Advanced Architectural Variations of nnUNet

Niccolò McConnell[a,b,*], Nchongmaje Ndipenoch[a], Yu Cao[a], Alina Miron[a], Yongmin Li[a]

[a]*Department of Computer Science, Brunel University London, Kingston Lane, London, UB8 3PH, UK*
[b]*Institute of Health Informatics, University College London, Euston Road, London, NW1 2DA, UK*

Abstract

The nnUNet is a state-of-the-art deep learning based segmentation framework which automatically and systematically configures the entire network training pipeline. We extend the network architecture component of the nnUNet framework by newly integrating mechanisms from advanced U-Net variations including residual, dense, and inception blocks as well as three forms of the attention mechanism. We propose the following extensions to nnUNet, namely Residual-nnUNet, Dense-nnUNet, Inception-nnUNet, Spatial-Single-Attention-nnUNet, Spatial-Multi-Attention-nnUNet, and Channel-Spatial-Attention-nnUNet. Furthermore, within Channel-Spatial-Attention-nnUNet we integrate our newly proposed variation of the channel-attention mechanism. We demonstrate that use of the nnUNet allows for consistent and transparent comparison of U-Net architectural modifications, while maintaining network architecture as the sole independent variable across experiments with respect to a dataset. The proposed variants are evaluated on eight medical imaging datasets consisting of 20 anatomical regions which is the largest collection of datasets on which attention U-Net variants have been compared in a single work. Our results suggest that attention variants are effective at improving performance when applied to tumour segmentation tasks consisting of two or more target anatomical regions, and that segmentation performance is influenced by use of the deep supervision architectural feature.

Keywords:
Biomedical Image Segmentation, nnUnet, Residual, Dense, Inception, Attention

1. Introduction

Medical imaging is a valuable tool utilised by clinicians to assess a variety of anatomical structures [1]. Medical image segmentation supports clinicians during anatomical diagnosis by assigning anatomical labels to pixels in an image and thereby transforms raw images into spatially meaningful evidence [2]. At present, clinicians manually segment images which is a time consuming process with intra and inter observer variability [3]. Automatic segmentation provides an opportunity for significant impact as clinical adoption would increase reproducibility and improve clinical workflows, which is topical due to increased healthcare demands and clinician shortages [4].

Deep Learning (DL) methods are the state-of-the-art approach for tackling automatic medical image segmentation tasks, with the U-Net [5] being the most widely adopted network variation [6]. Currently, the DL based medical image segmentation literature focuses predominantly on network architecture and architectural modifications, such as the integration of residual, dense, or inception blocks, for achieving performance improvements with evaluation commonly conducted on a single dataset or restricted number of datasets [7, 8, 9, 10, 11, 12, 13, 14, 15]. However, in addition

*Corresponding Author. Email: niccolo.mcconnell.17@ucl.ac.uk

to network architecture, DL based automatic segmentation performance depends on further network training pipeline components and hyperparameters, for example, image resampling strategy, input image patch size, augmentation strategy etc. In fact, algorithms with identical architectures can display performance gaps due to differences in training pipeline component selections [6]. Furthermore, DL segmentation pipelines require dataset-dependent tailoring, which is subject to inter-expert variability, and could negatively impact a networks's potential performance if not optimally configured. Consequently, for a standardised and transparent evaluation of U-Net network modifications we propose that architectural modifications be integrated within nnUNet which is the state-of-the-art automatic segmentation framework for biomedical image segmentation [16].

Isensee et al. [16] developed the nnUNet framework which, for any segmentation task, self-configures the network training pipeline while considering computer-hardware capabilities and dataset specific properties. Its use of a systematic method for pipeline component selection resulted in best automatic segmentation performance on 33 of the 53 anatomical structures nnUNet was evaluated on while it otherwise achieved performance in line with the top challenge participants. Additionally, nnUNet utilises a standard U-Net type architecture which self-configures its topology, and therefore allows researchers considerable freedom to experiment with integrating more advanced U-Net modifications in order to evaluate performance within the state-of-the-art framework.

We believe that widespread adoption of nnUNet as a base framework in which to implement architectural modifications allows an opportunity for increased reproducibility, consistency and transparency for optimal model selection in DL based biomedical image segmentation. The main contributions of this article are summarised as follows:

- Demonstrate that utilisation of the nnUNet framework allows for comparison of U-Net architectural modifications while maintaining network architecture as the sole independent variable across experiments and allowing fixed training pipeline components across all comparisons with respect to a dataset.
- Extension of the nnUNet framework via the integration of advanced U-Net architectural components for performance gains, and thereby we propose the following six variations with source code provided [1]: Residual-nnUNet, Dense-nnUNet, Inception-nnUNet, Spatial-Single-Attention-nnUNet, Spatial-Multi-Attention-nnUNet, Channel-Spatial-Attention-nnUNet.
- In the Channel-Spatial-Attention-nnUNet we integrate our newly designed variation of the channel-attention mechanism.
- Evaluate our novel nnUNet extensions on eight 3D medical imaging datasets consisting of 20 anatomical regions which is the largest collection of datasets on which attention U-Net variants have been compared in a single work.

2. Background

2.1. nnUNet Extensions

The nnUnet automatic segmentation framework has been shown to attain segmentation performance in line or close to the state-of-the-art [16]. Furthermore, while nnUNet's impressive

[1]Source code available via: `https://github.com/niccolo246/Advanced_nnUNet.git`

performance was achieved utilising a standard U-Net within its architecture component, research regarding integration of advanced architectural features within the nnUNet architecture component currently remains limited. Isensee et al. [16] investigated use of a fully residual U-Net encoder, however the nnUNet model extension failed to outperform standard nnUNet on eight of the ten investigated datasets. Isensee et al. [17] extended nnUNet for optimal performance on 2020 Brain Tumour Segmentation (BraTS) Challenge [18, 19]. The proposed nnUNet extension achieved first place, however the architecture-specific extensions were limited to the inclusion of batch normalisation layers and hence no advanced architectural modificaions were investigated. Luu et al. [20] extended nnUNet for BraTS 2021 and achieved first place on the unseen test data, with architecture alterations including a deeper U-Net encoder, axial attention mechanism in the decoder, and replacement of batch normalisation with group normalisation. The authors hence demonstrated that extending nnUNet via the inclusion of advanced architectural can improve performance, however their work was limited to single dataset evaluation. Finally, to the best of our knowledge, our previous work [21] is the only instance in which nnUNet was utilised to compare several advanced network modifications, specifically, we integrated residual, dense, and inception blocks into the nnUNet. An important distinction and extension of our current article is the controlling of training pipeline hyperparameters which were previously not constant meaning that network architecture was not the only modified variable during comparisons; furthermore, we now also investigate the use of three variations of the attention mechanism.

2.2. Residual, Dense, and Inception U-Net Architectures

While the nnUNet framework has been able to achieve impressive segmentation performance by utilising a standard U-Net architecture, various works in the literature have reported that the use of advanced U-Net architectures attain improvements in biomedical image segmentation tasks relative to standard U-Net, with all comparisons conducted via expertly customised network training pipelines.

Residual connections were originally developed by He et al. [22], and their proposed network, ResNet, achieved top performance on the ImageNet challenge [23]. Consequently, residual connections have been integrated into U-Net, in both the downsampling and upsampling paths of the network architecture, and have been found to have positive impact on biomedical image segmentation performance. Various works have investigated the integration of residual connections in U-Net in order to tackle segmentation tasks on target anatomical regions including the lung (CT) [24, 25], brain tumours (MRI) [26], and white matter hyperintensities (MRI) [27, 13] with proposed modifications outperforming the standard U-Net.

DenseNet, proposed by Huang et al. [28], was designed with the aim to improve ResNet by instead making use of dense connections, with DenseNet surpassing ResNet on the ImageNet challenge. H-DenseUNet was proposed by Li et al. [29] which achieved top performance on the 2017 Liver Tumour Segmentation (LiTS) challenge [30] relative to other expert solutions. Their proposed model consisted of a hybrid densely connected U-Net consisting of both a 2D and 3D DenseUNet which were jointly optimized via a hybrid feature fusion (HFF) layer.

Inception-blocks were originally integrated within Google's Inception-Net [31]. Inception-blocks avoid choosing a fixed convolutional filter size, and instead make use of multiple filter sizes without incurring a high computational costs. Chen et al. [32] proposed S3D-UNet which made use of 3D inception blocks in order to learn richer feature representations for the 2018 BRATS challenge, with their model shown to outperform a standard 3D U-Net. Li et al. [33] proposed a modified U-Net which integrated inception blocks, while modifying the skip connection between upsampling and down-sampling paths, and utilising a cascaded training strategy. Their method was applied on the

2015 and 2016 BRATS segmentation dataset and surpassed alternative baseline 2D and 3D U-Net models. In their work, Rad et al. [34] proposed four novel U-Net inspired segmentation models, which were applied on trophectoderm segmentation in human embryo images. Their proposed Inception-U-Net model achieved the highest performance relative to alternative U-Net models, outperforming the state-of-the-art solution by 9.3% in Dice coefficient score.

Various works have proposed U-Net inspired architectures which make use of a combination of the aforementioned advanced architectural features. Dolz et al. [35] proposed a Dense multi-path U-Net for multi-modal ishemic stroke lesion segmentation. Their U-Net inspired model featured the use of multiple encoders to process each input modality individually, and also made use of inception modules in order to help handle varying lesion sizes. Furthermore, the authors ensured each of the input paths are densely connected to allow for learning of the scale at which modalities should be processed and combined. Meng et al. [36] proposed a U-Net inspired architecture for cerebrovascular segmentation in Digital Subtraction Angiography (DSA) images. The authors reported that their proposed architecture, which made use of a modified dense block for improved feature extraction and an inception inspired multiscale atrous convolutional module designed to capture multiscale vessel structure, attained superior performance relative to other advanced U-Net architectures. The DRI-Net was proposed by Chen et al. [37] which makes use of dense, residual, and inception modules integrated within a U-Net inspired architecture with the proposed model outperforming standard U-Net on multi-class segmentation of cerebrospinal fluid on brain CT images, multi-organ segmentation on abdominal CT images, and multi-class brain tumor segmentation on MR images. Ziang et al. [38] proposed a U-Net inspired architecture which made use of Inception-Res blocks and Dense-Inception blocks which combined inception with residual and dense features, respectively. Their model was evaluated on three tasks including retina blood vessel segmentation, CT lung segmentation, and the BraTS challenge with the proposed model achieving superior performance relative to alternative state-of-the-art U-Net variations.

