International Journal on Semantic Web and Information Systems
Volume 20 • Issue 1

Semantic-Based Optimization of Deep Learning for Efficient Real-Time Medical Image Segmentation

Zhenkun Wei, Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu, China & University of Chinese Academy of Sciences, Beijing, China

Jia Liu, China Mobile Industry Research Institute, China

Yu Yao, Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu, China & University of Chinese Academy of Sciences, Beijing, China*

ABSTRACT

In response to the critical need for advanced solutions in medical imaging segmentation, particularly for real-time applications in diagnostics and treatment planning, this study introduces SM-UNet. This novel deep learning architecture efficiently addresses the challenge of real-time, accurate medical image segmentation by integrating convolutional neural network (CNN) with multilayer perceptron (MLP). The architecture uniquely combines an initial convolutional encoder for detailed feature extraction, MLP module for capturing long-range dependencies, and a decoder that merges global features with high-resolution CNN map. Further optimization is achieved through a tokenization approach, significantly reducing computational demands. Its superior performance is confirmed by evaluations on standard datasets, showing interaction times drastically lower than comparable networks—between 1/6 to 1/10, and 1/25 compared to SOTA models. These advancements underscore SM-UNet's potential as a groundbreaking tool for facilitating real-time, precise medical diagnostics and treatment strategies.

KEYWORDS

Computational Efficiency, Convolutional Neural Networks, Deep Learning Optimization, Healthcare System, Magnetic Resonance Imaging, Multilayer Perceptron

Medical imaging stands as a pivotal component within the tapestry of the modern healthcare system (Atutornu & Hayre, 2018). It provides indispensable insights, pivotal in steering the course of diagnosis and treatment, thereby markedly impacting the outcomes of patient care. In an era marked by rapid technological evolution, the healthcare landscape has been reshaped by advancements in artificial intelligence, blockchain (Nguyen et al., 2021), and cloud computing (Onyebuchi et al., 2022; Alamer et al., 2022; Xu et al., 2021). These innovations have fostered significant enhancements in the realms of data security (Li et al., 2019; Yu et al., 2018; Kaushik & Gandhi, 2022), retrieval (Wang et al.,

DOI: 10.4018/IJSWIS.340938 *Corresponding Author

2020; Nhi et al., 2022), and utilization (Yu & Reiff-Marganiec, 2022). The cumulative research in this field highlights the criticality of amalgamating these cutting-edge technologies to augment the efficiency, security, and efficacy of healthcare delivery systems (Nhi et al., 2022; Hidalgo et al., 2022; Sarivougioukas & Vagelatos, 2022; Sun et al., 2023).

Deep learning-based segmentation techniques, exemplified by models like U-Net (Ronneberger et al., 2015) and VNet (Abdollahi et al., 2020), have showcased remarkable proficiency in the segmentation of anatomical structures and the detection of abnormalities. These models prove invaluable in tasks such as tumor segmentation, organ delineation, and lesion detection. Furthermore, the utility of deep learning extends across various disease domains (Gupta et al., 2024). For instance, in the realm of mammography, deep learning models excel in the identification and analysis of suspicious regions, potentially indicative of breast cancer (Yoon & Kim, 2021). In lung cancer screening, these models prove instrumental in detecting lung nodules and other anomalies within chest CT scans (Guo et al., 2021). Deep learning's versatility is further demonstrated in identifying lesions within brain images (Li et al., 2020), diagnosing oral cancer (Huang et al., 2023), detecting fractures in musculoskeletal imaging (Karanam et al., 2023), and facilitating the analysis of cardiac images (Litjens et al., 2019). This comprehensive application of deep learning underscores its pivotal role in the contemporary landscape of medical imaging and diagnosis.

Despite the remarkable advancements in segmentation accuracy brought about by deep learning, the major purpose of today's popular segmentation algorithms is to enhance segmentation outcomes (e.g., dice score). While beneficial for accuracy, the pursuit of segmentation performance often leads to increased complexity in network architectures. Modern networks often have an enormous number of parameters, large memory consumption, and long calculation time, all of which result in high-performance hardware requirements, making them difficult to deploy in real-world applications.

The creation of efficient, real-time image analysis systems is crucial in the digital era, such innovations can transform diagnostics and treatment planning to make advanced medical services widely accessible. (Gao et al., 2024; Srivastava et al., 2022). These systems transform diagnostics and treatment planning, making advanced medical services accessible to remote areas. Quick, accurate image analysis is vital for emergency diagnostics, enhancing decision-making in critical conditions such as strokes or traumas. In surgeries, real-time segmentation improves precision and safety. For large-scale screenings (Sedik et al., 2021), such as for lung or breast cancer, the ability to process images quickly and accurately increases diagnostic efficiency and could save lives through early detection. Accurate analysis also supports personalized treatment planning and telemedicine, extending diagnostic capabilities to resource-limited settings. Overall, advancements in image analysis technology are pivotal for improving patient care, efficiency, and healthcare access.

In recent years, significant progress has been made in the development of efficient network architectures for real-time medical image segmentation. These advancements not only aim at maintaining high accuracy but also at ensuring the practical applicability of real-time medical image segmentation. Several studies have focused on addressing the challenges of improving segmentation accuracy, computational efficiency, and real-time performance. Qin et al. (2021) introduced a knowledge distillation-based approach, reducing model complexity while maintaining high segmentation accuracy and real-time performance. Furthermore, Chen et al. (2019) proposed a 3D dilated multi-fiber network specifically designed for real-time brain tumor segmentation in MRI, capturing local and global contextual information and achieving real-time performance. Zamzmi et al. (2021) proposed a trilateral attention network that improved segmentation accuracy and real-time performance.

This paper proposes SM-UNet, an efficient medical image segmentation framework that combines the advantages of CNNs and multilayer perceptron (MLP) to establish global features. SM-UNet has a CNN encoder block to compile CNN features, a decoder block to generate segmentation results, and several MLP blocks that capture global features. By combining CNN with MLP, SM-UNet can take advantage of both the global context captured by MLP and the finely detailed high-resolution spatial

information provided by CNN. Additionally, this study presents separable tokenization and separable MLP to reduce computational consumption and improve interaction speed. This work evaluates the performance of SM-UNet on several medical image segmentation tasks, including lung segmentation and brain tumor segmentation. The results show that SM-UNet achieves competitive segmentation performance while requiring fewer parameters, lower memory usage, and quicker computation times than existing methods.

