

RESEARCH ARTICLE

OrgUNETR: Utilizing Organ Information and Squeeze and Excitation Block for Improved Tumor Segmentation

SANGHYUK ROY CHOI¹, JUNGRO LEE², AND MINHYEOK LEE^{1,2}, (Member, IEEE)

¹Department of Intelligent Semiconductor Engineering, Chung-Ang University, Seoul 06974, Republic of Korea

²School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 06794, Republic of Korea

Corresponding author: Minhyeok Lee (mlee@cau.ac.kr)


This work was supported in part by the National Research Foundation of Korea (NRF) through the Korean Government (MSIT) under Grant RS-2024-00337250; and in part by Korea Institute for Advancement of Technology (KIAT) through the Korean Government, Advanced Training Program for Smart Sensor Engineers (MOTIE) under Grant P0020967.

ABSTRACT Convolutional Neural Networks (CNNs) have demonstrated remarkable performance in medical image segmentation tasks, with the U-Net architecture being a prominent example. The UNet Transformer (UNETR), an advanced variant of U-Net, incorporates a transformer architecture to effectively capture long-range dependencies in Computed Tomography (CT) scans. However, the application of deep learning models for tumor segmentation remains limited due to the challenges posed by the small size and unpredictable locations of tumors. To address this issue, we propose a novel approach that leverages organ information to improve tumor localization. Our model, named OrgUNETR, incorporates organ context by utilizing the fact that tumors typically exist within specific organs. By integrating organ information, OrgUNETR successfully detects tumors in CT scans with enhanced accuracy. Experimental results demonstrate that OrgUNETR surpasses the performance of a baseline model by achieving a 40.86% improvement in Dice score on the KiTS19 dataset and a 32.69% improvement on the Prostate158 dataset. Furthermore, we optimize the computational efficiency of UNETR by replacing the Multi-Head Self-Attention (MHSA) layers with Squeeze and Excitation (SE) layers, which perform attention in a similar manner. This modification reduces the computational cost by 13.9% while maintaining comparable performance. The proposed OrgUNETR model represents a significant advancement in tumor segmentation, combining the benefits of organ context and efficient attention mechanisms to achieve promising results. This research has the potential to assist medical professionals in accurate tumor localization and improve patient outcomes in clinical settings.

INDEX TERMS Organ segmentation, tumor segmentation, medical segmentation, deep learning, squeeze and excitation network, transformer.

I. INTRODUCTION

Convolutional Neural Networks (CNNs) have significantly advanced the field of image segmentation in recent years [1], [2], [3], [4]. The U-Net architecture, in particular, has proven highly effective for medical semantic segmentation by connecting a series of CNN layers [5]. The contracting path of the U-Net extracts spatial information from the input, while

The associate editor coordinating the review of this manuscript and approving it for publication was Alex James .

the expanding path reconstructs the segmented output using this spatial information via skip connections.

Several variations of the U-Net have been proposed to further improve its performance. The UNET++ features a nested encoder-decoder design with dense connections between layers, allowing it to capture detailed spatial features that may be lost during downsampling [6]. The KiU-NET combines the U-Net and Kite-Net architectures in parallel, enabling it to detect small, indistinct anatomical structures more effectively [7].

CNN-based architectures are highly effective at learning representations but struggle with learning global feature due to their localized receptive fields [8]. To extract global feature using a local range receptive field, it is necessary to employ multiple CNN operations. However, this approach causes computational inefficiency. Furthermore, the increased number of parameters imposes significant difficulties on optimization. These limitations can lead to inadequate semantic segmentation, particularly with multiple objects that have diverse boundaries in an image.

To overcome these challenges, dilated convolution and deformable convolution have been proposed to expand the receptive fields and capture a wider range of information [9], [10]. The dilated convolution enables the convolutional network to expand its receptive field by introducing different dilation. The deformable convolution adopts flexibility by utilizing 2D offset to the regular grid kernel. Since these 2D offsets are learned directly from the data, deformable convolution networks perform effectively in vision tasks requiring fine localization.

However, even with these improvement, the receptive fields in various convolutional layers remain constrained by the kernel window, which still involves significant computational complexity [11]. While developed CNNs can indeed capture global features, they often face limitations such as high computational costs and accuracy issues. There is a growing demand to capture relationships between distant parts of an input image without sacrificing accuracy. Therefore, the task should request for new models.

To address these issues, researchers have explored combining the CNN based U-Net architecture with transformers, which have shown great promise in Natural Language Processing (NLP) tasks. Vision Transformers (ViT) split the input image into patches and compute the connections between them, enabling the extraction of global spatial features through the use of Multi-Head Self Attention (MHSA) layers [6], [13], [16]. The UNet Transformer (UNETR) incorporates a ViT backbone as its encoder, allowing it to learn global spatial features and compress spatial information effectively.

While numerous deep learning models have been proposed for organ segmentation, comparatively fewer have been developed specifically for tumor segmentation. This is due in part to the challenges posed by the small size and unpredictable locations of tumors in medical images. Many existing models have therefore focused on organ segmentation, which inherently includes tumor information.

To address these issues, we propose the Organ UNETR (OrgUNETR), a modified version of the UNETR architecture designed specifically for kidney and prostate tumor segmentation. Our model leverages both organ and tumor information to improve tumor segmentation performance compared to a baseline model that uses only tumor information. We achieve this by predicting organs and tumors through distinct channels, each with its own loss function.

Through backpropagation, the model learns to locate both organs and tumors more accurately.

We further optimize the OrgUNETR architecture by replacing the MHSA layers with Squeeze and Excitation (SE) layers [17], [18], [19], [20], [21], [22]. The SE layers efficiently compute attention among feature channels, reducing the model's computational cost while maintaining comparable performance. They also enhance the model's ability to prioritize important features, making them particularly effective for medical segmentation tasks.

To validate the effectiveness of our approach, we compare the performance of OrgUNETR to a baseline model trained only on tumor information. The proposed model is evaluated with multiple tumor segmentation datasets with CT images, KiTS19 and Prostate158, for the generalization of the performance of the model.

The main contributions of this paper are as follows:

- 1) We demonstrate that including organ information enables more accurate tumor prediction compared to a baseline model, as organ and tumor information are inherently related.
- 2) By substituting MHSA layers with SE layers, we reduce the computational cost of OrgUNETR while maintaining tumor segmentation accuracy, making it more practical for real-world applications.

II. RELATED WORKS

A. TUMOR SEGMENTATION TASK

The tumor segmentation task is an essential process in the analysis of medical diagnosis, facilitating the precise identification of anatomical structures [23]. The tumor segmentation involves classification of every pixel, distinguishing tumor region from tissues. The objective of tumor segmentation is to accurately define the boundaries of tumors across the medical images such as CT scans, Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET) scans. This task is critical not only for diagnosing diseases but also for the strategic treatment plan, enabling personalized medicine approaches.

The CT scan and MRI scan are favored for segmentation tasks due to their ability to provide detailed images that facilitate the precise identification of abnormalities like tumors, making them efficient and crucial for accurate disease diagnosis [24]. Traditionally, this tumor segmentation from these scan images has been performed manually by physicians. This process that is not only time consuming but also demands significant human resources [25]. To conserve these resources and enhance the efficiency of the diagnostic process, the deep learning model is introduced [26]. The deep learning model is trained on medical image datasets that include detailed annotation of tumor locations, enabling the model to learn and subsequently perform tumor segmentation task autonomously. The integration of deep learning into tumor segmentation significantly promotes efficiency [27]. By providing the information of tumor location, the model assists physician in making decision regarding diagnosis.

