Universal Medical Image Segmentation using 3D Fabric Image Representation Encoding Networks

Siyu Liu¹, Wei Dai¹, Craig Engstrom¹, Jurgen Fripp², Stuart Crozier¹, Jason A. Dowling², Shekhar S. Chandra¹

¹School of Information Technology and Electrical Engineering, University of Queensland, Australia
²Australian e-Health Research Centre, CSIRO, Australia

Abstract—Data scarcity is a common issue for deep learning applied to medical image segmentation. One way to address this problem is to combine multiple datasets into a large training set and train a unified network that simultaneously learns from these datasets. This work proposes such a network, Fabric Image Representation Encoding Network (FIRENet), for simultaneous 3D multi-dataset segmentation. As medical image datasets can be extremely diverse in size and voxel spacing, FIRENet uses a 3D fabric latent module, which automatically encapsulates many multi-scale sub-architectures. An optimal combination of these sub-architectures is implicitly learnt to enhance the performance across many datasets. To further promote diverse-scale 3D feature extraction, a 3D extension of atrous spatial pyramid pooling is used within each fabric node to provide a finer coverage of rich-scale image features. In this study, FIRENet was first applied to 3D universal bone segmentation involving multiple musculoskeletal datasets of the human knee, shoulder and hip joints. FIRENet exhibited excellent universal bone segmentation performance across all the different joint datasets. When transfer learning is used, FIRENet exhibited both excellent single dataset performance during pre-training (on a prostate dataset), as well as significantly improved universal bone segmentation performance. In a following experiment which involves the simultaneous segmentation of the 10 Medical Segmentation Decathlon (MSD) challenge datasets, FIRENet produced good multi-dataset segmentation results and demonstrated excellent inter-dataset adaptability despite highly diverse image sizes and features. Across these experiments, FIRENet’s versatile design streamlined multi-dataset segmentation into one unified network. Whereas traditionally, similar tasks would often require multiple separately trained networks.

Deep learning methods for medical image segmentation is a rapidly evolving field with the potential to enhance disease diagnosis [36] and treatment planning [44]. The effectiveness of deep learning can be largely attributed to its data-driven nature. However, this reliance on data can also be a major limitation in medical image analysis as data scarcity remains a problem. Unlike the abundance of publicly available large-scale datasets [34], [18], [9] used in 2D computer-vision tasks, (expert) labelled medical image datasets are much smaller in quantity due to several factors:

- Data acquisition challenges: The acquisition of medical images such as magnetic resonance (MR) scans is highly specialised and resource-intensive. Voxel-wise manual annotation of 3D volumes (as required by training 3D image segmentation models) is expertise and time-intensive and subject to variable operator error. In addition, careful planning and expert contouring protocols are typically required to minimise intra- and inter-rater segmentation variability.
- Data fragmentation: Clinical studies involving medical imaging are typically highly focused, relatively small investigations due to high imaging costs. Datasets from different studies can exhibit considerable inter-dataset variations (for example, different imaging fields of view, and contrasts) associated with different acquisition sequences and protocols. Most current deep learning methods in medical image analysis lack versatility as they are typically trained on individual domain-specific datasets.
- Access to image datasets: Collecting large-scale medical imaging datasets with expert annotations is difficult as explicit consent, strict adherence to ethics and systematic coordination are required, as the publication of patient data, even when de-identified, is a highly sensitive matter.

Multi-dataset learning (including transfer learning [53], [7] and simultaneous multi-dataset learning [23]) has been shown to improve segmentation performance. However, most current deep learning methods used for medical image analysis are designed and optimised for single-dataset applications. Even highly adaptive architectures like nnUNet [25] are trained on a per dataset basis. Currently, using one model instance to segment data distributions (datasets) remains a challenge. Previous works in this field include 3D MDUNet and 3DU²Net, which both simultaneously applied a U-Net-like architecture to several MSD challenge datasets. However, they were only evaluated on a selection of the MSD datasets. To extend the coverage to drastically more datasets, a model with more focus on diverse-scale feature extraction and architecture versatility is desirable. Fabric-shaped architectures [58], [19], [43] are an idea candidate as they create a superposition of many multi-scale sub-architectures.

This work proposes a versatile end-to-end 3D fabric network, Fabric Image Representation Encoding Network (FIRENet), which self-adapts simultaneous to multi-dataset medical image datasets for medical image segmentation. The central features of FIRENet are summarised as follows:

1) Multi-dataset and multi-scale image feature representations are learnt and encoded in the dense residual fabric (DRF) (Figure 1) latent module. DRF’s diverse-scale fabric structure is designed to emulate a superposition...
of many multi-scale sub-architectures. The major advantage of DRF is architectural generalisability. This work demonstrates its application to multiple medical image datasets simultaneously without dataset-specific hyper-parameter tuning.

2) The nodes in DRF are connected via weighted residual summation (WRS) (weighted connections) for improved adaptability. These connections are trainable for automatic architecture-level adaptation to different medical image datasets. (Figure 1).

3) A 3D extension of atrous spatial pyramid pooling (ASPP) [5] is employed in each node of DRF to provide finer coverage of different-sized (image) features.