The contributions of this paper can be summarized as follows:

- We propose a new tokenization architecture that greatly reduces computational complexity while maintaining performance.
- Leveraging the tokenization architecture, we introduce SM-UNet, which synergizes the global contextual insights of MLP with the detailed, high-resolution spatial capabilities of CNN to offer a lightweight, real-time medical image segmentation solution.
- We present comprehensive testing on biomedical datasets, showing SM-UNet's competitive edge in accuracy, computational efficiency, and inference speed, which make it a cost-effective choice for practical applications.

This paper is organized as follows: The next section details the methodology, including the architecture of SM-UNet and its components. This is followed by a section presenting the experimental setup, datasets used, and performance evaluation criteria. We then discuss the results, demonstrating the efficacy of SM-UNet in various medical image segmentation tasks. Finally, the concluding section suggests directions for future research.

MATERIALS AND METHODS

Datasets

To evaluate the performance and generalizability of proposed method, experiments were conducted on the widely used Medical Segmentation Decathlon (MSD) dataset (Simpson et al., 2019) and a private rectal cancer magnetic resonance imaging (MRI) dataset (informed consent was obtained from all participants). In addition, to cover the ultrasound database, SM-UNet was also tested on DDTI dataset (Pedraza et al., 2015). To protect participant privacy and confidentiality, any identifying features were cropped out of all images, and no patient details or identifiers have been included in any scans or photographs presented in this paper. The authors confirm that the use of these datasets was in compliance with ethical guidelines and patient privacy and confidentiality were protected.

DDTI Dataset

The DDTI dataset (Pedraza et al., 2015) consists of a set of brightness-mode ultrasound images of the thyroid, including a complete annotation and diagnostic description of suspicious thyroid lesions by expert radiologists. While these lesions include thyroiditis, cystic nodules, adenomas, and thyroid cancer, we only include images with thyroid nodules for our experiments. After preprocessing to remove irrelevant regions and data cleaning using Wang's method (Wang, 2022), this study collected 637 images along with the corresponding thyroid nodule segmentation maps which were resized to a resolution of $512 * 512$.

Medical Segmentation Decathlon

The MSD dataset consists of 10 segmentation tasks from various organs and imaging modalities. These tasks are designed to simulate situations often encountered in medical im- ages, such as small training sets, unbalanced classes, multi-modality data, and small objects (Simpson et al.,

2019). SM-UNet and other CNN-based and ViT-based approaches were trained and evaluated on two of these tasks: *Task02_Heart* (Left atrium segmentation) and *Task10_Colon* (Colon Cancer segmentation). The modality and resolution of each task after preprocessing and data cleaning can be found in Table 1.

Rectal Cancer Dataset

Our private dataset consists of 101 magnetic resonance images (MRI) from patients diagnosed with rectal cancer with corresponding target regions of interest (ROIs) annotated by board-certified radiologists. The ROIs include mesocolic lymph node, mesocolon, and tumor region. This study only used images of tumor region, and only native T2-weighted (T2w) was used. This study collected 1249 images with a resolution of 512 * 512 after preprocessing and data cleaning.

Method

Given an input image $X \in R^{H*W*C}$ with a resolution of H*W and C channels, our goal is to output the corresponding segmentation mask with a size of H*W . Unlike previous approaches, our method has an encoder-decoder architecture with three stages: Encoder stage, Tokenized MLP stage, and Decoder stage, with a shortcut between encoder and decoder like UNet (see Figure 1). The encoder has 4 encoder blocks, each one doubles the number of channels while halves the feature resolution. A tokenized block was fit before MLP block to tokenize the feature, then six MLP blocks was stacked. Every step in the decoder path consists of a depth-wise separable convolution block that halves the number of feature channels followed by an interpolation with the scale factor of two by two. Note that these numbers are actually less than the number of filters of UNet and its variants, contributing to the first change to reduce parameters and computation.

Encoder Stage and Decoder Stage

The encoder comprises four encoder blocks, and each one doubles the number of channels while halving the feature resolution. It is worth noting that the initial encoder block solely expands channels without performing downsampling. These blocks contain the same modules: a pointwise convolution (PWConv) layer to expand the number of channels, a depthwise convolution (DWConv) layer and another pointwise convolution (PWConv) layer to shrink the number of channels. This establishes a three-layer (PWConv-DWConv-PWConv) inverted bottleneck structure, which is utilized as the basic encoder and decoder convolution unit. This unit is based on MobileNetV2's inverted residual bottleneck structure (Sandler et al., 2018) but with two differences: a) this work uses h-swish (Howard et al., 2019) (see Equation [1]) as the non-linearity instead of ReLU6 and b) the residual path was removed.

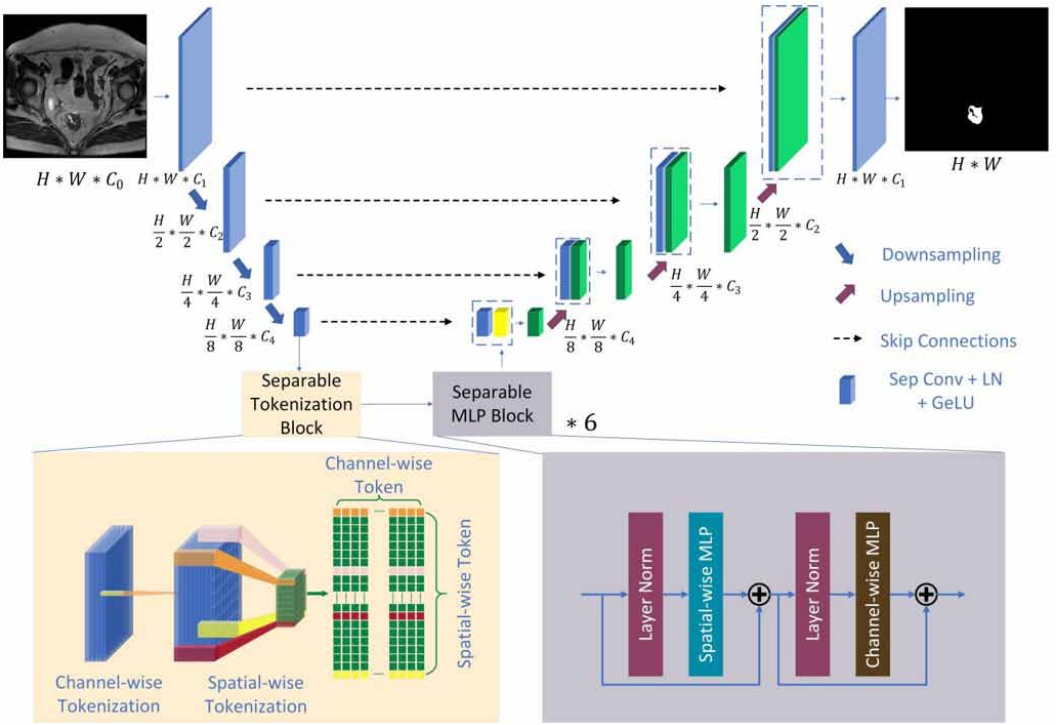
$$hswish\left(x\right) = x\frac{ReLU6\left(x+3\right)}{6} \tag{1}$$