Therefore, the deep learning model facilitates the diagnostic process and improves the accuracy of tumor detection and segmentation, thereby enhancing the efficiency and effectiveness of medical treatments.

There have been numerous attempts to develop a deep learning model that performs tumor segmentation task. For example, Swin U-Net Transformer (SwinUNETR) detects tumors by using input features at different resolutions that are extracted from SwinUNETR encoder utilizing shifted windows to compute self-attention [28]. Also, nnFormer successfully segments tumors by utilizing the cooperation of convolution and self-attention mechanism. By employing local and global self-attention mechanisms, more precise prediction is feasible [29].

B. U-NET ARCHITECTURE

The U-Net architecture, initially proposed for biomedical image segmentation, has become a widely adopted CNN-based model known for its effectiveness [5]. The U-Net consists of two primary paths: the contracting path and the expansive path. The contracting path is responsible for capturing the contextual information of the input image by gradually reducing its spatial dimensions while simultaneously increasing the depth of the feature channels. On the other hand, the expansive path, which is symmetrical to the contracting path, focuses on reconstructing the segmented image. It utilizes a series of upsampling operations to expand the feature maps back to the original input dimensions. During this reconstruction process, the expansive path receives feature maps from the corresponding levels of the contracting path via skip connections. These skip connections play a crucial role in transferring detailed spatial information, enabling more precise localization of features in the segmented output. By employing this dual-path architecture, the U-Net effectively captures both the global context and local details, making it highly suitable for biomedical image segmentation tasks.

Several variants of the U-Net have been introduced to further enhance its performance. One notable example is the UNET++, which features a nested encoder-decoder design with dense connections between the layers. These dense connections allow the UNET++ to capture fine-grained spatial features that might otherwise be lost during the downsampling process. By preserving these detailed features, the UNET++ is able to generate more accurate segmentation results, particularly in scenarios where small or intricate structures are present.

Another variant, the KiU-NET, combines the strengths of the U-Net and Kite-Net architectures by arranging them in parallel. This unique configuration enables the KiU-NET to effectively detect small and indistinct anatomical structures. The Kite-Net component of the KiU-NET expands the feature maps before downsampling them, which is in contrast to the traditional U-Net approach. By employing these two complementary networks, the KiU-NET is able to capture both the global context and the fine details of the input image,

resulting in improved segmentation performance for tiny and blurred objects.

Despite the remarkable success of CNN-based architectures like the U-Net and its variants, their ability to learn global features is inherently limited by the localized nature of their receptive fields. This limitation can lead to suboptimal semantic segmentation results, particularly when dealing with images containing multiple objects with diverse boundaries. To address this issue, researchers have explored various techniques, such as dilated convolutions and deformable convolutions, which aim to enlarge the receptive fields of the CNN layers. However, these approaches still face constraints in terms of the kernel window size and computational complexity, limiting their effectiveness in capturing long-range dependencies.

C. VISION TRANSFORMER

Transformers, which were originally introduced in the field of machine translation, have revolutionized the way sequence-to-sequence tasks are approached. These models replace the traditional recurrent and convolutional operations with self-attention mechanisms, enabling them to effectively capture long-range dependencies and achieve state-of-the-art performance [30], [31]. The success of transformers quickly spread beyond machine translation, finding applications in a wide range of NLP tasks. As a result, transformers have become the go-to architecture for many NLP applications, such as text classification, question answering, and language generation.

Another notable work in this direction is the Vision Transformer (ViT) [14], [32], [34], which represents a significant departure from traditional CNN-based architectures. Unlike CNNs, which primarily focus on local features, the ViT model excels at capturing global context by comparing patches of the input image. The ViT divides the image into fixed-size patches and linearly projects them into a sequence of embeddings. These embeddings are then processed by a stack of transformer layers, which utilize self-attention mechanisms to model the relationships between the patches. By attending to the entire sequence of patches, the ViT is able to capture long-range dependencies and global context, making it particularly effective for tasks beyond image classification, such as object detection and semantic segmentation.

D. UNETR FOR 3D MEDICAL SEGMENTATION

Building upon the success of transformers in computer vision, researchers have begun exploring their potential for medical image segmentation tasks. One notable example is the UNETR [35], [36], [37], which is specifically designed for 3D medical image segmentation. The UNETR architecture draws inspiration from the ViT and incorporates its self-attention mechanisms into the U-Net framework.

The UNETR consists of two main components: an encoder and a decoder. The encoder is responsible for extracting rich feature representations from the input 3D medical image. It achieves this by first dividing the image into uniform 3D patches and projecting them into a sequence of embeddings

using a linear projection layer. These embeddings are then processed by a series of transformer layers, which utilize MHSA and multi-layer perceptron (MLP) blocks to capture the relationships between the patches. By attending to the entire sequence of patches, the encoder is able to extract global contextual information and compress the spatial dimensions of the input image.

The decoder of the UNETR is designed to reconstruct the segmented output from the compressed feature representations generated by the encoder. It consists of a series of 3D deconvolution and convolution layers that progressively upsample the feature maps to the original input dimensions. To ensure that the decoder has access to the rich feature representations learned by the encoder, skip connections are employed between corresponding levels of the encoder and decoder. These skip connections allow the decoder to leverage both the global context captured by the encoder and the local details preserved in the higher-resolution feature maps.

One of the key advantages of the UNETR architecture is its ability to capture long-range dependencies and global context, which is particularly important for medical image segmentation tasks. By incorporating self-attention mechanisms, the UNETR is able to effectively model the relationships between different regions of the input image, enabling it to generate more accurate and coherent segmentation results. Moreover, the UNETR is designed to handle 3D medical images directly, without the need for slice-by-slice processing, which is common in CNN-based approaches. This allows the UNETR to exploit the inherent 3D structure of medical images and capture valuable volumetric information.

E. SQUEEZE AND EXCITATION NETWORK

The SE network is an architectural unit that aims to improve the representational power of CNNs by explicitly modeling the interdependencies between the channels of its convolutional features. The SE block is designed to adaptively recalibrate the feature maps generated by a CNN, allowing the network to emphasize informative features and suppress less useful ones.

The SE block consists of two main operations: squeeze and excitation. The squeeze operation aims to aggregate the spatial information of each feature map into a single numeric value, effectively capturing the global context of the feature map. This is typically achieved through global average pooling, which computes the average value of each feature map across its spatial dimensions. The resulting vector, often referred to as the channel descriptor, provides a compact representation of the global distribution of the feature map.

The excitation operation, on the other hand, aims to capture the interdependencies between the channels of the feature maps. It takes the channel descriptor as input and generates a set of channel-wise weights through a small neural network. This neural network typically consists of a dimensionality reduction layer, followed by a non-linearity (e.g., ReLU) and a dimensionality increasing layer. The output of the excitation

operation is a set of channel-wise weights that can be used to scale the original feature maps.

The scaled feature maps are obtained by element-wise multiplication of the original feature maps with the channel-wise weights generated by the excitation operation. This allows the SE block to adaptively recalibrate the feature maps, emphasizing the channels that are most informative for the task at hand and suppressing the less relevant ones. By doing so, the SE block enhances the representational power of the CNN and enables it to capture more discriminative features.