The experiment results show FIRENet’s excellent I) single dataset performance for semantic prostate segmentation II) simultaneous universal (support for multi-anatomy) bone segmentation with and without transfer learning performance, and III) simultaneous multi-object segmentation performance for all the 10 image datasets in the MSD challenge. Notably, FIRENet also exhibits multi-dataset adaptability without tailored training procedures to reach convergence.

I. RELATED WORK

A. Convolutional Neural Networks (CNNs) for image segmentation

The most common class of deep learning models for (medical) image segmentation are CNNs. Initially, CNNs were developed for image classification [29], [46] and were later found suitable for image segmentation tasks. Typically, CNNs use consecutive convolution (image filtering) and hierarchical down-sampling to extract features. However, since image segmentation outputs dense predictions (one for every pixel/voxel), image segmentation models require feature learning at both global (coarse) and local (fine-detailed) scales. Excessive down-sampling, as used in classic CNNs [46], [22] can degrade segmentation accuracy due to loss of image resolution. In response, later works, including UNet [42], incorporated an encoder-decoder architecture which first extracts large-context features using an encoder CNN, then re-constructs the full-resolution image using a decoder CNN (reversing the down-sampling performed by the encoder). Another notable contribution of UNet is the use of shortcuts. During the encoding phase, a shortcut temporarily stores high-resolution features. In the decoding phase, the stored high-resolution features are added back to fill in the lost fine-detailed features. There have been several recent UNet-based networks [59], [56], [39], [60] successfully applied to medical image segmentation tasks.

For 3D medical image analysis, most 2D segmentation methods can be readily extended to 3D [8], [37] to process volumetric medical image data. The primary issue in model building is the extremely large memory consumption of 3D convolution. Therefore, workarounds such as reducing model complexity are often required to train 3D CNNs. However, overly simplifying a CNN architecture can limit its learning capacity. In response, 2.5D CNNs [49], [47] and patch-
based methods [10], [12], [14], [13] have been proposed as alternative solutions to end-to-end 3D methods. However, both 2.5D and patch-based methods are unable to capture the full 3D context of an image. For patch-based methods, there are also additional hyper-parameters (such as fixed patch size) and pre or post-processing steps that can limit the model’s applicability to different datasets.

B. Multi-scale feature extraction

Multi-scale feature extraction [17], [38], [2], [57], [31] has been increasing in popularity for extract rich image feature representation. Typically, multi-scale feature extraction divides the input into several branches, each with a different receptive field. For example, the atrous spatial pyramid pooling (ASPP) in DeepLabV3 [5], [6] is a powerful multi-scale feature extractor using parallel dilated convolutions. These networks use interlaced multi-scale convolutional blocks, and all the blocks are trained end-to-end using gradient descent. Recently, AutoDeepLab [35] demonstrated explicit architecture. It employed weighted trainable connections between cells during training. After training, weak connections can be pruned to reveal a compact architecture for the specific training set. The resulting architecture was found to be as successful as many hand-crafted architectures.

1) Dense residual fabric latent module: Fabric structures present a general multi-scale architecture solution that inherently aligns with the nature of multi-dataset medical image segmentation. Hence, this work employs a 3D fabric latent representation module, dense residual fabric (DRF) (Figure 3a), in anticipation of diverse-sized medical images. The DRF consists of inter-weaved 3D feature extractors (denoted $\psi$). Each feature extractors has three major components: input size equaliser (ISE), weighted residual summation (WRS) and Atrous Spatial Pyramid Pooling 3D (ASPP3D).

ISE: operation for automatically resizing and voxel-aligning incoming 3D feature maps. ISE is required for feature map summation and concatenation.

WRS: operation for fusing the aligned feature maps from ISE. As Figure 3b shows, each input is multiplied by its associated sigmoid-gated weight before adding with others. The weights are trainable and are uniformly sampled from $\{-0.03, 0.03\}$. WRS gives DRF the flexibility to optimise connection strengths between inputs.

ASPP3D: Most multi-scale networks can only extract features at a limited number of scales (using different branches). ASPP3D gives DRF additional node-level multi-scale feature extraction capabilities. For instance, a DRF with three branches each using ASPP3D nodes with dilation rates 1, 2 and 4 would yield nine unique receptive field sizes (3, 5, 6, 7, 10, 12, 14, 20 and 28). This would otherwise require nine dedicated branches without ASPP3D.

1) Medical image segmentation involving multiple datasets: While deep learning models for 3D medical image segmentation have become more and more sophisticated [21], [32], most of them are still limited to one study (dataset) at a time. Incorporating multiple datasets can be beneficial for medical image datasets in tackling data scarcity. Works such as [41], [28], [50], [45], [7] have all demonstrated that pre-trained weights (transfer learning) can generally lead to improved accuracy and convergence. However, applying a deep learning model at a larger scale (and to multiple medical image datasets) is still an under-explored application with many practical benefits.

