert annotation and predicted segmentation masks. The DSC can be computed as the ratio of shared true positive pixels to the sum of the total pixels in both the sets. A high DSC means a good match between expert annotation and predicted masks, while a low score indicates potential errors in the segmentation. HD measures the greatest distance from any point in one set to the closest point in the other set, providing a quantitative assessment of segmentation accuracy.

To evaluate the performance/identify the weaknesses of the different deep learning models and to assess the resistance to adversarial attacks, input perturbation experiments were also performed as discussed in Section III. These experiments also aid in model interpretability, revealing decision-making processes and reducing biases for improved reliability and performance. Two types of perturbation experiments are performed: noise perturbation and weight perturbation. To evaluate the robustness of the different models, speckle noise with different mean values was added to the test data, and the corresponding noisy data were tested using the models trained on the original data (noise perturbation study). Precision errors in the rounding can cause weight perturbations. When the model is sensitive to small weight changes, rounding errors can accumulate and affect its behavior and performance. To evaluate the robustness of the different models in the presence of weight perturbation, the model weights W were perturbed using different η values as

$$W_{\text{perturbed}} = (1 + \eta)W. \tag{8}$$

E. Implementation

All the models were trained using PyTorch [49] along with Adam optimizer [50]. For VisTR (ResNet-101 and ResNet-50 as backbone), pretrained DETR [27] models' weights from the COCO dataset were used. These pretrained weights [27] used a d value of 384. Alternatively, VisTR (ResNet-18) was trained from scratch with the same d value. This work used a Linux workstation with Intel i9 9900X CPU, 128-GB RAM, and two NVIDIA RTX $A6000$ GPUs with 48-GB memory each for all computations, including deep learning model training. For all the U-Net-based models, the logcosh dice loss function was adopted [51]. This specific loss function was selected based on extensive experimentation and analysis, which consistently demonstrated its superior performance across a spectrum of figure-of-merit metrics considered in the evaluation. The logcosh dice loss function, known for its smoothness and robustness, has emerged as the preferred choice owing to its ability to strike a balance between capturing fine-grained details and effectively handling outliers. In contrast, for the ViSTR model, a custom cost function was composed of a linear combination of several distinct components, including binary cross-entropy (BCE), GIoU, L1-loss, focal loss, and dice loss, each serving a specific purpose [46] as discussed in Section II-B. Consideration of BCE penalizes uncertain predictions, whereas GIoU and dice loss contribute to accurate object localization and segmentation. In addition, L1-loss and focal loss were integrated to mitigate model sensitivity to class imbalance and focus on challenging regions within the ultrasound video frames.

For the U-Net-based models, the initial learning rate was set to 0.001 with lr schedular "REDUCELRONPLATEAU" with a patience level of 5. For VisTR, the transformer's initial learning rate was set to 10^{-4} and the backbone learning rate was set to 10^{-5}. Both the learning rates were decreased by 0.1 every 12 epochs. Because pretrained weights

from the COCO dataset were used in the VisTR backbone, a low learning rate was used to facilitate the fine-tuning of the model to adapt to the median nerve segmentation task. This approach allowed the model to gradually adjust its parameters in response to the intricacies of the median nerve dataset, promoting smoother convergence and preventing abrupt weight updates that could destabilize the training process. Conversely, for all other models trained from scratch, a relatively high learning rate was used to initialize the models. Using a higher learning rate for these models, the early stages of training were expedited, such that the model established a fundamental understanding of the dataset and began to refine its features.

The number of epochs for each model was chosen as that corresponding to the minimum loss for the validation dataset. The specific parameters used for training of models are summarized in Table I. Since the same parameters were used for all the U-Net-based models (U-Net, U-Net++, Siam U-Net, Attention (Attn) U-Net, LSTM U-Net, and Trans U-Net), they are listed in the first row of Table I.

III. Results

A. Performance Comparison With Expert Annotation

The models used in this study, the corresponding parameter size, FLOPs/frame, training time, and processed frames/s are summarized in Table II. To evaluate the segmentation capability of the proposed algorithm, in the case where the expert annotation is available, different metrics described in Section II-D were deployed. Table III summarizes the recall, precision (abbreviated as Prec), and DSC averaged over the frames for ten subjects obtained using the models considered in this study. All the methods exhibited a relatively high degree of accuracy and effectiveness in segmenting the median nerve at the wrist region. This can be attributed to the relatively straightforward nerve's consistent course and the absence of obstructing tissues that might obscure its visibility. However, as one moves further away from the wrist toward the elbow, the performance of U-Net-based methods notably deteriorates. This is because the region toward the elbow introduces substantial variability in its course, unlike the wrist region. Moreover, the surrounding anatomical structures such as muscles and blood vessels can hinder its clear visualization. The VisTR model, with its improved architectural design and capacity to capture contextual information efficiently, demonstrated a superior ability to handle the complexities associated with localizing the median nerve in regions further from the wrist.

Individual subject results (averaged over all frames of that subject) for the proposed VisTR are provided in Table IV. Example segmented images obtained using the models discussed in this work are shown in Fig. 3 for subject-4 and Fig. 4 for subject-6. Computed CSA of the median nerve has been provided at the bottom of every image in these figures. From the results, it is evident that the proposed VisTR provides superior performance in segmenting the median nerve compared with the existing deep learning models in these ultrasound videos.

B. Performance Comparison With Less Training Data

To evaluate the performance of the different models with less training data, all the models were trained 25% data. The corresponding results are summarized in Table V. As evident from the table, the VisTR approach performed better than all other methods in the wrist-to-elbow region even with 25% training data.

C. Evaluation Across Frames

Although tracking is not performed in this study, the performance of the proposed VisTR method was evaluated temporally using two metrics: precision and success rate [57]. The precision metric provides the percentage of frames in which the distance between the centers of the expert annotated segmentation and the predicted result falls below a specified threshold. The threshold varied from 0 to 50 pixels. The success rate metric measures the percentage of frames in which the intersection over union (IoU) ratio between the expert annotated median nerve and predicted median nerve exceeds a specified threshold. The threshold ranged from 0.5 to 1. Plots of precision and success rates averaged over all the test subjects in the wrist-to-elbow region are shown in Fig. 5. The proposed VisTR was found to show better performance in terms of these temporal metrics in all the ultrasound video frames.