III. MATERIALS AND METHODS

A. ORGUNETR: INCORPORATING ORGAN CONTEXT FOR ENHANCED TUMOR SEGMENTATION

The proposed OrgUNETR architecture is designed to incorporate organ context information for improved tumor segmentation. Figure 1 presents an overview of the OrgUNETR model. The input to the model is a 3D CT scan, which is first processed by a patch embedding layer. This layer divides the input image ($x \in R^{H \times W \times D}$) into uniform 3D patches ($x_p \in R^{N \times P^3}$), where $N = HWD/P^3$ represents the total number of patches. These patches are then projected into a sequence of tokens using a linear projection layer.

The encoder of the OrgUNETR model consists of a series of SE blocks, which are connected successively. The architecture employs 2, 4, and 6 SE blocks in the encoder, with each block downsampling the spatial dimensions by a factor of two. This allows the encoder to compress the spatial information while extracting relevant features at different scales.

At various stages of the network, the extracted features are upsampled using deconvolution layers and further enhanced by convolution layers, followed by batch normalization and ReLU activation functions. The upsampled features are then concatenated with the corresponding features from earlier blocks via skip connections. This process is repeated until the feature maps reach the same spatial dimensions as the input image.

One of the key challenges in tumor segmentation is the small size and unpredictable location of tumors within the CT scans. To address this issue, the OrgUNETR model incorporates both organ and tumor information for precise tumor localization. The output of the model consists of two channels, one for organ segmentation and the other for tumor segmentation. By sharing weights between the tumor prediction channel and the organ prediction channel, the model leverages organ information during the tumor prediction process. This dual-channel approach effectively incorporates organ context, enabling more accurate segmentation of tumor locations.

To achieve a balance between computational efficiency and performance, the OrgUNETR model replaces the MHSA layers, commonly used in transformer-based architectures, with SE layers. Unlike self-attention mechanisms, which require the computation of attention maps across all patch sequences, SE layers focus on modulating the feature channels based on global information obtained through a global

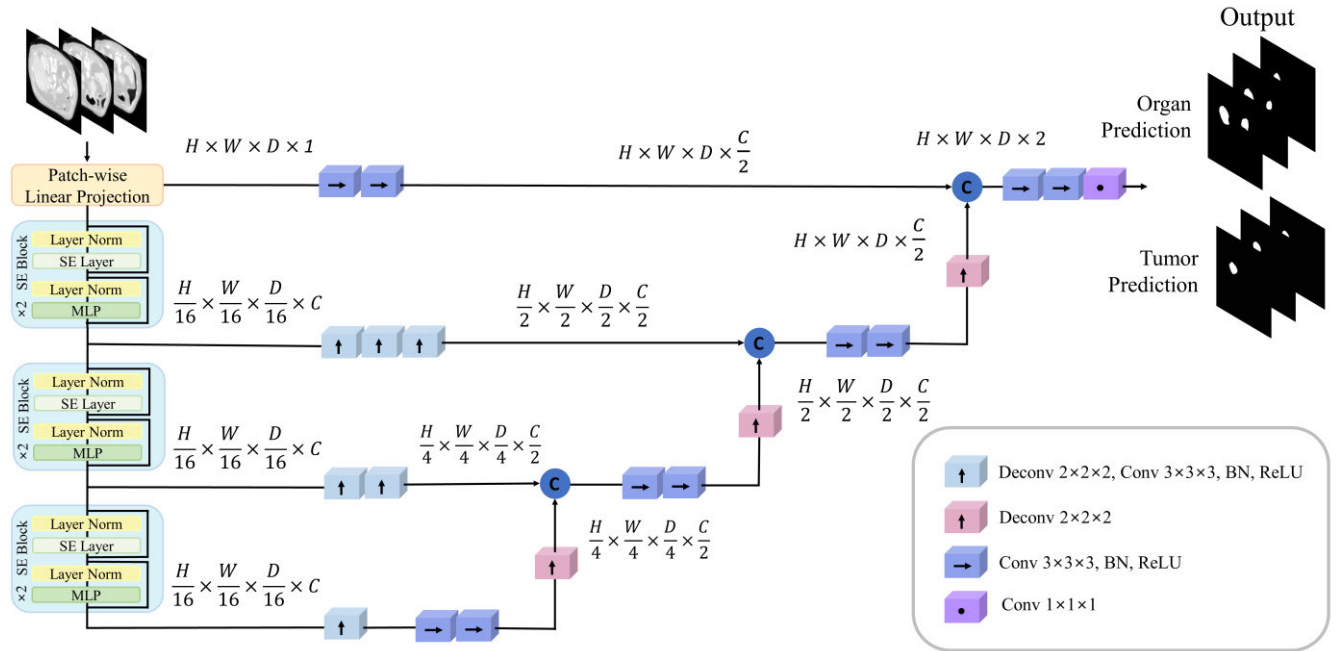


FIGURE 1. The overall architecture of OrgUNETR. It illustrates the comprehensive architecture of the OrgUNETR model, detailing its various layers and connections. A 3D CT scan is divided into a uniform 3D patches and projected into a token sequence by linear projection. The sequence is used as an input of SE Blocks. The encoded feature maps of different SE Blocks are extracted and integrated by the decoder. The final output dimension is $H \times W \times D \times 2$ and the two channels indicate organ and tumor prediction outputs. The number of feature map channel C is 32 and patch resolution P is 16.

averag pooling layer. By replacing self-attention layers with SE layers, the OrgUNETR model reduces computational complexity while maintaining segmentation performance.

B. PREPROCESSING AND PATCH EMBEDDING

Preprocessing the input CT scans is a crucial step in the OrgUNETR pipeline. Directly processing every pixel in a 3D CT scan would result in high computational complexity, making it impractical for real-world applications. To alleviate this issue, the OrgUNETR model adopts a patch embedding layer, inspired by the original UNETR architecture.

The patch embedding layer divides the input image ($x \in R^{H \times W \times D}$) into non-overlapping 3D patches ($x_p \in R^{N \times P^3}$), where $N = HWD/P^3$ represents the total number of patches and P denotes the patch size. Each patch is then transformed into a token using a linear projection layer. However, this process does not preserve the positional information of the patches, which is essential for the model to understand the spatial relationships between them.

To address this issue, a learnable positional encoding vector (E_{pos}) is added to the projected patch tokens [38], [39]. The positional encoding vector captures the spatial information of each patch, allowing the model to distinguish between patches at different positions. The embedding process can be represented by the following equation:

$$z_{tokens} = [p_1E; p_2E; \dots; p_NE] + E_{pos}, \tag{1}$$

where z_{tokens} represents the token vector; p_n represents the n -th patch vector; E represents the embedding matrix, and; E_{pos} represents the positional encoding vector.

C. SQUEEZE AND EXCITATION BLOCKS

The SE blocks form a crucial component of the OrgUNETR architecture. These blocks are designed to adaptively recalibrate the feature maps by explicitly modeling the interdependencies between channels [17], [18], [19], [20], [21], [22]. By doing so, the SE blocks enable the model to prioritize informative features and suppress less relevant ones.

In the OrgUNETR model, each SE block consists of an SE layer followed by an MLP layer. A normalization layer is appended after each layer to stabilize the weight values during training. The SE layer assesses the significance of each feature channel by generating a channel-wise attention vector.