where x represents the input. The hswish activation function is an innovation introduced in MobileNetV3 (Howard et al., 2019) as an update to V2. The authors of swish argued that this function possesses characteristics such as being unbounded, having a lower bound, being smooth, and being

Table 1. Preprocess for MSD

Task	Modality	Resolution
Task02_Heart	MRI	320*320
Task10_Colon	CT	512*512

Figure 1. Overview of the proposed SM-UNet architecture



non-monotonic, making it superior to ReLU in deep models (Ramachandran et al., 2017). However, due to the computational complexity of the sigmoid function, MobileNetV3 approximates swish using an approximate function to make it computationally more efficient, which in turn makes it harder (referred to as 'hard'). Concretely, the computation in the basic unit can be summarized as:

$$\begin{aligned} X_{layer_1} &= hswish\left(BN\left(PWConv\left(X_{input}\right)\right)\right) \\ X_{layer_2} &= hswish\left(BN\left(DWConv\left(X_{layer_1}\right)\right)\right) \\ X_{output} &= BN\left(PWConv\left(X_{layer_2}\right)\right) \end{aligned} \tag{2}$$

where X_{input} denotes the input, X_{output} denotes the output, $PWConv$ denotes point- wise convolution, DWConv denotes depthwise convolution, $hswish$ denotes non-linearity and BN denotes batch normalization.

In the encoder, this study used DWConv (stride of 2 and padding of 1) to downsample, and PWConv double the number of feature channels. For every step in the decoder, a bilinear interpolation layer was utilized to upsample the feature maps and, likewise, PWConv controls the number of feature channels. Note that the upsampled feature maps from the decoder are concatenated with the corresponding feature map from the encoder through skip-connection, and an additional basic convolution unit was added to halve the number of channels. The number of channels across each step is a hyperparameter denoted as C_1 to C_4. For the experiments using SM-UNet architecture, this work followed $C_1 = 16$, $C_2 = 32$, $C_3 = 64$, and $C_4 = 128$ unless stated otherwise.

Tokenized MLP Stage

Following the encoder stage, channel tokenization and spatial tokenization were performed on the extracted feature map $X_{feature} \in R^{\frac{H}{L}*\frac{W}{L}*C_4}$, Subsequently, a flatten operation is performed, mapping $X_{feature}$ into $X_{token} \in R^{S_T*S_T*C_4}$, which is represented in a latent 2-dimensional embedding space, denoted as $R^{S_T^2*C_4}$, $L = 2^3$, $S_T = 8$, and $C_4 = 128$ was followed unless stated otherwise.

The MLP stage consists of L layers of MLP blocks (see Equation [3]). Each block consists of two sub-MLP-blocks, each sub-block consists of a fully connected layer, a GELU layer, and another fully connected layer successively. Therefore, the output of the ℓ-th layer $z\ell$ can be expressed as follows:

$$x_\ell^{'} = MLP_s\left(LN\left(x_{\ell-1}\right)\right) + x_{\ell-1}$$
$$x_\ell = MLP_c\left(LN\left(x_\ell^{'}\right)\right) + x_\ell^{'} \tag{3}$$

where LN indicates the layer normalization operator, MLP_s denotes spatial-wise MLP, and MLP_c denotes channel-wise MLP. To recover the spatial order, the size of the encoded attention map must first be reshaped from $S_T^2 * C_4$ to $S_T * S_T * C_4$, then a bilinear interpolation layer with a scale factor of $\frac{H}{L * S_T} * \frac{W}{L * S_T}$ was used to upsample from $S_T * S_T * C_4$ to $\frac{H}{L} * \frac{W}{L} * C_4$. The feature map was multiplied with the reshaped attention map was multiplied with the reshaped attention map as the input to the decoder:

$$E_{input} = ReLU\left(X_{attention} \cdot X_{feature}\right) \tag{4}$$

where E_{input} denotes the input to the decoder, $X_{attention}$ denotes the reshaped output of MLP stage, and $X_{feature}$ is the extracted feature map from the encoder stage. Figure 1 shows the MLP stage's detailed architecture along with the tokenization operation.

Evaluation

Our goal is not to demonstrate state-of-the-art results, but to show that, remarkably, a simple model is competitive with today's best convolutional and attention-based models when considering the efficiency and performance together. We evaluate the efficiency with computational complexity (in FLOPs) and inference time (in milliseconds). For performance, this study used mean Dice coefficient (mDice), mean Jaccard Index (mJA) and 95% Hausdorff Distance (95HD) as criteria. The representation of these metrics are given in Equations (5) to (7):

$$Dice\left(A,B\right) = \frac{2\left|A \cap B\right|}{\left|A\right| + \left|B\right|} \times 100\% \tag{5}$$

$$Jaccard\left(A,B\right) = \frac{\left|A \cap B\right|}{\left|A \cup B\right|} \times 100\% \tag{6}$$

$$HD\left(A,B\right) = max\left(\max_{a\in A}\min_{b\in B} d\left(a,b\right), \max_{b\in B}\min_{a\in A} d\left(a,b\right)\right) \tag{7}$$

where A is the sets of pixels in the annotation and B represents the corresponding sets of the segmentation result, respectively. The maximin function defines the maximum Hausdorff distance from set A to set B. 95% Hausdorff distance and is similar to the maximum Hausdorff distance. However, it is calculated based on the 95th percentile of distances between boundary points in sets A and B. The purpose of using this metric is to mitigate the impact of a small number of outliers.