D. Perturbation Study

1) Noise Perturbation Study: The performance of the different models considered in this study on input perturbation (speckle noise mean) is summarized in Table VI. The better performance of U-Net-based models under noisy regimes is attributed to better regularization capabilities of the model. Specifically, it comprises an encoder and a decoder, a contracting-expansive pathway. The encoder extracts high-level features and spatial hierarchies from the input image, while the decoder reconstructs the segmentation map with finer details. This architecture ensures that the model is regularized, preventing it from learning overly complex features that may be specific to noise. Moreover, U-Net incorporates skip connections that facilitate the direct transfer of high-resolution information from the contracting (encoding) path to the expansive (decoding) path. This feature helps retain fine details and spatial information, which can be crucial for denoising noisy images. U-Net also uses a relatively simple architecture compared with the complex attention mechanisms of VisTR. In scenarios with high noise, this simplicity can be an advantage as it reduces the risk of overfitting and allows U-Net to focus on essential features.

2) Weight Perturbation Study: Table VII summarizes the performance averaged over ten subjects' data for different models with weight perturbation. It is evident that VisTR was less sensitive to rounding errors.

E. Clinical Evaluation

CSA is the most reliable imaging parameter for diagnosing carpal-tunnel-syndrome (CTS), inflammation, and edema in nerves, tendons, and ligaments. For severe CTS diagnosis,

TABLE I
SUMMARY OF TRAINING PARAMETERS FOR THE MODELS USED IN THIS WORK

Model	Loss function	Optimizer	# Epochs	Initial l_r	l_r update rule
U-Net based models	Logcosh Dice Loss	Adam	50	0.001	Reduce LR On Plateau
VisTR	BCE+L1+GIoU+Focal+Dice Loss	AdamW [52]	50	0.0001	Decreased by 0.1 every 12 epochs

TABLE II
SUMMARY OF MODELS USED IN THIS WORK. THE LAST COLUMN PROVIDES THE NUMBER OF FRAMES PROCESSED IN A SECOND

Model	Parameters	FLOPs/ frame	Training time (hrs/epoch)	Processed frames per second
U-Net [15]	17.27M	91.95G	0.24	117
U-Net++ [53]	36.63M	317.04G	1.05	37
Siam U-Net [54]	31.05M	113.71G	0.23	77
Attention U-Net [55]	34.88M	152.85G	0.33	77
LSTM U-Net [56]	45.78M	274.25G	1.12	32
Trans U-Net [26]	103.58M	73.82G	0.38	66
VisTR (ResNet-18) [46]	43.75M	7.60G	3.47	197
VisTR (ResNet-50) [46]	56.75M	15.00G	4.83	146
VisTR (ResNet-101) [46]	75.71M	26.15G	6.30	121

TABLE III
AVERAGED FIGURES OF MERIT OBTAINED FOR TEN SUBJECTS' TEST DATA USING THE DISCUSSED MODELS FOR THE WRIST REGION AND FOR THE REGION FROM WRIST-TO-ELBOW USING THE FULL TRAINING DATA. THE BEST PERFORMING METHODS ARE SHOWN IN BOLD. THE SUMMARY OF THE MODELS IS PROVIDED IN TABLE II

Models	Wrist Region				Wrist-to-Elbow Region			
	Recall	Prec	DSC	HD	Recall	Prec	DSC	HD
U-Net [15]	0.936	0.950	0.939	6.010	0.610	0.774	0.709	17.273
U-Net++ [53]	0.943	0.953	0.945	4.981	0.687	0.759	0.702	16.241
Siam U-Net [54]	**0.956**	0.950	**0.952**	**4.142**	0.558	0.562	0.555	47.282
Attn U-Net [55]	0.944	0.955	0.947	4.895	0.691	0.767	0.708	15.828
LSTM U-Net [56]	0.951	0.954	0.951	4.670	0.748	0.799	0.759	12.768
Trans U-Net [26]	0.952	0.948	0.949	4.974	0.773	0.823	0.783	19.318
VisTR (ResNet-18) [46]	0.911	**0.969**	0.938	5.426	0.791	0.861	0.814	15.543
VisTR (ResNet-50) [46]	0.943	0.960	0.950	4.408	**0.835**	0.850	0.834	13.403
VisTR (ResNet-101) [46]	0.939	0.962	0.949	4.396	0.833	**0.866**	**0.841**	**12.592**

TABLE IV
SUBJECTWISE SUMMARY OF FIGURES OF MERIT FOR THE PROPOSED VISTR (RESNET-101) MODEL. THE AVERAGED RESULTS ACROSS ALL TEN SUBJECTS ARE GIVEN IN THE LAST COLUMN OF TABLE III

Subject	Wrist Region				Wrist-to-Elbow Region			
	Recall	Prec	DSC	HD	Recall	Prec	DSC	HD
1	0.925	0.970	0.947	4.83	0.723	0.748	0.720	25.915
2	0.975	0.964	0.970	3.15	0.822	0.938	0.870	10.722
3	0.893	0.993	0.940	5.125	0.850	0.928	0.881	9.984
4	0.957	0.937	0.946	4.311	0.789	0.847	0.808	15.568
5	0.956	0.953	0.954	2.728	0.756	0.791	0.764	16.367
6	0.958	0.962	0.959	3.062	0.902	0.906	0.901	6.843
7	0.866	0.959	0.909	8.329	0.900	0.879	0.885	8.632
8	0.937	0.977	0.957	5.702	0.913	0.914	0.910	6.967
9	0.962	0.941	0.951	2.834	0.798	0.872	0.826	15.049
10	0.962	0.966	0.964	3.915	0.876	0.835	0.846	9.878

cutoff in terms of CSA is $> 12\text{mm}^2$ [58]. To evaluate the performance of the model in computing the CSA of the nerve, experiments were conducted on normal subjects and patients with mild and severe CTS, and the corresponding figure of merit are summarized in Table VIII.