The operation of the SE block can be represented by the following equation:

$$S_l = LN(MLP(LN(v_{l-1} \times S_{l-1}))), \tag{2}$$

where S_l represents the input to the l -th SE layer; LN denotes the layer normalization operation; v_{l-1} represents the channel-wise attention vector from the previous SE layer, and S_{l-1} represents the output of the previous SE layer.

The channel-wise attention vector v_{l-1} serves to inject attention into the feature map channels, enabling the model to focus on relevant features. The attention vector is generated by the SE layer through a series of operations. First, the SE layer applies global average pooling to the input feature maps, reducing their spatial dimensions and producing a channel descriptor. This descriptor captures the global information of each feature channel.

Next, the channel descriptor undergoes a dimensionality reduction operation, typically implemented using a fully connected layer with a smaller number of neurons compared to the number of channels. This step helps to reduce the computational complexity and prevent overfitting. The reduced channel descriptor is then passed through a non-linear activation function, such as ReLU, to introduce non-linearity into the attention mechanism.

Finally, the activated channel descriptor is expanded back to the original number of channels using another fully connected layer. This expansion operation generates the channel-wise attention vector (v_{l-1}), which contains a weight for each feature channel. The attention vector is then element-wise multiplied with the input feature maps (S_{l-1}) to produce the recalibrated feature maps.

The recalibrated feature maps are further processed by the MLP layer, which learns to combine the attended features effectively. The output of the MLP layer is then normalized using layer normalization to stabilize the training process.

By employing SE blocks, the OrgUNETR model can adaptively recalibrate the feature maps, emphasizing informative channels and suppressing less relevant ones. This mechanism enhances the model's ability to capture discriminative features and improves its segmentation performance. Moreover, by using a single attention vector for each SE block, the computational complexity is significantly reduced compared to the MHSA layers used in transformer-based architectures.

D. METRICS

To train and evaluate the performance of the OrgUNETR model, we employ a combination of the Dice coefficient and Cross-Entropy Loss, which are widely used metrics in segmentation tasks [40], [41], [42].

The Dice coefficient measures the overlap between the predicted segmentation and the ground truth. It is calculated using the following equation:

$$S_{dice} = \frac{2 \times (P_{true} \times P_{pred})}{P_{true} + P_{pred}}, \quad (3)$$

where P_{true} and P_{pred} are binary matrices representing the ground truth and predicted locations of the organ or tumor, respectively [41]. The Dice coefficient ranges from 0 to 1, with a higher value indicating better segmentation performance.

In addition to the Dice coefficient, we also employ the Cross-Entropy Loss, which quantifies the dissimilarity between the predicted probabilities and the ground truth labels. The Cross-Entropy Loss is calculated as follows:

$$CE = \sum \sum T \times \log(p_{truth}), \quad (4)$$

where T is number value indicates organ or tumor, p_{truth} denotes the probability of predicted value compared to truth.

To combine the Dice coefficient and Cross-Entropy Loss, we introduce the DiceCELoss, which is a weighted sum of

the two metrics. The DiceCELoss is defined as:

$$DiceCELoss = \alpha \times S_{dice} + \beta \times CE, \quad (5)$$

where α and β are hyperparameters that control the relative importance of the Dice coefficient and Cross-Entropy Loss, respectively. These hyperparameters are adjusted during the training process to optimize the model's performance.

Given the critical importance of accurately segmenting both the organ and the tumor, we further extend the DiceCELoss by introducing a weighted variant:

$$DL_{total} = 0.65 \times DL_{organ} + 0.35 \times DL_{tumor}, \quad (6)$$

where DL_{total} represents the overall Dice coefficient, Cross entropy loss, it is computed as a weighted sum of the DiceCELoss for organ (DL_{organ}) and tumor (DL_{tumor}), with respective weightings of 65% and 35%. This weighted approach prioritizes the organ segmentation slightly more than the tumor, reflecting the model emphasis on using organ context for improved tumor localization. The *DiceCELoss* is minimized by backpropagation algorithm [43].

E. KIDNEY TUMOR SEGMENTATION DATASET

To evaluate the performance of the OrgUNETR model on kidney tumor segmentation, we utilize the KiTS19 dataset. This dataset serves as a cornerstone for our study, providing CT scans accompanied by annotations for both the right and left kidneys, as well as kidney tumors.

The KiTS19 dataset comprises 544 CT scans, which were annotated by medical students under the supervision of expert radiologists. Each CT scan has a consistent resolution of 512×512 pixels, with the number of slices ranging from a minimum of 29 to a maximum of 1,059. However, due to computational constraints, we rescale the dataset to a uniform size of $128 \times 128 \times 128$ using linear interpolation, adjusting the number of slices accordingly.

It is important to note that the resizing process can introduce a challenge: small tumor pixels may merge with adjacent pixels, potentially leading to the elimination of tumor pixels in some cases. To mitigate this issue, we carefully examine the resized CT scans and exclude 54 scans that lack tumor pixels after resizing. This step ensures that the training dataset contains sufficient tumor information for the model to learn from. The KiTS19 dataset is publicly accessible and can be downloaded from the official repository (<https://github.com/neheller/kits19>) with the consent of the organizers.

F. PROSTATE TUMOR SEGMENTATION DATASET

In addition to the KiTS19 dataset, we also utilize the Prostate158 dataset to evaluate the performance of the OrgUNETR model on prostate tumor segmentation. The Prostate158 dataset is a comprehensive collection of high-quality 3 Tesla MRI scans specifically designed for prostate segmentation tasks.

The dataset includes scans of both anatomical zones and cancerous lesions within the prostate, making it a valuable

resource for prostate MRI image analysis. The inclusion of both healthy and cancerous tissue annotations enables the development of models that can accurately segment the prostate gland and identify tumors simultaneously [44].

The Prostate158 dataset consists of 139 training samples and 19 validation samples. Each MRI scan has a native resolution of $270 \times 270 \times 24$ voxels. However, due to computational limitations, we resize the scans to a uniform size of $128 \times 128 \times 128$ voxels using linear interpolation. During the resizing process, we carefully monitor the presence of tumor pixels and exclude 72 samples that lose tumor pixel information as a result of the resizing operation. This ensures that the training dataset maintains a sufficient representation of tumors for effective learning.

To normalize the intensity values of the MRI scans, we apply min-max scaling to each scan, bringing the pixel values into a consistent range. This normalization step helps to mitigate the influence of variations in scanner settings and acquisition protocols, making the dataset more suitable for training deep learning models.

The Prostate158 dataset is accompanied by expert annotations, which serve as ground truth labels for training and validating the segmentation models. These annotations were carefully curated by experienced radiologists, ensuring their reliability and accuracy.

The Prostate158 dataset is publicly available and can be accessed from the official repository (<https://zenodo.org/record/6481141>) with the consent of the organizers. This dataset has been widely used in the research community for developing and evaluating prostate segmentation algorithms, contributing to the advancement of prostate cancer diagnosis and treatment planning.

IV. RESULTS

A. SEGMENTATION RESULTS WITH KITS19 DATASET

To evaluate the performance of our proposed OrgUNETR model on kidney tumor segmentation, we conduct experiments using the KiTS19 dataset. The dataset consists of 490 CT volumes, each annotated with both kidney and kidney tumor labels. We partition the dataset into training and validation sets using a 70:30 ratio, ensuring a fair evaluation of the model's generalization ability.