Implementation Details

This study used the Pytorch framework to develop SM-UNet. To train SM-UNet, this work combined binary cross entropy (BCE) with dice loss. The loss L between prediction $\hat{y}$ and target y is expressed as follows:

$$L = 0.5BCE\left(\hat{y}, y\right) + Dice\left(\hat{y}, y\right) \tag{8}$$

For all experiments, this study applied augmentation from Albumentations (Buslaev et al., 2020); the operations and their probabilities this study used are shown in Table 2. To train the model, this study employed the Adam optimizer (Kingma & Ba, 2014) with a cosine annealing scheduler (Loshchilov & Hutter, 2016) of 500 iterations. Figure 2 shows the dice and loss obtained by SM-UNet over 500 epochs (axis) on the MSD dataset. Similar results were obtained for the other datasets considered in this study. The batch size and learning rate were set separately for different datasets and models. All of the experiments employed a five-fold cross-validation technique to train models. Each model was trained for a total of 500 epochs and reported the mean evaluation metrics. All training and experiments were run on a standard workstation equipped with 16 GB of memory, an AMD Ryzen 5 2600 Six-Core Processor working at 3.85GHz, and a NVIDIA GeForce RTX 2070 SUPER with 8 GB of video memory. All experiments are implemented with Pytorch1.10.0.

RESULTS

This study compare SM-UNet to well-known and cutting-edge medical image segmentation frameworks, including convolutional baselines, UNet, UNet++ (Zhou et al., 2018) and Res-UNet (Zhang et al., 2018), as well as vision transformer baselines, AttnUNet (Schlemper et al., 2019), Res-AttnUNet (Li et al., 2021), MedT (Valanarasu et al., 2021), TransUNet (Chen et al., 2021) and

Figure 2. Convergence of Training and Validation

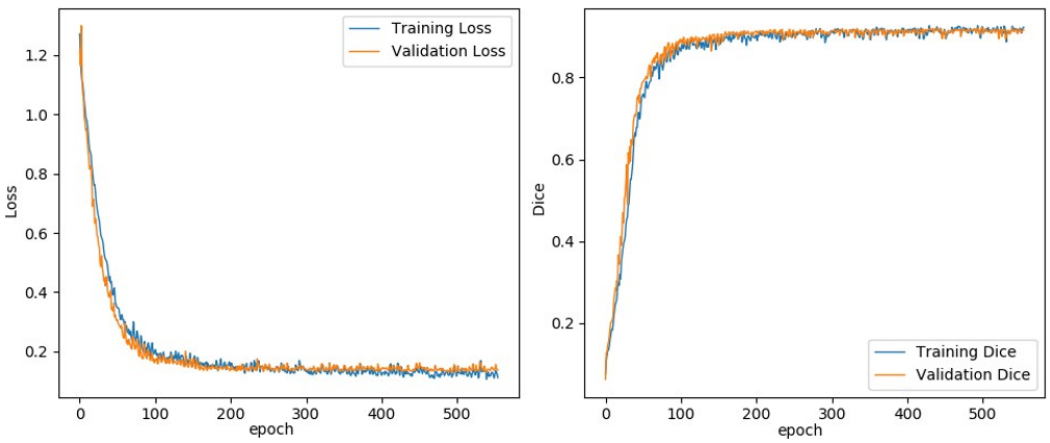


Table 2. Augmentations

Transforms	Hyperparameters
Cutout	*p=0.5*
VerticalFlip	*p=0.3*
HorizontalFlip	*p=0.3*
RandomBrightnessContrast	*brightness_limit=0.2*
	contrast_limit=0.2 p=0.3
	p=0.3
GridDistortion	*p=0.3*
HueSaturationValue	*p=0.3*
Downscale	*p=0.3*
GaussNoise	*p=0.3*

LeViT (Xu et al., 2023). This study ran the major experiments on four different dataset;, the results in terms of mean Dice, mean JA and 95HD are reported in Table 3.

It can be observed that SM-UNet obtains better segmentation performance than the basic convolutional baseline UNet on both datasets. It is a bit worse than the best model, but not by much. However, when considering the number of parameters and GFLOPs, our proposed method shows great advantages. SM-UNet is less computationally expensive than Res-AttnUNet (Li et al., 2021) because a) SM-UNet does not have any attention blocks and b) it uses separable convolution instead of traditional convolution. In comparison to all baselines, it is the most lightweight network. We should point out that SM-UNet contains just 0.582 M parameters compared to TransUNet's 67.097 M and Res-UNet's 39.091 M parameters. We also observed that SM-UNet has the fewest floating-point operators (GFLOPs): 1.94, compared to Res-AttnUNet's 442 and Res-UNet's 438. LeViT performs well on datasets with lower resolution but shows inferior performance on higher resolution datasets, likely due to its input size being constrained to 224×224. In contrast, our model demonstrates excellent performance across datasets of varying resolutions. Additionally, this study showed the typical inference time while using a CPU. It should be noted that this work particularly benchmarked the inference time in CPUs rather than GPUs because GPUs are typically more expensive and healthcare systems commonly run on a low budget with low compute power. It can be seen that the interaction time of SM-UNet is only $1/6$ to $1/10$ compared with the network with similar performance and only $1/25$ compared with the SOTA result.

Figure 3 illustrates a comparative analysis of efficiency (such as the number of parameters, GFLOPs, inference time, and fps) and performance (for instance, mean Dice score) between our proposed SM-UNet and other CNN-based and ViT-based methods in two distinct medical image segmentation tasks: MRI-based Rectal Cancer Tumor segmentation (Row 1) and Left atrium segmentation from the MSD dataset (Row 2). The graphical representation clearly shows that SM-UNet outshines its counterparts in all four categories. Particularly, in the first three graphs, SM-UNet's closer proximity to the upper left corner signifies its higher efficiency in terms of fewer parameters, reduced inference time, and lower computational complexity. Moreover, in the fourth graph, its position near the upper right corner highlights its outstanding performance in achieving a balance between computational demands and accuracy. This underscores SM-UNet's edge in delivering high-quality segmentation while excelling in aspects vital for real-time applications. It is noteworthy that the networks represented within the green area in the figure are capable of real-time operation, with LeViT and SM-UNet being the only two networks fully equipped to achieve this functionality. However, LeViT operates at an input resolution of 224×224, while SM-UNet is

Table 3. Performance Comparison on Four Datasets (Mean Dice Score %, Mean JA Score % and 95HD)

