applied sciences

MDPI

Article

Detecting Small Anatomical Structures in 3D Knee MRI Segmentation by Fully Convolutional Networks

Mengtao Sun [1], Li Lu [1], Ibrahim A. Hameed [1,*], Carl Petter Skaar Kulseng [2] and Kjell-Inge Gjesdal [2]

[1] Department of ICT and Natural Sciences, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, 6009 Ålesund, Norway; mengtao.sun@ntnu.no (M.S.); lulinw2018@gmail.com (L.L.)
[2] Sunnmøre MR-Klinikk, 6010 Ålesund, Norway; carlkulseng@gmail.com (C.P.S.K.); k.i.gjesdal@medisin.uio.no (K.-I.G.)
* Correspondence: ibib@ntnu.no; Tel.: +47-41315695

Abstract: Accurately identifying the pixels of small organs or lesions from magnetic resonance imaging (MRI) has a critical impact on clinical diagnosis. U-net is the most well-known and commonly used neural network for image segmentation. However, the small anatomical structures in medical images cannot be well recognised by U-net. This paper explores the performance of the U-net architectures in knee MRI segmentation to find a relative structure that can obtain high accuracies for both small and large anatomical structures. To maximise the utilities of U-net architecture, we apply three types of components, residual blocks, squeeze-and-excitation (SE) blocks, and dense blocks, to construct four variants of U-net, namely U-net variants. Among these variants, our experiments show that SE blocks can improve the segmentation accuracies of small labels. We adopt DeepLabv3plus architecture for 3D medical image segmentation by equipping SE blocks based on this discovery. The experimental results show that U-net with SE block achieves higher accuracy in parts of small anatomical structures. In contrast, DeepLabv3plus with SE block performs better on the average dice coefficient of small and large labels.

Keywords: medical image segmentation; convolutional neural networks; SE block; U-net; DeepLabV3plus

check for updates

Citation: Sun, M.; Lu, L.; Hameed, I.A.; Kulseng, C.P.S.; Gjesdal, K.-I. Detecting Small Anatomical Structures in 3D Knee MRI Segmentation by Fully Convolutional Networks. *Appl. Sci.* **2022**, *12*, 283. https://doi.org/10.3390/app12010283

Academic Editors: Manuel Armada and Fabio La Foresta

Received: 20 September 2021
Accepted: 27 December 2021
Published: 28 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Knee osteoarthritis is the most common musculoskeletal disease in the world [1]. Lesions commonly induce structural changes within the small articular cartilage [2]. MRI (magnetic resonance imaging) technologies provide the means to characterise structural alterations in different joint tissues affected by osteoarthritis. Patients who undergo knee MRI for presumed musculoskeletal disease can also have unexpected vascular findings or pathology in the imaged field. [3]. In general, osteoarthritis is categorised by the progressive degradation of joint tissues with various abnormalities [4] and has been a severe issue in recent years. However, some small anatomical structures around the joint are hardly detected. For example, the veins and ligaments can show critical early alarms in musculoskeletal lesions [5]. This study will explore the performance of small structure segmentation in knee MRI by using deep learning.

The accurate segmentation of structures is helpful in clinical workflows in multiple domains such as diagnostic intervention and treatment planning. However, manual segmentation of anatomical structures is often time-consuming, labour-intensive, and prone to errors; therefore, it has prompted studies on automated segmentation [6]. The knee joint is one of the most important joints in the human body and is frequently injured in sports and other accidents. Automated knee segmentation can assist orthopaedists in examining and treating various kinds of knee lesions. Deep convolutional neural networks (CNNs) have been used for medical image segmentation since the rapid development of deep

learning techniques in recent decades. They have improved the segmentation accuracy, and decreased the time and manual labour.

The definition of semantic segmentation is that each pixel or voxel of an image is marked with a specific label. A well-performed model can accurately classify each pixel or voxel. Traditionally, global features or structural features are beneficial for classification. They can be acquired through stacked convolutions with strides and various pooling operations, but the spatial information can be lost gradually during this process. Fully convolution networks (FCNs) [7] solved this problem, to some extent, by using three techniques: (1) replacing fully connected layers with convolutions to output images instead of probabilities; (2) using deconvolution to implement upsampling to reconstruct the output image; and (3) using skip connections to fuse information from layers with different strides to improve segmentation details.

Medical image segmentation requires higher accuracy than natural image segmentation and requires neural networks to adjust the architecture to yield more precise segmentation. U-net [8] was developed based on FCN. It uses a symmetrical encoder–decoder architecture and has succeeded in medical image segmentation [9,10]. In U-net, the encoder is a downsampling path that extracts features from input images through the combination of convolutions. The decoder uses upsampling operations instead of pooling, many stacked deconvolutions, and skip connections to combine the features with those in the encoder. These components enable the network to exploit multi-scale context information to obtain high-accuracy segmentation results.

Most medical images are 3D, such as MRI and CT (computed tomography) [11]. 3D medical image segmentation is computationally expensive; therefore, systems are commonly trained patch-wise. When patch-wise training is adopted, the receptive field is reduced initially, which will hinder the ability to capture global features and larger context. To address this problem, we adopt networks that can gain the same receptive field with fewer parameters and enable us to use a larger patch size to feed the network. DeepLab series [12–15] is designed for 2D image segmentation with high accuracy and detailed segmentation maps. DeepLabv1 [12] introduces atrous convolution and fully connected conditional random fields (CRF) to solve the problems brought by reduced resolution. DeepLabv2 [13] proposes atrous spatial pyramid pooling (ASPP), which uses multiple parallel atrous convolution layers with different sampling rates. DeepLabv3 [14] discusses four types of FCNs and improved ASPP. DeepLabv3plus [15] improves ASPP with atrous depthwise separable convolution to get the same receptive field with fewer parameters than standard convolutions. It uses an asymmetrical encoder–decoder architecture. The encoder is comprised of improved Xception and ASPP. The decoder uses bilinear upsampling concatenated with the corresponding low-level features from the network backbone.

Medical datasets are usually imbalanced, which will reduce the performance of machine learning models [16–18]. For the annotated MRI knee images used in this work, there are multiple classes where the frequencies of voxels of each class are extraordinarily different. For example, the size of bone is much larger than blood vessel (see Section 4). When the networks with fewer parameters are selected, their representation capacity may not be sufficient to maintain the required details to obtain accurate segmentation results for small objects/organs. However, patch-wise training cannot be avoided if complicated networks are chosen, damaging captured structural features that are beneficial for the segmentation of larger structures. For example, bones will be carelessly separated into different patches, making it difficult for the model to learn their structural features. Therefore, structural features need to be better extracted to assist the segmentation. We need to consider the results between the required segmentation details of both small and large structures and the computational cost.