A deep learning model specifically developed for multi-dataset medical analysis requires excellent dataset generalisability. Several works have developed models that can self-adapt to different medical image datasets. For example, nnUNet [25] automatically configures the model’s hyper-parameters according to the geometry of the training dataset. Neural Architecture Search (NAS) [55], [35], [61] is another popular class of methods to create neural networks that best suit the data. However, methods like nnUNet and NAS are not suitable for simultaneous multi-dataset processing as the models are configured (or optimised) on a per dataset basis. The resulting model has limited generalisability to new datasets, especially in medical image analysis.

Apply one model for multi-dataset medical image segmentation is more challenging and there are fewer studies in the literature. Methods such as 3D U^2Net [23], [51], [24], [33] showed that it is highly desirable, and indeed possible, to segment multiple datasets (organs) using a unified methods. However, methods such as [51] rely on and complex components developed for a limited scope (CT lesion detection) making them unsuitable for more general multi-dataset medical image analysis. [23], [33] and [24] demonstrated more flexibility by incorporating domain adapters throughout the model. However, non of these methods have demonstrated simultaneous multi-dataset segmentation at a large-scale: most of them were only evaluated on a small subset of the MSD challenge datasets. Moreover, these works used U-Net-like backbones which lack diverse-scale feature coverage. Hence their applicability to diverse image sizes can be limited.
using strided-convolution. The resulting scale $s_j$ of each branch satisfies $s_j = \frac{s}{2^j}$, $j \in \{1, ..., w\}$. Each branch is then separately processed through $N$ feature extractors to progressively learn features at different scales. To enable intermediate sharing of rich-scale features, the output of an feature extractor $\psi_{i,j}$, where $i \in \{1, ..., n\}$ and $j \in \{1, ..., w\}$, is fed into subsequent feature extractors $\psi_{i+1,j}$, $\psi_{i+1,j+1}$, and $\psi_{i+1,j-1}$ using WRS. Strided-convolution and bi-linear up-sampling are used to resize feature maps to their target sizes as needed. We avoid transpose convolution as it has been shown to produce “checkerboard” artefacts [40]. The depths $\{c_{i,j}\}i \in \{1, ..., n\}$ and $j \in \{1, ..., w\}$ of the feature extractors are distributed following the geometry of a pyramid - increasing towards the mid-point of the lowest resolution branch $\psi_{i=n/2, j=w}$ of the fabric. Let $C$ be DRF’s input channels, then the number of channels of any feature extractor in the first half of the fabric (from $i = 0$ to $i = \frac{n}{2}$) can be defined as $c_{i,j} = \min(C \times 2^j - 1, C \times 2^{j-1})$. The number of channels of the feature extractors in the second half of the fabric then gradually shrinks along $N$, mirroring the first half. At the end of the fabric, the different-scaled parallel branches are merged using WRS to form an output of the original scale $s$.

3) Dense residual connections: [22] showed that network depth is positively correlated to training difficulty. We include supplementary residual shortcuts (Figure 3c) to densely connect the feature extractors in the fabric. That is, in addition to the different-scaled features from the immediate previous layers, each feature extractor $\psi_{i,j}$ receives shortcut signals from all other preceeding feature extractors with compatible channel sizes $\{\psi_{i,j}| i \in \{0, ..., i-2\}, c_{i,j} = c_{i,j}\}$ where $c$ stands for the number of channels) as illustrated in Figure 3c.

4) Encoder-decoder backbone: As maintaining high-resolution features through a fully 3D network is not feasible for current generation GPUs, the DRF is embedded in limited encoder-decoder base (Figure 1), with WRS acting as shortcuts to passing features from the encoder to the decoder. The encoder and the decoder have the same number of convolutional blocks. Each block is a residual unit [22] with two convolutional layers followed by max-pooling. In addition, a convolutional layer is added to each encoder-to-decoder shortcut to reduce semantic gaps [59].

5) Instantiation parameters: The encoder contains two convolutional blocks of 32 and 64 channels, respectively. Then, the encoded representation of the input is passed into a DRF instantiated with $W = 3$, $N = 4$ and $C = 64$. Each feature extractor has three parallel branches with dilation rates of 1, 2 and 4, respectively. Finally, the fabric output is passed through two decoder blocks with 64 and 32 channels, respectively, to arrive at the network’s output. The shortcut convolutional layers used for semantic gap reduction have the same depths as their corresponding encoder or decoder blocks.

B. Training

Various works have shown that deep supervision [30] substantially increases the performance of deep learning for image segmentation [26], [33]. For the training of FIRENet, a similar concept is used where each decoder block produces an auxiliary segmentation output through point-wise $(1 \times 1 \times 1)$ convolution. The loss for each output (auxiliary and main) is the sum of a categorical cross-entropy loss and a Dice similarity coefficient (DSC) loss, and they were minimised using the Adam optimiser [27]. The hardware used for training consisted of an NVIDIA Tesla V100 (32GB), and the training duration was capped to 1 day as the performance plateaus.