2.3. Attention U-Net Architectures

With regards to integration of the attention mechanism, the original spatial-attention gate was proposed by Oktay et al. [39] who evaluated performance of their Attention U-Net for pancreas segmentation on two CT abdominal datasets consisting of 150 and 82 volumes, respectively. The authors focused on the use of multi-attention gates which aimed to highlight target structures of an input image. The results indicated attention gates showed an increase in the performance metrics relative to standard U-Net. Huang et al. [40] explored the use of several U-Net variations which included a U-Net with VGG-16 inspired encoder [41]. The authors experimented with residual connections and attention gates, with the resulting variants trained on a dataset of 910 median nerve images and tested on 207 images, with the VGG inspired Attention U-Net achieving best performance. Maji et al. [42] proposed Attention Res-UNet with Guided Decoder (ARU-GD). The architecture utilises spatial attention gates, deep supervision [43], and residual blocks in both the encoder and decoder. The model was applied to the 2019 BraTS challenge dataset and performed favourably relative to standard U-Net. Wu et al. [44] proposed the use of MSA-UNet which makes use of inception inspired blocks in the encoder, an Attention Atrous Spatial Pyramid Pooling (AASPP) module in the bottleneck layer, and a multi scale attention module in the skip connections which aim to increase the model's ability to learn spatially relevant contextual information from the encoder. The model was evaluated on a publicly available dataset of liver images in which it able to achieve top performance relative to other U-Net inspired architectures. Amer et al. [45] proposed Multi-Scale Dilated Attention U-Net. The authors proposed a channel attention module variation, inspired by Woo et al. [46], which encouraged selection of meaningful contextual

information along the channel dimension. Additionally, a multi-scale spatial attention module utilised a dilated convolutional module for capturing multi-scale contextual information. The authors evaluated performance on a lung segmentation dataset with 2628 axial CT images and an echocardiographic dataset containing 2000 images, and in both cases achieved superior performance relative to Attention U-Net. Wang et al. [47] proposed HDA-ResUNet which improves U-Net via the integration of residual blocks, a channel attention mechanism, and dilated convolutions. The authors claim that the channel attention block improves upon attention gate by taking advantage of the interdependencies between the feature map channels. The resulting model is evaluated on a liver, lung, nuclear, and neuron segmentation datasets, and is shown to attain better performance compared baseline U-Net. Wu et al. [48] proposed an Attention-based glioma grading network for MRI which made use of both spatial and channel attention mechanism which allowed for highlighting of key modalities and locations in feature maps. While the proposed network was not U-Net based, its use of attention allowed for robustenss and generalizability in the relative to other advanced models, thereby further demonstrating the potential for performance increase via the use of attention.

2.4. Synthesis

Overall, as discussed in sections 2.2, 2.3 there are numerous works in the literature which have evidenced that performance improvements for biomedical image segmentation tasks can be attained by integrating advanced architectural components into U-Net. However the aforementioned works report performance improvements resulting from network modifications using expertly tailored network training pipelines evaluated on a limited number of segmentation datasets and hence reported performance increases due to network extensions may not be reproducible or generalisable to new datasets and are also time-consuming to investigate as they require expert deep learning training pipeline adjustments. Meanwhile, the nnUNet fully automates the network training pipeline component selection, and has been shown to attain SOTA segmentation performance through utilising a standard U-Net [16], however, as discussed in section 2.1 there are limited works investigating the use of advanced architectures within the framework. In this work we therefore integrate advanced architectural features into the architectural component of the nnUNet framework, and demonstrate that nnUNet can be utilised for fair and transparent comparison of U-Net architectural modifications by maintaining the network training pipeline consistent, with resulting comparisons conducted on eight medical imaging datasets - the largest number of datasets on which Attention U-Net variants have been compared. Furthermore, due to the automated nature of the framework, network extensions can be swiftly evaluated on a range of datasets without requiring time-consuming network training pipeline tailoring.

3. Methods

We propose to implement U-Net architectural modifications into the architecture component of the nnUNet framework in order to allow consistent and transparent comparison of network variations on a range of datasets. In section 3.1 we provide a brief overview of nnUNet's systematic automated pipeline design process. In section 3.2 we discuss the standard architecture utilised by nnUNet, while in sections 3.3 - 3.8 we provide details of the architectural modifications we integrated within the framework whilst preserving its self-configuring nature.

3.1. nnUNet Overview

The nnUNet framework automatically self-configures the full network training pipeline by employing the systematic process extensively described in the original article by Isensee et al. [16].

In summary, once the framework is provided with training data, it obtains a "data-fingerprint" with information on the dataset's key properties such as modality, shape, and spacing. Based on data-fingerprint and GPU memory constraints, heuristic decisions are conducted to determine "rule-based parameters" which include the network topology, image resampling methods, and input-image patch sizes. Once training is complete the framework determines "empirical parameters" for output post-processing. Parameters which remain constant during training irrespective of the specific segmentation task, named "fixed parameters", include the use of Cross Entropy plus Dice loss function, ADAM optimiser [49], and type of data augmentation techniques done on the fly during training. Training duration is fixed at 1000 epochs with learning rate starting at 0.01 and decreasing according to the following: $(1 - epoch\_num/1000)^{0.9}$.

3.2. Standard-nnUNet

The network component of the Standard-nnUNet framework consists of a 3D U-Net inspired architecture which makes use of skip connections to allow feature maps from the encoder to be taken into account during the decoder's reconstruction process. The network consists of 3D convolutions for feature extraction, transposed convolutions for upsampling, and strided convolutions for downsampling. Furthermore, the network utilises deep supervision [43] in all but the two deepest levels in order to ameliorate gradient flow within the network. nnUNet will automatically configure the network topology in order to attain a feature map of size $4 \times 4 \times 4$ in the bottleneck layer. Topological modifications include network depth, kernel sizes, and stride parameters depending on the aforementioned dataset/ hardware-specific rule-based parameters. A Standard-nnUNet architecture representation is illustrated in Fig. 1. The network is taking as input $x \in \mathbb{R}^{C_x \times D_x \times H_x \times W_x}$, and outputting $y \in \mathbb{R}^{C_y \times D_x \times H_x \times W_x}$ in which C_y corresponds to the number of foreground classes, C_x is the input channel, D_x, H_x and W_x are the depth, height and width of the input image, respectively.

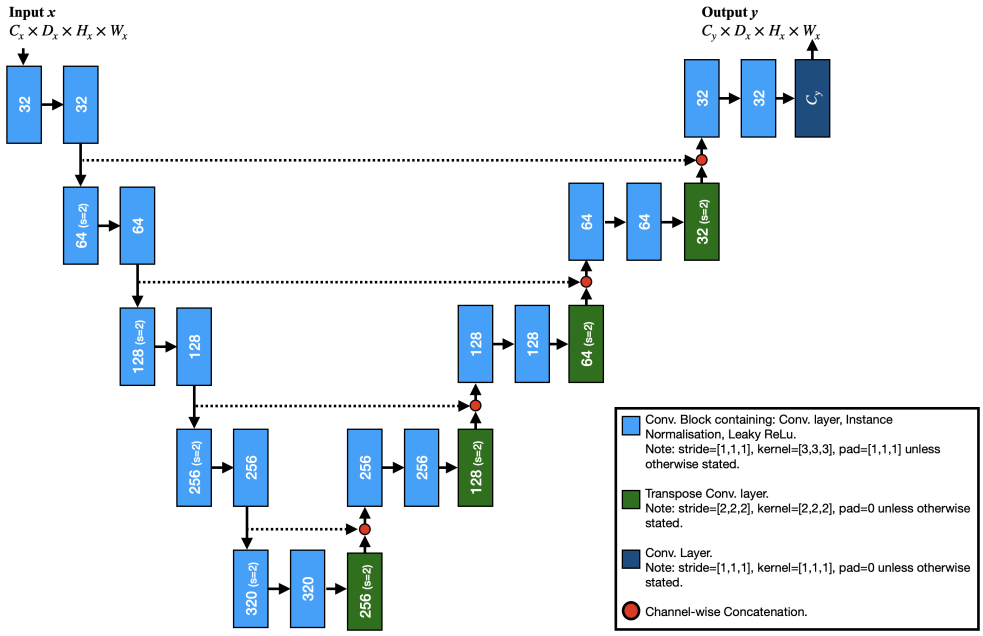

Figure (1) Baseline-nnUNet architecture representation.

3.3. Residual-nnUNet

We propose Residual-nnUNet for which we integrated a fully residual U-Net with the architecture component illustrated by the template in Fig. 2. The residual connection performs addition

of a convolutional block's input to its output. Residual connections allow for increased network depth without degradation in performance, which in turn allows networks to learn more discriminative features. Training of deep networks is improved due to enhanced gradient flow during backpropagation, which thereby helps alleviate potential vanishing gradient issues [50].

Our implementation integrates residual connections at depth level which implies that the residual connection will operate over the two convolutional blocks contained at each depth level. Therefore, for depth level $l \in \mathbb{Z}^+$ and input $x_{(l)} \in \mathbb{R}^{C_{x_{(l)}} \times D_{x_{(l)}} \times H_{x_{(l)}} \times W_{x_{(l)}}}$ our implementation of the residual-block may be denoted by $residual\_block_{(l)} = conv_{(l\_2)}(conv_{(l\_1)}(x_{(l)})) + x_{(l)}$.

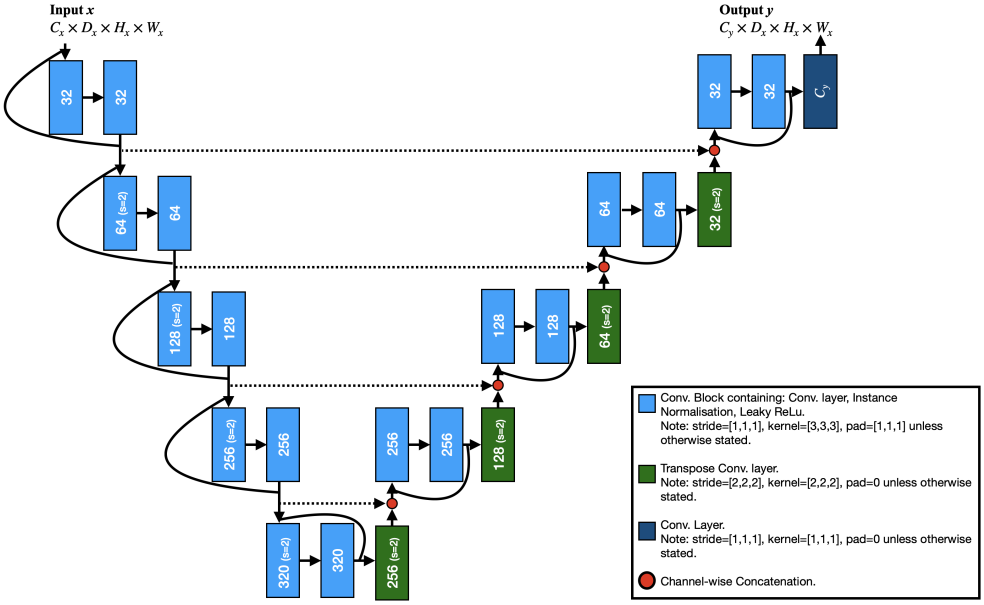

Figure (2) Residual-nnUNet architecture representation.