	Inference Time (ms)	Number of Parameters (M)	GFLOPs	DDTI Dataset			*Task02_Heart*			*Task10_Colon*			Rectal Cancer Dataset		
				mJA	mDice	95HD	mJA	mDice	95HD	mJA	mDice	95HD	mJA	mDice	95HD
UNet	57	31.391	224	53.24	63.59	19.86	87.22	93.03	6.87	43.64	49.79	5.59	69.66	81.51	13.08
UNet++	48	9.163	139	54.25	65.12	20.86	91.04	95.29	6.29	39.47	49.56	25.43	73.23	84.30	11.09
Res-UNet	205	39.091	438	56.06	68.68	14.43	91.27	95.43	9.49	48.36	53.53	3.67	**74.23**	**85.02**	10.53
AttnUNet	89	34.879	267	57.19	67.23	17.89	82.54	90.35	8.67	**49.92**	53.69	4.65	60.53	72.95	32.03
Res-AttnUNet	197	39.443	442	53.61	65.30	20.65	91.19	**95.76**	**6.23**	48.58	53.37	3.11	73.89	84.78	**6.67**
MedT	340	1.564	15	58.06	68.34	14.58	**92.19**	95.07	10.43	43.65	49.62	3.54	71.47	83.03	11.58
TransUNet	67	67.097	135	**59.60**	**71.47**	**13.99**	86.81	92.93	7.22	48.26	50.67	3.23	70.53	82.66	10.40
LeViT	12	52.17	25	56.78	70.92	14.39	86.32	92.33	7.62	47.71	50.18	2.93	69.98	82.06	11.85
SM-UNet	8	0.582	1.94	55.48	67.43	19.12	90.00	94.73	6.72	49.50	**54.56**	**2.54**	73.26	84.39	10.99

evaluated at a higher resolution of 320×320. Consequently, their actual FPS and network operation speed should be considered for reference purposes only. Despite this, SM-UNet outpaces LeViT in terms of speed and also boasts superior performance.

This study presents qualitative comparative results from multiple datasets, as shown in Figure 4. From left to right: (a) Input, (b) Ground Truth, (c) UNet, (d) UNet++, (e) AttnUNet, (f) MedT, (g) TransUNet, and (h) SM-UNet. Row 1 is the DTTI dataset (Ultrasound), Row 2 is the Left atrium segmentation task from the MSD dataset (*Task02_Heart*, MRI), Row 3 is the Colon Cancer segmentation task from the MSD dataset (*Task10_Colon*, CT), and Row 4 is the Rectal Cancer Tumor segmentation (MRI). It is clear that:

- The ViT-based method AttnUNet performs poorly, this is because transformer-based models, such as AttnUNet, require more data to overcome the limitations of lack of translation invariance and locality, so the performance on small datasets is not satisfactory.
- The CNN-based approach produces smoother edges than the ViT-based method. This may be caused by transformer layer operations on lower-resolution high-level feature maps, and its information is not well restored to the original resolution.
- The results in the last column show that our SM-UNet predictions are decent compared to other CNN-based methods, but better than AttnUNet and MedT, especially considering that they are only 1/10th to 1/100th the scale of other models. This shows that SM-UNet outperforms other ViT-based methods in terms of data utilization. Our design allows SM-UNet to enjoy the benefits of both high-level detailed global information and low-level localization context.

In summary, our approach combines information at different resolutions, reduces computational complexity through the use of separable convolutions, and modifies the MLP layer. On the basis of ensuring real-time segmentation, SM-UNet obtains performance close to SOTA.

Ablation Study

Several ablation studies were conducted to properly assess the proposed SM-UNet framework and validate the performance in various contexts.Table 4 shows the comparisons of different placements of the MLP stage. SM-UNet with the MLP stage at the deepest layer yields an average +0.27% mDice, +0.43% mJA, and 0.93 95HD on our test datasets in relation to those with position encoding

Figure 3. Efficiency and Performance Comparison of SM-UNet With CNN and ViT Models

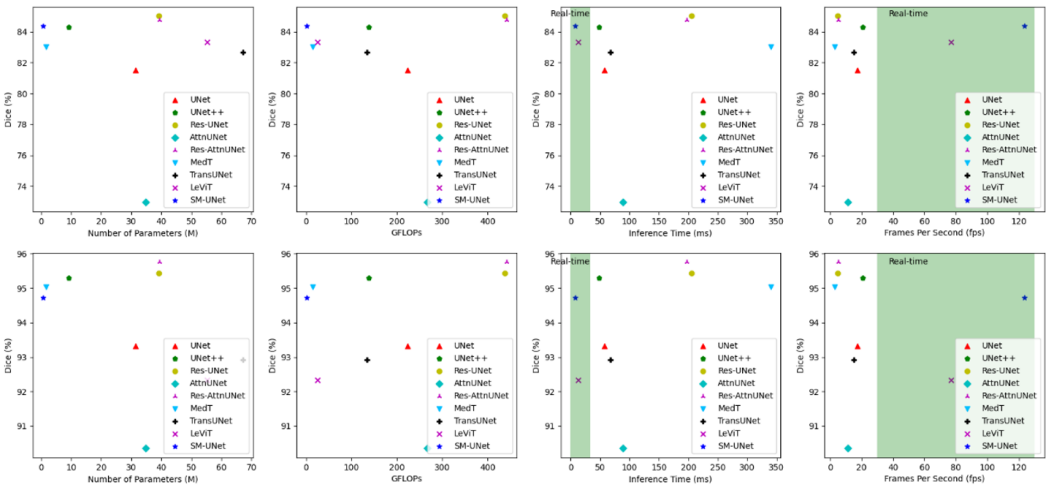
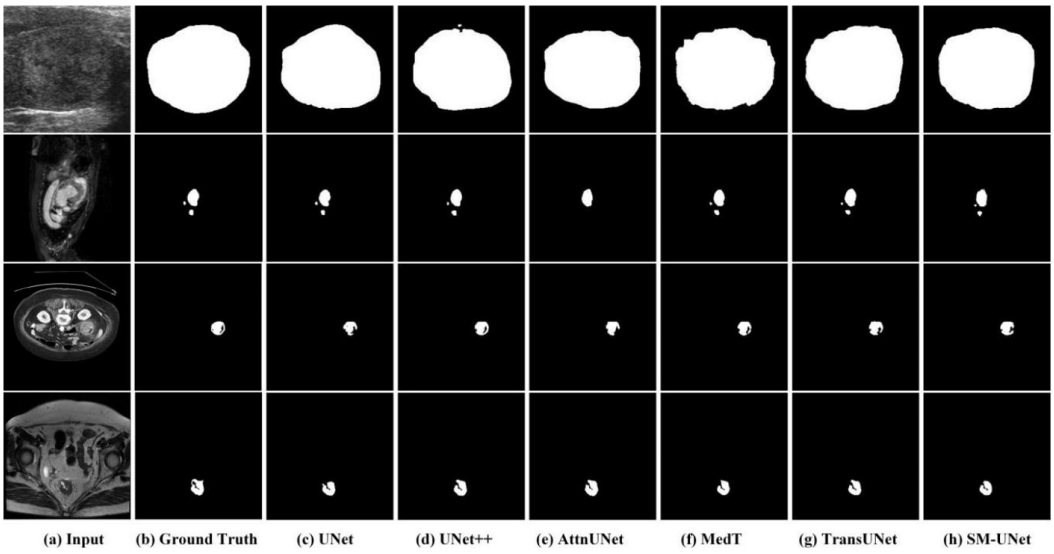