This paper proposes different U-net variants to check the performance between diverse knee anatomical structures. Inspired by DeepLabv3plus, we further develop DeepLabv3plus with ASPP (atrous spatial pyramid pooling). We test the effectiveness of DeepLabv3plus variants to balance the capability of global feature extraction and required

resolution. As a result, DeepLabv3plus variants perform better than U-net variants in terms of the average dice coefficient of all structures. The main contributions include:

(1) We propose four types of 3D U-net variants aimed at small anatomical structure segmentation in MRI images. We found that SE block performs well in small anatomical detection;
(2) Based on the success of SE block in U-net variants, we apply the DeepLabv3plus with SE block and transfer the 2D DeepLabv3plus into a 3D version for anatomical segmentation of real MRI images;
(3) In experiments, we improve the results from the small anatomical structure segmentation on the knee MRI images provided by Sunnmøre MR-Klinikk. Based on the experiments, it is concluded that DeepLabv3plus variants could achieve relatively high segmentation accuracy of small structures without decreasing accuracy for large structures.

The rest of this paper is organised as follows: Section 2 introduces the related works on image segmentation. Section 3 describes our applied neural networks, U-net and the recent version of DeepLab. Section 4 discuss the experimental results on the MRI dataset. The conclusions are described in Section 5.

2. Related Works

Machine learning techniques are widely applied in recent medical applications, such as disease detection [19,20] and medical robots [21]. However, typical data-driven models perform differently under diverse requirements [22–24]. For medical image segmentation, researchers proposed dozens of neural networks with encoder–decoder architectures, such as SegNet [25], RefineNet [26], and DecovNet [27]. The architectures mainly consisted of an encoder, a decoder, and fusion techniques. The encoder is responsible for extracting features from input images, where the dimension of feature maps is reduced. It can be seen as a classification neural network without fully connected layers. For example, SegNet uses VGG16 [28] in the encoder, DeepLabv3plus uses improved Xception [29] as a part of its encoder. In the decoder, the spatial dimension of feature maps is recovered through various upsampling operations, which is the opposite of the encoder. The third important part is the fusion technique, which can utilise multi-scale features from the encoder to recover the spatial resolution in the decoder. For example, U-net uses skip connection to fuse the features in the encoder with the features in the decoder.

The fusion of features in different scales is beneficial in improving the accuracy of semantic segmentation. DeepLabv3 [14] discussed four types of FCNs to capture multi-scale context, including image pyramid, encoder–decoder, context module (e.g., using atrous convolution as an approach), and spatial pyramid pooling. The first one is usually applied during the inference stage [30]. For the other three approaches, DeepLabv3plus takes advantage of them to propose a network that employs the architecture of encoder–decoder with atrous separable convolution.

ASPP (atrous spatial pyramid pooling) is one of the most important techniques used in the DeepLab series. It was proposed in DeepLabv2 [13] and developed based on spatial pyramid pooling, which was proposed by [31] to capture the context of images in different strides. DeepLabv2 employs a combination of spatial pyramid pooling with atrous convolution and named it ASPP, which is computationally efficient compared with the original method. DeepLabv3 improves ASPP by adding 1×1 convolution in the first layer and global average pooling in the last layer. In addition, DeepLabv3plus exploits atrous depthwise separable convolution to replace atrous convolution, which reduces the number of parameters to implement ASPP again. DeepLabv3plus is designed for 2D natural image segmentation. In this work, we modified the ASPP technique to 3D image segmentation.

3D image segmentation is very computationally expensive, and therefore related techniques that can make it more efficient are highly appreciated. Due to limited computing resources, volumetric inputs are usually cut into patches to feed them into 3D networks, enabling the use of a large input image size to feed the networks if the network can be

simplified. Other methods to decrease the computing resource are involved. For example, a 3D volume image is comprised of 2D slices, so researchers such as [32] attempted to use 2D networks to segment 2D slices and then fuse the results into a 3D volume image again. However, there are spatial information losses that have an adverse influence on volume image segmentation results. In addition, there are multiple labels in our datasets, including large structures and small structures. FocusNet [33] proposes a method that uses different networks to segment large structures and small structures, fusing the results to form the final segmentation and achieve high segmentation accuracy.

Researchers have proposed several MRI segmentation models aimed at various tissues for knee joint treatment. Reference [34] applied using U-net for capturing the complex morphology and texture of thigh muscle and adipose tissue. The results showed a good clinical effect compared with contralateral knees without knee pain and comparable effect size to manual segmentation. However, muscle and adipose tissue are large structures that can be more easily perceived. The authors did not research any other small issues that may lead to potential lesions. Reference [35] performed conditional Generative Adversarial Networks (cGANs) as a robust and potentially improved method for semantic segmentation than U-Net. Their works showed better results. However, they only consider three types of anatomical structures. The target objects are still apparent in MRI segmentation. Moreover, the attainment of cGANs is generally harder than U-net because the model requires more hyperparameters, and more training attempts were involved in adversary. Reference [36] analysed how 2D U-net functioned on cartilage and meniscus segmentation of knee MR imaging data for morphology and relaxometry compared with manual segmentation. They found that U-Net demonstrates efficacy and precision in quickly generating accurate segmentations that can be used to monitor and diagnose osteoarthritis. However, their model only considers cartilage and meniscus, and their experiments did not achieve 3D modelling. To our knowledge, it is very difficult to detect many small targets simultaneously in one model, and there is no existing research performing many small target segmentation in the knee joint MRI dataset.

As per the biomedical aspect, especially MRI segmentation, researchers focus on improving the segmentation accuracy of an anatomical structure. Awan et al. proposed ResNet-14 CNN, which achieves higher performance on knee ligament segments [37]. Simantiris et al. utilised a Dilated CNN to construct a cardiac MRI segmentation network that has produced the most satisfactory evaluation on dice coefficient [38]. Coupé et al. introduced AssemblyNet, a large ensemble CNN network for whole-brain MRI segmentation [39]. Their networks defeated the U-Net baseline, which is also one of our baseline models. However, although these models generally give better performance on testing images, they may ignore recognising specific small anatomical structures. We also found that CNN-based models have played a significant role in, and positively influenced, real-world medical applications [37–40].

This paper takes advantage of the above-mentioned encoder–decoder architecture to build improved networks, i.e., U-net variants and DeepLabv3plus variants. They showed a greater capability in experiments to segment small structures without losing the accuracy of large structures.

3. Methods

To explore the segmentation performances of different encoder–decoder architectures on our datasets, we developed two types of networks, i.e., U-net variants and DeepLabv3plus variants. Their architectures are shown in Figure 1.