3.4. Dense-nnUNet

Our proposed Dense-nnUNet, inspired by DenseNet [28], was designed by integrating dense-blocks into the nnUNet architecture component via the template illustrated in Fig. 3, in which the orange blocks represent the dense-blocks visualised in Fig. 4. A dense-block contains a series of convolutional sub-blocks, with each subsequent sub-block receiving as input channel-wise concatenated feature maps outputted from the preceding sub-blocks. Each convolutional sub-block outputs K channels, where K denotes the growth rate. Dense connections theoretically improve gradient flow within deep networks, allowing for greater preservation of information between layers, and implicit deep supervision [28].

The dense-block we integrated contains four convolutional sub-blocks each with a growth rate K=10. Importantly, we utilise a $1 \times 1 \times 1$ kernel within the dense-block's final layer, in order for the block to output the nnUNet-predetermined number of channels to be passed on to the rest of the network - this allows for the framework's self-configuration abilities to be preserved. The stride, kernel, and padding parameters utilised in our dense-block are such that at depth level $l \in \mathbb{Z}^+$, and for dense-block input $x_{(l)} \in \mathbb{R}^{C_{x_{(l)}} \times D_{x_{(l)}} \times H_{x_{(l)}} \times W_{x_{(l)}}}$, the dense-block's output is $dense\_block(x_{(l)}) = y \in \mathbb{R}^{C_{y_{(l)}} \times D_{x_{(l)}} \times H_{x_{(l)}} \times W_{x_{(l)}}}$ i.e. the dimensions of the input and output are equivalent except for the channel dimension as shown in Fig. 4.

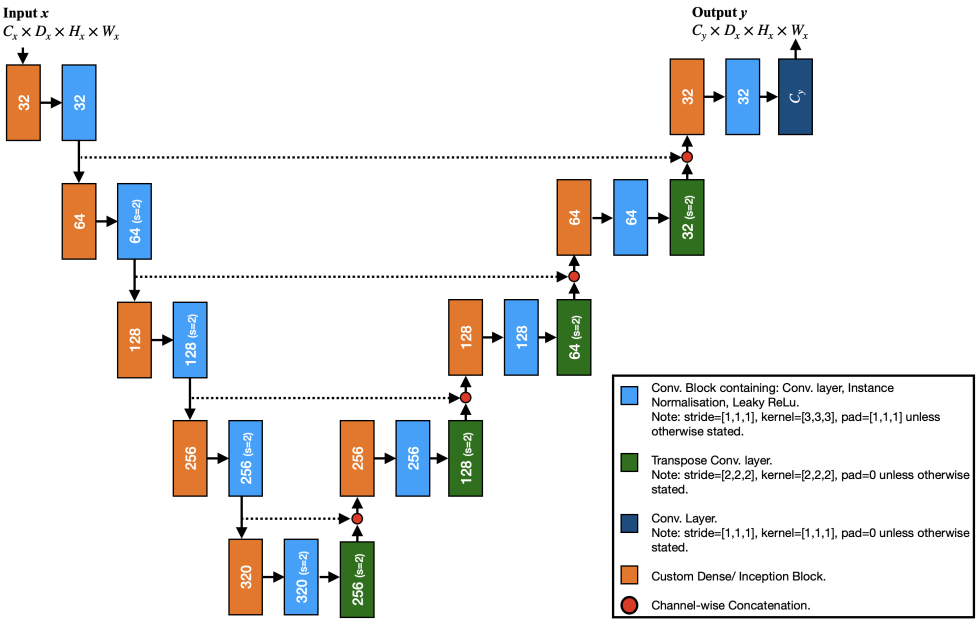

Figure (3) Dense-nnUNet or Inception-nnUNet architecture representation depending whether custom blocks (represented in orange) replaced with Dense-block (Fig. 4) or Inception-block (Fig. 5), respectively.

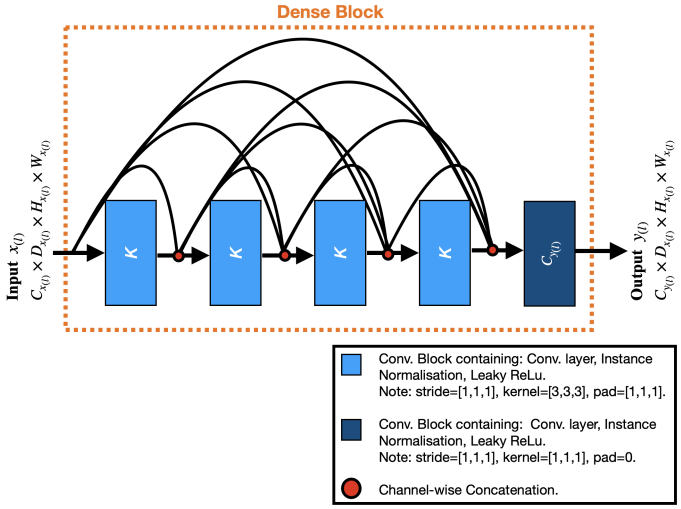

Figure (4) Dense-block representation.

3.5. Inception-nnUNet

We propose Inception-nnUNet, which makes use of the inception-block illustrated in Fig. 5. The block's input is passed to four branches which include kernel size $1 \times 1 \times 1$, kernel size $3 \times 3 \times 3$, kernel size $5 \times 5 \times 5$ and an average pool operation. The outputs of each branch are concatenated meaning the final output contains the nnUNet self-determined number of channels to be passed on to the rest of the network. To attain the correct number of channels for the block's final output, the penultimate layer of each sub-branch outputs a quarter of the desired number of final output channels, $C_{y_{(l)}}$ - alternatively one could use the $1 \times 1 \times 1$ convolutional layer, as was done with the dense-block in Fig. 4, and thereby have an arbitrary number of channels in individual branches, although our current design is more memory efficient. As with the dense-block (Section 3.4), stride, kernel, and padding parameters are automatically selected such that the inception-block's input and output are equivalent in dimensions except for the channel dimension.

Our proposed Inception-nnUNet integrates the custom inception-block into the nnUNet architecture component via the template illustrated in Fig. 3, where orange blocks represent the aforementioned inception-blocks shown in Fig. 5.

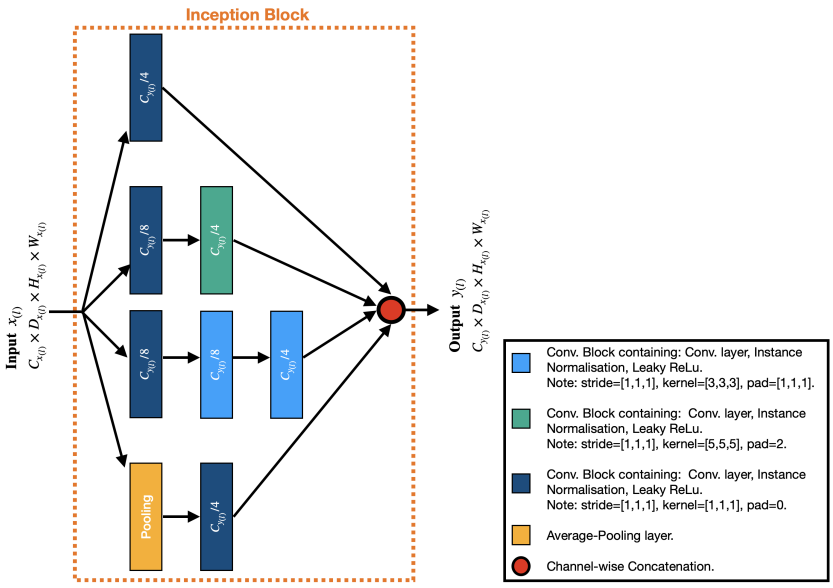

Figure (5) Inception-block representation.

3.6. Spatial-Single-Attention-nnUNet

Our proposed Spatial-Single-Attention-nnUNet was designed by integrating the spatial-single-attention gate illustrated in Fig. 7 into the nnUNet architecture component as illustrated by the template in Fig. 6, in which the purple blocks are replaced with the spatial-single-attention gates. The spatial-attention-gate was originally proposed by Oktay et al. [39]. For depth level, $l \in \mathbb{Z}^+$, the attention-block takes as input a gating signal, $g_{(l+1)} \in \mathbb{R}^{C_{g_{(l+1)}} \times D_{g_{(l+1)}} \times H_{g_{(l+1)}} \times W_{g_{(l+1)}}}$, and the feature map, $x_{(l)} \in \mathbb{R}^{C_{x_{(l)}} \times D_{x_{(l)}} \times H_{x_{(l)}} \times W_{x_{(l)}}}$, forwarded from the encoder via the skip connection, with the theoretical aim being to down-weight non-relevant spatial regions within $x_{(l)}$. Signal, $g_{(l+1)}$, originates from a depth level $l+1$, and is therefore at a coarser scale relative to $x_{(l)}$ which originates from depth level l. Therefore, $g_{(l+1)}$ originates from deeper in the decoder and provides increased contextual information which the gating mechanism can use to determine which spatial regions from the encoder's feature maps, $x_{(l)}$, forwarded via the skip connections are most relevant for the decoder's reconstruction process. The gate performs initial convolutional operations on both $x_{(l)}$ and $g_{(l+1)}$ with the resulting output channels set to $K \in \mathbb{Z}^+$. Input, $x_{(l)}$ is downsampled such that $\{conv_{x_{(l)}}(x_{(l)}), conv_{g_{(l+1)}}(g_{(l+1)})\} \in \mathbb{R}^{K \times D_{g_{(l+1)}} \times H_{g_{(l+1)}} \times W_{g_{(l+1)}}}$ which allows for the subsequent addition operation - for nnUNet compatibility this was achieved by utilising the same kernel, padding, and stride parameters for $conv_{x_{(l)}}(\cdot)$ that the encoder utilised for downsampling at corresponding depth level l. The sigmoid operation will output a grid of weights, which is upsampled, via trilinear interpolation, to be the same dimension as initially inputted $x_{(l)}$ with weight $\omega_i \in [0, 1]$. The upsampled grid, $\Omega \in \mathbb{R}^{1 \times D_{x_{(l)}} \times H_{x_{(l)}} \times W_{x_{(l)}}}$ is then multiplied with broadcasting along each of the channels in $x_{(l)}$, and therefore an increased weight is given to spatial regions of interest.