Figure 4. Example Results of Segmentation Comparison on Multi-Dataset



at other layers, respectively; networks with deeper mlp are superior in terms of number of parameters, inference time and computational complexity, indicating the effect of features from different layers on the MLP stage. The results also indicate the effectiveness of using the MLP stage, as they show a +3.23%/5.01% boost of mDice and mJA in contrast to those without the MLP stage. Additionally, note that the MLP stage at the first layer also achieved similar performance, the author speculate that it may be due to less information loss at shallower layers.

The effectiveness analysis of separable convolution and separable tokenization are shown in Table 5. Consider Conventional Conv as the foundational network, employing classical convolution operations, featuring four downsampling and four upsampling operations. In contrast, the DepSep Conv entails the substitution of conventional convolutions with Depthwise Separable Convolutions. Upon integrating the Sep MLP module, the process involves not a mere addition of an embedded module, but rather the replacement of the original network's lowest-level stratum (comprising Conv + downsample + upsample module) with the Sep MLP module. Results show that: a) the use of separable convolution greatly reduces the number of parameters, computation cost, and interaction time; b) the use of separable tokenization reduces the number of parameters, but due to its more complex structure, its inference time is not reduced; and c) using either alone increases the performance of

Table 4. Ablation Study on the MLP Stage's Placement

	Inference Time(ms)	*Number of Parameters(M)*	*GFLOPs*	*Task02_Heart*		
				mJA	*mDice*	*95HD*
On Layer 0	*13*	*1.596*	*2.842*	*85.25*	*92.02*	*5.91*
On Layer 1	*10*	*0.812*	*2.272*	*84.66*	*91.66*	*6.64*
On Layer 2	*9*	*0.619*	*2.062*	*85.24*	*92.01*	*6.27*
On Layer 3	*8*	*0.578*	*1.976*	*85.21*	*91.99*	*6.30*
On Layer 4	*8*	*0.582*	*1.938*	*85.52*	*92.58*	*5.35*
No MLP stage	*6*	*0.578*	*1.874*	*81.44*	*89.68*	*8.24*

SM-UNet, and the network works best when separable convolution and separable tokenization used at the same time. The depth of SM-UNet is also one major hyper-parameter that affects the number of parameters, complexity, and performance of the network. As can be seen in Table 6, with the increase of the network depth, the number of parameters, the computation cost, and the interaction time all increase, but mDice or mJA has not been significantly improved, and when the depth is too deep, the performance will be affected.

DISCUSSION

SM-UNet sets itself apart from existing models by introducing a unique combination of techniques. It employs a convolutional encoder to capture essential local details for precise segmentation, enhancing the U-Net design with multilayer perceptrons in deeper layers to merge global and high-resolution features, thereby enhancing segmentation results. Furthermore, SM-UNet introduces a tokenization architecture, inspired by the advancements in natural language processing, to reduce computational costs significantly. This innovative approach aims to elevate efficiency in medical image segmentation without sacrificing accuracy.

Our experiments, as detailed in Table 3 and visualized in Figures 3 and 4, demonstrate the superiority of SM-UNet over traditional CNN models and ViT-based methods in medical image segmentation tasks. SM-UNet not only achieves notable segmentation accuracy but also requires substantially fewer computational resources, marked by a significant decrease in parameters and GFLOPs. SM-UNet achieves this efficiency by leveraging a novel tokenization architecture and separable convolutions, enables this efficiency, making it exceptionally lightweight and fast without compromising on performance. Figures 3 and 4 highlight SM-UNet's balance between computational efficiency and segmentation accuracy, outshining other models in efficiency metrics while maintaining competitive accuracy. The ViT-based methods, like AttnUNet, perform poorly on smaller datasets due to their inherent limitations, such as the lack of translation invariance and locality. CNN-based methods (such as UNet) also perform poorly due to their lack of global information. In contrast, SM-UNet capitalizes on both global information and local details through its innovative architecture, explaining its ability to outperform others significantly. This unique blend of features, alongside the

Table 5. Ablation Study on MSD's Task02_Heart

	Inference Time(ms)	*Number of Parameters(M)*	*GFLOPs*	*Task02_Heart*		
				mJA	*mDice*	*95HD*
ConventionalConv	*14*	*3.371*	*8.763*	*81.16*	*89.59*	*8.96*
DepSepConv	*8*	*1.125*	*1.984*	*84.59*	*91.65*	*6.65*
SepMLP	*7*	*2.342*	*8.765*	*82.51*	*90.41*	*7.96*
DepSepConv+SepMLP	*8*	*0.582*	*1.938*	*85.52*	*92.58*	*5.35*

Table 6. Ablation Study on the Effect of Different Network Depths

	Inference Time(ms)	**Number of Parameters(M)**	**GFLOPs**	**Task02_Heart**		
				mJA	**mDice**	**95HD**
3	*7*	*0.146*	*0.682*	*86.09*	*92.50*	*5.76*
4	*8*	*0.582*	*1.938*	*85.52*	*92.58*	*5.35*
5	*10*	*1.533*	*2.347*	*86.03*	*92.46*	*5.80*

novel tokenization architecture, underscores SM-UNet's distinct advantage in real-time, accurate medical image segmentation.

