2023 26th International Conference on Computer and Information Technology (ICCIT)
13-15 December, Cox's Bazar, Bangladesh

2023 26th International Conference on Computer and Information Technology (ICCIT) | 979-8-3503-5901-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCIT60459.2023.10441092

Attention-refined U-Net with Skip Connections for Effective Brain Tumor Segmentation from MRI Images

A. F. M. Minhazur Rahman*, and Md. Ali Hossain†
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology, Rajshahi-6204, Bangladesh
Email- *m.r.saurov@gmail.com, †ali.ruet@gmail.com,

Abstract—**Brain tumor segmentation in MRI scans is a crucial yet challenging task due to the high variability in tumor shape, size, and location. Accurate and efficient segmentation is of paramount importance for timely diagnosis and effective treatment planning in brain tumor patients. In response to these challenges, this paper presents an enhanced U-Net architecture named SC-SE U-Net. Our proposed model integrates Squeeze-and-Excitation channel attention blocks and skip connections similar to residual network architecture, into the traditional U-Net framework. The Squeeze-and-Excitation blocks enhance the model's capability to focus on more relevant feature maps and suppress less pertinent ones, thereby improving the overall segmentation accuracy. Simultaneously, the skip connections facilitate the flow of gradients during the training process, leading to more stable and faster convergence, and improved performance. We conducted an extensive evaluation of our proposed model on the TCGA-LGG brain tumor dataset. The results show that the SC-SE U-Net outperforms several existing segmentation methods, achieving an Intersection over Union score of 84.22% and a Dice score of 91.43%. The impressive performance of SC-SE U-Net underscores its potential to significantly improve diagnostic efficiency in clinical settings, highlighting the importance of further research in this direction.**

Index Terms—**Magnetic Resonance Imaging, Medical Image Segmentation, Brain Tumor, Skip Connection, Attention Mechanism, Semantic Segmentation, Low-Grade Glioma**

I. Introduction

A brain tumor refers to an abnormal mass or growth of cells that develops within the brain. The growth of brain tumors can vary in pace, often displaying slow growth and limited spread to other parts of the body. However, cancerous brain tumors, characterized by uncontrolled and rapid growth with a high propensity to spread, are rare but extremely lethal [1]. According to the statistics, more than 300 thousand people were diagnosed with a primary brain and spinal tumor in 2020 [2].

Effective treatment and prediction of the growth rate of cancerous brain tumors are essential for improving patient survival probability. Achieving this goal involves the segmentation of brain tumor tissues from normal tissues. Segmentation requires detailed images of the brain with soft tissue contrast, possible via Magnetic Resonance Imaging (MRI) technique [3]. Compared to alternative imaging modalities such as Computed Tomography (CT), MRI sensors excel in capturing the contrast between different brain tissues.

Manual brain tumor segmentation is a laborious and challenging task performed by medical experts. It involves outlining the tumor's boundaries within MRI scans, a process made difficult by the natural variations in brain structures, and the heterogeneity of tumors. This manual process is time-consuming and susceptible to inter-observer variability, where different experts may segment the same tumor differently, hampering consistency and scalability. Automated methods are increasingly sought after to streamline and improve the accuracy of this critical task.

Computer-automated segmentation of brain tumors evolved from traditional algorithmic approaches to highly feature engineering dependent machine learning approaches. Machine learning approaches require intensive and manual feature engineering to ensure adequate features from MRI scans can be extracted for semantic segmentation i.e. pixel-wise classification of input image. Deep learning approaches, on the other hand, require far less feature engineering. Automatic pattern learning and feature extraction of deep learning-based approaches have led to a revolution in computer vision tasks such as semantic segmentation, instance segmentation, object detection, classification, image restoration etc. However, deep learning based approaches require a complicated model training process compared to traditional computer vision tasks. They may also encounter vanishing gradient problem due to poor flow of gradient in the network during training. These approaches are also susceptible to less-than-optimal feature extraction. In computer vision applications, particularly in tasks like semantic segmentation, it is customary to employ wider neural network architectures. Nonetheless, the increased width often necessitates additional feature refinement to attain the desired level of performance excellence. Therefore, developing a model for accurately segmenting brain MRI images that avoids the vanishing gradient problem and can refine features for enhanced performance holds significant importance.