F. Ablation Study

The ablation study of loss terms in deep learning models involves systematically removing or modifying specific loss terms used during training to understand their impact on the model's performance. This study helps analyze the contribution of individual loss terms toward the overall optimization process and the resulting model's behavior. As discussed earlier, this work used five different losses, BCE for classification, L1 and GIoU for the bounding box, and focal and dice loss for the segmentation task. To evaluate the impact of the losses for a given task, each one is eliminated while keeping all other loss terms unaffected. From Table IX, it can be seen that combining losses for the bounding box and segmentation increases the overall DSC. By combining the focal loss and dice loss, one can leverage the focal loss's ability to handle class imbalance and the dice loss's ability to evaluate segmentation similarity. The dice loss penalizes dissimilarities between the predicted and expert annotation sets more heavily, making it suitable for tasks where precise segmentation boundaries are essential. Thus removing dice loss from the equation, a heavy drop in DSC and other metrics was observed, as the shape of the median nerve is an essential factor in segmentation.

TABLE V
AVERAGED FIGURES OF MERIT OBTAINED USING DIFFERENT MODELS (TRAINED ON 25% TRAINING DATA) FOR THE WRIST REGION AND FOR THE REGION FROM WRIST-TO-ELBOW. THE BEST PERFORMING METHODS ARE SHOWN IN BOLD

Models	Wrist Region				Wrist-to-Elbow Region			
	Recall	Prec	DSC	HD	Recall	Prec	DSC	HD
U-Net [15]	**0.935**	0.954	0.941	6.860	0.557	0.651	0.572	22.783
U-Net++ [53]	0.915	0.962	0.932	6.440	0.550	0.660	0.571	21.327
Siam U-Net [54]	0.932	0.940	0.935	5.836	0.660	0.669	0.658	28.441
Attn U-Net [55]	0.930	0.951	0.937	8.099	0.568	0.646	0.579	25.186
LSTM U-Net [56]	0.932	**0.964**	**0.947**	**4.595**	0.659	0.730	0.670	**19.436**
Trans U-Net [26]	0.927	0.929	0.925	12.307	0.719	0.747	0.718	32.918
VisTR (ResNet-18) [46]	0.878	0.902	0.885	17.498	0.786	**0.792**	0.774	28.035
VisTR (ResNet-50) [46]	0.914	0.946	0.928	10.174	0.785	0.783	0.772	30.822
VisTR (ResNet-101) [46]	0.934	0.946	0.938	9.095	**0.808**	0.783	**0.783**	28.236

IV. DISCUSSION

The proposed VisTR architecture enables the model to extract detailed contextual information from an entire image using advanced attention mechanisms. VisTR excels in comprehending the global context of ultrasound images in contrast to conventional approaches that focus primarily on local features or handcrafted heuristics. This comprehensive awareness improves the ability to recognize fine nerve structures in complex anatomical backgrounds. The flexibility to adapt to various anatomical forms and structures is a distinguishing feature of VisTR. This versatility is crucial, particularly for nerve segmentation, because the median nerve manifests differently in various individuals and imaging scenarios. Many current techniques struggle to handle this unpredictability, necessitating extensive fine-tuning or specialized architectures for each scenario. In contrast, VisTR has self-attention mechanisms that enable it to dynamically adjust to different anatomical features without requiring a great deal of modification. This attention mechanism plays a pivotal role in enhancing the capacity of the model to

TABLE VI

PERFORMANCE METRICS OF DEEP LEARNING MODELS CONSIDERED IN THIS WORK AVERAGED OVER TEN SUBJECTS WITH INPUT PERTURBATION (THE FIRST COLUMN INDICATES THE MEAN OF SPECKLE NOISE). THE BEST PERFORMING METHODS ARE SHOWN IN BOLD

