Capstone Project on
Recalibrating Fully Convolutional Networks With Spatial and Channel "Squeeze and Excitation" Blocks

Aryaman Vikram Todi
20EE38033

Aman Soni
23EE65E20

Satyabrata Pradhan
23EE65R02

***Abstract*— Recalibrating activations based on spatial and channel importance allow the network to learn smaller image details which will improve semantic segmentation performance. The spatial and channel squeeze and excitation (scSE) blocks can be easily added into a network architecture. [1].**

***Index Terms*— Fully convolutional networks, image segmentation, squeeze & excitation.**

I. Introduction

We demonstrate the methods behind the scSE module as explained in [2]. For image segmentation, fully convolutional neural networks have set the benchmark performance in medical imaging and computer vision. The basic building block for all these architectures is the convolutional layer, which learns filters capturing local spatial patterns along all the input channels and generates feature maps jointly encoding the spatial and channel information. Channel squeeze and excitation (SE) blocks introduced in [3] highlight useful channels to improve classification accuracy. We demonstrate the spatial analogue of this block called the spatial squeeze and excitation, combined with the channel analogue, to create the spatial and channel squeeze and excitation (scSE) block. We demonstrate its effectiveness at improving segmentation performance on the CHAOS dataset [4] for organ segmentation in axial CT and MR images of the abdomen. This report is arranged as follows. In Section II, we will explain the Spatial and Channel Squeeze and Excitation block. In Section III, the Combined Healthy Abdominal Organ Segmentation (CHAOS) dataset is briefly discussed. In section IV the pre-processing and training details are outlined. Section V provides an overview of our results and compares the segmentation performance before and after incorporating the scSE module.

II. Squeeze and excitation block

In this method, denoted as F-CNN, we aim to approximate a complex mapping function $F_{\text{seg}}(\cdot)$ that takes an input image I and produces a segmentation map S. This mapping involves a sequence of functions $F_{\text{tr}}(\cdot)$, each corresponding to either an encoder or a decoder block. These blocks are connected in a cascaded manner, with max-pooling layers in the encoder path and upsampling layers in the decoder path.

For any given intermediate input feature map $\mathbf{X}$, represented as a tensor with dimensions $H \times W \times C_0$, passing it through an encoder or decoder block $F_{\text{tr}}(\cdot)$ yields an output feature map $\mathbf{U}$ with dimensions $H \times W \times C$. This process involves convolutional layers and nonlinear activations that blend spatial and channel information effectively.

To enhance the feature map $\mathbf{U}$, we employ SE blocks $F_{\text{SE}}(\cdot)$, which re-calibrate $\mathbf{U}$ to produce $\hat{\mathbf{U}}$. We propose three distinct variants of SE blocks, each aimed at refining the feature representation for better segmentation performance.

A. Spatial squeeze and channel excitation (cSE)

We describe the spatial squeeze and channel excitation block, which was proposed in [3]. We consider the input feature map $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_C]$ as a combination of channels $\mathbf{u}_i \in \mathbb{R}^{H \times W}$. Spatial squeeze is performed by a global average pooling layer, producing vector $\mathbf{z} \in \mathbb{R}^{1 \times 1 \times C}$ with its k-th element

$$z_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_k(i, j).$$

This operation embeds the global spatial information in vector $\mathbf{z}$. This vector is transformed to $\hat{\mathbf{z}} = \mathbf{W}_1(\delta(\mathbf{W}_2\mathbf{z}))$, with $\mathbf{W}_1 \in \mathbb{R}^{C \times C/r}, \mathbf{W}_2 \in \mathbb{R}^{C/r \times C}$ being weights of two fully-connected layers and the ReLU operator $\delta(\cdot)$. The parameter C/r indicates the bottleneck in the channel excitation, which encodes the channel-wise dependencies. Foreshadowing some of our results, the best performance is obtained by r = 2. The dynamic range of the activations of $\hat{\mathbf{z}}$ are brought to the interval $[0, 1]$, passing it through a sigmoid layer $\sigma(\hat{z})$. The resultant vector is used to re-calibrate or excite $\mathbf{U}$ to $\mathbf{U}_{\text{cSE}} = [\sigma(\hat{\mathbf{z}}_1)u_1, \sigma(\hat{\mathbf{z}}_2)u_2, \ldots, \sigma(\hat{\mathbf{z}}_C)u_C]$.

The activation $\sigma(\hat{\mathbf{z}}_i)$ indicates the importance of the i-th channel, which is either scaled up or down. As the network learns, these activations are adaptively tuned to ignore less important channels and emphasize the important ones.