A. Related Works

Early deep learning architectures for semantic segmentation were incapable of being trained end-to-end using the full input image. These architectures relied on patch-wise training, a method that divided the input data into smaller, non-overlapping patches upon which the network was trained [4].

979-8-3503-5901-5/23/$31.00 ©2023 IEEE

Long et al. [5] introduced a significant deviation from the conventional method of patch-wise training with Fully Convolutional Networks (FCN). FCNs employ the entire input image for end-to-end training, departing from the widely adopted patch-wise training approach. A critical feature of FCNs is their ability to upsample lower-stage features and concatenate them with higher-stage features. This operation enhances the precision of the output segmentation mask. Ronnenberger et al. [6] further refined this concept by proposing the U-Net, a symmetric encoder-decoder architecture. The U-Net architecture utilizes skip connections, referred to as long skip connection, from the encoder network to supplement the decoder by providing high-resolution encoder features. These features play a critical role in restoring the full spatial resolution at the network's output layer.

In addition to long skip connections from the encoder to the decoder, skip connections around non-linear layers, comparable to the approach used in Residual Networks [7] has demonstrated to improve model convergence speed and overall performance [8]. The concept of short skip connections between layers was the inspiration for Wu et al. [9] to propose the Skip Connection Unet (SC U-net) architecture. SC U-net introduced a novel skip connection between the up-convolution and down-convolution components of the U-Net to enhance convergence speed and improve segmentation performance compared to the standard U-Net architecture.

In encoder-decoder architectures for semantic segmentation, the regulation of feature map flow from the encoder to the decoder plays an essential role. This can be effectively managed via attention mechanisms. Attention mechanisms in deep learning have the capability to refine feature maps by assigning more weight to features that are crucial for segmentation or classification tasks, while simultaneously reducing the emphasis on less relevant features. The adoption of the attention mechanism has demonstrated promising results in various studies [10]. For instance, Oktay et al. [11] proposed an attention-gating mechanism that leverages attention gates (AGs) to filter the features transmitted through the long skip connections in U-Net. This method uses information extracted from a coarse scale for gating to distinguish between irrelevant and noisy responses in skip connections. However, this approach did not take into account the attention in the channel dimension, which is a crucial factor for localizing objects accurately.

It is evident from the literature that semantic segmentation of MRI images could greatly benefit from an enhanced U-Net architecture, incorporating skip connections and feature refinement through channel attention. This research is motivated by the potential for improvements in both the segmentation task and, consequently, the probability of enhancing patient survival in cases of brain cancer.

B. Our Contributions

In this research work, we introduce a modified U-Net model named SC-SE U-Net, which combines skip connections similar to residual networks and channel attention-based refinement of features. For attention-based refinement, our model integrates Squeeze and Excitation (SE) blocks, designed to selectively enhance or suppress feature maps within convolution and transposed convolution outputs. We conducted experiments using a dataset of low-grade glioma brain tumors to showcase the effectiveness of our attention mechanism and skip connections in the context of semantic segmentation.

II. Methodology

This section provides an in-depth exploration of the foundational elements underpinning our proposed model, encompassing the U-Net architecture, the Skip Connection U-Net, and the integration of the Squeeze-and-Excitation network for channel attention. Within each following subsection, we describe the fundamental concepts and articulate our model's adaptation or modification to these architectural concepts.

A. U-Net Architecture

We use the U-Net [6] architecture as a starting point for our proposed model SC-SE U-Net. U-Net is a fully convolutional neural network designed for medical image segmentation. It has a symmetrical U-shape with an encoder, decoder, and a bottleneck layer. The encoder extracts spatial features via convolutional layers, ReLU activation, and max-pooling. The bottleneck captures high-level semantic data, linking the encoder and decoder. The decoder restores spatial resolution using upsampling, concatenation with encoder feature maps, and convolutional layers with ReLU activations. Skip connections (also called long skip connections) between encoder and decoder preserve details and reduce information loss. The final 1×1 convolution layer assigns pixel-wise class probabilities using softmax (multi-class segmentation) or sigmoid (binary segmentation), producing a segmented image.