mean	Models	Wrist Region				Wrist-to-Elbow Region			
		Recall	Precision	DSC	HD	Recall	Precision	DSC	HD
0.10	U-Net	0.926 ± 0.0004	0.950 ± 0.0007	0.930 ± 0.0007	6.425 ± 0.1942	0.625 ± 0.0008	0.718 ± 0.0017	0.645 ± 0.0006	16.427 ± 0.2062
	U-Net++	0.946 ± 0.0005	0.953 ± 0.0005	0.948 ± 0.0005	4.884 ± 0.2108	0.635 ± 0.0006	0.719 ± 0.0019	0.653 ± 0.0005	15.422 ± 0.1756
	Siam U-Net	**0.950 ± 0.0003**	0.951 ± 0.0003	0.949 ± 0.0000	**4.354 ± 0.0207**	0.598 ± 0.0121	0.613 ± 0.0145	0.600 ± 0.0128	45.768 ± 3.8922
	Attn U-Net	0.939 ± 0.0005	0.958 ± 0.0006	0.944 ± 0.0005	5.549 ± 0.1426	0.621 ± 0.0007	0.708 ± 0.0015	0.641 ± 0.0004	15.434 ± 0.1986
	LSTM U-Net	0.944 ± 0.0004	0.962 ± 0.0004	**0.951 ± 0.0003**	4.569 ± 0.0755	0.697 ± 0.0005	0.770 ± 0.0008	0.718 ± 0.0007	**11.62 ± 0.0762**
	Trans U-Net	0.913 ± 0.0017	0.929 ± 0.0015	0.918 ± 0.0015	10.762 ± 0.4148	0.762 ± 0.0009	0.818 ± 0.0007	0.773 ± 0.0006	22.026 ± 0.3431
	VIsTR (ResNet-18)	0.849 ± 0.0030	0.946 ± 0.0017	0.891 ± 0.002	16.915 ± 0.5718	0.729 ± 0.0016	0.848 ± 0.0029	0.77 ± 0.0019	20.887 ± 0.3872
	VisTR (ResNet-50)	0.943 ± 0.0005	0.952 ± 0.0014	0.946 ± 0.0012	6.114 ± 0.1464	**0.805 ± 0.0013**	0.828 ± 0.0018	0.805 ± 0.0016	18.28 ± 0.3097
	VisTR (ResNet-101)	0.910 ± 0.0009	**0.967 ± 0.0004**	0.936 ± 0.0008	6.751 ± 0.6023	0.797 ± 0.0012	**0.863 ± 0.0012**	**0.818 ± 0.0010**	15.469 ± 0.1205
0.20	U-Net	0.886 ± 0.0009	0.936 ± 0.0033	0.895 ± 0.0009	8.731 ± 0.2037	0.536 ± 0.0008	0.654 ± 0.0014	0.562 ± 0.0012	17.52 ± 0.3188
	U-Net++	**0.937 ± 0.0005**	0.957 ± 0.0006	**0.944 ± 0.0005**	**4.806 ± 0.0887**	0.560 ± 0.0013	0.662 ± 0.0020	0.583 ± 0.0015	15.873 ± 0.1866
	Siam U-Net	0.934 ± 0.0007	0.956 ± 0.0005	0.943 ± 0.0003	6.012 ± 0.0805	0.579 ± 0.0365	0.626 ± 0.0415	0.594 ± 0.0383	38.33 ± 4.1474
	Attn U-Net	0.910 ± 0.0015	0.958 ± 0.0013	0.923 ± 0.0019	7.051 ± 0.225	0.535 ± 0.0012	0.636 ± 0.0024	0.559 ± 0.0010	15.719 ± 0.2572
	LSTM U-Net	0.929 ± 0.0006	**0.966 ± 0.0034**	**0.944 ± 0.0021**	8.563 ± 1.0679	0.608 ± 0.0005	0.719 ± 0.0012	0.642 ± 0.0008	**12.498 ± 0.2192**
	Trans U-Net	0.860 ± 0.0028	0.895 ± 0.0029	0.874 ± 0.0027	17.709 ± 0.7002	**0.748 ± 0.0015**	0.808 ± 0.0022	0.761 ± 0.0014	23.469 ± 0.4609
	VisTR (ResNet-18)	0.711 ± 0.0048	0.842 ± 0.0082	0.754 ± 0.0058	46.227 ± 1.8961	0.623 ± 0.0021	0.774 ± 0.003	0.672 ± 0.002	38.04 ± 0.5075
	VisTR (ResNet-50)	0.623 ± 0.0055	0.590 ± 0.0038	0.596 ± 0.0045	52.883 ± 0.6464	0.634 ± 0.0014	0.666 ± 0.0018	0.637 ± 0.0016	50.436 ± 0.2979
	VisTR (ResNet-101)	0.822 ± 0.0023	0.896 ± 0.0046	0.845 ± 0.0029	63.55 ± 2.5272	0.730 ± 0.0016	**0.834 ± 0.0016**	**0.764 ± 0.0014**	27.862 ± 0.5458
0.30	U-Net	0.770 ± 0.0017	0.882 ± 0.0024	0.804 ± 0.0016	10.882 ± 0.3955	0.411 ± 0.0012	0.545 ± 0.0025	0.442 ± 0.0012	20.33 ± 0.4054
	U-Net++	0.872 ± 0.0019	0.945 ± 0.0036	0.897 ± 0.0016	8.827 ± 0.2301	0.463 ± 0.0010	0.585 ± 0.0018	0.492 ± 0.0010	17.649 ± 0.2039
	Siam U-Net	**0.895 ± 0.0028**	**0.957 ± 0.0020**	**0.919 ± 0.0023**	9.261 ± 0.2814	0.408 ± 0.0596	0.477 ± 0.0715	0.433 ± 0.0637	61.583 ± 9.4385
	Attn U-Net	0.864 ± 0.0008	0.928 ± 0.0015	0.888 ± 0.0008	**6.309 ± 0.2767**	0.434 ± 0.0011	0.551 ± 0.0015	0.464 ± 0.0010	18.054 ± 0.2818
	LSTM U-Net	0.830 ± 0.0005	0.899 ± 0.0004	0.859 ± 0.0005	8.494 ± 0.5435	0.472 ± 0.0006	0.603 ± 0.0019	0.512 ± 0.0005	**13.63 ± 0.1873**
	Trans U-Net	0.838 ± 0.0029	0.891 ± 0.0014	0.860 ± 0.0023	21.014 ± 0.5784	**0.731 ± 0.0017**	**0.801 ± 0.0020**	0.748 ± 0.0017	25.878 ± 0.2121
	VisTR (ResNet-18)	0.397 ± 0.0092	0.530 ± 0.0173	0.436 ± 0.0108	89.507 ± 2.1207	0.463 ± 0.0027	0.617 ± 0.0063	0.511 ± 0.0035	67.204 ± 0.7939
	VisTR (ResNet-50)	0.314 ± 0.0048	0.295 ± 0.0038	0.301 ± 0.0043	88.159 ± 0.7271	0.364 ± 0.0017	0.395 ± 0.0021	0.370 ± 0.0018	96.27 ± 0.4716
	VisTR (ResNet-101)	0.645 ± 0.0041	0.739 ± 0.0038	0.677 ± 0.0037	127.954 ± 4.0446	0.725 ± 0.0015	0.787 ± 0.0018	**0.750 ± 0.0008**	58.286 ± 0.6931

TABLE VII

PERFORMANCE METRICS OF DEEP LEARNING MODELS CONSIDERED IN THIS WORK AVERAGED OVER TEN SUBJECTS WITH WEIGHT PERTURBATION (THE FIRST COLUMN INDICATES THE VALUE OF PERTURBATION). THE BEST PERFORMING METHODS ARE SHOWN IN BOLD