C. Experiment setups

1) Experiment I: Multi-dataset Transfer Learning: FIRENet was tested for transfer learning involving several 3D medical imaging datasets. For pre-training, FIRENet was first trained on a recently released 3D prostate magnetic resonance (MR) dataset [15] (the prostate dataset). Then, the trained FIRENet instance was transferred for simultaneous multi-dataset bone segmentation on a composite bone dataset (the multi-bone dataset). Elastic deformation was used for data augmentation [42]. In line with other methods applied to this dataset, the evaluation metrics used were DSC, Hausdorff distance (HD) and mean surface distance.

The prostate dataset contains 211 3D MR examinations of the pelvic region with manual segmentation labels for
(a) Structure of the Dense Residual Fabric module. Feature extractors are denoted $\psi_{i,j}$ where $i$ and $j$ enumerate the number of feature extractor along the network’s depth (N) and the number of branches (W) of the fabric, respectively. Feature extractors are coloured based on the number of channels.

(b) Detailed feature extractor configuration. The number of channels is governed by the location $(i, j)$ of the extractor in the fabric.

(c) Example dense shortcut arrangement in the Residual Fabric. A shortcut connection is only established if the layers involved have the same number of channels. This arrangement is applied to every branch of the fabric.

Fig. 3: DRF feature representation model. (a) Structure of DRF. (b) Detailed feature extractor architecture. (c) Example residual connections used in the fabric.
five foreground classes: body, bone (pelvic spine and girdle, proximal femur), urinary bladder, rectum and prostate.

The multi-bone dataset is composed of four smaller datasets: three 3T MR imaging musculoskeletal (MSK) datasets (knee [20], shoulder [52], hip [4]) and the OAI knee [1] dataset. The main difficulty of segmenting this experiment is the diverse image sizes and imbalanced numbers of training examples in each dataset (62, 25, 53 and 507 MR examinations, respectively).

2) Experiment II: Simultaneous multi-dataset segmentation on MSD: Medical Segmentation Decathlon (MSD) is a well-known 10-dataset segmentation challenge targeted at assessing the generalisability of machine learning models applied to medical image segmentation. Most previous works ([23] and [24]) on simultaneous multi-dataset segmentation were performed on a subset of the MSD datasets only, which excluded important and also challenging tasks such as HepaticVessel and Lung. To provide a complete performance assessment, the current work evaluated FIRENet on all 10 MSD constituent datasets. These multi datasets are highly diverse in image size and voxel spacing with the smallest dimension being 11 voxels and the largest dimension being 751 voxels. For preprocessing, each image was re-sampled to the same voxel spacing of [1, 1, 1] and patches were extracted from the re-sampled images. To limit memory usage, the size for patch extraction was set to \( \min(d, 160) \) for each dimension (where \( d \) is the number of voxels of that dimension). Finally, voxel intensity standardisation was applied before entering the network. The dataset was divided into training and validation sets according to an 80%-20% split ratio as per previous work [33]. Because the datasets contain different numbers of classes, and a universal categorical output for all the classes would be unrealistically memory intensive, each dataset was paired with a designated up-sampling decoder after DRF to produce segmentation with the desired number of classes. However, the DRF, which contains most of the learnt features, was shared across the 10 datasets. The segmentation for each dataset was predicted using the shared DRF and its designated decoder. The evaluation metric was the averaged per-class (excluding the background) DSC which alleviates bias towards majority classes.

Due to the lack of available comparison methods for this multi-bone segmentation task, nnUNet was chosen, and one instance was trained to establish a baseline. All the datasets were pooled into single dataset, during which the smaller datasets were duplicated to ensure data balance. All the preprocessing, training and evaluation steps were carefully followed as per nnUNet’s official instructions and performed using the official script.

III. RESULTS AND DISCUSSION

A. Experiment I: Multi-dataset Transfer Learning

1) FIRENet pre-training evaluation: Table I shows the subject-level, 3-fold validation segmentation results from the prostate MR pre-training experiment for the body, bone, urinary bladder, rectum and prostate classes. FIRENet was compared to four other contemporary 3D deep learning baseline methods (3D UNet [8], improved UNet [26], VNet [37] and CAN3D [11]). FIRENet produced better results than the baseline methods in DSC, HD and mean surface distance values across the different classes. In terms of the performance on outlier cases (min DSC), FIRENet was amongst the most resilient models producing fewer segmentation errors.

Table II shows the baseline results published by [16], [3] using traditional methods. These baseline results are comparable even to the more recently deep learning results (UNet and VNet) in Table I FIRENet and CAN3D’s median results significantly exceed these traditional methods, especially across the most challenging classes (prostate and rectum).

2) Transfer learning: For transfer learning, two FIRENet instances were applied to the multi-bone dataset: the FIRENet (denoted FIRENet-T) previously pre-trained on the prostate dataset and a randomly initialised FIRENet (denoted FIRENet-R). The bone segmentation DSC results of FIRENet-R and FIRENet-T are provided in Tables III and Fig C. Overall, FIRENet-R and T exhibited the ability to simultaneously segment diverse medical image data despite relying on only one shared set of weights: the results on the OAI dataset are comparable to other methods [11], [1] which were specifically designed for one dataset. Comparing the mean DSC values of FIRENet-R and FIRENet-T, it can be seen that FIRENet could leverage pre-training to improve the segmentation performance across all the constituent bone datasets. Moreover, as Fig C FIRENet-T’s lowest DSC results were also noticeably improved over FIRENet-R, indicating reduced critical segmentation errors. In terms of convergence, FIRENet-T was also more stable and faster (Fig. 5) supporting the benefits of transfer learning for simultaneous multi-dataset segmentation.