3.7. Spatial-Multi-Attention-nnUNet

The spatial-multi-attention block is illustrated in Fig. 8, with the block's inputs, $x_{(l)}$ and $g_{(l+1)}$, being equivalent to the inputs of the spatial-single-attention gate described in section 3.6. Our

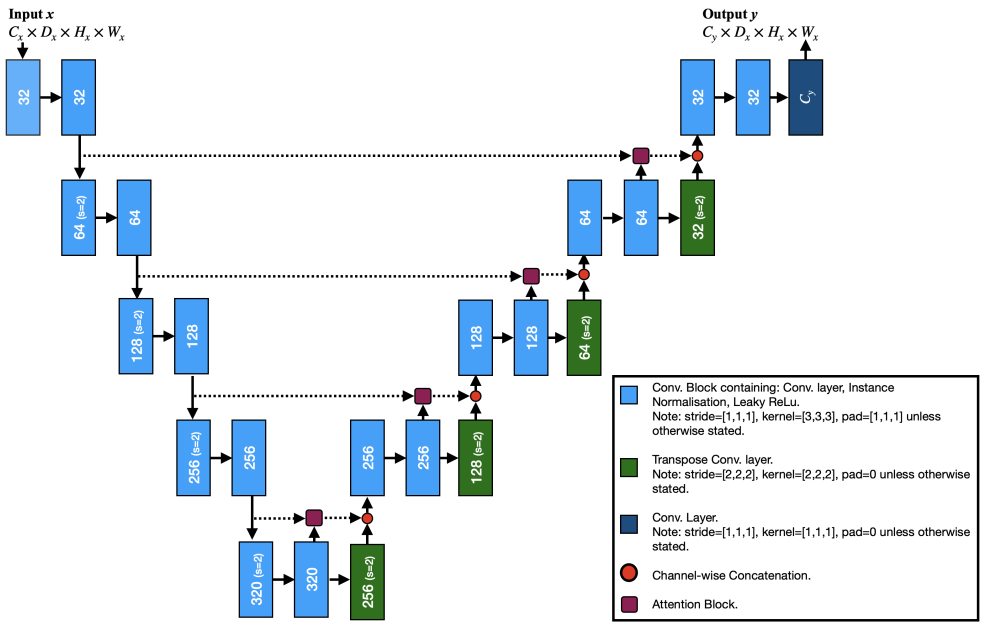

Figure (6) Spatial-Single-Attention-nnUNet, Spatial-Multi-Attention-nnUNet, or Channel-Spatial-Attention-nnUNet architecture representation depending o whether attention blocks (represented in purple) are replaced with Spatial-Single-Attention block (Fig. 7), Spatial-Multi-Attention block (Fig. 8), or Channel-Spatial-Attention block (Fig. 9), respectively.

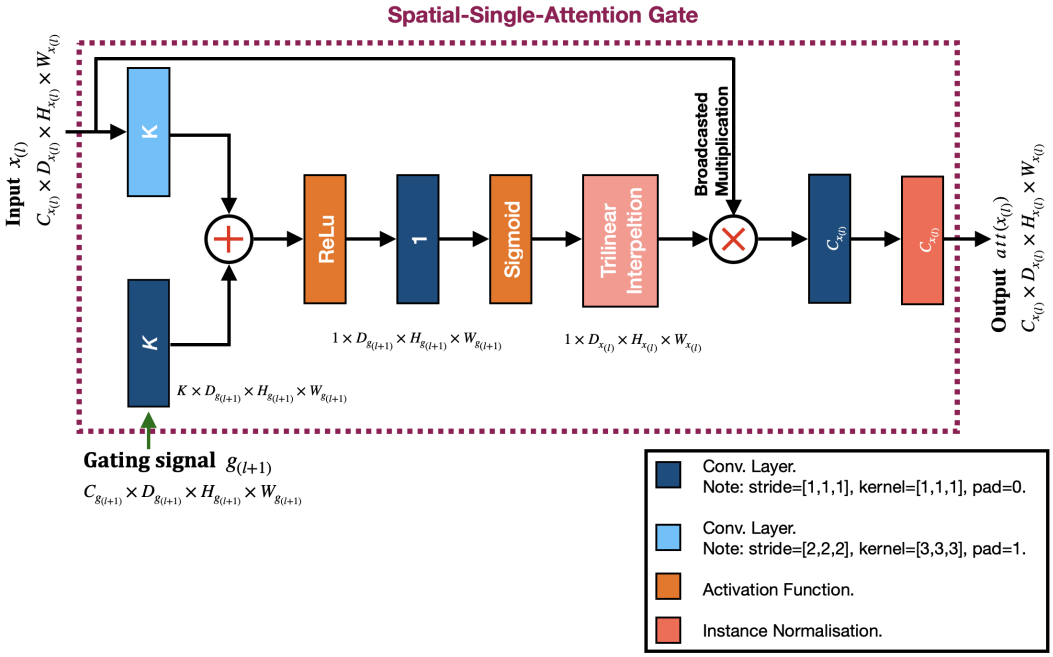

Figure (7) Spatial-Single-Attention block representation.

multi-attention block contains two spatial-single-attention gates α and β with $x_{(l)}$ and $g_{(l+1)}$ being inputted to each. The theoretical aim is for each respective attention gate, α and β, to focus on highlighting different spatial regions of interest from the feature maps forwarded from the encoder to the decoder via the skip connections - this was demonstrated empirically by Oktay et al. [39]. As illustrated in Fig. 8, in spatial-multi-attention block the outputs of the two spatial-single-attention gates are concatenated and passed to a convolutional block with $1 \times 1 \times 1$ kernel in order for the final output to be of the required dimension $\mathbb{R}^{C_{x_{(l)}} \times D_{x_{(l)}} \times H_{x_{(l)}} \times W_{x_{(l)}}}$. Our proposed Spatial-Multi-

Attention-nnUNet was designed by integrating the spatial multi-attention block illustrated in Fig. 8 into the nnUNet architecture component as illustrated by the template in Fig. 6, in which the purple blocks therefore represent the spatial-multi-attention gates.

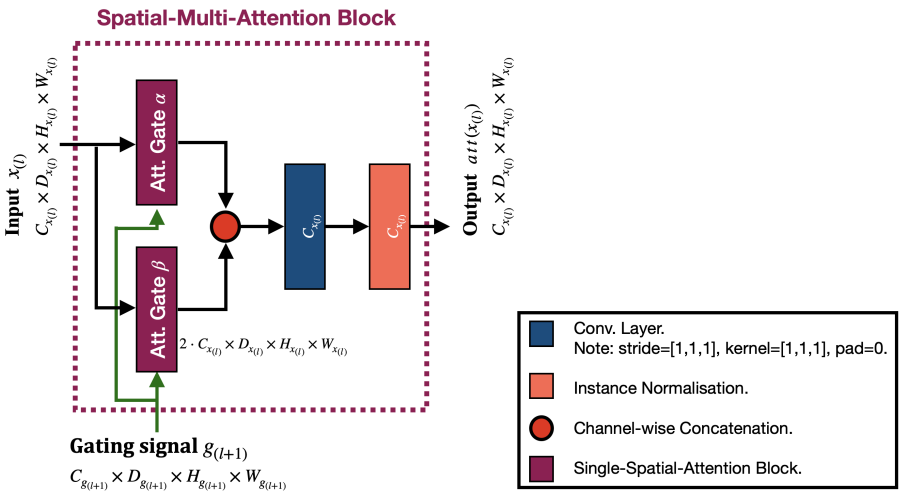

Figure (8) Spatial-Multi-Attention block representation.

3.8. Channel-Spatial-Attention-nnUNet

Our proposed channel-attention-block takes equivalent inputs as the spatial-single-attention gate discussed in section 3.6, however it utilises spatial-attention to highlight most-relevant spatial regions while also making use of channel-attention which aims to highlight which channels of the input feature map, $x_{(l)} \in \mathbb{R}^{C_{x_{(l)}} \times D_{x_{(l)}} \times H_{x_{(l)}} \times W_{x_{(l)}}}$, forwarded from encoder, are most relevant to the decoder by assigning a weight to each individual channel of $x_{(l)}$ via weight tensor $\Gamma \in \mathbb{R}^{C_{x_{(l)}} \times 1 \times 1 \times 1}$. In contrast, the spatial-attention block gives an equal weighting to all channels of $x_{(l)}$ via weight tensor $\Omega \in \mathbb{R}^{1 \times D_{x_{(l)}} \times H_{x_{(l)}} \times W_{x_{(l)}}}$. We designed our Channel-Spatial-Attention-nnUNet to take advantage of both channel and spatial attention by sequentially utilising a channel-attention-block and spatial-single-attention block; we refer to the overall block as the channel-spatial-attention block illustrated in Fig. 6. Channel-Spatial-Attention-nnUNet integrates our proposed channel-spatial-attention block illustrated in Fig. 9 into the nnUNet architecture component as illustrated by the template in Fig. 6, where the purple blocks represent channel-spatial-attention blocks. The design of our channel-attention-block is inspired by Woo et al. [46] and Amer et al. [45]. We added key modifications which consist in the replacement of fully connected layers with $1 \times 1 \times 1$ convolutional layers, and the replacement of an addition operation after the fully connected layers with a concatenation before the convolutions for improved information preservation and to maintain numerical distinction between max and average output for use by the subsequent convolutional layers [28]. A computationally inexpensive maximum and mean operation is utilised for channel-wise spatial aggregation - outputs are concatenated and passed to the fully convolutional layers, with the resulting outputs being summed and passed to a sigmoid function. Finally, the original input $x_{(l)}$ is multiplied with the channel weights $\Gamma \in \mathbb{R}^{C_{x_{(l)}} \times 1 \times 1 \times 1}$.