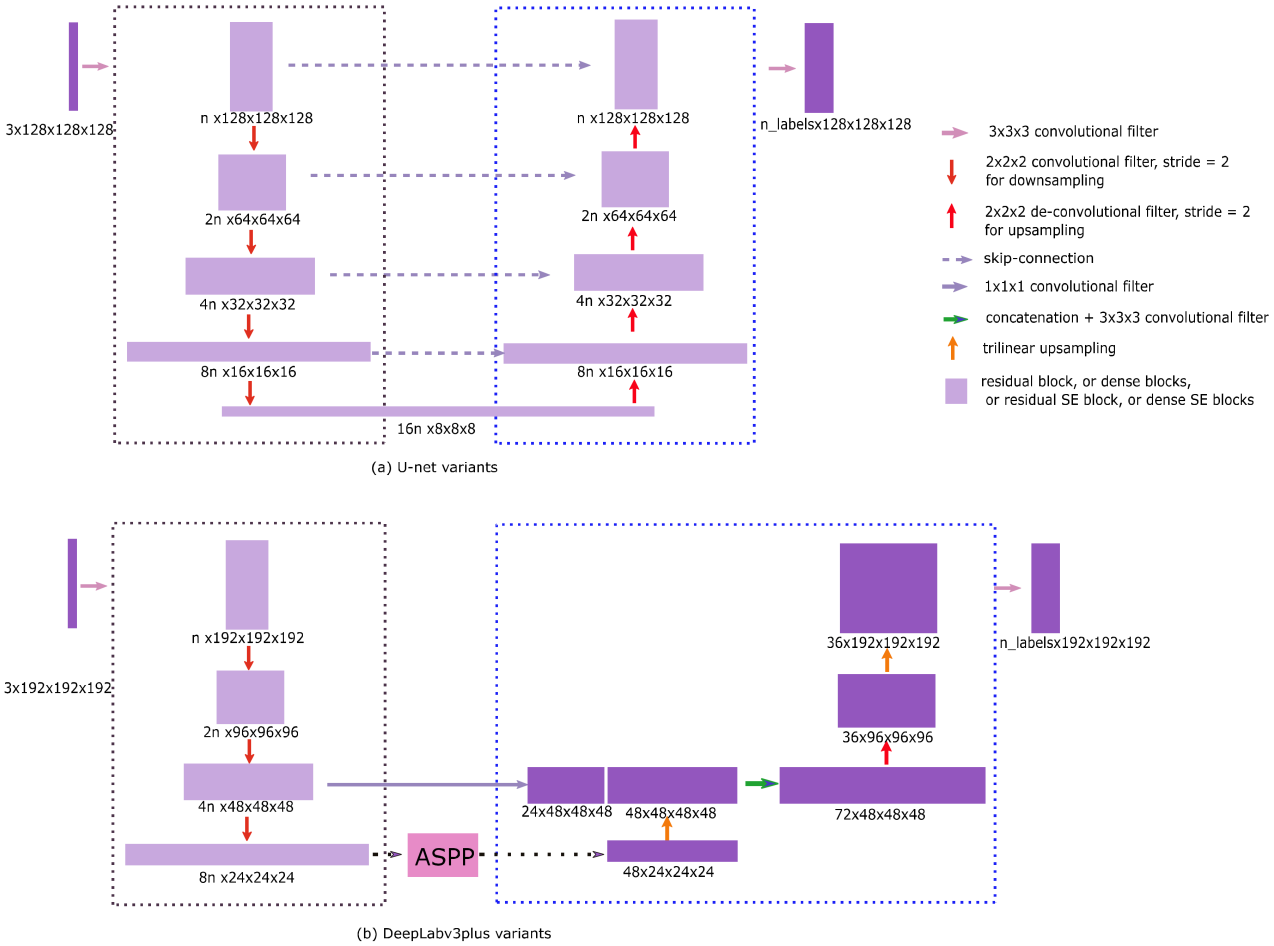


Figure 1. Architectures of U-net variants and DeepLabv3plus variants.

3.1. U-Net Variants

Figure 1a shows the architecture of U-net variants (n = 24), where n is the number of channels. The downsampling path on the left (encoder) extracts features using convolutions from the input images. The upsampling path on the right (decoder) uses deconvolutions to reconstruct the details for the final segmentation results. The skip connections are used to fuse features of different layers obtained in the downsampling path with those in the upsampling path to improve segmentation accuracy.

To maximise the performance of U-net architecture, we used SE blocks with residual and dense structures to construct four types of blocks, namely residual blocks, residual SE blocks, dense blocks, and dense SE blocks. Their structures are shown in Figure 2. SE blocks can be conveniently added in a residual structure and a non-residual structure. For residual structures with SE blocks, several convolution blocks are stacked first. Then, we use SE blocks to strengthen essential features. SE blocks are introduced by Squeeze-and-Excitation Networks (SENet) [11,41]. It improves a prediction accuracy through modelling the correlations between channels and adaptively strengthening important features.

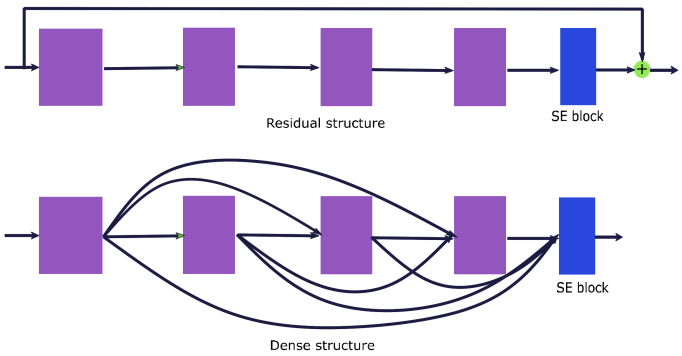


Figure 2. Residual structure and dense structure.

A dense convolutional network (DenseNet) [42] considers that the residual connections combining the input with the output of stacked convolutions by summation impede the information flow in the network. Then, a different connectivity pattern was proposed: concatenation rather than summation. The structure is also shown in Figure 2, where four-layer dense blocks precede SE block.

Each layer takes all preceding feature maps as the input to reuse these features to take advantage of features gained in different layers to exploit the network's potential. For dense structures, we added SE block in a non-residual block (i.e., dense block). Thus, four types of blocks are formed, including residual blocks, residual SE blocks, dense blocks, and dense SE blocks. For the blocks in the encoder path and the decoder path, we fill them using these four types of blocks, respectively, to form four U-net variants.