The results on this nnUNet’s bone segmentation task indicated it is unsuitable for simultaneous multi-dataset processing - the best DSC results were substantially lower at 0.612, 0.851, 0.633 and 0.986 for the MSK hip, knee, shoulder and OAI, respectively. As the visualisation in Fig 4, the results from nnUNet were inconsistent with several segmentation failures each one of the smaller datasets (MSK knee, shoulder and hip). It was also observed that the training process of nnUNet was highly unstable, and early stopping was required to obtain usable results. nnUNet’s inability to produce satisfactory results could be attributed to the data-dependent nature of its self-configuring procedure, as well as the lack of multi-scale feature exchange in the UNet architecture. In the case of FIRENet-R and FIRENet-T, aside from the deliberate lack of dataset-specific configurations, there is also a strong emphasis on multi-dataset feature learning. Figure 7 A, B and D show visualisations of how the multi-scale features are exchanged within the DRF of FIRENet. Four representative pairs of feature maps were captured before and after WRS. Noticeably, WRS is shown to merge features to create more prominent activation. It also seems to “clean up” the activation.
TABLE I: Mean and minimum prostate dataset segmentation results using FIRENet and the selected baseline deep learning methods: UNet, improved UNet, VNet and CAN3D.

| Class  | Metrics       | 3D UNet [42] | 3D improved UNet [26] | VNet [37] | CAN3D [11] | FIRENet |
|--------|---------------|--------------|-----------------------|-----------|------------|---------|
| Body   | DSC (mean ± sd) | 0.96 ± 0.08  | 0.98 ± 0.01           | 0.97 ± 0.01 | 0.98 ± 0.01 | 0.99 ± 0.01 |
|        | HD (mean ± sd)  | 42.73 ± 37.42 | 35.80 ± 12.02         | 29.00 ± 9.34 | 30.57 ± 11.36 | 13.08 ± 6.48 |
|        | DSC (min)      | 0.65         | 0.91                  | 0.92       | 0.92       | 0.93     |
| Bone   | DSC (mean ± sd) | 0.91 ± 0.02  | 0.91 ± 0.02           | 0.85 ± 0.04 | 0.91 ± 0.02 | 0.92 ± 0.02 |
|        | HD (mean ± sd)  | 36.35 ± 23.70 | 41.05 ± 28.78         | 51.33 ± 17.70 | 39.73 ± 24.48 | 13.19 ± 11.09 |
|        | MSD (mean ± sd) | 1.65 ± 0.42  | 1.52 ± 0.39           | 2.54 ± 1.15 | 1.50 ± 0.38 | 0.85 ± 0.22 |
|        | DSC (min)      | 0.80         | 0.85                  | 0.84       | 0.84       | 0.85     |
| Bladder| DSC (mean ± sd) | 0.87 ± 0.21  | 0.93 ± 0.10           | 0.80 ± 0.16 | 0.92 ± 0.13 | 0.93 ± 0.10 |
|        | HD (mean ± sd)  | 30.53 ± 40.06 | 27.94 ± 35.66         | 39.61 ± 36.36 | 33.89 ± 37.92 | 9.50 ± 16.49 |
|        | MSD (mean ± sd) | 4.53 ± 12.69 | 2.10 ± 3.99           | 3.91 ± 3.01 | 2.51 ± 5.44 | 1.35 ± 3.37 |
|        | DSC (min)      | 0.00         | 0.27                  | 0.06       | 0.06       | 0.23     |
| Rectum | DSC (mean ± sd) | 0.78 ± 0.11  | 0.87 ± 0.05           | 0.76 ± 0.09 | 0.85 ± 0.05 | 0.87 ± 0.05 |
|        | HD (mean ± sd)  | 27.74 ± 28.71 | 18.53 ± 25.34         | 38.28 ± 45.40 | 33.30 ± 36.65 | 14.14 ± 20.76 |
|        | MSD (mean ± sd) | 3.16 ± 2.93  | 1.71 ± 1.10           | 3.62 ± 5.95 | 2.08 ± 1.45 | 1.24 ± 1.29 |
|        | DSC (min)      | 0.18         | 0.67                  | 0.29       | 0.66       | 0.64     |
| Prostate| DSC (mean ± sd) | 0.75 ± 0.18  | 0.84 ± 0.08           | 0.74 ± 0.13 | 0.81 ± 0.10 | 0.86 ± 0.06 |
|        | HD (mean ± sd)  | 28.23 ± 69.67 | 8.91 ± 9.17           | 25.38 ± 29.83 | 14.28 ± 16.33 | 7.36 ± 10.10 |
|        | MSD (mean ± sd) | 15.85 ± 65.41 | 1.92 ± 0.87           | 3.54 ± 2.89 | 2.17 ± 0.99 | 1.13 ± 0.48 |
|        | DSC (min)      | 0.00         | 0.45                  | 0.08       | 0.39       | 0.51     |