4. Experiments and Discussions

In section 4.1 we describe our approach for maintaining pipeline component consistency across the architectural comparisons with respect to an anatomical dataset. We then present the datasets

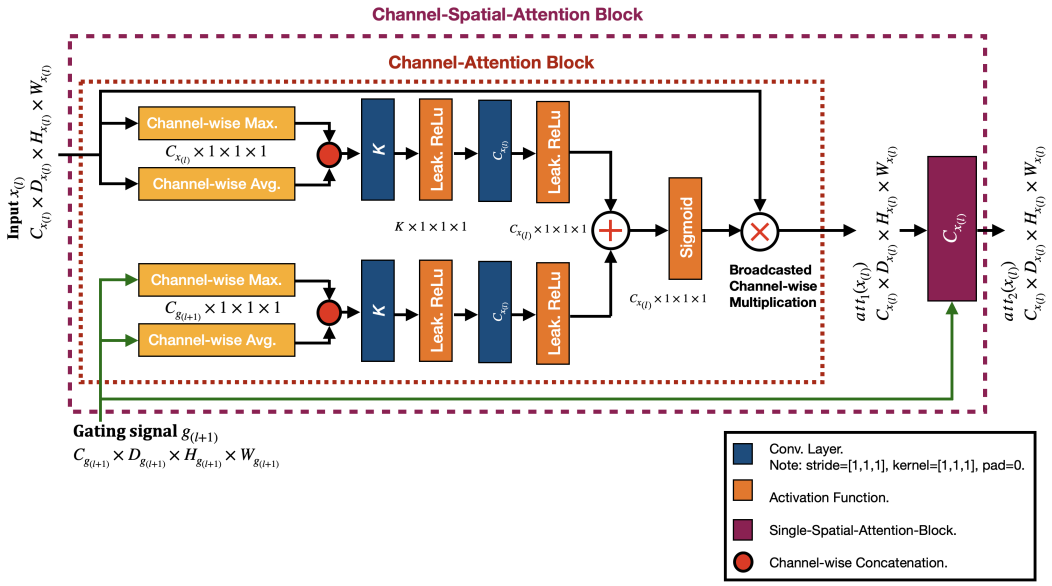

Figure (9) Channel-Spatial-Attention block representation.

utilised for the nnUNet variation comparison in section 4.2. Finally, results are presented and discussed in sections 4.3 and 4.4, respectively.

4.1. Pipeline Component Consistency and Network Training:

The training of each nnUNet variation was executed on an NVIDIA A6000 GPU, with training times ranging from 24-72 hours, depending on dataset properties. For reproducibility, we ensured training was completed with nnUNet deterministic training activated.

As explained at the start of Section 3, the automated design of the nnUNet training pipeline takes into account hardware memory constraints and, therefore, utilising a network architecture with increased memory requirements will alter other components in the training pipeline, such as the input image patch size, in order to keep overall memory required by the pipeline equal to the available memory provided by the hardware. In this article, we aim to compare architectural variations while maintaining training pipeline consistency with respect to a specific anatomical dataset, and therefore we hard-coded nnUNet network architecture additions as requiring zero additional memory. Our approach ensures pipeline components including input image patch size and base network topology (depth, kernel, and stride parameters) are all allowed to remain consistent across all network variations applied to a specific anatomical dataset. Hence, network architecture is the only variable changing between experiments. Importantly, we note that controlling training pipeline hyperparameters is one of this article's distinctions and extensions relative to our previous work [21].

We also note that since our nnUNet extensions maintain the framework's ability to self-configure the training pipeline, the network visualisations in Fig. 1, 2, 3 and 6 are meant to serve as conceptual illustrations. The depth of the neural networks will vary according to the specific network modifications based on the training dataset.

Table 1 provides a summary of the explored nnUNet variations.

4.2. Overview of Explored Datasets

Table 2 shows parameters for the evaluation datasets, with datasets D1-D7 derived from the Medical Segmentation Decathlon Challenge [51, 52] and dataset D8 originating from the 2021

Fetal Brain Tissue Annotation and Segmentation Challenge (FeTA) [53]. We conducted train and test splits locally using the challenge train sets since the challenge test sets did not have segmentation ground truths publicly available. For datasets D1-D8 we maintained a consistent train to test ratio in which the test set accounted for approximately 33% of the total cases for each respective dataset. Example visualisations of both the grayscale volume and the respective clinician annotated ground truth segmentation, for example cases from datasets D1-D8, are shown in Fig. 10. Table 2 shows that all the datasets have a restricted training dataset size, with D3, D4, D6 and D8 having under 100 training observations. Restricted dataset sizes are a common challenge in medical image segmentation due to the sensitive nature of the data, however, nnUNet utilises a 3D patch based approach for training in conjunction with on-the-fly augmentation which helps to address this issue (please refer to [16] for further details).

4.3. Results

As described in Section 4.1, our approach ensures pipeline components are consistent across all network variations applied to a specific anatomical dataset, which means network architecture is the independent variable in our experiments. Table 3 presents the nnUNet automatically selected hyperparameters relating to input image patch size, spacing of resampled image, and network depth which were kept constant across network variations with respect to each anatomical dataset. For our performance evaluation metric we utilise Dice score due to its widespread adoption in medical volume segmentation evaluation [55]. Given a segmentation prediction represented by set A, and a ground truth represented by set B, Dice score is computed as:

$$DSC(A, B) = \frac{2|A + B|}{|A| + |B|}$$

The average Dice scores for the explored nnUNet variations discussed in Section 3 are presented in Tables 4, 5, 6, 7, and the results are visualised in Fig. 11. We denote Standard-nnUNet to refer to the original nnUNet described in section 3.2 which utilises deep supervision, while Baseline-nnUNet is identical to Standard-nnUNet except for the removal of deep supervision. Residual-nnUNet, Dense-nnUNet, and Inception-nnUNet refer to our proposed nnUNet extensions which make use of residual (section 3.3), dense (section 3.4), and inception (section 3.5) blocks, respectively. Spatial-Single-Attention-nnUNet, Spatial-Multi-Attention-nnUNet, and Channel-Spatial-Attention-nnUNet refer to our proposed nnUNet variants which make use of spatial-single-attention (section 3.6), spatial-multi-attention (section 3.7), and our designed channel-spatial-attention block (section 3.8). Deep supervision is utilised in all the proposed nnUNet variations.

The results indicate that Standard-nnUNet and Baseline-nnUNet achieved best performance on at least one anatomical region in three and two of the eight datasets, respectively. Our proposed Residual-nnUNet, Dense-nnUNet, and Inception-nnUNet variants achieved best performance on at least one anatomical region in two, two, and zero, datasets, respectively. Our proposed attention nnUNet variants, namely Spatial-Single-Attention-nnUNet, Spatial-Multi-Attention-nnUNet, and Channel-Spatial-Attention-nnUNet achieved best performance on at least one anatomical region in two, four, and four, datasets, respectively. Overall, as shown in Fig. 12, Standard-nnUNet, Baseline-nnUNet, Residual-nnUNet, Dense-nnUNet, Inception-nnUNet, Spatial-Single-Attention-nnUNet, Spatial-Multi-Attention-nnUNet, and Channel-Spatial-Attention-nnUNet attained top performance on four, three, two, two, zero, five, five, and four of the 20 overall anatomical regions, respectively.

It is observed that on datasets with a single target region, specifically D4 and D6, none of the proposed nnUNet variants outperformed the Baseline-nnUNet (D4) and the Standard-nnUNet

Table (1) Summary of investigated nnUNet variants. Note that in all experiments with respect to a dataset, only U-Net architecture changes. Hence, network training pipeline components including, patch-size, resampling spacing, and network depth all remain constant - Refer to Table 2 for specific values of α, β, γ for each respective dataset. Relative memory gives an indication of the relative memory consumption of the network variations relative to Baseline-nnUNet memory μ - we note that since network topology changes for each dataset, a fixed memory value cannot be provided but this serves as an indication of relative memory proportions.

nnUNet Variant	**Component Blocks**	**Deep-supervision**	**Patch-Size**	**Spacing**	**Network-Depth**	**Relative Memory**
Standard-nnUNet	Standard (Fig. 1)	Yes	α	β	γ	1.01μ
Baseline-nnUNet	Standard (Fig. 1)	No	α	β	γ	μ
Residual-nnUNet	Residual (Fig. 2)	Yes	α	β	γ	1.20μ
Dense-nnUNet	Dense (Fig. 4)	Yes	α	β	γ	2.58μ
Inception-nnUNet	Inception (Fig. 5)	Yes	α	β	γ	1.55μ
Spatial-Single-Attention-nnUNet	Spatial-Single-Attention (Fig. 7)	Yes	α	β	γ	1.17μ
Spatial-Multi-Attention-nnUNet	Spatial-Multi-Attention (Fig. 8)	Yes	α	β	γ	1.41μ
Channel-Spatial-Attention-nnUNet	Channel-Spatial-Attention (Fig. 9)	Yes	α	β	γ	1.17μ

Table (2) Summary of explored datasets. MRI—magnetic resonance imaging, FLAIR—fluid-attenuated inversion recovery, T1w—T1 weighted image, T1w_Gd—post-Gadolinium (Gd) contrast T1-weighted image, T2w—T2 weighted image, CT—computed [51].

Dataset	**Modality**	**Regions of Interest**	**Median Volume Size (Voxel)**	**Median Volume Spacing (mm)**	**No. Cases (train/valid/test)**
D1 - Brain	Multi-modal MRI (FLAIR, T1w, T1w_Gd, T2w)	Edema, non-enhancing tumour, enhancing tumour	[137, 169, 138]	[1.00, 1.00, 1.00]	484 (257/ 65 / 162)
D2 - Hippocampus	Mono-modal MRI	Anterior hippocampus, posterior hippocampus	[40, 56, 40]	[1.00, 1.00, 1.00]	260 (137/ 35 / 88)
D3 - Liver	Portal venous phase CT	Liver, liver tumour	[482, 512, 512]	[1.00, 0.76, 0.76]	129 (68/ 18 / 43)
D4 - Lung	CT	Lung cancer	[253, 512, 512]	[1.25, 0.79, 0.79]	63 (33/ 9 / 21)
D5 - Pancreas	Portal venous phase CT	Pancreas, pancreatic tumour mass	[96, 512, 512]	[2.50, 0.79, 0.79]	280 (148/ 38 / 94)
D6 - Colon	CT	Colon cancer primaries	[152, 512, 512]	[3.00, 0.78, 0.78]	126 (67/ 17 / 42)
D7 - Hepatic Vessels	CT	Hepatic vessels, hepatic tumour	[150, 512, 512]	[1.50, 0.80, 0.80]	303 (161/ 41 / 101)
D8 - Fetal Brain	Mono-modal MRI	External cerebrospinal fluid, grey matter, white matter, ventricles, cerebellum, deep grey matter, brainstem/ spinal-cord	[256, 256, 256]	[0.50, 0.50, 0.50]	80 (44/ 12/ 24)