B. Channel squeeze and spatial excitation (sSE)

We introduce the Channel Squeeze and Spatial Excitation (sSE) block, which compresses the feature map $\mathbf{U}$ across channels and enhances spatially, a crucial aspect for detailed image segmentation. Here, we adopt an alternative organization of the input tensor $\mathbf{U} = [\mathbf{u}_{1,1}, \mathbf{u}_{1,2}, \ldots, \mathbf{u}_{i,j}, \ldots, \mathbf{u}_{H,W}]$, where $\mathbf{u}_{i,j} \in \mathbb{R}^{1 \times 1 \times C}$ corresponds to the spatial location (i, j) with $i \in \{1, 2, \ldots, H\}$ and $j \in \{1, 2, \ldots, W\}$. Spatial compression is achieved through a convolution operation $\mathbf{q} = \mathbf{W}_{\text{sq}} \star \mathbf{U}$ with weights $\mathbf{W}_{\text{sq}} \in \mathbb{R}^{1 \times 1 \times C \times 1}$, producing a projection tensor $\mathbf{q} \in \mathbb{R}^{H \times W}$. Each $q_{i,j}$ represents the linearly combined representation across all channels for the

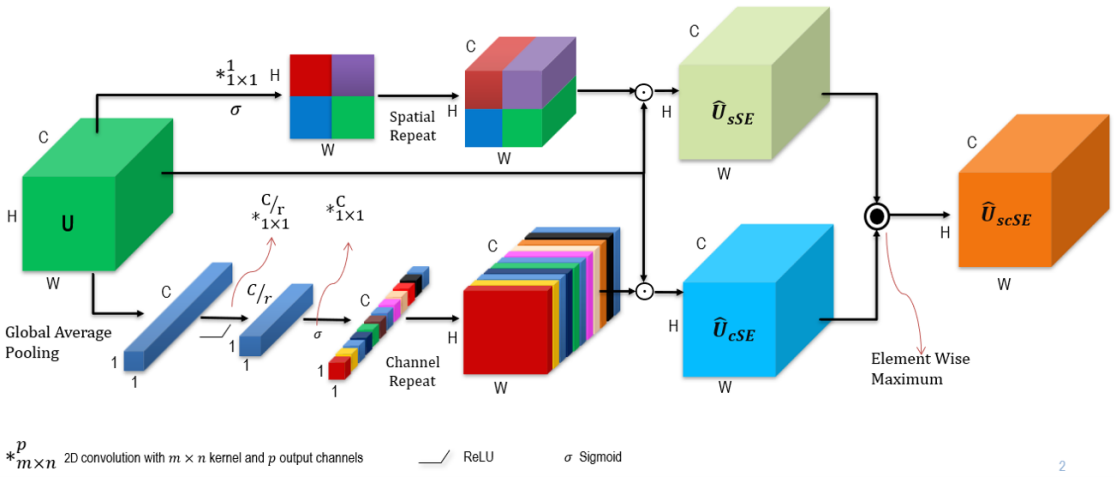


Fig. 1. Spatial and Channel Squeeze and Excitation

spatial location (i, j). This projection is passed through a sigmoid layer $\sigma(\cdot)$ to scale activations to $[0, 1]$, which is then used to recalibrate or enhance $\mathbf{U}$ spatially:

$$\mathbf{U}_{\text{sSE}} = [\sigma(q_{1,1})\mathbf{u}_{1,1}, \ldots, \sigma(q_{i,j})\mathbf{u}_{i,j}, \ldots, \sigma(q_{H,W})\mathbf{u}_{H,W}].$$

Each $\sigma(q_{i,j})$ signifies the relative importance of spatial information at (i, j) within the feature map. This recalibration prioritizes relevant spatial locations while disregarding irrelevant ones.

C. Spatial and channel squeeze and excitation block (scSE)

Each of the cSE and sSE blocks described above possesses distinct characteristics. The cSE blocks adjust channel weights by integrating global spatial information. Utilizing global average pooling layers, these blocks provide a receptive field covering the entire spatial extent at each stage of the FCNN, thereby assisting the segmentation process. On the contrary, sSE blocks maintain the same receptive field size since channel compression is achieved through a 1×1 convolution layer. Instead, they function as spatial attention maps, indicating where the network should concentrate more to aid segmentation. We propose a fusion of the complementary information from both types of SE blocks, simultaneously recalibrating the input $\mathbf{U}$ spatially and channel-wise. We perform the MaxOut operation. The architecture of the combined scSE block is in Fig 1.