In Figure 2, the standard U-Net architecture is depicted as the base configuration. Notably, when we omit the SE blocks and skip connections (emphasized by bold green lines), and element-wise summation of feature maps (shown in $\oplus$ symbol), the resultant representation corresponds to the traditional U-Net architecture.

We build a leaner model by utilizing fewer feature maps compared to those in the papers [6], [9]. In each convolution operation, we employ 2^k filters of size 3×3 (excluding the output layer, where we use one 1×1 filter), where $k \in \{4, 5, 6, 7, 8\}$. The transposed convolution operations utilize 2^k filters of size 2×2, where $k \in \{4, 5, 6, 7\}$. Using a lower network width compared to the original U-Net architecture both decreases training time and reduces overfitting in small datasets.

The standard U-Net architecture can be improved upon by introducing short skip connection in conjunction with existing long skip connection [8].

B. Skip Connection U-Net

Inspired by Skip Connection U-net (SC U-net) architecture [9], we have enhanced the standard U-Net architecture through the introduction of short skip connections between the encoder and decoder pathways. In Fig. 2, the bold green lines highlight these short skip connections. The output feature maps from the 2D-Maxpooling operation in the encoder are added element-wise, with the outputs from the second convolution layer in each decoder stage. These skip connections, similar to residual connections in ResNet [7], facilitate the flow of gradients through the network during training, thus improving optimization and convergence speed. Moreover, they help the decoder to more effectively utilize encoder features, resulting in enhanced localization and detection accuracy.

C. Attention Mechanism: Squeeze-and-Excitation Networks

In addition to the advantages gained from the inclusion of skip connections, we have incorporated Squeeze-and-Excitation (SE) [12] blocks into our proposed network to further augment the model's capacity to learn valuable representations.

Squeeze-and-Excitation is a channel attention mechanism that improves the representational power of convolutional neural networks (CNN) by explicitly modelling interdependencies between channels. In Standard CNNs, the convolution operation fuses spatial and channel information within local receptive fields at each layer. However, dependencies between channels are only implicitly embedded in the learned filters and there is no mechanism to explicitly model channel relationships.

The SE block, as shown in Fig. 1, introduces a lightweight gating mechanism that performs feature recalibration by modeling channel interdependencies. It consists of a squeeze operation that uses global average pooling to produce channel-wise statistics, capturing the global receptive field. Subsequently, an excitation operation using fully connected layers and sigmoid activation generates channel-specific gains. The SE block adaptively recalibrates channel-wise responses by rescaling feature maps with these gains, thereby exploiting global information.

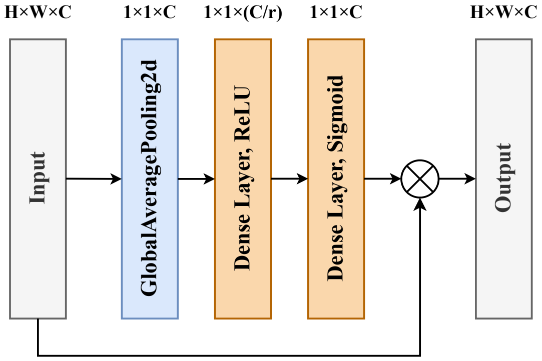

Fig. 1. Squeeze-and-Excitation (SE) block. SE block takes a 3D tensor as input and outputs an attention rescaled version of the input tensor.

SE block performs two key operations— Squeeze and Excitation.

Squeeze Operation: This step aggregates information across spatial dimensions to generate a channel-wise descriptor. It typically involves global average pooling to obtain a single value for each channel. Let, X be the input feature tensor with dimensions $H \times W \times C$, where H is the height, W is the width and C is the number of channels. Z is the output of the squeeze operation, which results in a C-dimensional vector. The squeeze operation can be mathematically represented as:

$$Z_c = \text{GlobAvgPool}\left(X_c\right) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{c,i,j} \tag{1}$$

Where Z_c is the value for channel c in the output vector Z.

Excitation Operation: This step learns how to weight each channel's importance and generates scaling factors for each channel. These scaling factors are applied to the original feature maps to enhance informative channels and suppress less informative ones. Let S be the scaling factor vector of size C, where each element S_c is a learned weight for channel c. Y is the output feature tensor after the excitation operation. The excite operation can be mathematically represented as:

$$S_c = \sigma\left(W_2 \delta\left(W_1 Z_c\right)\right) \tag{2}$$

$$Y_c = S_c \cdot X_c \tag{3}$$

W_1 and W_2 are learnable weights implemented as fully connected layers. δ represents an activation function like ReLU. Number of units in first fully connected layer is C/r where r is the reduction ratio. We choose $r = 16$ according to the suggestion of the original paper. σ represents a sigmoid activation function to ensure the scaling factors are in the range between 0 and 1.