Values of η	Models	Wrist Region				Wrist-to-Elbow Region			
		Recall	Precision	DSC	HD	Recall	Precision	DSC	HD
-0.10	U-Net	0.0	0.0	0.0	inf	0.0	0.0	0.0	inf
	U-Net++	0.0	0.0	0.0	inf	0.0	0.0	0.0	inf
	Siam U-Net	0.0	0.0	0.0	inf	0.0	0.0	0.0	inf
	Attn U-Net	0.0	0.0	0.0	inf	0.0	0.0	0.0	inf
	LSTM U-Net	0.0	0.0	0.0	inf	0.0	0.0	0.0	inf
	Trans U-Net	0.255	0.925	0.366	30.380	0.168	0.584	0.230	28.806
	VisTR (ResNet-18)	**0.910**	**0.962**	**0.933**	**6.272**	**0.788**	**0.827**	**0.795**	**18.810**
	VisTR (ResNet-50)	0.622	0.801	0.672	52.918	0.335	0.404	0.350	143.871
	VisTR (ResNet-101)	0.584	0.610	0.593	69.845	0.644	0.665	0.641	43.503
-0.04	U-Net	0.600	0.774	0.648	20.877	0.264	0.369	0.289	17.090
	U-Net++	0.276	0.496	0.307	37.946	0.174	0.270	0.189	20.741
	Siam U-Net	0.775	0.955	0.831	19.346	0.398	0.503	0.431	39.078
	Attn U-Net	0.116	0.413	0.146	60.161	0.147	0.253	0.164	28.276
	LSTM U-Net	0.459	0.667	0.516	33.814	0.275	0.427	0.309	20.561
	Trans U-Net	0.932	0.960	0.943	5.208	0.681	0.836	0.725	16.234
	VisTR (ResNet-18)	0.914	**0.967**	0.939	5.421	0.800	**0.853**	0.815	15.927
	VisTR (ResNet-50)	**0.937**	0.964	**0.949**	**4.485**	0.815	0.831	0.813	15.191
	VisTR (ResNet-101)	0.933	0.964	0.947	4.512	**0.834**	0.839	**0.829**	**13.901**
0.04	U-Net	0.908	0.718	0.788	265.641	0.503	0.54	0.495	230.484
	U-Net++	0.904	0.955	0.926	39.760	0.512	0.655	0.549	119.189
	Siam U-Net	0.695	0.615	0.647	45.407	0.222	0.204	0.209	99.285
	Attn U-Net	0.947	0.850	0.892	128.966	0.585	0.604	0.576	167.586
	LSTM U-Net	0.943	0.933	0.935	46.373	0.643	0.726	0.664	74.606
	Trans U-Net	**0.952**	0.945	0.947	20.202	0.802	0.800	0.790	85.232
	VisTR (ResNet-18)	0.905	**0.971**	0.936	5.582	0.770	0.864	0.802	16.767
	VisTR (ResNet-50)	0.945	0.953	0.948	4.716	**0.818**	0.852	**0.825**	**14.381**
	VisTR (ResNet-101)	0.939	0.961	**0.949**	**4.705**	0.790	**0.871**	0.818	14.870
0.10	U-Net	0.525	0.817	0.618	252.733	0.205	0.276	0.223	286.583
	U-Net++	0.230	0.035	0.059	405.843	0.123	0.023	0.038	322.925
	Siam U-Net	0.049	0.097	0.056	398.316	0.034	0.041	0.033	322.031
	Attn U-Net	**0.994**	0.051	0.097	404.855	0.877	0.029	0.056	326.964
	LSTM U-Net	0.750	0.212	0.326	286.535	0.340	0.060	0.097	303.044
	Trans U-Net	0.938	0.864	0.895	250.472	**0.795**	0.597	0.666	278.430
	VisTR (ResNet-18)	0.883	**0.972**	**0.924**	**6.296**	0.673	**0.816**	**0.720**	**28.515**
	VisTR (ResNet-50)	0.888	0.918	0.900	19.549	0.663	0.746	0.689	93.503
	VisTR (ResNet-101)	0.626	0.813	0.689	72.275	0.371	0.560	0.419	105.826

effectively process input data. In particular, the attention mechanism empowers VisTR to focus on each distinct element present in incoming data. Unlike standard convolutional techniques, which may miss subtle relationships between distant pixels, this capability enables VisTR to establish intricate connections among all the input components. This comprehensive awareness is particularly valuable in scenarios in which objects of interest exhibit variable positions and scales within the images. Moreover, this attention mechanism excels in capturing long-range dependencies in data. This allows VisTR to recognize and leverage relationships that span considerable distances across images. This proves invaluable in image segmentation tasks where the context provided by distant regions can profoundly impact the

TABLE VIII

PERFORMANCE METRICS OF DIFFERENT MODELS AVERAGED OVER NORMAL SUBJECTS, MILD, AND SEVERE CTS PATIENTS AT THE WRIST REGION. EACH CLASS HAD TEN SUBJECTS. HERE, PRED CSA CORRESPONDS TO THE PREDICTED CROSS SECTION AREA OF THE MEDIAN NERVE, IN MM^2. THE BEST PERFORMING METHODS ARE SHOWN IN BOLD

Models	Normal Subjects (Expert Annotated CSA = 8.274 mm^2)					Mild CTS Patients (Expert Annotated CSA=10.423 mm^2)					Severe CTS Patients (Expert Annotated CSA=14.689 mm^2)				
	Recall	Prec	DSC	HD	Pred CSA	Recall	Prec	DSC	HD	Pred CSA	Recall	Prec	DSC	HD	Pred CSA
U-Net	0.997	0.922	0.875	**3.645**	9.158	0.993	0.944	0.9	10.697	11.119	0.962	0.988	0.925	6.417	14.215
U-Net++	0.993	0.928	0.879	10.749	9.031	0.98	0.957	0.9	4.038	10.845	**0.966**	0.989	**0.929**	**5.836**	14.213
Siam U-Net	**0.999**	0.899	0.863	6.144	9.457	**0.996**	0.922	0.885	4.67	11.514	0.901	0.898	0.848	16.479	13.925
Attn U-Net	0.997	0.923	0.876	12.604	9.183	0.992	0.95	0.904	4.191	11.019	0.957	0.989	0.926	24.182	14.074
LSTM U-Net	0.996	0.883	0.852	53.775	9.865	0.992	0.957	0.905	5.181	10.98	0.958	0.963	0.911	40.695	**14.565**
Trans U-Net	0.997	0.925	0.876	17.619	9.171	0.995	0.948	0.901	19.395	11.117	0.958	0.983	0.918	41.054	14.350
VisTR (ResNet-18)	0.975	**0.943**	**0.888**	5.490	**8.514**	0.965	**0.968**	**0.914**	4.876	10.175	0.886	**0.995**	0.899	10.507	12.565
VisTR (ResNet-50)	0.994	0.920	0.876	4.232	9.086	0.986	0.964	0.912	3.815	**10.632**	0.952	0.992	0.927	6.503	13.856
VisTR (ResNet-101)	0.996	0.929	0.881	8.891	9.024	0.992	0.965	0.912	**3.489**	10.793	0.955	0.992	**0.929**	8.791	13.926

