Enhancing-TransUNet:Hybrid UNet++ with Vision-Transformer Approach for Image Segmentation

GuanLin Liu, 20319045
School of Computer Science and Engineering, Sun Yat-sen University

***Abstract*—CNNs are among the most prevalent structures in computer vision, excelling in tasks such as image classification, object detection, and image segmentation. However, their effectiveness is limited by the size of the receptive field, hindering their ability to capture comprehensive global contextual information. Additionally, CNNs are challenged in discerning the spatial relationships between objects within images. In contrast, the Transformer model, recently gaining traction in computer vision, leverages a self-attention mechanism to recognize long-range dependencies within sequences, thus excelling in complex tasks such as those involving intricate contextual relationships in language models. Furthermore, Transformers are adept at understanding the positional dynamics between elements, enabled by their positional encoding capability. Nevertheless, their ability to recognize fine-grained local features in images remains suboptimal. UNet stands out as one of the most successful CNN implementations in image segmentation, characterized by its straightforward structure and robust segmentation capabilities. Its advanced iteration, UNet++, enhances performance further by integrating additional up-sampling nodes and skip-connections, leading to superior image segmentation outcomes. Recognizing the inherent strengths and limitations of CNNs and Transformers, and drawing inspiration from TransUNet, I propose the ”Enhancing-TransUNet,” a novel model that synergistically integrates the virtues of UNet++ and the Transformer. This fusion aims to enable precise handling of local details while also capturing extensive global contextual information, thereby addressing the individual limitations of the two architectures.**

***Index Terms*—Computer Vision, Image Segmentation, Transformer, TransUNet, Convolutional Neural Networks, UNet++.**

I. Introduction

IN the realm of computer vision, Convolutional Neural Networks (CNNs) have demonstrated remarkable proficiency over the past decade, excelling in tasks such as image classification, object detection, and image segmentation. By employing convolutional layers, CNNs autonomously and adaptively extract and identify hierarchical spatial features, enabling the detection of a diverse array of features and objects within images. Their formidable capacity to interpret visual data has established CNNs as a foundational technology in the domain of computer vision.

Nonetheless, a notable limitation arises from the convolutional layers' function as localized perceivers, which are confined to a subset of the input image, referred to as the receptive field. While adept at capturing local attributes like edges and textures, this localized view often falls short in apprehending complex structures and relationships over a broader range, thus limiting the acquisition of comprehensive contextual information. Attempts to mitigate this constraint through the augmentation of network depth or the enlargement of convolutional kernels typically incur substantial computational overhead. Furthermore, the propensity of CNNs to learn translation-invariant features renders the model relatively insensitive to the precise spatial positioning of objects, thereby hindering the recognition of spatial relationships among objects within the image.

Emerging as a transformative architecture in deep learning, the Transformer model was unveiled by Vaswani et al. in the seminal work ”Attention is All You Need” in 2017 [1]. Initially finding its application in Natural Language Processing (NLP), notably in sequence-to-sequence tasks such as text translation and text generation, the Transformer model has been recognized for its exceptional parallel processing capabilities and its adeptness in handling long-range dependencies. These attributes have facilitated the extension of the Transformer framework beyond NLP, finding impactful applications in fields including computer vision, as exemplified by the Vision-Transformer model [2].

At the heart of the Transformer architecture is the self-attention mechanism, a feature that empowers the model to dynamically allocate attention weights across various positions within a sequence, facilitating the model's ability to capture long-range dependencies. This mechanism proves particularly effective in managing complex contextual relationships, as seen in intricate language models.

Moreover, to compensate for the inherent inability of the self-attention mechanism to track positional order, Transformer models incorporate positional encoding. This technique is crucial as it ensures that the model comprehends not only the content of each sequence element but also the relative positional dynamics among them. However, the absence of a convolutional layer, akin to that found in CNNs, means that Transformers lack an intrinsic local receptive field, potentially undermining their performance in tasks demanding acute local feature discrimination and may lead to an underemphasis on proximate and local feature interactions.