3.2. DeepLabv3plus Variants

DeepLabv3plus is designed for 2D natural image semantic segmentation. The advanced components used in it provide us with an alternative method to extract features from the images. For example, ASPP with depthwise separable convolution enables us to capture the multi-scale features of images with fewer parameters. If we modify them according to the features of 3D medical images and use them in the network, it assists in the reduction of the required computing power and improves the segmentation results and overall performance.

The encoder of DeepLabv3plus comprises improved Xception and ASPP. However, the features in medical images are not as complicated as in natural images. It is not necessary to use such complex networks to recognise the pattern in the dataset. In addition, 3D volume image segmentation is computationally expensive, and adopting too deep networks might waste computing resources. Here, we used a similar network to U-net variants to replace Xception as the primary network in the encoder, as shown in Figure 1. Four types of blocks can be used here as well. ASPP reduced the number of filters and obtained a larger receptive field. In addition, three layers of downsampling are employed in DeepLabv3plus variants instead of four layers in the U-net variants, and half reduces the numbers of channels in each layer, so n = 12 in Figure 1. The last layer of the encoder is removed, and ASPP is added at the bottom of the encoder in the same way as it is in DeepLabv3plus.

ASPP is a powerful tool that enables us to capture multi-scale information on images and obtain larger receptive fields with fewer parameters and hence more efficient performance and more accurate segmentation results. The implementation details are shown in Figure 3, where we employ a $1 \times 1 \times 1$ convolution to get the first feature map. A $1 \times 1 \times 1$ convolution can select the important features from the input and is frequently used to reduce the dimension of feature maps (i.e., depth). Then, $3 \times 3 \times 3$ atrous depthwise separable convolutions, with dilatation rates of 4, 8, and 12, are used to get the following three feature maps. Different rates could effectively capture multi-scale information. Finally, global average pooling is applied to obtain the last feature map. We concatenate these

five feature maps and use a $1 \times 1 \times 1$ convolution to choose the important features from them and reduce the number of channels (i.e., depth). To be concatenated, these five output feature maps must have the same dimensions.

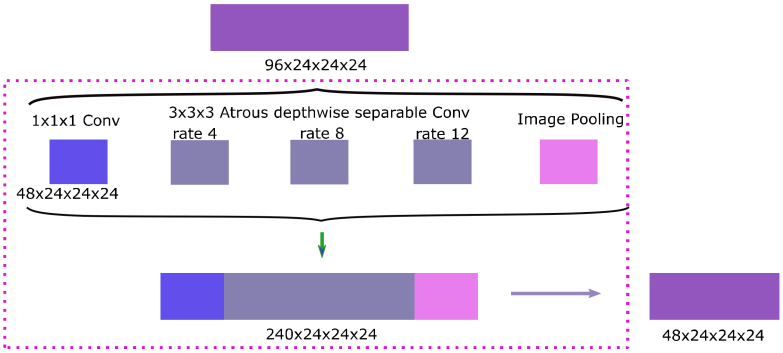


Figure 3. Structure of ASPP.

Atrous depthwise separable convolution is used to obtain the three feature maps in the middle. The calculation process is explained as follows. Each input channel is convolved with the convolution filter first, and then use $1 \times 1 \times 1$ convolution to choose feature maps from different channels. For the first step, we set the stride s to 1, the padding number p to same as the dilation rate r, and apply the $3 \times 3 \times 3$ convolution filter (k = 3) for each channel. As a result, the width of output feature maps, can be calculated by:

$$W_{out} = 1/s * (w_{in} + 2p - r(k-1) - 1) + 1 = w_{in} \quad (1)$$

The height and depth can be obtained in the same way. For atrous convolution, the voxels can be calculated according to the equation:

$$y(i, j, k) = \sum_{d=0}^{2} \sum_{h=0}^{2} \sum_{w=0}^{2} x(i + w * r, j + h * r, k + d * r) * W(w, h, d) \quad (2)$$

After this process, the size of feature maps remains unchanged (C_in, D, H, W). Then, we use $1 \times 1 \times 1$ convolution to choose the features from different channels, and the size of the feature maps is changed to (C_out, D, H, W).

For depthwise separable convolution, the number of parameters used in the process above is:

$$3 \times 3 \times 3 \times chan_{in} + 1 \times 1 \times 1 \times chan_{in} \times chan_{out} = 27 \times chan_{in} + chan_{in} \times chan_{out} \quad (3)$$

If we use standard convolution, the number of parameters should be:

$$3 \times 3 \times 3 \times chan_{in} \times chan_{out} = 27 \times chan_{in} \times chan_{out} \quad (4)$$

Thus, it reduces the number of parameters when the number of channels used in the network is large.

The decoder part in DeepLabv3plus uses bilinear interpolation for 2D images, and the factor of upsampling is 4. For our DeepLabv3plus variants, we utilise trilinear interpolation to implement upsampling instead of deconvolution used in U-net architecture, which reduces the number of parameters again. However, we use 2 as the upsampling factor, because medical image segmentation requires higher accuracy than natural image segmentation. Trilinear interpolation is beneficial for saving computing resources, but it damages the representative capacity of the decoder compared with deconvolution. We add one deconvolution upsampling block between the trilinear interpolation operations to introduce more flexibility in the upsampling path. To take advantage of the features

extracted by the encoder, we also concatenate the features obtained before ASPP in the encoder with the first trilinear interpolation feature map. The result is used as the input of the 3 × 3 × 3 deconvolution block.

3.3. Loss Functions

Small object/organ segmentation is always a challenge in semantic segmentation. In our case, the smallest structure takes up less than 0.01% of the whole volume of the MRI images. Dice loss was introduced by V-net [43]. It was developed based on the Dice coefficient to address the problem that the learning process is trapped in the local minimum when the predictions are strongly biased towards the background. We also use weighted dice loss as the loss function. Weighted dice loss uses weighting to adjust the importance of different categories in training. This is commonly used to address the imbalance problem in samples by adding weight on the category whose proportion is small. It is calculated according to the equation:

$$\text{Loss} = 1 - \frac{1}{n} \times \sum_{i=1}^{n} (w_i \times \frac{2 \times \sum_{j=1}^{m} t_{ij} p_{ij} + \text{smooth}}{\sum_{j=1}^{m} (t_{ij} + p_{ij}) + \text{smooth}}) \quad (5)$$

where n is the number of classes, w_i is the weight of class i, m is the number of voxels of class i, t_{ij} is the j-th voxel of class i using one-hot encoding in the truth, and p_{ij} is the corresponding voxel in the prediction. To set the weights, we use the percentages of classes in the dataset.