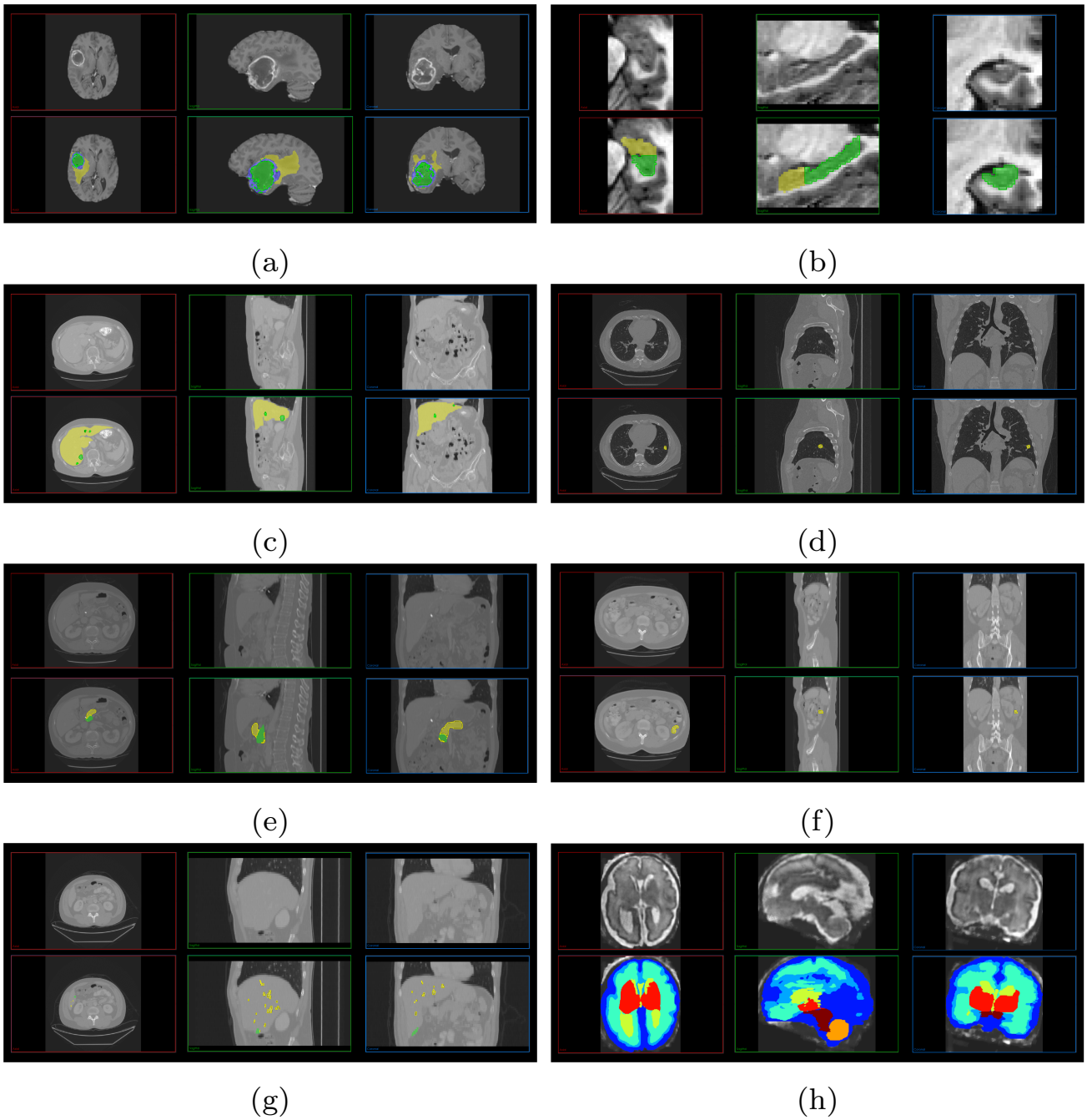

Figure (10) Example visualisations of the explored datasets. In each subfigure, axial (red), sagittal (green), and coronal (blue) projections of the 3D volume are presented, with the raw greyscale scan presented in the upper portion and its respective ground truth segmentation presented below, for each respective example case.
(a) D1-Brain: Edema (yellow), non-enhancing tumour (green), enhancing tumour (violet). (b) D2-Hippocampus: Anterior hippocampus (yellow), posterior hippocampus (green). (c) D3-Liver: Liver (yellow), liver tumour (green). (d) D4-Lung: Lung cancer (yellow). (e) D5-Pancreas: Pancreas (yellow), pancreatic tumour mass (green). (f) D6-Colon: Colon cancer primaries (yellow). (g) D7-Hepatic Vessel: Hepatic vessels (yellow), hepatic tumour (green). (h) D8-Fetal Brain: External cerebrospinal fluid (dark blue), grey matter (pastel blue), white matter (neon blue), ventricles (yellow), cerebellum (orange), deep grey matter (red), brainstem (maroon). Note all visualisations created via The Medical Imaging Interaction Toolkit [54]

Table (3) Table presenting selection of hyperparameters automatically chosen by nnUNet with respect to each dataset - these hyperparameters were kept constant across each of the experimented nnUNet architecture variants. Patch size is the image patch size inputted to network during training/inference; resampled spacing is the spacing selected for the resampled input image; network depth is the depth of the U-Net type architecture component.

Dataset	nnUNet Patch Size	nnUNet Spacing (mm)	nnUNet Network Depth
D1 - Brain	[96, 128, 96]	[1.00, 1.00, 1.00]	6
D2 - Hippocampus	[40, 56, 40]	[1.00, 1.00, 1.00]	4
D3 - Liver	[96, 112, 112]	[2.99, 2.27, 2.27]	5
D4 - Lung	[64, 128, 128]	[1.23, 0.79, 0.79]	6
D5 - Pancreas	[40, 224, 224]	[2.50, 0.80, 0.80]	6
D6 - Colon	[64, 224, 224]	[3.00, 0.78, 0.78]	6
D7 - Hepatic Vessels	[48, 160, 160]	[1.50, 0.80, 0.80]	6
D8 - Fetal Brain	[96, 128, 96]	[0.50, 0.50, 0.50]	6

(D6). Relative to Standard-nnUNet, Baseline-nnUNet attained a lower average Dice score on 14 of the 20 investigated anatomical regions which suggests that removal of deep supervision generally resulted in decreased performance. Overall Residual-nnUNet and Dense-nnUNet attained marginal differences in performance relative to Standard-nnUNet and Baseline-nnUNet. The residual and dense variants attained a performance improvement of up to 2.83% and 8.10%, respectively, compared to Standard-nnUNet; which compares to an increased performance improvement of up to 5.35% and 13.36%, respectively, when Residual-nnUNet and Dense-nnUNet variants are, respectively, compared to Baseline-nnUNet.

The results indicate that spatial-multi-attention and channel-spatial-attention nnUNet variants tended to attain their best relative performance on datasets which consisted of two target anatomical regions with minority region consisting of tumour, namely D2, D3, D5. On dataset D3 Channel-Spatial-Attention-nnUNet attained a 8.30% increase in average dice score for the tumour region relative to the worst performing variant, Inception-nnUNet; meanwhile for dataset D5 Spatial-Multi-Attention-nnUNet attained a 14.29% increase in average Dice for the tumour region relative to Baseline-nnUNet. Interestingly, we note that the Spatial-Single-Attention-nnUNet attained top performance on four of the seven brain regions in dataset D8.

In terms of performance variations from the use of different forms of attention, we observe that Spatial-Multi-Attention-nnUNet displayed a percentage difference in average Dice score compared to Spatial-Single-Attention-nnUNet ranging up to 2.57% in dataset D4 and down to -5.56% in dataset D3. Channel-Spatial-Attention-nnUNet displayed a percentage difference in average Dice score compared to Spatial-Single-Attention-nnUNet ranging up to 3.45% in dataset D6 and down to -1.43% in dataset D1. The results indicate that choice of the optimal attention variant is dataset dependent.

4.4. Discussions

In our experiments, we observed that the attention mechanism was most effective on datasets consisting of two or more anatomical regions. In particular, the attention mechanism improved the performance of the anatomical region, which was spatially a minority region within the original volume - in our explored datasets, this region was a tumour. We hypothesise the performance gain is due to the nature of the attention mechanism being to highlight spatially relevant regions, which can therefore increase the relative weighting of minority class voxels in a restricted region of

Table (4) Average Dice score for anatomical regions in datasets D1-D2. Top score presented in bold.

nnUNet Variant	D1 Edema	D1 Non-Enhancing Tumour	D1 Enhancing Tumour	D2 Anterior	D2 Posterior
Standard-nnUNet	0.798	**0.621**	**0.797**	0.893	0.879
Baseline-nnUNet	0.799	0.619	0.790	0.893	0.878
Residual-nnUNet	0.798	0.613	0.793	0.892	0.879
Dense-nnUNet	0.799	0.614	0.788	**0.894**	0.879
Inception-nnUNet	0.800	0.617	0.790	0.892	0.879
Spatial-Single-Attention-nnUNet	0.799	0.618	0.792	0.892	0.879
Spatial-Multi-Attention-nnUNet	**0.801**	0.617	0.788	0.893	**0.881**
Channel-Spatial-Attention-nnUNet	0.800	**0.621**	0.781	0.893	**0.881**

Table (5) Average Dice score for anatomical regions in datasets D3-D5. Top score presented in bold.

nnUNet Variant	D3 Liver	D3 Tumour	D4 Cancer	D5 Pancreas	D5 Tumour
Standard-nnUNet	0.947	0.663	0.716	0.823	0.486
Baseline-nnUNet	0.943	0.670	**0.743**	0.820	0.463
Residual-nnUNet	**0.956**	0.667	0.736	0.822	0.488
Dense-nnUNet	0.941	0.639	0.729	0.820	0.525
Inception-nnUNet	0.940	0.685	0.727	0.819	0.470
Spatial-Single-Attention-nnUNet	0.945	0.675	0.697	**0.824**	0.517
Spatial-Multi-Attention-nnUNet	0.946	0.638	0.715	**0.824**	**0.529**
Channel-Spatial-Attention-nnUNet	0.948	**0.691**	0.714	**0.824**	0.528

Table (6) Average Dice score for a anatomical regions in datasets D6-D8. Top score presented in bold.

nnUNet Variant	D6 Cancer	D7 Vessels	D7 Tumour	D8 External Cerebrospinal Fluid	D8 Grey Matter
Standard-nnUNet	**0.464**	0.636	0.709	0.789	0.737
Baseline-nnUNet	0.457	0.632	0.707	**0.791**	0.737
Residual-nnUNet	0.420	**0.640**	0.707	0.785	0.735
Dense-nnUNet	0.445	0.630	**0.718**	0.782	0.731
Inception-nnUNet	0.401	0.638	0.717	0.789	0.735
Spatial-Single-Attention-nnUNet	0.444	0.637	0.711	0.790	**0.739**
Spatial-Multi-Attention-nnUNet	0.445	0.639	0.716	0.787	0.735
Channel-Spatial-Attention-nnUNet	0.459	0.634	0.717	0.788	0.737