Integrating the SE blocks in U-Net architecture will allow the network to increase sensitivity to informative features and suppress less useful ones. We insert the SE blocks after every convolutional block in the encoder and before each transposed convolution block in the decoder. In encoder, SE blocks help the network to emphasize or suppress certain feature maps based on the information they contain, improve the quality of feature extraction. Placing SE blocks before each transposed convolution block in the decoder can help the network focus on more relevant features when reconstructing the output segmentation map. Our final SC-SE U-Net network with skip connection and SE attention blocks is shown in Fig 2.

III. Dataset

A. TCGA-LGG Dataset

We employed the TCGA-LGG Segmentation dataset [13] to train our segmentation model. This dataset comprises 3929 brain tumor images, each paired with a corresponding segmentation mask that outlines FLAIR image abnormalities. The images are sourced from 110 patients included in the glioma collection of The Cancer Genome Atlas (TCGA) lower-grade group. In total, there are 110 patients, with each patient contributing multiple brain MRI scans and expert-annotated masks delineating regions of FLAIR abnormality indicative of the presence and spatial extent of lower grade gliomas.

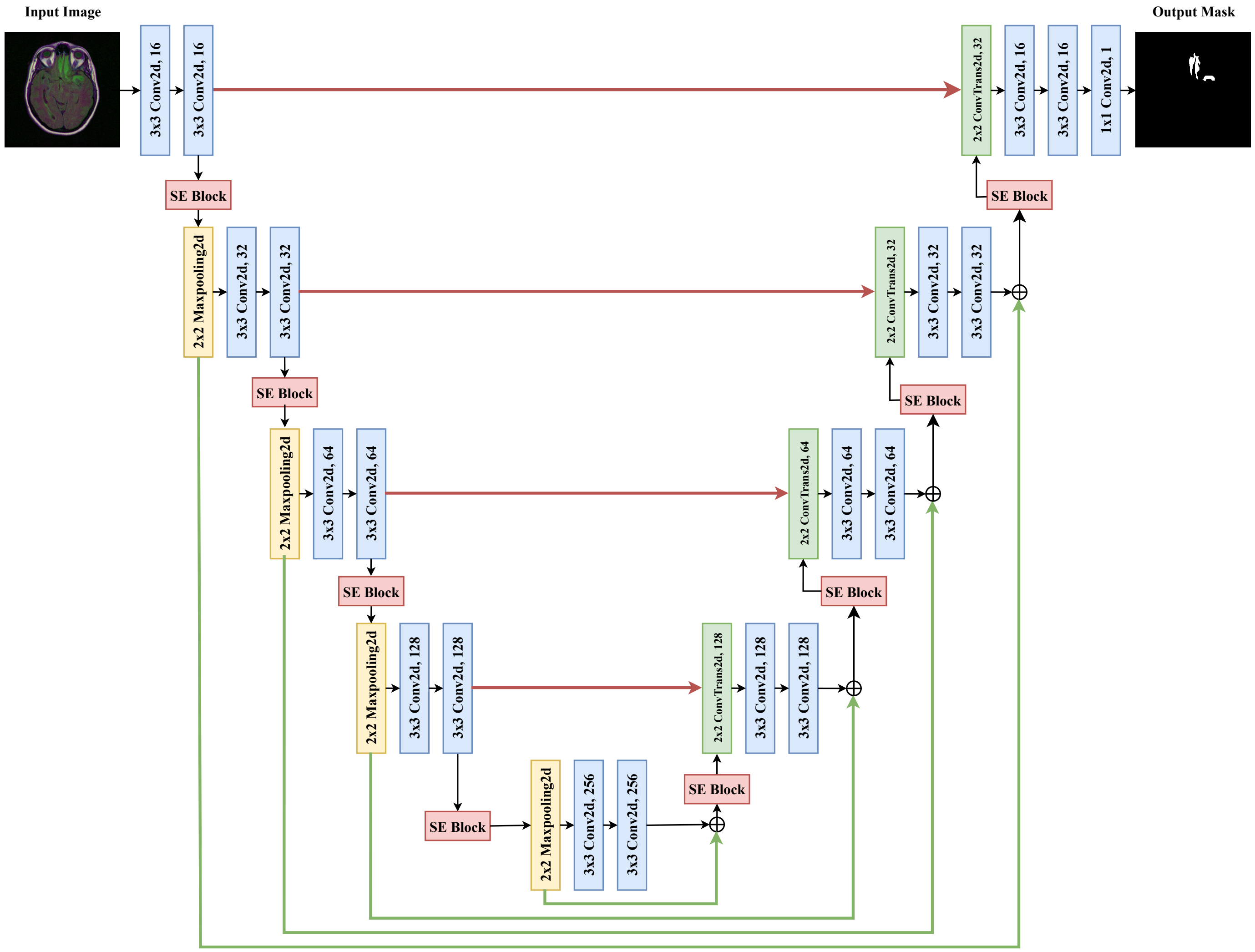

Fig. 2. SC-SE U-Net architecture. Red arrow lines indicate long skip connections from the encoder to the decoder, while green arrow lines indicate short skip connections.

The images in the dataset are RGB with dimensions of $256 \times 256 \times 3$. The corresponding segmentation masks are single channel images with dimensions of $256 \times 256 \times 1$, where each pixel has a value of either 0 or 255. A pixel value of 0 indicates a non-tumorous region, while a value of 255 indicates a tumorous region belonging to the abnormality.

B. Data Augmentation and Splitting

TCGA-LGG dataset contains a relatively small number of samples. 2556 images from the dataset have no tumor region. This imbalance can potentially skew the model's learning process, leading to less accurate predictions on unseen data. To overcome this limitation and enhance the robustness of our model, we artificially increase the number of images by augmenting the training set during the model training process. This augmentation process includes techniques such as random width and height shifts, rotation, and zooming.