In light of these distinct characteristics, a surge in research efforts has been observed, aiming to amalgamate the strengths of both CNN and Transformer architectures. A prominent example is TransUNet [3], an innovative hybrid that marries the capabilities of UNet [4] with those of the Transformer. This fusion results in a model that exhibits enhanced performance in image segmentation, leveraging the fine-grained spatial awareness of UNet [4] and the global contextual comprehension of the Transformer.

The UNet [4] architecture stands as a pivotal construct in the domain of image segmentation, distinguished by its encoder-decoder framework. The encoder progressively condenses the spatial dimensionality of data via convolution and pooling operations, whereas the decoder incrementally reconstructs the spatial attributes through up-sampling and convolutional processes. Originally conceptualized for biomedical image segmentation, UNet's utility has been extensively recognized, finding applications across a diverse array of image and video segmentation tasks. UNet++ [5] evolves this model further by incorporating deep supervision and nested residual connections. This enhancement not only refines the gradient flow and learning dynamics but also establishes a densely interconnected network through skip-connections across corresponding encoder and decoder layers. Such architectural refinements in UNet++ [5] have been empirically validated to substantially elevate performance [6], yielding superior segmentation precision and fostering expedited network convergence during training phases, courtesy of the deep supervision mechanism's role in reinforcing gradient flow.

Drawing inspiration from both TransUNet [3] and UNet++ [5], the novel **Enhancing-TransUNet** model emerges. This innovative framework synergistically harnesses the architectures of UNet++ [5] and Transformer, meticulously integrating the strengths of CNNs and Transformer models. **Enhancing-TransUNet** is architected to adeptly navigate the intricacies of local feature delineation while simultaneously harnessing a robust competency in assimilating global contextual nuances, thereby offering an advanced solution for complex segmentation challenges.

II. Related Work

UNet [4] stands as a seminal architecture in the domain of medical image segmentation, revered for its innovative integration of CNNs and Fully Convolutional Networks (FCNs). This integration melds CNNs' profound capacity for deep feature extraction with FCNs' precision in pixel-level segmentation, further augmented by the strategic use of skip-connections. These connections bridge the gap between low-level and high-level features, culminating in a segmentation model marked by heightened accuracy and robustness.

Introduced by Zhou et al. in 2018 [5], UNet++ advances the U-Net framework by weaving in dense connectivity. This advancement retains the foundational long skip-connections of U-Net while enriching the architecture with an intricate web of short skip-connections and up-sampling convolutional blocks. This innovation crafts a more nuanced encoder hierarchy, poised to enhance the model's segmentation prowess.

The Transformer model, originated by Vaswani et al. in 2017 [1], epitomizes the power of the self-attention mechanism. Initially making waves in NLP, especially in sequence-to-sequence tasks, its exceptional ability in parallel processing and capturing long-range dependencies has catalyzed its application across diverse domains. In computer vision, the Vision-Transformer [2] stands as a testament to the Transformer's adaptability and potency, underscoring its capability to deliver outstanding performance even in the intricate realm of visual data interpretation.

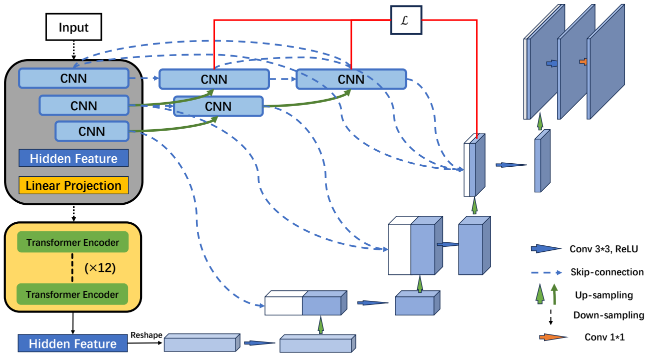

Fig. 1. Overview of the Enhancing-TransUNet framework.