D. Max-Out operation

In this aggregation method, any location (i, j, c) of the output feature map U_{scSE} has the maximal activation of $\mathbf{U}_{\text{cSE}}$ and $\mathbf{U}_{\text{sSE}}$. This corresponds to a location-wise max operator

$$\mathbf{U}_{\text{scSE}}(i, j, c) = \max(\mathbf{U}_{\text{cSE}}(i, j, c), \mathbf{U}_{\text{sSE}}(i, j, c)).$$

The max-out layer enforces an element-wise competitiveness between the two SE blocks, similar to [17]. This provides a selective spatial and channel excitation, such that the final segmentation is improved.

III. Explanation of Datasets

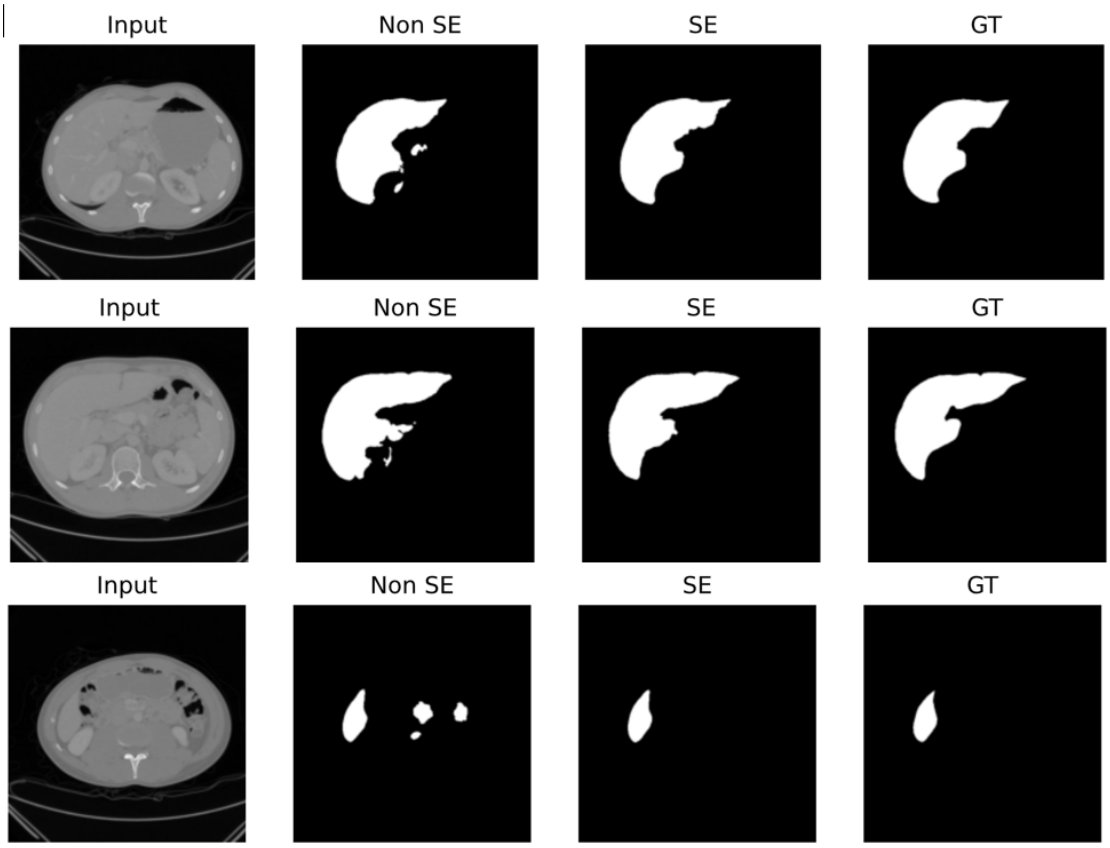
This is the dataset from the Combined (CT-MR) Healthy Abdominal Organ Segmentation (CHAOS) Challenge. This data consists of CT and MR images of the abdomen from different patients. There are 20 training and 20 testing cases (patients) in the CT dataset. The MRI dataset contains 20 training and 20 testing cases with T1-Dual and T2 SPIR sequences. Train data contains both DICOM images and the ground truth binary masks for each class. The testing set only contains DICOM images. In CT cases only livers were annotated. In MRI cases, livers, left/right kidneys, and the spleen was annotated. For CT we have 1507 training images and 377 test images. For MRI we have 560 training images and 63 testing images.