Images from the TCGA-LGG dataset were divided into distinct sets, with 80% reserved for training and 20% for testing. Furthermore, within the training set, an additional 20% subset was set aside specifically for validation purposes. This validation set will be used for fine-tuning hyperparameters and facilitating model selection. Data augmentation was applied after splitting the original dataset to avoid data leakage into the test set. Only the train set was augmented.

The primary reason for using augmentation is to diversify the training data, ensuring the model is exposed to a wide array of scenarios. This approach allows the model to better generalize to unseen data, reducing the likelihood of overfitting and improving overall performance.

IV. Experiments and Results

A. Hardware Configuration and Important Hyperparameters

For our deep learning experiments, we utilized a hardware configuration consisting of an Intel Core i9 11900 CPU, 32GB of RAM, and an NVIDIA RTX A4500 GPU. This setup allowed us to efficiently train our models and tune important hyperparameters without significant waiting time.

The model architectures were implemented using the Keras deep learning framework. The models were trained for a maximum of 150 epochs, optimized with the Adam optimizer, and utilized a batch size of 32. To mitigate overfitting and expedite training, an early stopping mechanism was implemented, with a patience threshold set at 30 epochs. We selected the Dice loss function due to its suitability for our task. For monitoring hyperparameter tuning, we selected the validation IoU score of the trained model.

B. Evaluation Metrics

The research utilizes four metrics to assess image segmentation quality: Intersection over Union (IoU), Dice Coefficient, Precision and Recall.

1) Intersection over Union (IoU): IoU, or the Jaccard index, measures the overlap between predicted and actual segmentations. It's calculated as the ratio of the intersection to the union of both segmentations.

$$\text{IoU Score} = \frac{\text{Prediction} \cap \text{Ground Truth}}{\text{Prediction} \cup \text{Ground Truth}} \tag{4}$$

A higher IoU score indicates better alignment between predicted and actual masks.

2) Precision and Recall: Precision is a measure of the accuracy of the positive predictions made by a model. It is the ratio of correctly predicted positive observations to the total predicted positive observations. Higher precision indicates fewer false positives.

Recall, also known as sensitivity or true positive rate, is the ratio of correctly predicted positive observations to all actual positive observations. Higher recall indicates fewer false negatives.