TABLE IX

ABLATION STUDY FOR LOSS TERMS USED IN THE PROPOSED VISTR. THE FIGURES OF MERIT REPORTED HERE ARE AVERAGED OVER TEN SUBJECTS DATA FOR THE WRIST-TO-ELBOW REGION

Loss Terms					Wrist-to-Elbow Region		
BCE	L1	GIoU	Focal	Dice	Recall	Prec	DSC
✓	✓	✓	✓	✓	**0.833**	**0.866**	**0.841**
✗	✓	✓	✓	✓	0.812	0.856	0.822
✓	✗	✓	✓	✓	0.803	0.847	0.812
✓	✓	✗	✓	✓	0.818	0.849	0.824
✓	✓	✓	✗	✓	0.816	0.843	0.819
✓	✓	✓	✓	✗	0.699	0.889	0.762

accurate delineation of objects, especially ones with complex boundaries.

Although capable, Trans U-net fell short of VisTR's depth and effectiveness in managing the global context. Owing to its self-attention mechanism, VisTR's architecture naturally processes all the input elements in parallel. Furthermore, owing to parallelization, training and inference are expedited, which increases computational efficiency and makes it suitable for real-time applications. However, Trans U-Net uses sequential processing, which can become a computational bottleneck.

In addition to the attention mechanism, transformers like VisTR incorporate positional encodings. By providing model knowledge of the spatial relationships between input components, these encodings act as essential supplements to the attention process. They encode information regarding the relative positions of pixels or regions, enabling the model to differentiate between features that might share similar characteristics but are located at distinct spatial coordinates. This positional information, combined with context-awareness facilitated by the attention mechanism, collectively enhances the ability of VisTR to comprehend the feature positions necessary for accurate median nerve segmentation. It ensures that the model can discriminate between median nerve and background, even when their appearances vary across different frames within an ultrasound video. The success of the proposed VisTR model will open the door to more transformer-based deep learning models and better generalization.

In essence, the VisTR model for median nerve segmentation is an effective tool. Its robustness, adaptability, and contextual awareness through attention mechanisms set it apart from the traditional ideas in the literature. This transformer-based model has the potential to have a substantial impact on the clinical practice of ultrasound-guided procedures involving nerves (such as nerve neuropathy and local anesthesia) and contribute to better patient care with accelerated identification of nerves.

A. Clinical Impact

Segmentation of the median nerve is important for clinical diagnosis and medical procedures. One critical application lies in the diagnosis of CTS, a prevalent neurological disorder characterized by compression of the median nerve at the wrist. Accurate segmentation of the median nerve enables measurement of its CSA. This quantification aids in the diagnosis, determining the severity of nerve compression, and guiding treatment decisions. Currently, the localization and delineation of the median nerve are performed manually by a sonographer. Furthermore, the frame for which the CSA needs to be computed is determined by visually analyzing the nerve in a qualitative manner. The proposed method enables automated segmentation of the nerve. The CSA was computed for every frame in an automated fashion, leading to quantitative measurements. To evaluate the performance of the model in computing the CSA of the nerve, experiments were conducted on normal subjects and patients with mild and severe CTS, and the corresponding figure of merit is summarized in Table VIII.

The accurate segmentation of the median nerve is vital to plan effective treatments that involve nerve compression or injury. In such cases, determining the precise location and extent of the damage is essential for selecting the best treatment approach, which may involve surgery, physical therapy, or other interventions. In surgical procedures, precise segmentation of the nerve helps the surgeon identify its position, size, and relationships with surrounding structures, ensuring a safer and more effective procedure. In chronic conditions affecting the median nerve, such as CTS or nerve entrapment, regular imaging and segmentation can be used to monitor disease progression. Changes in the shape, size, and location of the median nerve provide valuable information on disease severity and help guide treatment decisions. The exact dimensions and location of the median nerve are crucial for planning personalized rehabilitation and physical therapy programs. Quantitative measures of the nerve CSA ensure that exercises and treatments are tailored to the specific needs of the patient, thereby promoting optimal recovery.

The proposed model developed for median nerve segmentation in ultrasound images can find immediate applications in the segmentation of other nerves and structures in

ultrasound scans, facilitating the diagnosis and treatment of peripheral nerve disorders or musculoskeletal conditions such as segmentation of the sciatic nerve or ulnar nerve. However, anatomical variability, such as varying shapes, sizes, and locations in ultrasound images, can hinder direct application, and it is challenging to generalize the model's performance across diverse anatomical regions. Furthermore, the availability of annotated data for segmenting specific nerves or structures may be limited, which necessitates extensive data collection and annotation.

V. Conclusion

In this study, a modified version of the VisTR model was used to efficiently segment the median nerve from ultrasound videos. Transformers are particularly well-suited for handling long sequences because they are less susceptible to issues such as vanishing or exploding gradients when processing lengthy sequences, thereby making them more advantageous for tasks such as video segmentation. Furthermore, transformers can overcome the limitations of CNNs, that is, the limited restricted receptive field and, therefore, can effectively leverage the comprehensive temporal and spatial information present in continuous video frames. Thus, the proposed model aids sonographers in efficiently segmenting the median nerve in the wrist-to-elbow region and outperforms the existing methods considered in this study. This is also the first work on the application of transformers for segmenting a nerve in an ultrasound video and to show the efficacy and effectiveness of transformer-based models. The proposed VisTR model, along with other discussed models in this study, is made available here: https://github.com/karang2606/Median-Nerve-Segmentation.