During training, we employ the AdamW optimizer [45] with a learning rate of 0.0001. To enhance the model's robustness and prevent overfitting, we apply data augmentation techniques, specifically random rotation of the input images within a range of 0 to 10 degrees [46]. The loss function used for training is a combination of Dice Loss and Cross-Entropy Loss, referred to as DiceCELoss.

To assess the effectiveness of our proposal, which involves training models with organ information to enhance accuracy, across various state-of-the-art models, we estimate its applicability on KiTS19 using the conventional UNETR, SwinUNETR, nnFormer, and U-Net [5], [28], [29], [36], both in their original form and with our proposed modification. We evaluate the performance of each conventional model

trained solely with tumor information against each proposed model trained with both tumor and organ information using the Dice score metric, which calculates the overlap between the predicted segmentation and the ground truth.

The results presented in Table 1 indicate that our proposed models that are trained with organ and tumor information yields improved Dice score compared to the conventional models that are trained only with tumor information. The overall our proposed models demonstrated superior performance compared to the conventional models. The proposed UNETR outperforms the conventional UNETR by 34.9%. Also, in case of nnFormer, the proposed model surpasses the conventional model by 14.9%. Particularly noteworthy is the SwinUNETR model, where our proposed modification achieved a Dice score of 0.4786, representing an increase of 103%. The U-Net model proposed in our study demonstrates a 47.0% increase in accuracy compared to the conventional U-Net. These results clearly show that training models with organ information that is explicitly related to the tumor enhances the tumor segmentation ability.

TABLE 1. Performance evaluation with KiTS19 dataset.

Model	Proposed	Conventional
<i>UNETR</i> [36]	0.3282	0.2137
<i>nnFormer</i> [29]	0.4114	0.3501
<i>SwinUNETR</i> [28]	0.4786	0.2311
<i>U-Net</i> [5]	0.2442	0.1294

The Dice scores of OrgUNETR, nnFormer, U-Net and Swin UNETR on KiTS19 dataset. The Dice score is calculated for only tumor prediction. The second column indicates the proposed models with learning both organ and tumor information. The third column indicates the conventional models.

By examining the results from the various models, the approach of simultaneously training on both organ and tumor information is applicable to other models. This indicates that the approach is not only applicable to the models discussed in this paper but can also be extended to other deep learning models. Furthermore, this strategy can be expanded into a general methodology that to detect the target precisely, the related information that related to the target is required.

The results of our OrgUNETR experiments on the KiTS19 dataset are presented in Figure 2. Our OrgUNETR model achieves a Dice score that is 49.04% higher than the baseline model, demonstrating the significant impact of incorporating organ information in tumor localization. The dual-channel approach enables the model to leverage the contextual information provided by the organ labels, leading to more accurate tumor segmentation.

Figure 3 illustrates the training and validation loss curves for both OrgUNETR and the baseline model. Our model demonstrates a substantial reduction in DiceCELoss compared to the baseline model, with a decrease of 37.85%. This observation suggests that training the model with organ

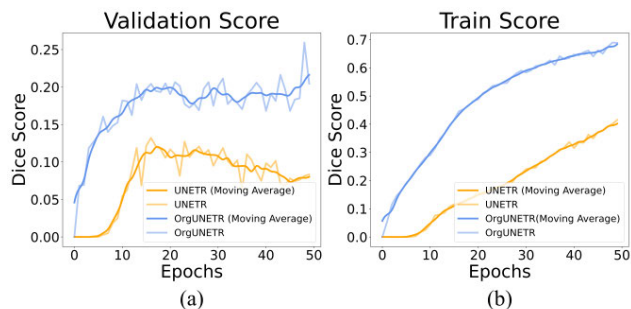


FIGURE 2. Dice score comparison for tumor segmentation on the KiTS19 dataset. (a) depicts the comparison of the OrgUNETR versus UNETR model using the validation dataset, and (b) presents the Dice scores of OrgUNETR versus UNETR on the training dataset. In both (a) and (b), the orange line shows the Dice score from OrgUNETR, while the blue lines show the Dice score from UNETR. The bold lines represent the application of a moving average to enhance clarity.

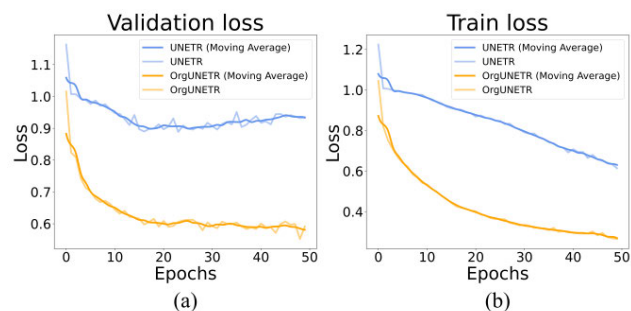


FIGURE 3. Comparison of DiceCELoss for tumor segmentation on the KiTS19 dataset. (a) illustrates the comparison between the OrgUNETR and UNETR models using the validation dataset, while (b) displays the loss for OrgUNETR compared to UNETR on the training dataset. In both (a) and (b), the orange line represents the loss for OrgUNETR, whereas the blue lines indicate the loss for UNETR. Bold lines signify the use of a moving average for clarity.

information enhances its learning capacity by providing additional supervision regarding organ location. The lower validation loss achieved by OrgUNETR indicates its superior performance and generalization ability compared to the baseline model.

One of the key contributions of our work is the replacement of MHSA layers with SE layers in the OrgUNETR architecture. By adopting SE layers that perform channel-wise attention, we achieve a notable reduction in computational complexity while maintaining segmentation accuracy. Specifically, our model exhibits a 13.9% reduction in computational cost compared to the original UNETR architecture, making it more efficient and suitable for practical applications.

To further illustrate the impact of incorporating organ information on tumor segmentation, we present a visual comparison of the segmentation results obtained by OrgUNETR and the baseline model in Figure 4. The first row shows the ground truth segmentation, while the second and third rows display the tumor predictions of OrgUNETR and the baseline model, respectively. The pink pixels represent the

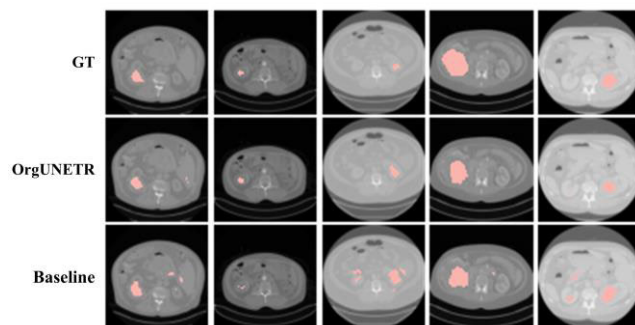


FIGURE 4. Tumor prediction from OrgUNETR and the baseline model on KiTS19 dataset. The first row indicates the ground truth. The second row illustrates the tumor prediction of OrgUNETR. The third row shows the tumor prediction performed by the baseline model. The pink pixels throughout the image represents the tumor pixels, whereas the grayscale pixels indicate the background.

tumor regions, while the grayscale pixels correspond to the background.

In the third column of Figure 4, we observe a significant difference between the predictions of OrgUNETR and the baseline model. For instance, in the second row, OrgUNETR accurately predicts the tumor in the right kidney, whereas the baseline model incorrectly predicts a non-existent tumor in the left kidney. This observation highlights the inferior performance of the baseline model in detecting tumors accurately from CT scans.