Recent years have witnessed a burgeoning interest among scholars in fusing the Transformer architecture with U-shaped models. A notable contribution in this field is the TransUNet, conceptualized by Chen et al. in 2021 [3]. This hybrid model, one of the pioneering forays in medical image segmentation, innovatively pairs a Transformer encoder with a cascaded convolutional decoder.

In a progressive development, the Swin-UNet, introduced by Cao et al. in 2023 [7], represents an evolution in this domain. This architecture, distinctively outlined in Figure 4, diverges from TransUNet [3] by substituting the convolutional blocks in the U-Net encoder with Swin Transformer blocks, thereby harnessing Swin Transformer's [8] capability to methodically extract layered features from the input image. Swin-UNet [7] stands as a pioneering instance of a fully Transformer-based U-shaped framework, underscoring a significant shift from traditional approaches.

In a similar vein, models such as UNETR [9] and Swin UNETR [3] have emerged, integrating Transformers and convolutional decoders in their encoder design to create segmentation maps. These models, akin to TransUNet and Swin-UNet, exemplify the ongoing innovation in merging Transformer capabilities with the structural merits of U-shaped architectures, paving the way for advanced segmentation methodologies.

III. Method

In this section, I will elucidate the intricate process inherent to the newly developed model, forged by the integration of UNet++ and Transformer architectures. Drawing inspiration from TransUNet and advancing its foundation, the UNet model has been upgraded to UNet++, leading to the denomination of this enhanced version as **Enhancing-TransUNet**. Furthermore, this segment will detail the operational methodologies and integration techniques employed for the UNet++ and Vision-Transformer(ViT) models, respectively.

A. The General Process

Given an image $x \in \mathbb{R}^{H \times W \times C}$, where $H \times W$ denotes the image's dimensions and C represents the number of channels, our objective is to perform pixel-wise classification on the

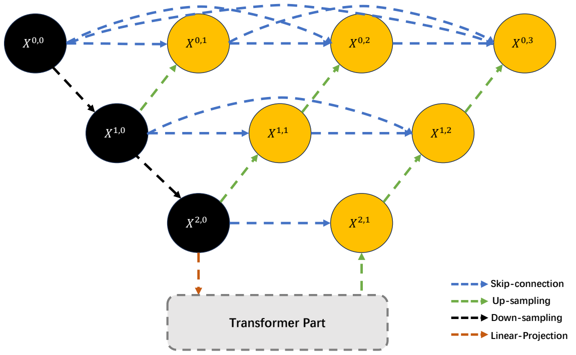

Fig. 2. The framework of the UNet++ component within the Enhancing-TransUNet model.

image. The task of image segmentation involves discerning the distinct category each pixel within the image pertains to. UNet and UNet++, developed on the foundation of CNNs, have demonstrated commendable efficacy in this task, with TransUNet exhibiting even more pronounced performance. To synergize the UNet++ and Transformer architectures, I consulted and adapted the architectural paradigm of TransUNet. This adaptation entailed the introduction of additional up-sampling nodes and skip-connections between the contracting (left side) and expansive (right side) pathways of TransUNet, thereby endowing the model with the capability to assimilate features extracted across diverse depth levels. The comprehensive architectural schema of Enhancing-TransUNet is depicted in Fig. 1.

B. UNet++

The upper segment of Enhancing-TransUNet incorporates the UNet++ architecture, with its overview framework delineated in Fig. 2. Within this architecture, the leftmost column consists of down-sampling nodes (depicted as black nodes), whereas the nodes situated to the right (illustrated as yellow nodes) function as up-sampling nodes. Constituting each node is a convolutional block from a CNN. Within every CNN node, the image is subjected to two successive 3×3 convolutions, each succeeded by a Rectified Linear Unit (ReLU). Nodes within the same hierarchical layer are interconnected through skip-connections. Each down-sampling node, denoted as $X^{i,0}$, administers a 2×2 max pooling operation with a stride of 2 on its subsequent node $X^{i+1,0}$ (if existent), effectuating the down-sampling process. Furthermore, each node $X^{i,j}$ executes a 2×2 convolution, termed "up-convolution," aimed at halving the number of feature channels for its diagonally adjacent upper-right node $X^{i-1,j+1}$ (if existent)