IV. Training and Dataset Preparation

Binary ground truth masks were available for the CT and MR datasets. These were used directly for the CT images. The binary masks were stacked across the channel dimension and reduced from pixel wise one-hot encodings to integer labels to generate the segmentation ground truths. The base architecture was the U-Net with four double convolution blocks each in the encoder and decoder. Batch Normalization was used after every convolution layer. [5].The models were trained using the Adam optimizer [6]. The initial learning rate was set to 0.004, reducing by half after every 4 epochs. We used a batch size of 20. The loss function used was the sum of weighted categorical cross entropy plus Dice loss. The class weights were computed as the inverse frequency of occurence in the training dataset. The models were trained on an Nvidia GeForce GTX 1650Ti.

V. Results

U-Net CT	SE-U-Net CT	U-Net MR	SE-U-Net MR
0.793	**0.873**	0.683	**0.757**



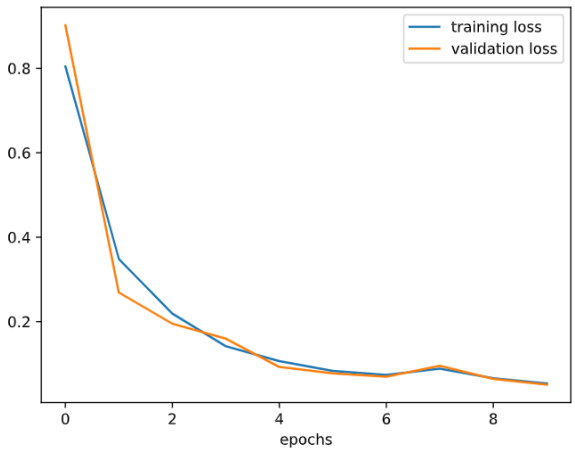


Fig. 2. Loss curves with scSE blocks on CT images

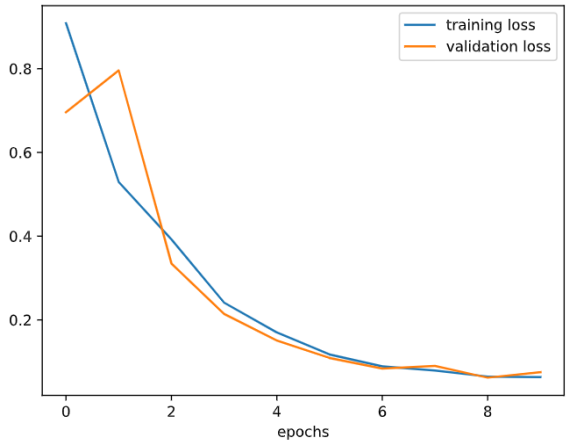


Fig. 3. Loss curves without scSE blocks on CT images

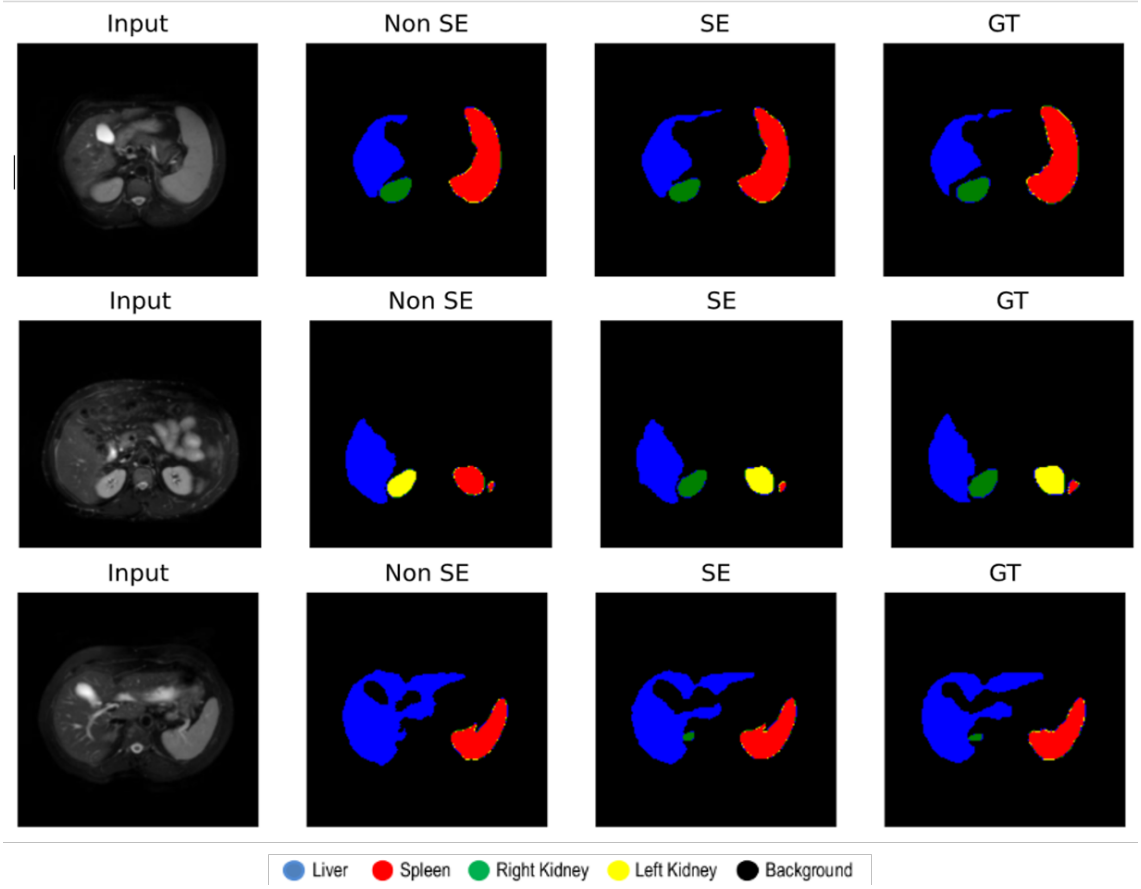
In this section, we compare the results of our experiments. We have implemented the UNet on two datasets as mentioned before. On each dataset, we trained it with and without the addition of scSE block. The loss curves and the segmentation results are shown here.