Table (7) Average Dice score for a subset of anatomical regions in dataset D8. Top score presented in bold.

nnUNet Variant	D8 White Matter	D8 Ventricles	D8 Cerebellum	D8 Deep Grey Matter	D8 Brainstem
Standard-nnUNet	**0.914**	0.887	0.881	0.870	**0.833**
Baseline-nnUNet	**0.914**	0.885	**0.884**	0.870	0.831
Residual-nnUNet	0.913	0.883	0.877	0.870	0.830
Dense-nnUNet	0.913	0.880	0.871	0.871	0.828
Inception-nnUNet	0.912	0.882	0.880	0.870	0.828
Spatial-Single-Attention-nnUNet	**0.914**	**0.888**	0.882	**0.874**	0.831
Spatial-Multi-Attention-nnUNet	**0.914**	0.885	0.882	0.873	0.831
Channel-Spatial-Attention-nnUNet	0.913	0.884	0.882	0.873	0.831

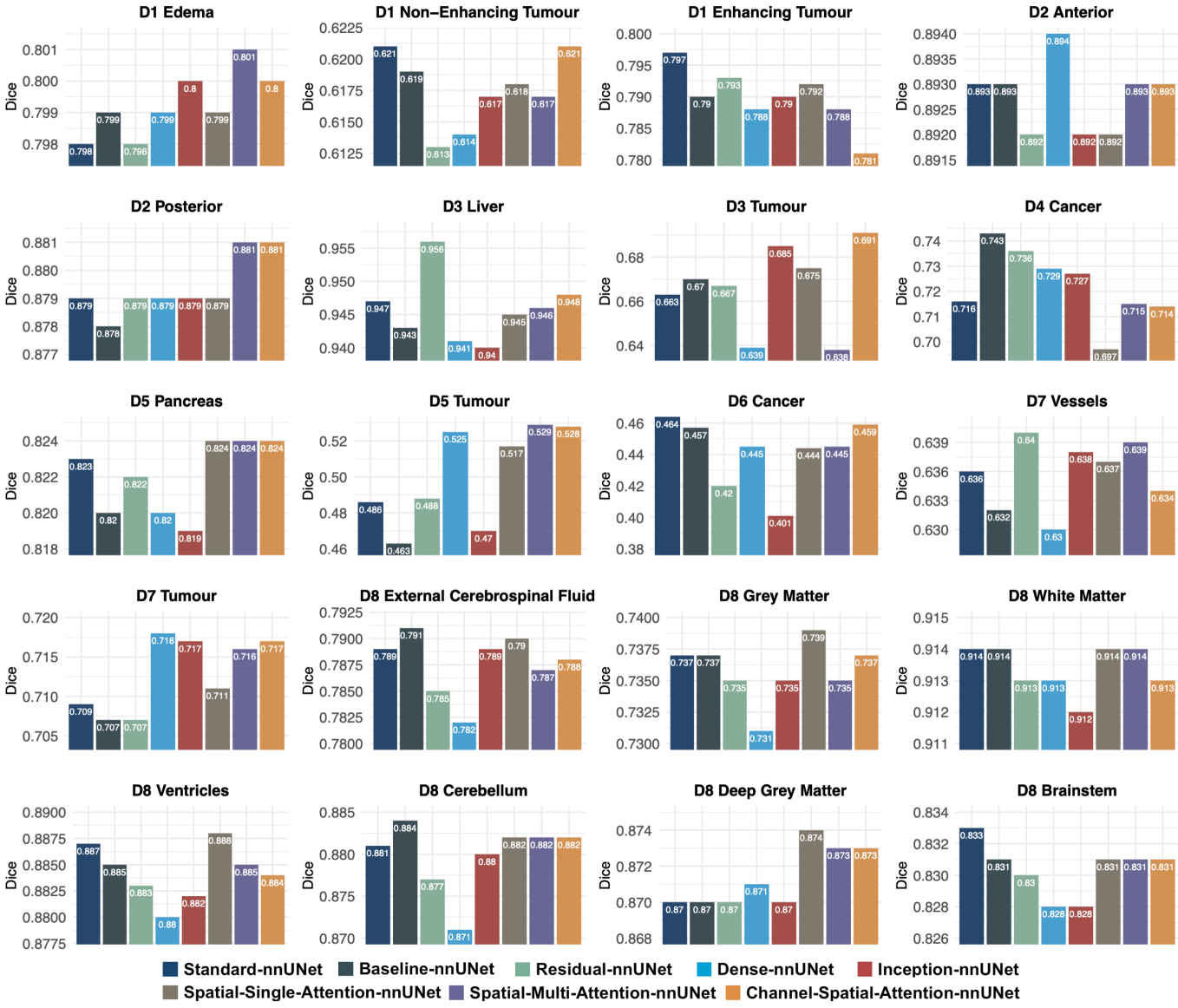

Figure (11) Graph showing average Dice scores attained by each nnUNet variation for the anatomical regions from datasets D1 - D8.

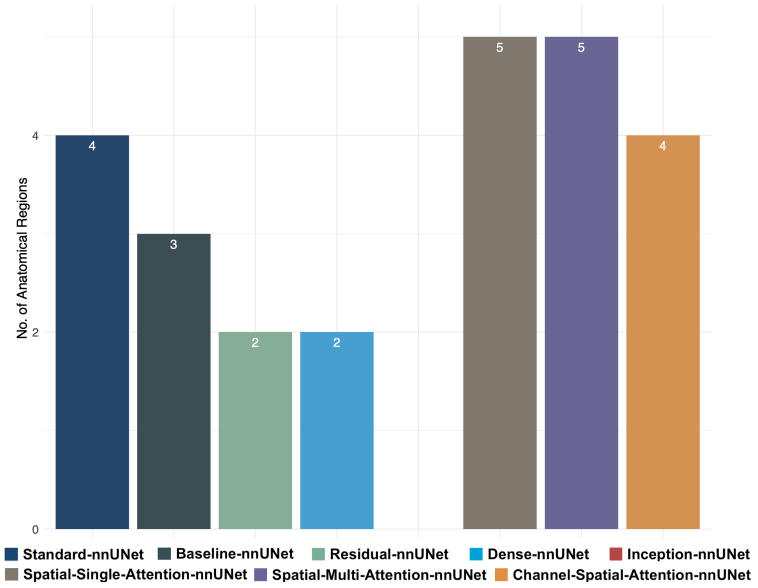

Figure (12) Graph showing number of anatomical regions for which each respective nnUNet variation achieved top performance.

the initial volume. Results on dataset D8, consisting of seven anatomical regions, suggested that Spatial-Single-Attention-nnUNet was the relatively best performing model, which was contrary to our expectation as, according to theory, spatial-multi-attention gates were expected to outperform spatial-single-attention gates on datasets with multiple target regions. We hypothesise that the use of spatial-single-attention gates outperformed the use of spatial-multi-attention gates due to their reduced trainable parameter number, which leads to decreased risk of overfitting on the training set, which is especially relevant for dataset D8 with 33 training cases.

In fact, from the results we notice that Standard-nnUNet and Baseline-nnUNet outperformed the architecturally more advanced nnUNet variations on the D4 (Lung) and D6 (Colon) datasets, with both datasets consisting of a single target anatomical region. Observing Table 2, both datasets D4 and D6 had relatively small training sets consisting of 33 and 67 cases, respectively - even though early-stopping was employed, we note that the size of the validation sets were also restricted at 9 and 17, respectively. We hypothesise that since Standard-nnUNet and Baseline-nnUNet did not contain advanced architectural components, they were therefore less prone to overfitting to the small number of training cases and, thereby, able to generalise better to the respective test sets. Dataset D8, as mentioned, consisted of 44 (12) training (validation) set cases, with Standard-nnUNet and Baseline-nnUNet achieving top performance on one and two regions, respectively out of the 7 total regions. As discussed, Single-Attention-nnUNet achieved best performance on the other four regions, and we believe this was due to the spatial-single-attention gate being able to highlight spatially relevant regions while benefiting from less trainable parameters relative to spatial-multi-attention gate, which led to decreased risk of overfitting on the restricted training set and, therefore, improved test set generalisation ability.

It was shown that Standard-nnUNet tended to outperform Baseline-nnUNet likely due to the reduced gradient flow during backpropagation, resulting from the removal of deep-supervision; consequently, the performance gap suggests that deep supervision is beneficial to segmentation performance. Dense-nnUNet and Residual-nnUNet generally resulted in marginal performance differences relative to Standard-nnUNet, and this may be due to all three variations making use of deep supervision. Residual and dense connections both aim to improve gradient flow during

network training, and so the effectiveness of these connection types may be limited due to gradient flow already being enhanced via the incorporation of deep supervision. Furthermore, it was observed that Inception-nnUnet achieved second best performance on three of the 20 anatomical regions, while it failed to achieve top performance on any of the investigated datasets. This therefore suggest use of the standard $3 \times 3 \times 3$ convolutional kernel tended to result in superior performance relative to utilising a combination of larger and smaller kernel sizes as done by the inception block. Nevertheless, use of Inception-nnUnet on an alternative dataset may result in performance gains.

The results suggest that while no single architectural variant performs optimally on all datasets, specific network modifications may result in marginal or considerable gains in performance. The marginal performance differences were expected due to the nnUNet generally performing in line with the state-of-the-art, with performance differences for top-ranking segmentation challenge submissions also tending to be marginal. Therefore, given a dataset, selecting the optimal nnUNet architectural variation may improve performance enough to establish a state-of-the-art result.

A current limitation of our work is the restricted number of datasets investigated. While this is the most extensive collection of datasets on which the attention mechanism has been evaluated, increasing the dataset number would allow for more generalisable conclusions.

Potential future avenues of exploration include investigating performance increases from combining the different network architectures explored in this work, and thereby taking advantage of the nnUnet built-in ensembling mechanism of averaging softmax predictions from several of the best performing proposed models. Finally, we have demonstrated the integration of advanced U-Net components however, Visual Transformers [56] have started to achieve SOTA results in several image analysis tasks, and hence investigating the integration of the visual transformer for segmentation within the automated nnUNet framework may be a promising avenue for exploration especially when provided with large medical imaging datasets.