1) Network Connectivity: Let $x^{i,j}$ denote the output of node $X^{i,j}$ where i indexes the down-sampling layer along the encoder and j indexes the convolution layer of the dense block along the skip connection. The stack of feature maps represented by $x^{i,j}$ is computed as

$$x_j^i = \begin{cases} \mathcal{H}(\mathcal{D}(x_j^{i-1})), & \text{if } j = 0 \\ \mathcal{H}\left(\left[[x^{i,k}]_{k=0}^{j-1}, \mathcal{U}(x^{i+1,j-1})\right]\right), & \text{if } j > 0 \end{cases} \tag{1}$$

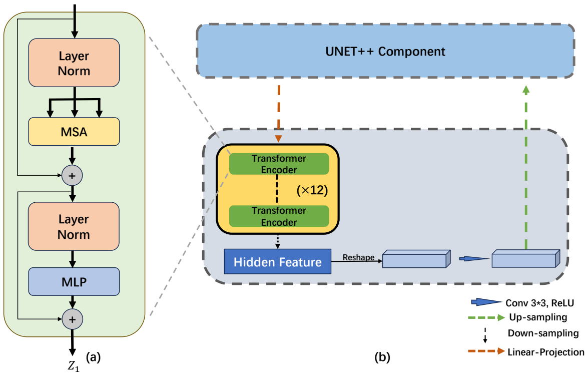

Fig. 3. The framework of the Transformer component within the Enhancing-TransUNet model.(a)The architectural configuration of the Transformer encoder.(b)Process depiction of reshape and up-sampling

where function $\mathcal{H}(\cdot)$ is a convolution operation followed by an activation function, $\mathcal{D}(\cdot)$ and $\mathcal{U}(\cdot)$ denote a down-sampling layer respectively, and [ ] donates the concatenation layer.

2) Deep Supervision: Deep Supervision constitutes a pivotal characteristic of the UNet++ architecture. For this purpose, Deep Supervision append a 1×1 convolution with $\mathcal{C}$ kernels followed by a $Sigmoid$ activation function to the outputs from node $X^{0,1}, X^{0,2}, X^{0,3}$ where $\mathcal{C}$ is the number of classes observed in the given dataset. UNet++ define a hybrid segmentation loss consisting of pixel-wise cross-entropy loss and soft dice-coefficient loss for each semantic scale.Mathematically, the hybrid loss is defined as:

$$\mathcal{L}(Y, P) = -\frac{1}{N} \sum_{c=1}^{C} \sum_{n=1}^{N} \left(y_{n,c} \log p_{n,c} + \frac{2 y_{n,c} p_{n,c}}{y_{n,c}^2 + p_{n,c}^2} \right) \tag{2}$$

where $y_{n,c} \in Y$ and $p_{n,c} \in P$ denote the target labels and predicted probailities for class c and n^{th} pixel in the batch, N indicates the number of pixels within one batch. The overall loss function for UNet++ is then defined as the weighted summation of the hybrid loss from each individual decoders: $\mathcal{L} = \sum_{i=1}^{d} \eta_i \cdot \mathcal{L}(Y, P^i)$, where d indexes the decoder.

C. Transformer Component

Enhancing-TransUNet is conceptualized by amalgamating the distinctive features of UNet++ with the foundational structure of TransUNet. Pertaining to the Transformer component of the model, it remains largely consistent with the implementation in TransUNet, and the technical specifics are delineated below.

1) Image Sequentialization: We first perform tokenization by reshaping the input x into a sequence of flattened 2D patches $\{X_p^i \in \mathbb{R}^{P^2 \cdot C} | i = 1, ..., N\}$, where each patch is of size $P \times P$ and $N = \frac{HW}{P^2}$ is the number of image patches.