TABLE II: Median DSC, mean surface distance and HD values for the prostate dataset segmentation results using FIRENet and the published baseline traditional methods.

| Method       | Median Body DSC | Median Bone DSC | Median Bladder DSC | Median Rectum DSC | Median Prostate DSC | Median Prostate MSD (mm) | Median Prostate HD (mm) | Mean Prostate DSC |
|--------------|-----------------|-----------------|-------------------|------------------|--------------------|--------------------------|----------------------|------------------|
| Dowling et al.[16] | 0.94            | 0.81            | 0.87              | 0.79             | 0.81               | 2.08                     | 9.60                 | 0.79             |
| Chandra et al.[3] | 0.99            | 0.92            | 0.96              | 0.89             | 0.87               | 1.00                     | 4.24                 | 0.86             |
| FIRENet      | 0.99            | 0.92            | 0.96              | 0.89             | 0.87               | 1.00                     | 4.24                 | 0.86             |

Fig. 4: Examples bone segmentation results of FIRENet-R and nnUNet on MSK hip, knee and shoulder datasets. Both methods used identical data splits, and nnUNet was trained following and official documentation using the official scripts. nnUNet was unable to consistently produce satisfactory results when simultaneously applied to multiple datasets.

Fig. 5: Early convergence (50 epochs) plots of FIRENet-R and FIRENet-T. The DSCs for the bones were averaged across the four MSK bone datasets. FIRENet-T consistently showed accelerated convergence, especially in the first 30 epochs.
**TABLE III:** Mean validation bone DSC results of 3D MSK hip, knee, shoulder datasets and the OAI dataset. Some traditional methods reported per class results.

| Method   | Hip   | Knee  | Shoulder | OAI    | Overall |
|----------|-------|-------|----------|--------|---------|
| FIRENet-R | 0.954 | 0.961 | 0.940    | 0.985  | 0.952   |
| FIRENet-T | **0.962** | **0.965** | **0.947** | **0.966** | **0.965** |
| nnUNet    | 0.612 | 0.851 | 0.633    | 0.96   | -       |
| CAN3D     | -     | -     | -        | 0.986  | -       |
| SSM+CNN   | -     | -     | -        | 0.985  | -       |
| SSM       | 0.950 | -     | -        | -      | -       |
| SSM       | -     | 0.922 | -        | -      | -       |
| SSM       | -     | -     | 0.8815   | -      | -       |

**C. Limitations**

As FIRENet universally applies to different imaging datasets, its design and training methodology purposefully lack domain-specific optimisations. Compared to public challenge results produced by highly specialised and single-dataset methods, a general architecture does not often yield optimal numerical results across all the different datasets. Although FIRENet’s architecture is free of overhead in handling multiple datasets, there is an increase in total model size with each additional decoder (required for different output formats). Finally, as a CNN, FIRENet faces the current limitations of deep learning. For example, the lack of clinical explain-ability and difficulty extrapolating to unseen data distributions.

**D. Future work**

As new labelled medical imaging datasets and deep learning accelerators become available, FIRENet’s size and training set composition can continue to expand. It also would be beneficial to train multiple instances of FIRENet to specialise in different imaging modalities. In terms of architectural development, FIRENet could be effectively re-purposed for classification, regression and multi-task learning by adding downstream prediction heads. For example, a classification head could be added for 3D medical image classification based on the features extracted from FIRENet.

**IV. Conclusion**

In response to the issue of data scarcity in deep learning for medical image analysis, his work proposes FIRENet, a versatile 3D neural network architecture geared towards simultaneous multi-dataset segmentation. To ensure maximum flexibility when learning features from multiple datasets, FIRENet uses a generally inclusive fabric structure to encapsulate a superposition of many sub-networks, thus alleviating the need for dataset-specific architecture designs. In addition, each fabric node employs ASPP 3D for rich-scale feature extraction to ensure maximum coverage of different scaled features. The prostate, bone and MSD segmentation tasks showed that FIRENet is well-suited for multi-modal, multi-size and multi-target segmentation simultaneously.
TABLE IV: FIRENet’s validation DSC results on the full MSD dataset compared against 3D MDUNet [33]. These DSC results were computed without the background class.

| Task Code      | FIRENet | 3D MDUNet | 3D MDUNet(3D DUnet baseline) | 3D DUnet(universal) |
|---------------|---------|-----------|-------------------------------|---------------------|
| Task01_BrainTumour | 0.571   | 0.921     | -                             | 0.919               |
| Task02_Heart   | 0.913   | -         | -                             | 0.935               |
| Task03_Liver   | 0.917   | -         | -                             | 0.882               |
| Task04_Hippocampus | 0.826   | 0.650     | 0.624                         | 0.789               |
| Task05_Prostate| 0.807   | -         | -                             | -                   |
| Task06_Lung    | 0.778   | -         | -                             | -                   |
| Task07_Pancreas| 0.742   | 0.622     | 0.534                         | 0.621               |
| Task08_HepaticVessel | 0.787   | -         | -                             | -                   |
| Task09_Spleen  | 0.895   | 0.833     | 0.815                         | -                   |
| Task10_Colon   | 0.437   | -         | -                             | -                   |

Fig. 8: Simultaneous 3D segmentation of the MSD challenge datasets using FIRENet.