Similarly, in the fifth column, OrgUNETR correctly shows the absence of a tumor in the left kidney, while the baseline model incorrectly predicts the presence of tumors in the left kidney. These examples demonstrate the effectiveness of incorporating organ information in improving tumor segmentation accuracy.

Overall, the experimental results on the KiTS19 dataset strongly support the efficacy of our proposed OrgUNETR model in kidney tumor segmentation. By leveraging organ information through a dual-channel approach and employing SE layers for efficient attention mechanisms, OrgUNETR achieves superior performance compared to the baseline model, both in terms of Dice score and visual quality of the segmentation results.

B. SEGMENTATION RESULTS WITH PROSTATE158 DATASET

To further validate the effectiveness of our proposed OrgUNETR model, we conduct experiments on the Prostate158 dataset, which consists of high-quality 3 Tesla MRI scans specifically designed for prostate segmentation tasks. The dataset includes annotations for both anatomical zones and cancerous lesions within the prostate, making it a comprehensive resource for evaluating prostate tumor segmentation models.

The Prostate158 dataset comprises 139 training samples and 19 validation samples. Each MRI scan has a native resolution of $270 \times 270 \times 24$ voxels. Due to computational limitations, we resize the scans to a uniform size

of $128 \times 128 \times 128$ voxels using linear interpolation. During the resizing process, we carefully examine the presence of tumor pixels and exclude 72 samples that lose tumor pixel information as a result of the resizing operation. This ensures that the training dataset maintains a sufficient representation of tumors for effective learning. Additionally, we apply min-max scaling to normalize the intensity values of the MRI scans, bringing them into a consistent range. Using the Prostate158 dataset, we evaluated the performance of each conventional model and the proposed model in the identical approach as the experiment conducted with KiTS19 dataset.

TABLE 2. Performance evaluation with Prostate158 dataset.

Model	PROPOSED	Conventional
UNETR [36]	0.1893	0.1461
nnFormer [29]	0.4256	0.4423
SwinUNETR [28]	0.2893	0.2255
U-Net [5]	0.2123	0.1880

The Dice scores of our OrgUNETR, nnFormer, and SwinUNETR on Prostate158 dataset. The Dice score is calculated for only tumor prediction. The second column indicates the proposed models with learning both organ and tumor information. The third column indicates the conventional models.

Table 2 presents a comparative analysis of the Dice scores achieved by various models on Prostate158 dataset. In case of U-Net, the proposed model surpasses the conventional model by 11.4%. For the UNETR and SwinUNETR, the proposed models show superior performance by 22.8% and 22.1% respectively. Contrary to other models, the nnFormer demonstrated better performance in the conventional model. However, the overall architectures of the proposed models show superior performance relative to the conventional models. Through experiments conducted with Prostate158 dataset, we confirm the applicability of our model across various models.

Especially for OrgUNETR, The training and validation sets are split in a ratio of 70 to 30. We train the OrgUNETR model using the AdamW optimizer with a learning rate of 0.0001. To enhance the model's robustness, we employ data augmentation techniques, specifically random rotation of the input images within a range of 0 to 30 degrees.

The performance of OrgUNETR is evaluated using the Dice score metric, and we compare it against a baseline model that focuses solely on tumor localization using a single channel. Figure 5 presents the comparison of Dice scores between OrgUNETR and the baseline model on the Prostate158 dataset. The left plot shows the Dice scores on the validation set, while the right plot displays the Dice scores on the training set.

Our OrgUNETR model achieves a Dice score that is 32.69% higher than the baseline model, demonstrating the significant impact of incorporating organ information in prostate tumor segmentation. Interestingly, we observe that

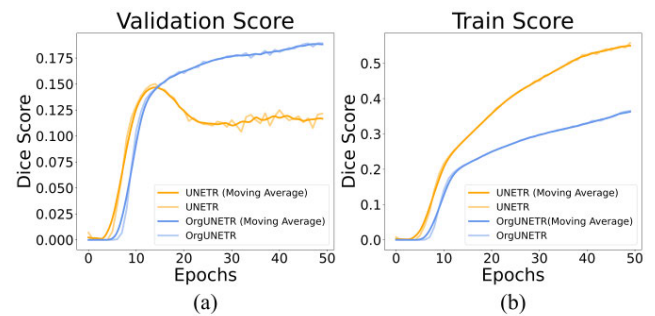


FIGURE 5. Dice score comparison for tumor segmentation on the Prostate 158 dataset. (a) depicts the comparison of the OrgUNETR versus UNETR model using the validation dataset, and (b) presents the Dice scores of OrgUNETR versus UNETR on the training dataset. In both (a) and (b), the orange line shows the Dice score from OrgUNETR, while the blue lines show the Dice score from UNETR. The bold lines represent the application of a moving average to enhance clarity.

the training Dice score of the baseline model surpasses that of OrgUNETR. However, when evaluated on the validation set, OrgUNETR consistently outperforms the baseline model. This observation suggests that OrgUNETR is more effective in generalizing to unseen data and is less prone to overfitting compared to the baseline model.

Figure 6 illustrates the training and validation loss curves for both OrgUNETR and the baseline model on the Prostate158 dataset. Our model demonstrates a substantial reduction in DiceCELoss compared to the baseline model, with a decrease of 43.39%. This observation underscores the benefit of incorporating organ information into the segmentation process, leading to more accurate tumor predictions. It is worth noting that the training DiceCELoss of both models shows only a 3.44% difference, indicating that both models are trained similarly. However, the superior performance of OrgUNETR on the validation and test sets highlights its ability to generalize well and make accurate predictions on unseen data.

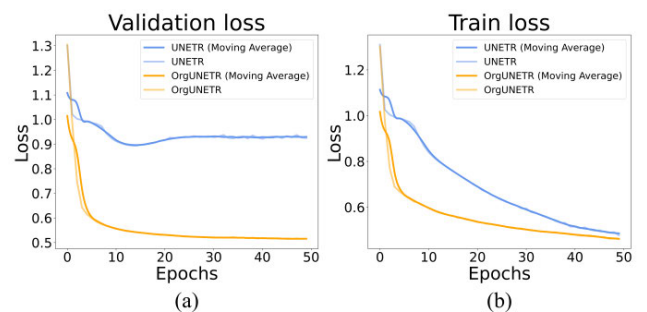


FIGURE 6. Comparison of DiceCELoss for tumor segmentation on the Prostate 158 dataset. (a) illustrates the comparison between the OrgUNETR and UNETR models using the validation dataset, while (b) displays the loss for OrgUNETR compared to UNETR on the training dataset. In both (a) and (b), the orange line represents the loss for OrgUNETR, whereas the blue lines indicate the loss for UNETR. Bold lines signify the use of a moving average for clarity.

To provide a qualitative assessment of the segmentation results, we present representative examples in Figure 7.