The evaluation of SM-UNet, as it stands, concentrates on a select range of medical image segmentation tasks, leading to queries regarding its adaptability and broad applicability across varying medical imaging modalities and segmentation challenges. To comprehensively ascertain its efficacy and utility, expanded research into its performance across a more diverse array of tasks and datasets is imperative. Moreover, while the current manuscript accentuates the inference speed and segmentation precision of SM-UNet, it omits an in-depth analysis of network resource consumption, including memory demands, latency, and hardware prerequisites. An elaborate exploration of these metrics is crucial to understand SM-UNet's operational efficiency and viability for real-world deployment, especially under computational constraints.

CONCLUSION

This article introduced SM-UNet, a pioneering deep learning architecture crafted for the efficient segmentation of medical images in real-time applications. By ingeniously merging CNN with MLP and incorporating a novel tokenization method, SM-UNet sets new benchmarks in computational efficiency and accuracy. The architecture's evaluation, conducted across various medical segmentation tasks, showcases its superior inference speed and segmentation accuracy, underscoring its potential to revolutionize real-time medical diagnostics and treatment planning. Although SM-UNet exhibits substantial performance enhancements over existing models, this study also delineates avenues for further investigation. Future studies should extend the evaluation of SM-UNet across a wider array of medical imaging tasks and technologies to thoroughly assess its versatility and effectiveness. Additionally, a deeper analysis of the model's computational resource demands is crucial for optimizing its application in diverse real-world environments, particularly where resources may be limited. Such future endeavors will be vital for advancing SM-UNet's contribution to the field of medical diagnostics and treatment planning, potentially leading to significant improvements in healthcare outcomes.

REFERENCES

Abdollahi, A., Pradhan, B., & Alamri, A. (2020). VNet: An end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data. *IEEE Access : Practical Innovations, Open Solutions*, *8*, 179424–179436. doi:10.1109/ACCESS.2020.3026658

Alamer, S. A., Ilyas, Q. M., Ahmad, M., & Irfan, D. A. (2022). A metaphoric design of electronic medical record (EMR) for periodic health examination reports: An initiative to cloud's medical data analysis. *International Journal of Cloud Applications and Computing*, *12*(1), 1–18. doi:10.4018/IJCAC.2022010110

Atutornu, J., & Hayre, C. M. (2018). Personalised medicine and medical imaging: Opportunities and challenges for contemporary health care. *Journal of Medical Imaging and Radiation Sciences*, *49*(4), 352–359. doi:10.1016/j.jmir.2018.07.002 PMID:30514550

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. *Information (Basel)*, *11*(2), 125. Advance online publication. doi:10.3390/info11020125

Chen, C., Liu, X., Ding, M., Zheng, J., & Li, J. (2019). 3D dilated multi-fiber network for real-time brain tumor segmentation in MRI. In Lecture notes in computer science: Vol. 11766. *Medical image computing and computer assisted intervention–MICCAI 2019* (pp. 184–192). Springer, doi:10.1007/978-3-030-32248-9_21

Gao, S., Cheng, Y., Mao, S., Fan, X., & Deng, X. (2024). SSVEP-enhanced threat detection and its impact on image segmentation. *International Journal on Semantic Web and Information Systems*, *20*(1), 1–20. doi:10.4018/IJSWIS.336550

Guo, Z., Xu, L., Si, Y., & Razmjooy, N. (2021). Novel computer-aided lung cancer detection based on convolutional neural network-based and feature-based classifiers using metaheuristics. *International Journal of Imaging Systems and Technology*, *31*(4), 1954–1969. doi:10.1002/ima.22608

Gupta, B. B., Gaurav, A., & Arya, V. (2024). Deep CNN based brain tumor detection in intelligent systems. *International Journal of Intelligent Networks*, *5*, 30–37. doi:10.1016/j.ijin.2023.12.001

Hassanpour, H., Samadiani, N., & Salehi, S. M. (2015). Using morphological transforms to enhance the contrast of medical images. *The Egyptian Journal of Radiology and Nuclear Medicine*, *46*(2), 481–489. doi:10.1016/j.ejrnm.2015.01.004

Hidalgo, A., Pérez, N., & Lemus-Aguilar, I. (2022). Factors determining the success of ehealth innovation projects. *International Journal of Software Science and Computational Intelligence*, *14*(1), 1–22. doi:10.4018/IJSSCI.309709

Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). *Searching for MobileNetV3. In 2019 IEEE/CVF international conference on computer vision*. ICCV. doi:10.1109/ICCV.2019.00140

Huang, Q., Ding, H., & Razmjooy, N. (2023). Optimal deep learning neural network using ISSA for diagnosing the oral cancer. *Biomedical Signal Processing and Control*, *84*, 104749. Advance online publication. doi:10.1016/j.bspc.2023.104749

Kaushik, S., Gandhi, C., & Gandhi, C. (2022). Capability-based access control with trust for effective healthcare systems. *International Journal of Cloud Applications and Computing*, *12*(1), 1–28. doi:10.4018/IJCAC.297107

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization* [Conference presentation]. 3rd International Conference for Learning Representations, San Diego, CA, USA. doi:10.48550/arXiv.1412.6980

Li, D., Deng, L., Gupta, B. B., Wang, H., & Choi, C. (2019). A novel CNN based security guaranteed image watermarking generation scenario for smart city applications. *Information Sciences*, *479*, 432–447. doi:10.1016/j.ins.2018.02.060

Li, L., Wei, M., Liu, B., Atchaneeyasakul, K., Zhou, F., Pan, Z., Kumar, S., Zhang, J., Pu, Y., Liebeskind, D. S., & Scalzo, F. (2020). Deep learning for hemorrhagic lesion detection and segmentation on brain CT images. *IEEE Journal of Biomedical and Health Informatics*, *25*(5), 1646–1659. doi:10.1109/JBHI.2020.3028243 PMID:33001810

Li, R., Zheng, S., Duan, C., Su, J., & Zhang, C. (2021). Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. doi:10.1109/LGRS.2021.3063381

Litjens, G., Ciompi, F., Wolterink, J. M., de Vos, B. D., Leiner, T., Teuwen, J., & Išgum, I. (2019). State-of-the-art deep learning in cardiovascular image analysis. *JACC: Cardiovascular Imaging*, *12*(8), 1549–1565. doi:10.1016/j.jcmg.2019.06.009 PMID:31395244

Loshchilov, I., & Hutter, F. (2017, April 24–26). *SGDR: Stochastic gradient descent with warm restarts* [Conference presentation]. ICLR 2017 5th International Conference on Learning Representations, Toulon, France. https://openreview.net/forum?id=Skq89Scxx