The performances of the networks in this paper are evaluated by dice coefficients for each class. To compare the predicted segmentation and the ground truth for each class, the percentage of class y predicted as class x was calculated according to:

$$P_{(x,y)} = \frac{n_{(x,y)}}{n_y} \quad (6)$$

where $n_{(x,y)}$ is the number of voxels predicted as class x but annotated as class y in the ground truth, n_y is the number of voxels annotated as class y in the ground truth. It is the recall rate for class x when x = y. Notably, it is not a confusion matrix. We call it a performance matrix for short in this paper. The data in the diagonal line are the recall rate for each class. They should be 1 if all voxels are segmented correctly. The data in the diagonal line will be precise if it replaces the denominator of (6) as the number of voxels predicted as class y.

3.4. Hyperparameter

The original size of the images is 400 × 400 × 400. We conduct the experiments of the four U-net variants on the dataset. The time of a single training epoch of U-net with residual blocks on the dataset of resolution 400 × 400 × 400 is about 7 h. Because of the constraints on GPU memory, we set the patch size to 128 × 128 × 128 and the batch size to 1. We found that the training with a larger patch size and a smaller batch size performs better than the training with a smaller patch size and larger batch size.

DeepLabv3plus variants utilise several techniques to reduce the number of parameters, enabling us to use a larger patch size to feed the networks. The patch size used for training DeepLabv3plus variants has increased from 128 × 128 × 128 to 192 × 192 × 192. Although we increase the patch size, the model parameters are still lower than U-Net, and the training time of an epoch for DeepLabv3plus variants is less than 10 min.

The MR images were obtained from 20 volunteers: 18 volunteers provide the training dataset and 2 volunteers provide the testing dataset. Each volunteer provides 30 images; the total number of images was 600, that is 90% (540 images) for training and 10% (60 images) for testing. The main hyperparameters of U-Net and DeepLabv3plus are displayed in Figure 1. We directly concatenate the results together when we receive all the outputs from patches.

All the networks are trained on GPU GeForce RTX 2080 Ti, with a GPU memory of 11 GB GDDR6 and by using an Adam optimiser. The initial learning rate is set to 0.01 for training with dice loss and is set to 0.001 for training with weighted dice loss. The learning rates were set to be reduced by a factor of 0.5 after two epochs if the validation loss is not decreasing. The training was stopped when the loss on the validation dataset had not decreased for at least three epochs.

4. Experiments and Results

The knee MRI images provided by Sunnmøre MR-Klinikk used in our experiments were manually annotated. The final dataset used to compare the architecture of U-net variants and DeepLabv3plus variants contains 13 labels. Figure 4 shows the extraordinary imbalance of the labels on the final dataset. The background accounts for 60.43% on average in all samples. The largest label is AD (adipose tissue), accounting for 19.17%, while the smallest is ACL (anterior cruciate ligament), accounting for 0.03%. Table 1 shows the abbreviations and values of classes (structures) in performance matrices. The red classes in Table 1 are the small organs in terms of the statistics of Figure 4.

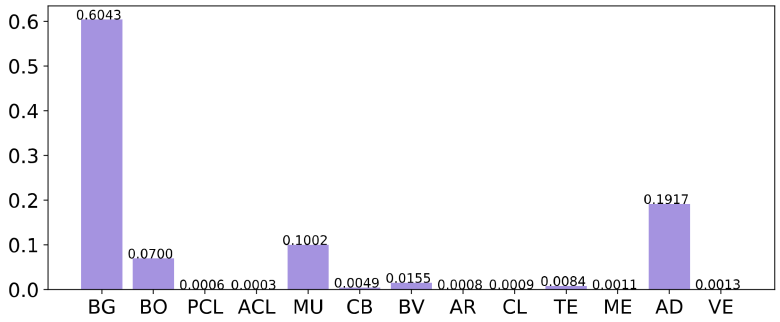


Figure 4. Frequencies of voxels for each class.

Table 1. Abbreviations and values of classes.

Classes	Abbreviations	Index
Background	BG	0
Bone	BO	1
Posterior cruciate ligament	PCL	2
Anterior cruciate ligament	ACL	3
Muscle	MU	4
Cortical bone	CB	5
Blood vessel (popliteal artery/vein++)	BV	6
Artery	AR	7
Collateral ligament	CL	8
Tendons	TE	9
Menisci	ME	10
Adipose tissue (fat)	AD	11
Veins	VE	12

The dataset contains annotated knee MRI images from 20 volunteers. For each knee image, three weighted volumes, including T1 (longitudinal relaxation time), PD (proton density), and FS (fat-saturation) are provided. Different weighted volumes provide different contrasts for different tissues. For example, T1-weighted images provide high contrast for fat, but low contrast for water. FS pulse sequences can improve the detection of musculoskeletal lesions. The sizes of these images is $400 \times 400 \times 400$. We cut each image into patches to feed them into the networks and use a large selection of them for training and the remaining for validation. (An example MRI can be found in Section 4.3).

4.1. Results on U-Net Variants

The performances of U-net variants are shown in Table 2.

Table 2. Performances of U-net variants.

Class	Dice Loss		Weighted Dice Loss			
	U-Net	**U-Net + Res**	**U-Net + Res**	**U-Net + Res + SE**	**U-Net + Dense**	**U-Net + Dense + SE**
BG	0.94	0.987	0.992	0.989	0.990	0.990
BO	0.95	0.917	0.831	0.938	0.932	0.948
PCL	0	0.002	0.153	0.347	0.338	0.228
ACL	0	0	0.083	0.096	0.084	0.041
MU	0.95	0.927	0.944	0.956	0.928	0.939
CB	0	0.567	0.500	0.457	0.495	0.515
BV	0	0.556	0.622	0.603	0.555	0.648
AR	0	0.057	0.198	0.270	0.152	0.194
CL	0	0.041	0.470	0.248	0.237	0.370
TE	0	0.726	0.701	0.723	0.641	0.726
ME	0	0.191	0.133	0.126	0.139	0.107
AD	0.92	0.919	0.876	0.919	0.904	0.925
VE	0	0	0.396	0.491	0.305	0.528
Avg-All organs	0.289	0.453	0.531	0.551	0.515	0.551
Avg-Small organs	0	0.238	0.362	0.373	0.327	0.373

For U-net and U-net + residual blocks [43] with dice loss, large structures bone (BO), muscle (MU), and adipose tissue (AD) are segmented correctly, but most of the small structures are missing (see Table 2). The main reason for this is the extremely imbalanced classes in the dataset.

To increase the importance of small structures, weighted dice loss is adopted. We set the weights according to the proportions of the classes in the dataset. From Table 2, we can see that the results for small structures are improved, but the accuracy of BO is decreased. One possible reason is that when patch-wise training is used, they are unavoidably cut into several parts for the bones since they account for a large proportion of the image and their positions are in the middle. It is difficult for the neural network to distinguish its pattern because the special structure of bones is scattered into different patches. To improve the network's performance, SE blocks are added, the results of small structures are improved compared with that of no SE blocks. Moreover, the accuracy of BO has a little increasement.