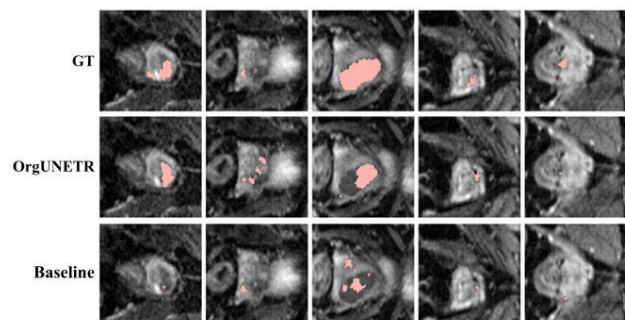


FIGURE 7. Tumor prediction from OrgUNETR and the baseline model on Prostate 158 dataset. The first row indicates the ground truth. The second row illustrates the tumor prediction of OrgUNETR. The third row shows the tumor prediction performed by the baseline model. The pink pixels throughout the image represents the tumor pixels, whereas the greyscale pixels indicate the background.

The first row shows the ground truth segmentation, while the second and third rows display the tumor predictions of OrgUNETR and the baseline model, respectively. The pink pixels represent the tumor regions, while the grayscale pixels correspond to the background.

In the third column of Figure 7, we observe significant differences between the predictions of OrgUNETR and the baseline model. OrgUNETR accurately predicts a large tumor in the middle of the prostate, closely resembling the ground truth. In contrast, the baseline model not only predicts the tumor in the middle of the prostate but also incorrectly identifies additional tumor regions. These examples demonstrate the superior performance of OrgUNETR in accurately segmenting prostate tumors by leveraging organ information.

The experimental results on the Prostate158 dataset further validate the effectiveness of our proposed OrgUNETR model in tumor segmentation tasks. By incorporating organ information through a dual-channel approach, OrgUNETR achieves significant improvements in Dice score and visual quality of the segmentation results compared to the baseline model. The model's ability to generalize well to unseen data and its robustness to overfitting make it a promising tool for prostate tumor segmentation in clinical practice.

C. ADDITIONAL EXPERIMENTS

To thoroughly evaluate the robustness and performance of our OrgUNETR model, we conducted an extensive series of experiments across various hyperparameter configurations. Initially, the number of channels in the decoder layer was set to 16, and the learning rate was fixed at 0.0001. We then explored different embedding dimensions for the input patches, specifically 8, 16, and 32, to analyze their impact on model performance.

Our experimental setup employed the KiTS19 and Prostate158 datasets, which are well-regarded benchmarks for assessing the efficacy of medical image segmentation models. The performance of our model was quantified using the Dice score, a standard metric for evaluating the accuracy of segmentation models.

TABLE 3. Performance evaluation with different embedding dimensions.

Embedding Dimension	KiTS19	Prostate158
8	0.1323	0.1449
16	0.2137	0.2102
32	0.1120	0.2195

The Dice scores of OrgUNETR on KiTS19 and Prostate158 datasets with embedding dimension of 8, 16, and 32.

Table 3 presents the Dice scores achieved by OrgUNETR across the different embedding dimensions. Notably, even with varying hyperparameters, our model consistently demonstrated robust performance. For instance, with an embedding dimension of 16, OrgUNETR attained a Dice score of 0.2137 on the KiTS19 dataset, which is the highest performance observed for this dataset. On the Prostate158 dataset, the model achieved its peak performance with an embedding dimension of 32, recording a Dice score of 0.2195. It is important to highlight that while the highest Dice scores were observed at embedding dimensions of 16 and 32, the scores at dimension 8 also showed commendable performance, indicating the model's stability and efficiency across different configurations.

The marginal convergence of Dice scores between dimensions 16 and 32 further underscores the model's robustness. Despite the variation in embedding dimensions, the performance remained consistently high, demonstrating the effectiveness and reliability of OrgUNETR in handling complex medical image segmentation tasks. This consistent performance, irrespective of the embedding dimension, attests to the superior design and implementation of our model.

In conclusion, the experimental results confirm that OrgUNETR performs exceptionally well across different hyperparameter settings. The consistent Dice scores across varying embedding dimensions indicate that our model maintains high performance regardless of specific parameter adjustments. This robustness highlights the potential of OrgUNETR as a reliable tool for medical image segmentation, capable of delivering accurate and consistent results.

V. CONCLUSION

In this study, we introduced OrgUNETR, an enhanced version of the UNETR architecture specifically designed for tumor segmentation in medical images. The proposed model incorporates organ context information to improve the accuracy and robustness of tumor localization. By leveraging the fact that tumors typically exist within specific organs, OrgUNETR effectively addresses the challenges posed by the small size and unpredictable locations of tumors in CT and MRI scans.

One of the key contributions of OrgUNETR is its ability to simultaneously segment both the organ and the tumor using a dual-channel approach. This approach significantly

improves tumor segmentation performance by allowing the model to learn the inherent relationships between organs and tumors. The experimental results on the KiTS19 and Prostate158 datasets demonstrate the effectiveness of incorporating organ information, with OrgUNETR achieving substantial improvements in Dice score compared to a baseline model that focuses solely on tumor segmentation.

On the KiTS19 dataset, which consists of CT scans of the kidney and kidney tumors, OrgUNETR achieved a remarkable 40.54% increase in Dice score for tumor segmentation when organ information was included. Similarly, on the Prostate158 dataset, which contains MRI scans of the prostate gland and prostate tumors, OrgUNETR outperformed the baseline model by 32.69% in terms of Dice score. These results provide strong evidence for the benefits of leveraging organ context in tumor segmentation tasks.

In addition to the performance gains, we also optimized the computational efficiency of OrgUNETR by replacing the MHSA layers with SE layers. The SE layers efficiently compute channel-wise attention, reducing the computational complexity of the model while maintaining its segmentation accuracy. By substituting MHSA layers with SE layers, we achieved a 13.9% reduction in computational cost, making OrgUNETR more practical for real-world applications with limited computational resources.

The superior performance of OrgUNETR can be attributed to its ability to capture both local and global contextual information. The encoder of OrgUNETR, which consists of a series of SE blocks, effectively compresses spatial information while extracting relevant features at different scales. The decoder, on the other hand, reconstructs the segmented output by integrating the extracted features through skip connections and upsampling operations. This architecture enables OrgUNETR to generate precise and coherent segmentation results, even for challenging cases with small and irregularly shaped tumors.

Furthermore, the inclusion of organ information in the segmentation process helps to mitigate the issue of false positives, where the model incorrectly identifies non-tumor regions as tumors. By learning the relationships between organs and tumors, OrgUNETR is able to distinguish between normal anatomical structures and abnormal growths more effectively. This is particularly important in clinical settings, where accurate tumor detection and delineation are crucial for treatment planning and patient management.

The experimental results also highlight the generalization ability of OrgUNETR. Despite the variations in tumor size, shape, and location across different patients and imaging modalities, OrgUNETR consistently outperforms the baseline model. This robustness is essential for the practical deployment of the model in real-world scenarios, where it may encounter a wide range of tumor characteristics.

ACKNOWLEDGMENT

(Sanghyuk Roy Choi and Jungro Lee contributed equally to this work.)