Overall, we have demonstrated that U-Net architectural variants can be implemented within the state-of-the-art nnUNet framework, allowing for the comparison of network architecture modifications evaluated on several datasets. We believe that the widespread adoption of nnUNet, within the deep-learning-based biomedical image segmentation community, as a base framework for implementing architectural modifications, allows for two key advantages, which we have evidenced in this article. The first advantage is improved transparency and consistency across all the training pipeline components of network variations with respect to a specific dataset, allowing network architecture to be the sole independent variable across experiments. The second key advantage regards nnUNet's automatic generalisability to new datasets, allowing researchers to rapidly evaluate and compare newly proposed U-Net architectural variations implemented within nnUNet without requiring training pipeline component alterations. Hence, any U-Net architecture variation, once integrated into nnUNet, is ready to be rapidly utilised and evaluated "out of the box" on any given dataset without needing dataset-dependent pipeline tailoring.

5. Conclusions

In this article we have extended the nnUNet framework via the integration of advanced architectural components including residual, dense, inception, and attention gates, resulting in six new nnUNet variations including the Residual-nnUNet, Dense-nnUNet, Inception-nnUNet, Spatial-Single-Attention-nnUNet, Spatial-Multi-Attention-nnUNet, and Channel-Spatial-Attention-nnUNet. We have demonstrated that the nnUNet framework allows for consistent and transparent compari-

son of advanced U-Net variations with respect to a given dataset. We have evaluated the proposed models on eight medical imaging datasets consisting of 20 anatomical structures.

Experimental results indicate that the optimal architectural variation is dataset dependent. While no single architectural variation performs dominantly on all datasets, the choice of certain variations may offer a marginal or significant performance gain. In particular, (1) on datasets consisting of a single target anatomical regions, the Standard-nnUNet and Baseline-nnUNet should be able do well, therefore no need to use the advanced network variants; (2) on datasets consisting of two or more anatomical regions and especially on minority regions in spatially imbalanced tasks, the attention nnUNet variants tend to perform best.

References

[1] E. S. of Radiology (ESR) communications@ myesr. org, Medical imaging in personalised medicine: a white paper of the research committee of the european society of radiology (esr), Insights into imaging 6 (2015) 141–155.

[2] L. Liu, J. Cheng, Q. Quan, F.-X. Wu, Y.-P. Wang, J. Wang, A survey on u-shaped networks in medical image segmentations, Neurocomputing 409 (2020) 244–258.

[3] I. R. I. Haque, J. Neubert, Deep learning approaches to biomedical image segmentation, Informatics in Medicine Unlocked 18 (2020) 100297.

[4] F. K. Lock, D. Carrieri, Factors affecting the uk junior doctor workforce retention crisis: an integrative review, BMJ open 12 (3) (2022) e059397.

[5] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical image analysis 42 (2017) 60–88.

[7] Y. Zhang, A. Chung, Deep supervision with additional labels for retinal vessel segmentation task, in: International conference on medical image computing and computer-assisted intervention, Springer, 2018, pp. 83–91.

[8] H. A. Leopold, J. Orchard, J. S. Zelek, V. Lakshminarayanan, Pixelbnn: augmenting the pixelcnn with batch normalization and the presentation of a fast architecture for retinal vessel segmentation, Journal of Imaging 5 (2) (2019) 26.

[9] W. Dai, N. Dong, Z. Wang, X. Liang, H. Zhang, E. P. Xing, Scan: Structure correcting adversarial network for organ segmentation in chest x-rays, in: Deep learning in medical image analysis and multimodal learning for clinical decision support, Springer, 2018, pp. 263–273.

[10] E. Giacomello, D. Loiacono, L. Mainardi, Brain mri tumor segmentation with adversarial networks, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–8.

[11] S. Wang, L. Yi, Q. Chen, Z. Meng, H. Dong, Z. He, Edge-aware fully convolutional network with crf-rnn layer for hippocampus segmentation, in: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), IEEE, 2019, pp. 803–806.

[12] P. F. Christ, F. Ettlinger, F. Grün, M. E. A. Elshaera, J. Lipkova, S. Schlecht, F. Ahmaddy, S. Tatavarty, M. Bickel, P. Bilic, et al., Automatic liver and tumor segmentation of ct and mri volumes using cascaded fully convolutional neural networks, arXiv preprint arXiv:1702.05970 (2017).

[13] D. Jin, Z. Xu, A. P. Harrison, D. J. Mollura, White matter hyperintensity segmentation from t1 and flair images using fully convolutional neural networks enhanced with residual connections, in: 2018 IEEE 15th International Symposium on biomedical imaging (ISBI 2018), IEEE, 2018, pp. 1060–1064.

[14] N. Ibtehaz, M. S. Rahman, Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation, Neural networks 121 (2020) 74–87.

[15] Q. Jin, Z. Meng, C. Sun, H. Cui, R. Su, Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans, Frontiers in Bioengineering and Biotechnology (2020) 1471.

[16] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, K. H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nature methods 18 (2) (2021) 203–211.

[17] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, K. H. Maier-Hein, nnu-net for brain tumor segmentation, in: International MICCAI Brainlesion Workshop, Springer, 2020, pp. 118–132.

[18] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge, arXiv preprint arXiv:1811.02629 (2018).

[19] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (brats), IEEE transactions on medical imaging 34 (10) (2014) 1993–2024.

[20] H. M. Luu, S.-H. Park, Extending nn-unet for brain tumor segmentation, arXiv preprint arXiv:2112.04653 (2021).

[21] N. McConnell, A. Miron, Z. Wang, Y. Li, Integrating residual, dense, and inception blocks into the nnunet, in: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2022, pp. 217–222.

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

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International journal of computer vision 115 (3) (2015) 211–252.

[24] A. Khanna, N. D. Londhe, S. Gupta, A. Semwal, A deep residual u-net convolutional neural network for automated lung segmentation in computed tomography images, Biocybernetics and Biomedical Engineering 40 (3) (2020) 1314–1327.

[25] H. Cheng, Y. Zhu, H. Pan, Modified u-net block network for lung nodule detection, in: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), IEEE, 2019, pp. 599–605.

[26] A. Kermi, I. Mahmoudi, M. T. Khadir, Deep convolutional neural networks using u-net for automatic brain tumor segmentation in multimodal mri volumes, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 37–48.

[27] R. Guerrero, C. Qin, O. Oktay, C. Bowles, L. Chen, R. Joules, R. Wolz, M. d. C. Valdés-Hernández, D. A. Dickie, J. Wardlaw, et al., White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks, NeuroImage: Clinical 17 (2018) 918–934.

[28] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[29] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, P.-A. Heng, H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes, IEEE transactions on medical imaging 37 (12) (2018) 2663–2674.

[30] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser, et al., The liver tumor segmentation benchmark (lits), arXiv preprint arXiv:1901.04056 (2019).

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[32] W. Chen, B. Liu, S. Peng, J. Sun, X. Qiao, S3d-unet: separable 3d u-net for brain tumor segmentation, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 358–368.

[33] H. Li, A. Li, M. Wang, A novel end-to-end brain tumor segmentation method using improved fully convolutional networks, Computers in biology and medicine 108 (2019) 150–160.

[34] R. M. Rad, P. Saeedi, J. Au, J. Havelock, Trophectoderm segmentation in human embryo images via inceptioned u-net, Medical Image Analysis 62 (2020) 101612.

[35] J. Dolz, I. Ben Ayed, C. Desrosiers, Dense multi-path u-net for ischemic stroke lesion segmentation in multiple image modalities, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 271–282.

[36] C. Meng, K. Sun, S. Guan, Q. Wang, R. Zong, L. Liu, Multiscale dense convolutional neural network for dsa cerebrovascular segmentation, Neurocomputing 373 (2020) 123–134.

[37] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, D. Rueckert, Drinet for medical image segmentation, IEEE transactions on medical imaging 37 (11) (2018) 2453–2462.

[38] Z. Zhang, C. Wu, S. Coleman, D. Kerr, Dense-inception u-net for medical image segmentation, Computer methods and programs in biomedicine 192 (2020) 105395.

[39] 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).

[40] A. Huang, L. Jiang, J. Zhang, Q. Wang, Attention-vgg16-unet: a novel deep learning approach for automatic segmentation of the median nerve in ultrasound images, Quantitative Imaging in Medicine and Surgery 12 (6) (2022) 3138.

[41] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[42] D. Maji, P. Sigedar, M. Singh, Attention res-unet with guided decoder for semantic segmentation of brain tumors, Biomedical Signal Processing and Control 71 (2022) 103077.

[43] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, in: Artificial intelligence and statistics, PMLR, 2015, pp. 562–570.

[44] J. Wu, S. Zhou, S. Zuo, Y. Chen, W. Sun, J. Luo, J. Duan, H. Wang, D. Wang, U-net combined with multi-scale attention mechanism for liver segmentation in ct images, BMC Medical Informatics and Decision Making 21 (1) (2021) 1–12.

[45] A. Amer, T. Lambrou, X. Ye, Mda-unet: A multi-scale dilated attention u-net for medical image segmentation, Applied Sciences 12 (7) (2022) 3676.

[46] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

[47] Z. Wang, Y. Zou, P. X. Liu, Hybrid dilation and attention residual u-net for medical image segmentation, Computers in Biology and Medicine 134 (2021) 104449.

[48] P. Wu, Z. Wang, B. Zheng, H. Li, F. E. Alsaadi, N. Zeng, Aggn: Attention-based glioma grading network with multi-scale feature extraction and multi-modal information fusion, Computers in Biology and Medicine 152 (2023) 106457.

[49] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[50] S. Hochreiter, The vanishing gradient problem during learning recurrent neural nets and problem solutions, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 6 (1998) 107–116. doi:10.1142/S0218488598000094.

[51] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, et al., A large annotated medical image dataset for the development and evaluation of segmentation algorithms, arXiv preprint arXiv:1902.09063 (2019).

[52] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken, et al., The medical segmentation decathlon, arXiv preprint arXiv:2106.05735 (2021).

[53] K. Payette, P. de Dumast, H. Kebiri, I. Ezhov, J. C. Paetzold, S. Shit, A. Iqbal, R. Khan, R. Kottke, P. Grehten, et al., An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset, Scientific Data 8 (1) (2021) 1–14.

[54] M. Nolden, S. Zelzer, A. Seitel, D. Wald, M. Müller, A. M. Franz, D. Maleike, M. Fangerau, M. Baumhauer, L. Maier-Hein, et al., The medical imaging interaction toolkit: challenges and advances, International journal of computer assisted radiology and surgery 8 (4) (2013) 607–620.

[55] A. A. Taha, A. Hanbury, Metrics for evaluating 3d medical image segmentation: Analysis, selection, and tool, BMC Medical Imaging 15 (08 2015). doi:10.1186/s12880-015-0068-x.

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