Acknowledgment

The authors are thankful to Aster-CMI Hospital, Bangalore, for enabling this research work.

References

[1] J. Avendaño-Coy, D. Serrano-Muñoz, J. Taylor, C. Goicoechea-García, and J. Gómez-Soriano, "Peripheral nerve conduction block by high-frequency alternating currents: A systematic review," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 6, pp. 1131–1140, Jun. 2018.

[2] A. Heuser et al., "Telerehabilitation using the Rutgers master II glove following carpal tunnel release surgery: Proof-of-concept," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 1, pp. 43–49, Mar. 2007.

[3] L. D. Hobson-Webb and L. Padua, "Median nerve ultrasonography in carpal tunnel syndrome: Findings from two laboratories," *Muscle Nerve*, vol. 40, no. 1, pp. 94–97, Jul. 2009.

[4] E. Ehler, "Median nerve ultrasonography in carpal tunnel syndrome," *Clin. Neurophysiol. Pract.*, vol. 2, pp. 186–187, Oct. 2017.

[5] P. Marhofer, M. Greher, and S. Kapral, "Ultrasound guidance in regional anaesthesia," *Brit. J. Anaesthesia*, vol. 94, no. 1, pp. 7–17, 2005.

[6] K. H. W. Lange, T. Jansen, S. Asghar, P. L. Kristensen, M. Skjønnemand, and P. Nørgaard, "Skin temperature measured by infrared thermography after specific ultrasound-guided blocking of the musculocutaneous, radial, ulnar, and median nerves in the upper extremity," *Brit. J. Anaesthesia*, vol. 106, no. 6, pp. 887–895, Jun. 2011.

[7] M. G. Hochman and J. L. Zilberfarb, "Nerves in a pinch: Imaging of nerve compression syndromes," *Radiologic Clinics*, vol. 42, no. 1, pp. 221–245, 2004.

[8] L. D. Hobson-Webb, L. Padua, and C. Martinoli, "Ultrasonography in the diagnosis of peripheral nerve disease," *Exp. Opinion Med. Diag.*, vol. 6, no. 5, pp. 457–471, Sep. 2012.

[9] J. I. Suk, F. O. Walker, and M. S. Cartwright, "Ultrasonography of peripheral nerves," *Current Neurol. Neurosci. Rep.*, vol. 13, pp. 1–9, Feb. 2013.

[10] B. C. Tsui, S. Suresh, and D. S. Warner, "Ultrasound imaging for regional anesthesia in infants, children, and adolescents: A review of current literature and its application in the practice of extremity and trunk blocks," *J. Amer. Soc. Anesthesiologists*, vol. 112, no. 2, pp. 473–492, 2010.

[11] O. Hadjerci, A. Hafiane, D. Conte, P. Makris, P. Vieyres, and A. Delbos, "Ultrasound median nerve localization by classification based on despeckle filtering and feature selection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 4155–4159.

[12] O. Hadjerci, A. Hafiane, D. Conte, P. Makris, P. Vieyres, and A. Delbos, "Computer-aided detection system for nerve identification using ultrasound images: A comparative study," *Informat. Med. Unlocked*, vol. 3, pp. 29–43, Oct. 2016.

[13] A. Hafiane, P. Vieyres, and A. Delbos, "Deep learning with spatiotemporal consistency for nerve segmentation in ultrasound images," 2017, *arXiv:1706.05870*.

[14] Y.-W. Wang, R.-F. Chang, Y.-S. Horng, and C.-J. Chen, "MNT-DeepSL: Median nerve tracking from carpal tunnel ultrasound images with deep similarity learning and analysis on continuous wrist motions," *Computerized Med. Imag. Graph.*, vol. 80, Mar. 2020, Art. no. 101687.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[16] R. T. Festen, V. J. M. M. Schrier, and P. C. Amadio, "Automated segmentation of the median nerve in the carpal tunnel using U-Net," *Ultrasound Med. Biol.*, vol. 47, no. 7, pp. 1964–1969, Jul. 2021.

[17] M.-H. Horng, C.-W. Yang, Y.-N. Sun, and T.-H. Yang, "DeepNerve: A new convolutional neural network for the localization and segmentation of the median nerve in ultrasound image sequences," *Ultrasound Med. Biol.*, vol. 46, no. 9, pp. 2439–2452, Sep. 2020.

[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[21] C.-H. Wu et al., "Automated segmentation of median nerve in dynamic sonography using deep learning: Evaluation of model performance," *Diagnostics*, vol. 11, no. 10, p. 1893, Oct. 2021.

[22] M. Di Cosmo et al., "A deep learning approach to median nerve evaluation in ultrasound images of carpal tunnel inlet," *Med. Biol. Eng. Comput.*, vol. 60, no. 11, pp. 3255–3264, Nov. 2022.

[23] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache, "A smart atlas for endomicroscopy using automated video retrieval," *Med. Image Anal.*, vol. 15, no. 4, pp. 460–476, Aug. 2011.

[24] F. W. Kremkau, *Diagnostic Ultrasound: Principles and Instrumentsa*, 6th ed. Philadelphia, PA, USA: W. B. Saunders Company, 2001.

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[26] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Comput. Vis.–ECCV 16th Eur. Conf.* Glasgow, U.K.: Springer, Aug. 2020, pp. 213–229.

[28] Z. Shen, R. Fu, C. Lin, and S. Zheng, "COTR: Convolution in transformer network for end to end polyp detection," in *Proc. 7th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2021, pp. 1757–1761.

[29] X. Ma, G. Luo, W. Wang, and K. Wang, "Transformer network for significant stenosis detection in CCTA of coronary arteries," in *Medical Image Computing and Computer Assisted Intervention–MICCAI*. Strasbourg, France: Springer, Sep. 2021, pp. 516–525.