3) Dice Score: The Dice Score, or F1-score, is the harmonic mean of precision and recall. It can also be calculated as twice the intersection area divided by the sum of the areas of both segmentations.

$$\text{Dice Score} = \frac{2(\text{Prediction} \cap \text{Ground Truth})}{\text{Prediction} + \text{Ground Truth}} \tag{5}$$

Like IoU, a higher Dice score indicates better segmentation performance.

C. Result Analysis

TABLE I
CONFUSION MATRIX FOR SC-SE U-NET ON TEST SET

		Predicted	
		Non-tumor	**Tumor**
True	**Non-tumor**	50958769	40850
	Tumor	46368	465309

We trained our model and three others on the TCGA-LGG dataset, following to the hyperparameters specified in Section IV-A. All models underwent training on the same training set and were subsequently tested on the same test set.

The performance of our SC-SE U-Net model is shown by the confusion matrix in Table I, derived from the aggregated predictions and ground truths of 786 test images. The model achieved an Intersection over Union (IoU) score of 99.82% for the non-tumor class and 84.21% for the tumor class, with the high IoU for non-tumor largely attributed to the predominance of non-tumorous pixels in the ground truths. The mean IoU score for both classes is 92.02%, indicating accurate pixel-wise classification across both classes. Precision of 91.93% in our binary segmentation task indicates that the model correctly identified most of the tumor areas and made few false positive errors. The Recall of 90.93% suggests that the model was successful in detecting a large majority of tumor areas in the images, missing only a small fraction. Dice score of the model is 91.43%. The scores highlight the model's robustness and high-quality performance. The IoU of the tumor class, being more relevant in binary segmentation tasks, will be compared with other methods in Section IV-D.

D. Comparison with Other Methods

Table II contrasts our model with three other methods. The standard U-Net [6], our baseline, achieved an IoU of 78.81% and a Dice score of 88.04%. The Attention U-Net [11], which is an advanced version of the U-Net that incorporates soft attention mechanisms, demonstrated a noticeable improvement over U-Net. The SC U-net [9], which leverages skip connections to facilitate the flow of gradients during training and capitalize on encoder features more effectively, achieved slightly better performance than the Attention U-Net with an IoU of 83.53% and a Dice score of 90.97%. Our proposed method, combining Squeeze-and-Excitation attention with skip connection, outperformed all methods with an IoU of 84.22% and a Dice score of 91.43%.

Figure 3 presents the performance of the four models on two sample images from the test set. A qualitative comparison of the predicted segmentation masks from the aforementioned approaches reveals that the SC-SE U-Net predicted masks are more precise in segmenting tumors with irregular and rough borders. This superior performance becomes evident when examining the qualitative similarity between the ground truth mask and the predicted mask.

TABLE II
QUANTITATIVE COMPARISON OF THE PERFORMANCE OF FOUR METHODS ON THE TCGA-LGG DATASET

Method	**IoU(%)**	**Dice(%)**
U-Net [6]	78.81	88.04
Attention U-Net [11]	83.31	90.83
SC U-net [9]	83.53	90.97
Ours	**84.22**	**91.43**