REFERENCES

[1] Felix Ambellan, Alexander Tack, Moritz Ehlke, and Stefan Zachow. Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the osteoarthritis initiative. *Medical Image Analysis*, 52:109 – 118, 2019.

[2] Pierre Buyssens, Abderrahim Elmoataz, and Olivier Lezoray. Multiscale convolutional neural networks for vision–based classification of cells. volume 7725, 41 2012.

[3] Shekhar S Chandra, Jason A Dowling, Peter B Greer, Jarad Martin, Chris Wratten, Peter Pichler, Jurgen Fritter, and Stuart Crozier. Fast automated segmentation of multiple objects via spatially weighted shape learning. *Physics in Medicine and Biology*, 61(22):8070–8084, oct 2016.

[4] Shekhar S. Chandra, Ying Xia, Craig Enngstrom, Stuart Crozier, Raphael Schwarz, and Jurgen Fritter. Focused shape models for hip joint segmentation in 3D magnetic resonance images. *Medical Image Analysis*, 18(3):567 – 578, 2014.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv e-prints*, page arXiv:1706.05587, June 2017.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 833–851, Cham, 2018. Springer International Publishing.

[7] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer Learning for 3D Medical Image Analysis. *arXiv e-prints*, page arXiv:1904.00625, April 2019.

[8] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 432–432, Cham, 2016. Springer International Publishing.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[10] Z. Cui, J. Yang, and Y. Qiao. Brain MRI segmentation with patch-based CNN approach. In *2016 35th Chinese Control Conference (CCC)*, pages 7026–7051, 2016.

[11] Wei Dai, Boyeong Woo, Siyu Liu, Matthew Marques, Fangfang Tang, Stuart Crozier, Craig Eggington, and Shekhar Chandra. Cca3d: Fast 3D knee MRI segmentation via Compact Context Aggregation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1505–1508, 2021.

[12] Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage*, 170:456 – 470, 2018.

[13] Suyu Dong, Gongning Luo, Clara Tam, Wei Wang, Kuanquang Wang, Shaodong Cao, Bo Chen, Henggui Zhang, and Shao Li. Deep atlas network for efficient 3D left ventricle segmentation on echocardiography. *Medical Image Analysis*, 61:101638, 2020.

[14] Qi Dou, Lequan Yu, Hao Chen, Yueming Jin, Xin Yang, Jing Qin, and Pheng-Ann Heng. 3D deeply supervised network for automated segmentation of volumetric medical images. *Medical Image Analysis*, 41:40 – 54, 2017. Special Issue on the 2016 Conference on Medical Image Computing and Computer Assisted Intervention (Analog to MICCAI 2015).

[15] Jason Dowling and Peter Greer. Labelled weekly MR images of the male pelvis, 2021.

[16] Jason A. Dowling, Jidi Sun, Peter Pichler, David Rivest-Hénault, Soumya Ghose, Haylea Richardson, Chris Wratten, Jarad Martin, Jameen Arm, Leah Best, Shekhar S. Chandra, Jurgen Fritter, Frederick W. Menk, and Peter B. Greer. Automatic substitute computed tomography external beam radiation therapy from standard MRI sequences. *International Journal of Radiation Oncology*Biology*Physics*, 93(5):1144 – 1153, 2015.

[17] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. pages 2650–2658, 2015.

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[19] Damien Fourure, Rémi Emonet, Elisa Fromont, Damien Muselet, Alain Tremieu, and Christian Wolf. Residual Conv-Deconv Grid Network
for Semantic Segmentation. arXiv e-prints, page arXiv:1707.07958, July 2017.

[20] Jurgen Frits, Stuart Crozier, Simon K Warfield, and Sébastien Ourselin. Automatic segmentation of the bone and extraction of the bone–cartilage interface from magnetic resonance images of the knee. Physics in Medicine and Biology, 52(6):1617–1631, feb 2007.

[21] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. Unet: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 574–584, January 2022.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. pages 770–778, 06 2016.

[23] G. Sharp, K. D. Fritscher, V. Pekar, M. Peroni, N. Shusharina, and Klaus H. Maier-Hein. nnU-net: A 3d universal u-net for multi-domain medical image segmentation. In Lecture Notes in Computer Science, pages 291–299. Springer International Publishing, 2019.

[24] Rui Huang, Yuanjie Zheng, Zhiqiang Hu, Haotian Han, Qingtong Yao, Shankuan Zhu, and S. Kevin. 3d u-net: A 3d universal u-net for multi-modal medical image segmentation. In International Conference on Learning Representations, 12. 2014.