Nguyen, G. N., Viet, N. H., Elhoseny, M., Shankar, K., Gupta, B., & Abd El-Latif, A. A. (2021). Secure blockchain enabled Cyber–physical systems in healthcare using deep belief network with ResNet model. *Journal of Parallel and Distributed Computing*, *153*, 150–160. doi:10.1016/j.jpdc.2021.03.011

Nhi, N. V., Le, T. M., & Van, T. T. (2022). A model of semantic-based image retrieval using C-tree and neighbor graph. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–23. doi:10.4018/IJSWIS.295551

Onyebuchi, A., Matthew, U. O., Kazaure, J. S., Okafor, N. U., Okey, O. D., Okochi, P. I., Taiwo, J. F., & Matthew, A. O. (2022). Business demand for a cloud enterprise data warehouse in electronic healthcare computing: Issues and developments in e-healthcare cloud computing. *International Journal of Cloud Applications and Computing*, *12*(1), 1–22. doi:10.4018/IJCAC.297098

Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., & Romero, E. (2015). An open access thyroid ultrasound image database. In E. Romero, & N. Lepore (Eds.), *Proceedings of the 10th international symposium on medical information processing and analysis* (Vol. 9287, pp. 188–193). SPIE. doi:10.1117/12.2073532

Qian, W., Li, H., & Mu, H. (2022). Circular LBP prior-based enhanced GAN for image style transfer. *International Journal on Semantic Web and Information Systems*, *18*(2), 1–15. doi:10.4018/IJSWIS.315601

Qin, D., Bu, J. J., Liu, Z., Shen, X., Zhou, S., Gu, J. J., Wang, Z. H., Wu, L., & Dai, H. F. (2021). Efficient medical image segmentation based on knowledge distillation. *IEEE Transactions on Medical Imaging*, *40*(12), 3820–3831. doi:10.1109/TMI.2021.3098703 PMID:34283713

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. Wells, & A. Frangi (Eds.), Lecture notes in computer science*:* Vol. 9351. *MICCAI: International Conference on Medical image computing and computer-assisted intervention* (pp. 234–241). Springer. doi:10.1007/978-3-319-24574-4_28

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510–4520). IEEE. doi:10.1109/CVPR.2018.00474

Sarivougioukas, J., & Vagelatos, A. (2022). Fused contextual data with threading technology to accelerate processing in home UbiHealth. *International Journal of Software Science and Computational Intelligence*, *14*(1), 1–14. doi:10.4018/IJSSCI.285590

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., & Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, *53*, 197–207. doi:10.1016/j.media.2019.01.012 PMID:30802813

Sedik, A., Hammad, M., Abd El-Samie, F. E., Gupta, B. B., & Abd El-Latif, A. A. (2021). Efficient deep learning approach for augmented detection of Coronavirus disease. *Neural Computing & Applications*, *34*(14), 11423–11440. doi:10.1007/s00521-020-05410-8 PMID:33487885

Srivastava, A. M., Rotte, P. A., Jain, A., & Prakash, S. (2022). Handling data scarcity through data augmentation in training of deep neural networks for 3D data processing. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–16. doi:10.4018/IJSWIS.297038

Sun, L., Wang, P., Liu, P., & Nie, Z. (2023). Image processing method of a visual communication system based on convolutional neural network. *International Journal on Semantic Web and Information Systems*, *19*(1), 1–19. doi:10.4018/IJSWIS.333063

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Keysers, D., Uszkoreit, J., Lucic, M., & Dosovitskiy, A. (2021). MLP-mixer: An all-MLP architecture for vision. *Advances in Neural Information Processing Systems*, *34*, 24261–24272. doi:10.48550/arXiv.2105.01601

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, & C. Essert (Eds.), Lecture notes in computer science*:* Vol.12901. *MICCAI: International Conference on Medical image computing and computer-assisted intervention*. (pp. 36–46). Springer. doi:10.1007/978-3-030-87193-2_4

Wang, H., Li, Z., Li, Y., Gupta, B. B., & Choi, C. (2020). Visual saliency guided complex image retrieval. *Pattern Recognition Letters*, *130*, 64–72. doi:10.1016/j.patrec.2018.08.010

Wang, M. (2022). *TNSCUI2020-SEG-rank1st: 1st place solution for segmentation task in MICCAI 2020 TN-SCUI Challenge.* GitHub. https://github.com/WAMAWAMA/TNSCUI2020-Seg-Rank1st

Xu, G., Zhang, X., He, X., & Wu, X. (2023). LeVit-UNet: Make faster encoders with transformer for medical image segmentation. In Q. Liu, H. Wang, Z. Ma, W. Zheng, H. Zha, X.Chen, L. Wang, & R. Ji (Eds.), *6th Chinese Conference on Pattern Recognition and Computer Vision (PRCV)* (pp. 42–53). Springer Nature Singapore. doi:10.1007/978-981-99-8543-2_4

Xu, Z., He, D., Vijayakumar, P., Gupta, B., & Shen, J. (2021). Certificateless public auditing scheme with data privacy and dynamics in group user model of cloud-assisted medical WSNs. *IEEE Journal of Biomedical and Health Informatics*, *27*(5), 2334–2344. doi:10.1109/JBHI.2021.3128775 PMID:34788225

Yoon, J. H., & Kim, E. K. (2021). Deep learning-based artificial intelligence for mammography. *Korean Journal of Radiology*, *22*(8), 1225–1239. doi:10.3348/kjr.2020.1210 PMID:33987993

Yu, C., Li, J., Li, X., Ren, X., & Gupta, B. B. (2018). Four-image encryption scheme based on quaternion Fresnel transform, chaos and computer generated hologram. *Multimedia Tools and Applications*, *77*(4), 4585–4608. doi:10.1007/s11042-017-4637-6

Yu, H. Q., & Reiff-Marganiec, S. (2022). Learning disease causality knowledge from the web of health data. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–19. doi:10.4018/IJSWIS.297145

Zhang, Z., Liu, Q., & Wang, Y. (2018). Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, *15*(5), 749–753. doi:10.1109/LGRS.2018.2802944

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested U-net architecture for medical image segmentation. In D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. Manuel, R. S. Tavares, A. Bradley, J. Paulo Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, & A. Madabhushi (Eds.), *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 3–11). Springer. doi:10.1007/978-3-030-00889-5_1