VI. Concluding Remarks

The addition of the spatial and channel squeeze and excitation blocks clearly results in a 15% increase in the segmentation performance as measured by the Jaccard coefficient, while the number of trainable parameters increase by only around 2%. This block can be easily integrated into existing architectures due to the same input and output size. The model's ability to segment smaller structures improves which is demonstrated with the examples shown above for both CT and MR Images.

VII. Accessing the code

The code can be found at `https://github.com/AryamanIIT20/Medical_Image_Analysis_Term_Project`. The entire repository can be loaded as a zip file by clicking the drop-down menu from "<> Code".



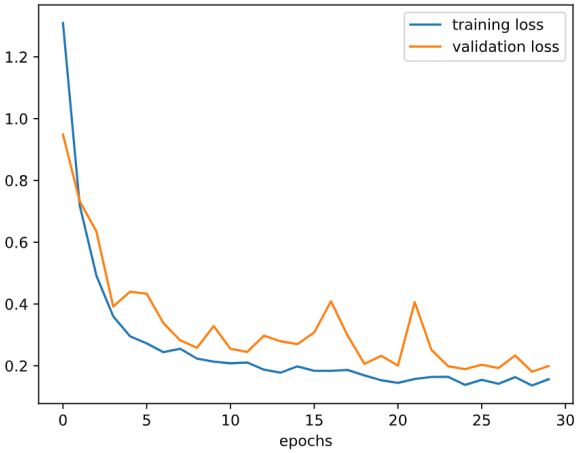


Fig. 4. Loss curves with scSE blocks on MRI images

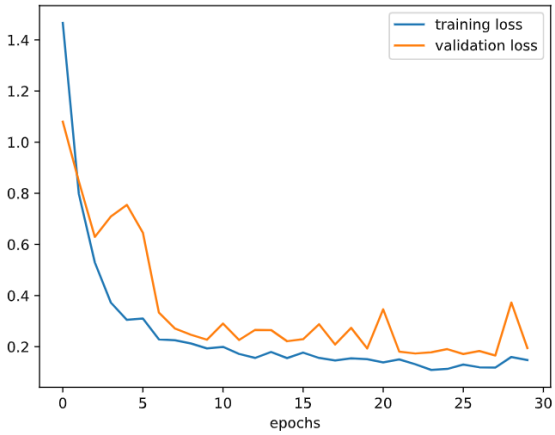


Fig. 5. Loss curves without scSE blocks on MRI images

REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[2] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger, "Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 540–549, 2019.

[3] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[4] Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer, "CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data," Apr. 2019.

[5] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.

[6] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2017.