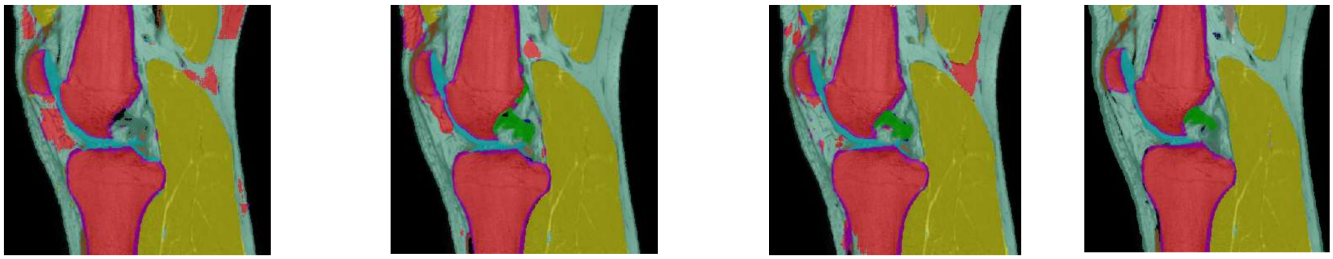
In conclusion, by using SE-based methods, U-Net + Res + SE or U-Net + Dense + SE, it can be pointed out that with the help of SE block, average evaluations are apparently increased. During our training, another valuable observation is that the bones are segmented more completely than those on the smaller patch size. The main reason is that a larger patch size is beneficial for the networks to capture the global features. When a larger patch size is used, more parts of bones will be in the same patch. As a result, it can be segmented more correctly.

4.2. Results with DeepLabv3plus Variants

Because it is confirmed that SE structure can improve the performance of the networks, we further train the networks with SE blocks for both residual structure and dense structure in DeepLabv3plus. As is shown in Table 3, two variants are trained first, including one in which the basic network uses residual block and one in which the basic network uses dense block. We found that adding SE block can improve the two DeepLabv3plus baselines without SE block. As shown in Table 3 and Figure 5, we also find that DeepLabv3plus variants can segment the structures better than U-net variants.

Table 3. Performances of DeepLabv3plus-based model.

Class	U-Net + Res	Deeplab + Dense	Deeplab + Res	Deeplab + Dense + SE	Deeplab + Res + SE
BG	0.987	0.977	0.984	0.986	0.983
BO	0.917	0.809	0.797	0.958	0.936
PCL	0.002	0.703	0.800	0.752	0.522
ACL	0	0.345	0.457	0.462	0.504
MU	0.927	0.926	0.950	0.969	0.976
CB	0.567	0.639	0.701	0.825	0.861
BV	0.556	0.553	0.627	0.813	0.781
AR	0.057	0.534	0.422	0.622	0.671
CL	0.041	0.449	0.536	0.681	0.589
TE	0.726	0.519	0.609	0.685	0.745
ME	0.191	0.706	0.794	0.844	0.819
AD	0.919	0.798	0.799	0.927	0.91
VE	0	0.227	0.344	0.290	0.265
Avg-All organs	0.453	0.629	0.678	0.755	0.735
Avg-Small organs	0.238	0.519	0.588	0.663	0.639

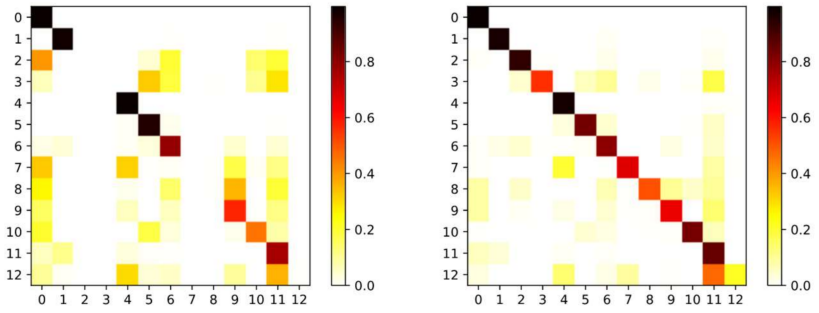


(**a**) (**b**) (**c**) (**d**)

Figure 5. A case study of segmentation experiments. (**a**): U-Net + Res with dice loss. (**b**): Deeplab + Dense + SE with dice loss. (**c**): Deeplab + Res + SE with dice loss. (**d**): Ground truth.

Where (a) shows the segmentation of the results of U-Net + Res with dice loss, (b) and (c) show the results of DeepLabv3plus variants with dice loss. We can see that, for example, the green part (PCL) in the middle is not segmented correctly in (a), but it is segmented in (b) and (c). The results are shown in Table 2. With dice loss, the accuracies of small anatomical segments are quite acceptable compared with the results on U-Net + Res. One reason could be that advanced components, such as ASPP, are used to obtain a larger receptive field without losing too much resolution. Generally speaking, Deeplab + Dense + SE performs better than Deeplab + Res + SE. Deeplab + Dense + SE has higher average accuracies on small structures, especially in PCL, BV, and CL. Deeplab + Res + SE has slightly higher accuracies on ACL, MU, and CB.

Figure 6b shows the performance matrix of Deeplab + Dense + SE, where we can see it achieved relatively high accuracies on all anatomical segments compared with U-Net + Res (as it is shown in Figure 6a). For U-Net + Res with dice loss, small structures, including PCL(2), ACL(3), AR(7), and CL(8) were missing, and predicted as other structures such as BG(0), CB(5), and BV(6). An example of the segmentation result of Residual SE Deeplab with dice loss is shown in Figure 5c, where we can see that there is a small part of AD(11) predicted as BO(1), similar to the results of Dense SE Deeplab with dice loss. A possible reason could be that patch-wise training that negatively influences the capture ability of structural features of bones. However, this problem has already been solved largely in DeepLabv3plus variants.



(**a**) (**b**)

Figure 6. (**a**) Performance matrix of U-Net-based model; (**b**) performance matrix of DeepLabv3plus-based model.

4.3. Discussion

We trained U-net and U-Net + Res with dice loss as the preliminary model. Then, we proposed new varieties to discover the small and low-resourced pixels in the MRI dataset, i.e., the structures with SE block. The proposed varieties showed a satisfactory performance on small organ segmentation.

For U-Net + Res trained with dice loss, from Table 3 we can see that some small anatomical structures such as PCL and ACL are not segmented correctly. For the U-net variant with residual SE blocks trained with weighted dice loss, the performances on the small structures are improved.