REFERENCES

- [1] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.
- [2] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. Int. Conf. Eng. Technol.*, 2017, pp. 1–6.
- [3] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *Proc. 13th Int. Conf. Control Automat. Robot. Vis.*, Dec. 2014, pp. 844–848.
- [4] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015, *arXiv:1511.08458*.
- [5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," pp. 234–241.
- [6] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. 4th Int. Workshop*, 2018, pp. 3–11.
- [7] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, and V. M. Patel, "KiU-Net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 4, pp. 965–976, Apr. 2022.
- [8] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Nov. 2018, pp. 7794–7803.
- [9] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–12.
- [10] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jul. 2017, pp. 764–773.
- [11] P. Freire, S. Srivallapanondh, B. Spinnler, A. Napoli, N. Costa, J. E. Prilepsky, and S. K. Turitsyn, "Computational complexity optimization of neural network-based equalizers in digital signal processing: A comprehensive approach," *J. Lightw. Technol.*, pp. 1–25, Apr. 10, 2024.
- [12] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2021, pp. 36–46.
- [13] J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du, "ViT-V-Net: Vision transformer for unsupervised volumetric medical image registration," 2021, *arXiv:2104.06468*.
- [14] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surveys*, vol. 54, no. 10, pp. 1–41, Jan. 2022.
- [15] S. Hoon Lee, S. Lee, and B. Cheol Song, "Vision transformer for small-size datasets," 2021, *arXiv:2112.13492*.
- [16] T. Yao, Y. Li, Y. Pan, Y. Wang, X.-P. Zhang, and T. Mei, "Dual vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 1–13, Sep. 2023.
- [17] X. Cheng, X. Li, J. Yang, and Y. Tai, "SESr: Single image super resolution with recursive squeeze and excitation networks," in *Proc. 24th Int. Conf. Pattern Recognit.*, Aug. 2018, pp. 147–152.
- [18] S. R. Choi and M. Lee, "Estimating the prognosis of low-grade glioma with gene attention using multi-omics and multi-modal schemes," *Biology*, vol. 11, no. 10, p. 1462, Oct. 2022.
- [19] J. Gu, X. Sun, Y. Zhang, K. Fu, and L. Wang, "Deep residual squeeze and excitation network for remote sensing image super-resolution," *Remote Sens.*, vol. 11, no. 15, p. 1817, Aug. 2019.
- [20] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
- [21] M. Lee, "An ensemble deep learning model with a gene attention mechanism for estimating the prognosis of low-grade glioma," *Biology*, vol. 11, no. 4, p. 586, Apr. 2022.
- [22] S. K. Roy, S. R. Dubey, S. Chatterjee, and B. Baran Chaudhuri, "FuSENet: Fused squeeze-and-excitation network for spectral-spatial hyperspectral image classification," *IET Image Process.*, vol. 14, no. 8, pp. 1653–1661, Jun. 2020.
- [23] R. Ranjbarzadeh, A. B. Kargari, S. J. Ghouschi, S. Anari, M. Naseri, and M. Bendechache, "Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images," *Sci. Rep.*, vol. 11, no. 1, pp. 1–17, May 2021.

- [24] T. Fujimaki, M. Matsutani, N. Funada, T. Kirino, K. Takakura, O. Nakamura, A. Tamura, and K. Sano, "CT and MRI features of intracranial germ cell tumors," *J. Neuro-Oncol.*, vol. 19, no. 3, pp. 217–226, 1994.
- [25] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Sci. Data*, vol. 4, no. 1, pp. 1–13, Sep. 2017.
- [26] M. Aggarwal, A. K. Tiwari, M. P. Sarathi, and A. Bijalwan, "An early detection and segmentation of brain tumor using deep neural network," *BMC Med. Informat. Decis. Making*, vol. 23, no. 1, p. 78, Apr. 2023.
- [27] H. Jiang, Z. Diao, and Y.-D. Yao, "Deep learning techniques for tumor segmentation: A review," *J. Supercomput.*, vol. 78, no. 2, pp. 1807–1851, Feb. 2022.
- [28] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *Proc. Int. MICCAI Brainlesion Workshop*, 2021, pp. 272–284.
- [29] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "NnFormer: Interleaved transformer for volumetric segmentation," 2021, *arXiv:2109.03201*.
- [30] K. Subramanyam Kalyan, A. Rajasekharan, and S. Sangeetha, "AMMUS: A survey of transformer-based pretrained models in natural language processing," 2021, *arXiv:2108.05542*.
- [31] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, and M. Funtowicz, "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2020, pp. 38–45.
- [32] S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," in *Proc. International Conf. Mach. Learn.*, Jul. 2296, pp. 2286–2296.
- [33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [34] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.
- [35] H. Chu, L. R. De la O Arevalo, W. Tang, B. Ma, Y. Li, A. De Biase, S. Both, J. A. Langendijk, P. van Ooijen, and N. M. Sijtsema, "Swin UNETR for tumor and lymph node segmentation using 3D PET/CT imaging: A transfer learning approach," in *3D Head Neck Tumor Segmentation PET/CT Challenge*. Cham, Switzerland: Springer, 2022, pp. 114–120.
- [36] 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 *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Nov. 2022, pp. 574–584.
- [37] H. Tao, K. Mao, and Y. Zhao, "DBT-UNETR: Double branch transformer with cross fusion for 3D medical image segmentation," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Dec. 2022, pp. 1213–1218.
- [38] P.-C. Chen, H. Tsai, S. Bhojanapalli, H. Won Chung, Y.-W. Chang, and C.-S. Ferng, "A simple and effective positional encoding for transformers," 2021, *arXiv:2104.08698*.
- [39] Y.-A. Wang and Y.-N. Chen, "What do position embeddings learn? An empirical study of pre-trained language model positional encoding," 2020, *arXiv:2010.04903*.
- [40] J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimizing the dice score and Jaccard index for medical image segmentation: Theory and practice," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2019, pp. 92–100.
- [41] S. Ghosal, A. Xie, and P. Shah, "Uncertainty quantified deep learning for predicting dice coefficient of digital histopathology image segmentation," 2021, *arXiv:2109.00115*.
- [42] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Proc. 3rd Int. Workshop*, 2017, pp. 240–248.
- [43] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.
- [44] L. C. Adams, M. R. Makowski, G. Engel, M. Rattunde, F. Busch, P. Asbach, S. M. Niehues, S. Vinayahalingam, B. van Ginneken, G. Litjens, and K. K. Bressen, "Prostate158—An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection," *Comput. Biol. Med.*, vol. 148, Sep. 2022, Art. no. 105817.
- [45] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
- [46] J. Shijie, W. Ping, J. Peiyi, and H. Siping, "Research on data augmentation for image classification based on convolution neural networks," in *Proc. Chin. Automat. Congr.*, 2017, pp. 4165–4170.



SANGHYUK ROY CHOI received the bachelor's degree from the School of Electrical and Electronics Engineering, Chung-Ang University, in 2021, where he is currently pursuing the Ph.D. degree with the Department of Intelligent Semiconductor Engineering. His research interests include medical image segmentation and medical image neural radiance fields.



JUNGRO LEE is currently pursuing the bachelor's degree in electrical engineering with Chung-Ang University, Seoul, South Korea. His research interests include applying various artificial intelligence models to medical data and implicit neural representation.



MINHYEOK LEE (Member, IEEE) received the bachelor's and Ph.D. degrees in electrical engineering from Korea University, in 2015 and 2020, respectively. He has been an Assistant Professor with the School of Electrical and Electronics Engineering, Chung-Ang University, since 2021, and concurrently holds a position with the Department of Intelligent Semiconductor Engineering, since 2023. Prior to these appointments, he was a Research Professor with Korea University, from 2020 to 2021. Over the past three years, he has published over 40 papers in international journals and conferences, primarily focusing on artificial intelligence and generative AI. His research interests include generative artificial intelligence, generative adversarial networks (GANs), diffusion models, neural radiance fields (NeRF), AI-based finance engineering, and bioinformatics.

...