[30] Q. Kong, Y. Wu, C. Yuan, and Y. Wang, "CT-CAD: Context-aware transformers for end-to-end chest abnormality detection on X-rays," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2021, pp. 1385–1388.

[31] R. Tao and G. Zheng, "Spine-transformers: Vertebra detection and localization in arbitrary field-of-view spine CT with transformers," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021*. Strasbourg, France: Springer, Sep. 2021, pp. 93–103.

[32] G. Li, D. Jin, Q. Yu, Y. Zheng, and M. Qi, "MultiIB-TransUNet: Transformer with multiple information bottleneck blocks for CT and ultrasound image segmentation," *Med. Phys.*, Aug. 2023, doi: 10.1002/mp.16662.
[33] Z. Tao, H. Dang, Y. Shi, W. Wang, X. Wang, and S. Ren, "Local and context-attention adaptive LCA-net for thyroid nodule segmentation in ultrasound images," *Sensors*, vol. 22, no. 16, p. 5984, Aug. 2022.
[34] Q. He, Q. Yang, and M. Xie, "HCTNet: A hybrid CNN-transformer network for breast ultrasound image segmentation," *Comput. Biol. Med.*, vol. 155, Mar. 2023, Art. no. 106629.
[35] F. Shamshad et al., "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102802.
[36] Z. Liu, Q. Lv, Z. Yang, Y. Li, C. H. Lee, and L. Shen, "Recent progress in transformer-based medical image analysis," *Comput. Biol. Med.*, vol. 164, Sep. 2023, Art. no. 107268.
[37] K. Al-Hammuri, F. Gebali, A. Kanan, and I. T. Chelvan, "Vision transformer architecture and applications in digital health: A tutorial and survey," *Vis. Comput. Ind., Biomed., Art*, vol. 6, no. 1, pp. 1–28, 2023.
[38] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[39] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.
[40] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.
[41] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization transformer," 2021, *arXiv:2102.04432*.
[42] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12299–12310.
[43] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6836–6846.
[44] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 23296–23308.
[45] T. Ferreira and W. Rasband, *ImageJ User Guide: IJ 1.42 R*. Bethesda, MD, USA: National Institute of Health, 2012.
[46] Y. Wang et al., "End-to-end video instance segmentation with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8737–8746.
[47] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.
[48] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[49] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 8026–8037.
[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[51] S. Jadon, "A survey of loss functions for semantic segmentation," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Oct. 2020, pp. 1–7.
[52] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[53] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-net architecture for medical image segmentation," in *Proc. Deep Learn. Med. Image Anal. (DLMIA) Workshop*. Granada, Spain: Springer, Jul. 2018, pp. 3–11.
[54] M. Dunnhofer et al., "Siam-U-net: Encoder–decoder Siamese network for knee cartilage tracking in ultrasound images," *Med. Image Anal.*, vol. 60, Feb. 2020, Art. no. 101631.
[55] O. Oktay et al., "Attention U-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.
[56] F. Xu, H. Ma, J. Sun, R. Wu, X. Liu, and Y. Kong, "LSTM multi-modal UNet for brain tumor segmentation," in *Proc. IEEE 4th Int. Conf. Image, Vis. Comput. (ICIVC)*, Jul. 2019, pp. 236–240.
[57] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.
[58] L. Bathala, P. Kumar, K. Kumar, A. B. Shaik, and L. H. Visser, "Normal values of median nerve cross-sectional area obtained by ultrasound along its course in the arm with electrophysiological correlations, in 100 Asian subjects," *Muscle Nerve*, vol. 49, no. 2, pp. 284–286, Feb. 2014.

Karan R. Gujarati received the M.Tech. degree from the Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India, in 2023.

He is currently working as a Data Scientist at Strand Life Sciences Ltd., Bangalore. His research interests include neuroimaging, medical image analysis using deep learning methods, and genomic data analysis using machine learning.

Mr. Gujarati was a recipient of the Wipro GE Healthcare M.Tech. Fellowship for the year 2021.

Lokesh Bathala received the M.B.B.S. and M.D. degrees in internal medicine, and the D.M. degree in neurology from Kasturba Medical College, Manipal, India, in 1997, 2001, and 2004, respectively.

Currently, he is the Lead Consultant Neurologist at Aster-CMI Hospital, Bangalore, India. He is the President of the Society of Neurosonology. He is actively involved in developing deep learning models for improved ultrasound imaging. His research interests include application of neuromuscular ultrasound on various peripheral nerve disorders.

Dr. Bathala has received certification on applied physics, carotid duplex, and transcranial Doppler from the American Society of Neuroimaging and on artificial intelligence and machine learning.

Vaddadi Venkatesh received the M.Tech. degree from the School of Computer and Information Sciences, University of Hyderabad, Hyderabad, India, in 2019. He is currently pursuing the Ph.D. degree with the Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India.

His primary areas of research include development of efficient model-based deep learning methods and automated medical image analysis using deep learning.

Raji Susan Mathew (Member, IEEE) received the M.Tech. degree in signal processing and the Ph.D. degree in medical image reconstruction from the Cochin University of Science and Technology, Cochin, India, in 2013 and 2021, respectively.

She is a Postdoctoral Fellow with the Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India. Her research interests include medical image reconstruction, image analysis, and computational methods of medical imaging.

Dr. Mathew was a recipient of the C. V. Raman Post-Doctoral Fellowship from 2021 to 2023, awarded by the Indian Institute of Science, Bangalore.

Phaneendra K. Yalavarthy (Senior Member, IEEE) received the M.Sc. degree in engineering from the Indian Institute of Science, Bangalore, India, in 2004, and the Ph.D. degree in biomedical computation from Dartmouth College, Hanover, NH, USA, in 2007.

He is a Professor of medical imaging with the Department of Computational and Data Sciences, Indian Institute of Science, Bangalore. His research interests include medical image computing, medical image analysis, and biomedical optics.

Dr. Yalavarthy is a Senior Member of SPIE and OSA. He serves as an Associate Editor for IEEE TRANSACTIONS ON MEDICAL IMAGING.