The main question is that the neural networks have a smaller field of view on the resolution $400 \times 400 \times 400$, although the patch sizes used on the two datasets are the same. The patch-wise training can be seen as reducing the receptive field of the images at the beginning. For the downsampling dataset, the view was reduced to $(128 / 400)^3 \approx 0.033$. The locations of anatomical structures are in the centre of the image and will be separated into different patches; therefore, the difficulty of recognising their patterns increases.

Tables 2 and 3 show the comparison of U-net variants and DeepLabv3plus variants. DeepLabv3plus variants enable us to segment all structures and achieve relatively higher average accuracy with dice loss. However, U-net variants cannot segment the small structures correctly, leading to lower average accuracy. Generally speaking, DeepLabv3plus variants perform better than U-net variants in terms of average dice coefficient on all labels with dice loss.

DeepLabv3plus-based models are overall better than U-net-based models. Compared with natural images, the MR image semantics are relatively simple, and the pattern is relatively fixed; both high-level and low-level semantic features appear to be very important. To utilise the features, U-Net applied skip connection and U-shaped structure to combine the high-level and low-level features. However, this network becomes too complex, leading to more parameters in the model. Because it is tricky to obtain medical imaging data, many studies only provide data for less than 100 cases. Therefore, the model we designed should not be too large. Too many parameters can easily lead to overfitting.

In U-net, the receptive field of pixels on the feature map depends on convolution and pooling operations. The receptive field of ordinary convolution can only increase two pixels at a time step, and the progress is too slow. The increase of the receptive field of the traditional convolutional network is generally done by pooling operation. The pooling operation will increase the receptive field while reducing the image's resolution, thus losing some information. Moreover, the upsampling of the pooled image will make it impossible to restore a lot of detailed information, limiting segmentation accuracy.

To address the problem, DeepLabv3plus utilises a new atrous convolution. It was performed in the ASPP structure (see Figure 3), which simultaneously satisfies:

1. Connecting high- and low- MRI features together;
2. Significantly reducing model parameters, and thus alleviating overfitting and shortening training time.

As a result, in our experiments, DeepLabv3plus variants achieve the best performance on average accuracy. However, the performance could be improved further by adopting more advanced components. These components should be able to obtain a larger receptive field with fewer parameters and have the ability to reserve details during the process.

5. Conclusions

This work attempted to use annotated knee MRI images by Sunnmøre MR-Klinikk to explore two types of encoder–decoder architecture FCNs, including U-net and DeepLabv3plus. Based on FCN, some neural network variants are proposed, which uses U-net as the basic network, ASPP to capture the multi-scale feature of images, and atrous depthwise separable convolutions to reduce the number of parameters in the encoder. The decoder utilises trilinear interpolation without parameters to implement upsampling instead of deconvolution. This architecture enables us to use a larger patch size and achieves relatively high segmentation accuracies on small structures without the sacrifice of accuracies on large structures. In addition, the training time is significantly reduced from hours to minutes on one epoch.

For U-net architecture, we use four types of blocks, including residual blocks, residual SE blocks, dense blocks, and dense SE blocks, to replace the standard convolution blocks in the original network. The experiments show that the segmentation accuracies of small structures are improved with SE structures. For DeepLabv3plus variants, SE structures also help improve the small structure detection and, compared with U-Net, they require fewer parameters, run faster, and perform better in terms of average accuracy. In conclusion, DeepLabv3plus networks with SE block better capture structural features in MRI segmentation.

Author Contributions: Conceptualization, M.S. and L.L.; methodology, M.S. and L.L.; software, L.L.; validation, M.S.; formal analysis, M.S.; investigation, M.S.; resources, C.P.S.K. and K.-I.G.; data curation, C.P.S.K. and K.-I.G.; writing—original draft preparation, M.S. and L.L.; writing—review and editing, M.S. and I.A.H.; visualization, L.L.; supervision, I.A.H.; project administration, I.A.H.; funding acquisition, I.A.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Norwegian University of Science and Technology.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Panfilov, E.; Tiulpin, A.; Klein, S.; Nieminen, M.T.; Saarakkala, S. Improving robustness of deep learning based knee mri segmentation: Mixup and adversarial domain adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019; pp. 450–459.
2. Nieminen, M.T.; Casula, V.; Nevalainen, M.T.; Saarakkala, S. Osteoarthritis year in review 2018: Imaging. *Osteoarthr. Cartil.* **2019**, *27*, 401–411. [CrossRef] [PubMed]
3. Gaetke-Udager, K.; Fessell, D.P.; Liu, P.S.; Morag, Y.; Brigido, M.K.; Yablon, C.; Jacobson, J. Knee MRI: Vascular pathology. *Am. J. Roentgenol.* **2015**, *205*, 142–149. [CrossRef] [PubMed]
4. Castañeda, S.; Roman-Blas, J.A.; Largo, R.; Herrero-Beaumont, G. Subchondral bone as a key target for osteoarthritis treatment. *Biochem. Pharmacol.* **2012**, *83*, 315–323. [CrossRef] [PubMed]
5. More, S.; Singla, J.; Abugabah, A.; AlZubi, A.A. Machine Learning Techniques for Quantification of Knee Segmentation from MRI. *Complexity* **2020**, *2020*, 6613191. [CrossRef]

6. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
7. Li, Y.; Zhao, H.; Qi, X.; Wang, L.; Li, Z.; Sun, J.; Jia, J. Fully Convolutional Networks for Panoptic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 214–223.
8. Calisto, M.B.; Lai-Yuen, S.K. AdaEn-Net: An ensemble of adaptive 2D–3D Fully Convolutional Networks for medical image segmentation. *Neural Netw.* **2020**, *126*, 76–94. [CrossRef] [PubMed]
9. Luo, X.; Zeng, W.; Fan, W.; Zheng, S.; Chen, J.; Liu, R.; Liu, Z.; Chen, Y. Towards cascaded V-Net for automatic accurate kidney segmentation from abdominal CT images. In Proceedings of the Medical Imaging 2021: Image Processing, Online, 15–19 February 2021; Volume 11596, p. 1159619.
10. Zhao, W.; Jiang, D.; Queralta, J.P.; Westerlund, T. MSS U-Net: 3D segmentation of kidneys and tumors from CT images with a multi-scale supervised U-Net. *Inform. Med. Unlocked* **2020**, *19*, 100357. [CrossRef]
11. Zeng, G.; Yang, X.; Li, J.; Yu, L.; Heng, P.A.; Zheng, G. September. 3D U-net with multi-level deep supervision: Fully automatic segmentation of proximal femur in 3D MR images. In Proceedings of the International Workshop on Machine Learning in Medical Imaging, Quebec City, QC, Canada, 10 September 2017; Springer: Cham, Switzerland; pp. 274–282.
12. Zhang, S.; Ma, Z.; Zhang, G.; Lei, T.; Zhang, R.; Cui, Y. Semantic image segmentation with deep convolutional neural networks and quick shift. *Symmetry* **2020**, *12*, 427. [CrossRef]
13. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
14. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv Preprint* **2017**, arXiv:1706.05587.
15. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 801–818.
16. Alam, T.M.; Shaukat, K.; Mahboob, H.; Sarwar, M.U.; Iqbal, F.; Nasir, A.; Hameed, I.A.; Luo, S. A Machine Learning Approach for Identification of Malignant Mesothelioma Etiological Factors in an Imbalanced Dataset. *Comput. J.* **2021**. [CrossRef]
17. Khushi, M.; Shaukat, K.; Alam, T.M.; Hameed, I.A.; Uddin, S.; Luo, S.; Yang, X.; Reyes, M.C. A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access* **2021**, *9*, 109960–109975. [CrossRef]
18. Yang, X.; Khushi, M.; Shaukat, K. Biomarker CA125 Feature Engineering and Class Imbalance Learning Improves Ovarian Cancer Prediction. In Proceedings of the 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 16–18 December 2020; pp. 1–6.
19. Latif, M.Z.; Shaukat, K.; Luo, S.; Hameed, I.A.; Iqbal, F.; Alam, T.M. Risk factors identification of malignant mesothelioma: A data mining based approach. In Proceedings of the 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), Istanbul, Turkey, 12–13 June 2020; pp. 1–6.
20. Alam, T.M.; Shaukat, K.; Hameed, I.A.; Khan, W.A.; Sarwar, M.U.; Iqbal, F.; Luo, S. A novel framework for prognostic factors identification of malignant mesothelioma through association rule mining. *Biomed. Signal Processing Control.* **2021**, *68*, 102726. [CrossRef]
21. Shaukat, K.; Iqbal, F.; Alam, T.M.; Aujla, G.K.; Devnath, L.; Khan, A.G.; Iqbal, R.; Shahzadi, I.; Rubab, A. The impact of artificial intelligence and robotics on the future employment opportunities. *Trends Comput. Sci. Inf. Technol.* **2020**, *5*, 050–054.
22. Shaukat, K.; Luo, S.; Varadharajan, V.; Hameed, I.A.; Xu, M. A survey on machine learning techniques for cyber security in the last decade. *IEEE Access* **2020**, *8*, 222310–222354. [CrossRef]
23. Shaukat, K.; Luo, S.; Varadharajan, V.; Hameed, I.A.; Chen, S.; Liu, D.; Li, J. Performance comparison and current challenges of using machine learning techniques in cybersecurity. *Energies* **2020**, *13*, 2509. [CrossRef]
24. Shaukat, K.; Luo, S.; Chen, S.; Liu, D. Cyber Threat Detection Using Machine Learning Techniques: A Performance Evaluation Perspective. In Proceedings of the 2020 International Conference on Cyber Warfare and Security (ICCWS), Islamabad, Pakistan, 20–21 October 2020; pp. 1–6.
25. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
26. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
27. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
28. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
29. POLAT, Ö. Detection of Covid-19 from Chest CT Images using Xception Architecture: A Deep Transfer Learning based Approach. *Sak. Univ. J. Sci.* **2021**, *25*, 813–823. [CrossRef]
30. Xu, Y.; Gong, M.; Chen, J.; Chen, Z.; Batmanghelich, K. 3D-BoxSup: Positive-Unlabeled Learning of Brain Tumor Segmentation Networks from 3D Bounding Boxes. *Front. Neurosci.* **2020**, *14*, 350. [CrossRef] [PubMed]

31. Peng, C.; Ma, J. Semantic segmentation using stride spatial pyramid pooling and dual attention decoder. *Pattern Recognit.* **2020**, *107*, 107498. [CrossRef]
32. Song, Y.; Yu, Z.; Zhou, T.; Teoh, J.Y.C.; Lei, B.; Choi, K.S.; Qin, J. Learning 3d features with 2d cnns via surface projection for ct volume segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; pp. 176–186.
33. Kaul, C.; Manandhar, S.; Pears, N. Focusnet: An Attention-Based Fully Convolutional Network for Medical Image Segmentation. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 455–458.
34. Kemnitz, J.; Baumgartner, C.F.; Eckstein, F.; Chaudhari, A.; Ruhdorfer, A.; Wirth, W.; Eder, S.K.; Konukoglu, E. Clinical evaluation of fully automated thigh muscle and adipose tissue segmentation using a U-Net deep learning architecture in context of osteoarthritic knee pain. *Magn. Reson. Mater. Phys. Biol. Med.* **2020**, *33*, 483–493. [CrossRef] [PubMed]
35. Kessler, D.A.; MacKay, J.W.; Crowe, V.A.; Henson, F.M.; Graves, M.J.; Gilbert, F.J.; Kaggie, J.D. The optimisation of deep neural networks for segmenting multiple knee joint tissues from MRIs. *Comput. Med. Imaging Graph.* **2020**, *86*, 101793. [CrossRef]
36. Norman, B.; Pedoia, V.; Majumdar, S. Use of 2D U-Net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry. *Radiology* **2018**, *288*, 177–185. [CrossRef] [PubMed]
37. Javed Awan, M.; Mohd Rahim, M.S.; Salim, N.; Mohammed, M.A.; Garcia-Zapirain, B.; Abdulkareem, K.H. Efficient detection of knee anterior cruciate ligament from magnetic resonance imaging using deep learning approach. *Diagnostics* **2021**, *11*, 105. [CrossRef] [PubMed]
38. Simantiris, G.; Tziritas, G. Cardiac mri segmentation with a dilated cnn incorporating domain-specific constraints. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 1235–1243. [CrossRef]
39. Coupé, P.; Mansencal, B.; Clément, M.; Giraud, R.; de Senneville, B.D.; Ta, V.-T.; Lepetit, V.; Manjon, J.V. AssemblyNet: A large ensemble of CNNs for 3D whole brain MRI segmentation. *NeuroImage* **2020**, *219*, 117026. [CrossRef] [PubMed]
40. Mohammed, M.A.; Abdulkareem, K.H.; Mostafa, S.A.; Ghani, M.K.A.; Maashi, M.S.; Garcia-Zapirain, B.; Oleagordia, I.; AlHakami, H.; Al-Dhief, F.T. Voice pathology detection and classification using convolutional neural network model. *Appl. Sci.* **2020**, *10*, 3723. [CrossRef]
41. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
42. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
43. Milletari, F.; Navab, N.; Ahmadi, S. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.