2) Patch Embedding: We map the vectorized patches x_p into a latent D-dimensional embedding space using a trainable linear projection. To encode the patch spatial information, we learn specific position embeddings which are added to the patch embeddings to retain position information as follows:

$$Z_0 = [X_p^1 E; X_P^2 E; ...; X_p^N E] + E_{pos} \tag{3}$$

where $E \in \mathbb{R}^{(P^2 \cdot C) \times D}$ is the patch embedding projection, and $E_{pos} \in \mathbb{R}^{N \times D}$ denotes the position embedding.

The architectural configuration of the Transformer encoder is illustrated in Fig. 3(a), which consists of L layers of Multihead Self-Attention(MSA) [1] and Multi-Layer Perceptron(MLP) Blocks. (Eq. (4)(5)) Therefore the output of the ℓ-th layer can be written as follows:

$$Z'_\ell = MSA(LN(Z_{\ell-1},)) + Z_{\ell-1}, \tag{4}$$

$$Z_\ell = MLP(LN(Z'_\ell,)) + Z'_\ell, \tag{5}$$

where $LN(\cdot)$ denotes the layer normalization operator and Z_L is the encoded image representation.

3) Reshape and Up-sampling: For segmentation tasks, a straightforward approach involves employing an up-sampling technique on the encoded feature representation $Z_L \in \mathbb{R}^{\frac{HW}{P^2} \times D}$, elevating it to full resolution to facilitate dense output prediction. Initially, the dimension of the encoded feature is reshaped from $\frac{HW}{P^2}$ to $\frac{H}{P} \times \frac{H}{P}$. This reshaping is accomplished by first applying a 1×1 convolution to diminish the channel size of the reshaped feature to the number of classes. Subsequently, the feature map undergoes direct bilinearly up-sampling to the node $X^{2,1}$ within the UNet++ framework, culminating in the generation of an image with dimensions $H \times W$, which serves as the basis for predicting the final segmentation result. Refer to Fig. 3(b) for the process depiction.

IV. Analysis

A. Hybrid CNNs with Transformer achieving balanced local-global performance

Convolutional layers in CNNs, serving as local receptive fields, are naturally restricted to processing only segments of the input image individually. This inherent characteristic leads UNet++, which is based on CNN, to exhibit constraints in explicitly modeling long-range dependencies due to the localized nature of convolution operations.

Although Transformers are adept at capturing long-range dependencies, their effectiveness may be compromised by a lack of detailed, low-level information, resulting in somewhat diminished localization capabilities.

Enhancing-TransUNet emerges as a potent alternative for image segmentation by synergizing the strengths of both Transformers and UNet++. It employs Transformers to encode tokenized image patches derived from CNN feature maps, forming an input sequence that captures global contexts. Concurrently, the decoder aspect of Enhancing-TransUNet up-sampling these encoded features. These up-sampled features are then integrated with high-resolution CNN feature maps, facilitating precise localization and effectively bridging the gap between local and global contextual understanding.

B. UNet++ enhances the performance of TransUNet

TransUNet, as a pioneer among the earliest models integrating CNN and Transformer architectures, has demonstrated superior performance compared to models solely based on CNN structures. UNet++, by extracting shallow network features on top of the UNet architecture, also exhibits enhanced image segmentation capabilities beyond those of the standard UNet. Consequently, Enhancing-TransUNet, leveraging UNet++ for the CNN component, is poised to capture shallow features inaccessible to TransUNet based on UNet, thereby potentially achieving significantly improved performance.

V. Conclusion

Enhancing-TransUNet capitalizes on the synergistic combination of UNet++ and Transformer architectures, effectively harnessing the strengths of both CNN and Transformer. This integration empowers the model to adeptly handle precise local information while also exhibiting robust capabilities in capturing comprehensive global contextual information. Moreover, Enhancing-TransUNet introduces refinements to the convolutional network architecture, building upon the foundation of TransUNet. This enhancement facilitates the extraction of shallow image features, thereby substantially augmenting the model's overall capabilities.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

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

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[5] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.

[6] W. Yao, J. Bai, W. Liao, Y. Chen, M. Liu, and Y. Xie, "From cnn to transformer: A review of medical image segmentation models," *arXiv preprint arXiv:2308.05305*, 2023.

[7] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[9] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.