[25] F. Knoll, K. Hammernik, E. Kobler, T. Pock, M. P. Recht, and D. K. Sodickson. Assessment of the generalization of learned image reconstruction and the potential for transfer learning. Magn Reson Med, 81(1):116–128, 2019.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 25. Curran Associates, Inc., 2012.

[27] Chen-Xi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Liu, Alireza Golkhou, and Li Fei-Fei. Hierarchical neural architecture search for semantic image segmentation. pages 82–92, 06 2019.

[28] I. Y. Maalood, Y. E. A. Al-Salhi, and S. Lu. Thresholding for Medical Image Segmentation for Cancer using Fuzzy Entropy with Level Set Algorithm. Open Med (Wars), 13:374–383, 2018.

[29] Fausun Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Parametric Medical Image Segmentation. In 2016 Fourth International Conference on 3D Vision (3DV), pages 565–571, 2016.

[30] Hafiz Tayyab Mustafa, Jie Yang, and Masoumeh Zareapoor. Multi-scale convolutional neural network for multi-focus image fusion. Image and Vision Computing, 85:26 – 35, 2019.

[31] Zhen-Liang Ni, Gui-Bin Bian, Xiao-Hu Zhou, Zeng-Guang Hou, Xiao-Liang Xie, Chen Wang, Yan-Jie Zhou, Rui-Qi Li, and Zhen Li. RAUNet: Residual Attention U-Net for Semantic Segmentation of Cataract Surgical Instruments. arXiv e-prints, page arXiv:1909.10360, September 2019.

[32] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. Distill, 2016.

[33] Hariharan Ravishankar, Prasad Sudhakar, Rahul Venkataramani, Shekhadi Thiruvengadam, Pavan Anangari, Narayanan Babu, and Vivek Vaidya. Understanding the mechanisms of deep transfer learning for medical images. In Putro Carneiro, Diana Malteus, Loïc Peter, Andrew Bradley, João Manuel R. S. Tavares, Vasileios Belagiannis, João Paulo Papa, Jacinto C. Nascimento, Marco Loog, Zhi Lu, Jaime S. Cardoso, and Julien Cornebise, editors, Deep Learning and Data Labeling for Medical Applications, pages 188–196, Cham, 2016. Springer International Publishing.

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pages 234–241, Cham, 2015. Springer International Publishing.

[35] Shreyas Saxena and Jakob Verbeek. Convolutional neural fabrics. In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS’16, page 4060–4068, Red Hook, NY, USA, 2016. Curran Associates Inc.

[36] G. Sharp, K. D. Fritscher, V. Pekar, M. Peroni, N. Shusharina, H. Veeraraghavan, and J. Yang. Vision 20/20: perspectives on automated image segmentation for radiotherapy. Med Phys, 41(5):050902, May 2014.

[37] H. C. Shin, H. R. Roth, M. Gao, L. L. Z. Xu, I. Noguez, J. Yao, D. Mollura, and R. M. Summers. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Trans Med Imaging, 35(5):1285–1298, 05 2016.

[38] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv e-prints, page arXiv:1409.1556, September 2014.

[39] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation based on cascaded convolutional networks with uncertainty estimation. Frontiers in Computational Neuroscience, 13:56, 2019.

[40] Fengdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–20, 2020.

[41] Yinzhe Xue, Fadi G. Farhat, Olga Boukrina, A.M. Barnett, Jeffrey R. Binder, Usman W. Roshan, and William W. Graves. A multi-path 2.5 dimensional convolutional neural network system for segmenting stroke lesions in brain mri images. NeuroImage: Clinical, 25:102118, 2020.

[42] Samir S. Yadav and Shivairao M. JadHAV. Deep convolutional neural network based medical image classification for disease diagnosis. Journal of Big Data, 6(1):13, Dec 2019.

[43] Ke Yuan, Zhentao Chen, Adam P. Harrison, Dakai Jin, Yunzhe Xue, Fadi G. Farhat, Olga Boukrina, A.M. Barrett, Jeffrey R. Binder, Usman W. Roshan, and William W. Graves. A multi-path 2.5 dimensional convolutional neural network system for segmenting stroke lesions in brain mri images. NeuroImage: Clinical, 25:102118, 2020.
architecture search for 3D medical image segmentation. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, June 2020.

[56] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, May 2018.

[57] Hengshuang Zhao, Jinming Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6230–6239, 2017.

[58] Yisu Zhou, Xiaolin Hu, and Bo Zhang. Interlinked convolutional neural networks for face parsing. *volume 9377, pages 222–231, 10 2015.*

[59] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jinming Liang. U-net++: A nested u-net architecture for medical image segmentation. In Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R.S. Tavares, Andrew Bradley, João Paulo Papa, Vasilieos Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayitt Greenspan, and Anant Madabhushi, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, Cham, 2018. Springer International Publishing.

[60] Hancan Zhu, Feng Shi, Li Wang, Sheng-Che Hung, Meng-Hsiang Chen, Shuai Wang, Weili Lin, and Dinggang Shen. Dilated dense u-net for infant hippocampus subfield segmentation. *Frontiers in Neuroinformatics*, 13:30, 2019.

[61] Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. 2017.