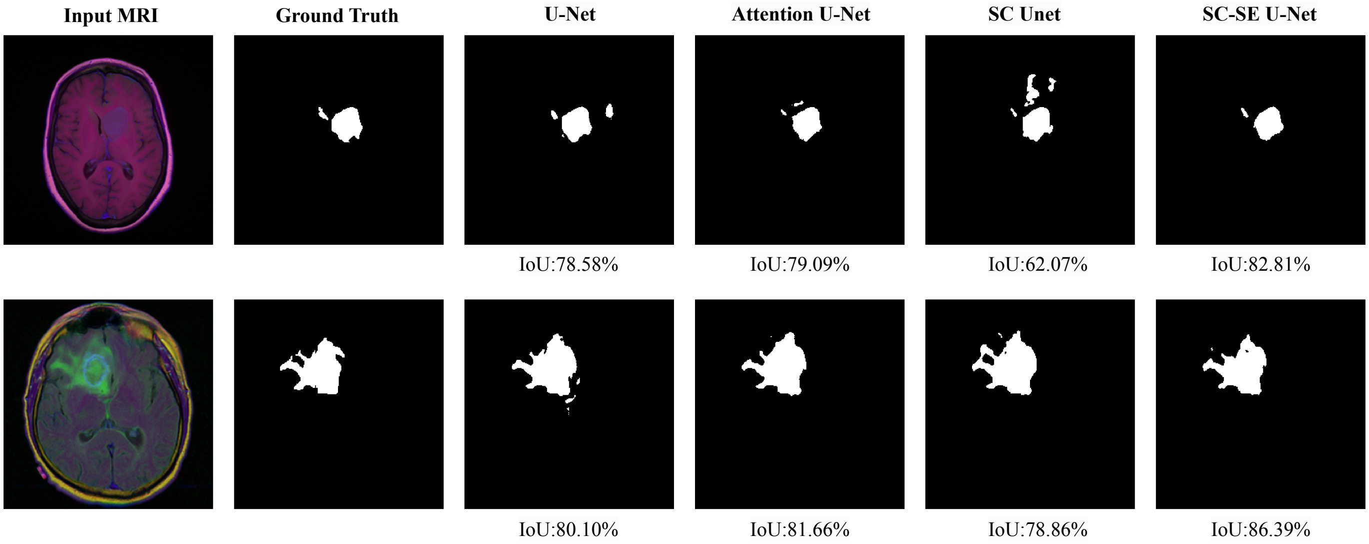

Fig. 3. Qualitative comparison of segmentation methods on samples from the test set.

V. Conclusion

This paper presents an effective semantic segmentation method for the detection of brain tumors in MRI scans. Our research modifies the widely recognized U-Net architecture, proposing a model named SC-SE U-Net. This model incorporates skip connections, similar to residual network architecture, which have proven to enhance the training process significantly.

The inclusion of Squeeze-and-Excitation blocks in the encoder and decoder stages further refines our model. These blocks enable the model to concentrate on relevant features during both the feature extraction and mask generation phases, while simultaneously diminishing the importance of less important features for segmentation. This dual action enhances the precision of the segmentation process.

Our proposed architecture has demonstrated superior performance, achieving an IoU score of 84.22%, an improvement upon the standard U-Net model. This outstanding performance indicates not only that our SC-SE U-Net model outperforms many existing approaches, but also suggests that it holds considerable potential for improving diagnostic efficiency. Further research and development of this model could lead to more precise and faster detection of brain tumors, ultimately contributing to more effective patient treatment and outcomes.

References

[1] K. D. Miller, Q. T. Ostrom, C. Kruchko, N. Patil, T. Tihan, G. Cioffi, H. E. Fuchs, K. A. Waite, A. Jemal, R. L. Siegel, and J. S. Barnholtz-Sloan, “Brain and other central nervous system tumor statistics, 2021,” *CA: A Cancer Journal for Clinicians*, vol. 71, no. 5, pp. 381–406, 2021, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21693. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21693

[2] “Brain Tumor - Statistics,” May 2023, [Online; accessed 21. Sep. 2023]. [Online]. Available: https://www.cancer.net/cancer-types/brain-tumor/statistics

[3] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, “Brain tumor segmentation with Deep Neural Networks,” *Medical Image Analysis*, vol. 35, pp. 18–31, Jan. 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1361841516300330

[4] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, “Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions,” *Journal of Digital Imaging*, vol. 30, no. 4, pp. 449–459, Aug. 2017. [Online]. Available: http://link.springer.com/10.1007/s10278-017-9983-4

[5] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[6] 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.

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, “The importance of skip connections in biomedical image segmentation,” in *International Workshop on Deep Learning in Medical Image Analysis, International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer, 2016, pp. 179–187.

[9] J. Wu, Y. Zhang, K. Wang, and X. Tang, “Skip Connection U-Net for White Matter Hyperintensities Segmentation From MRI,” *IEEE Access*, vol. 7, pp. 155 194–155 202, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8877824/

[10] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, “Attention mechanisms in computer vision: A survey,” *Computational visual media*, vol. 8, no. 3, pp. 331–368, 2022.

[11] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, “Attention u-net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.

[12] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[13] M. Buda, A. Saha, and M. A. Mazurowski, “Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm,” *Computers in biology and medicine*, vol. 109, pp. 218–225, 2019.