HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation

Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed

Abstract—Recently, dense connections have attracted substantial attention in computer vision because they facilitate gradient flow and implicit deep supervision during training. Particularly, DenseNet, which connects each layer to every other layer in a feed-forward fashion, has shown impressive performances in natural image classification tasks. We propose *HyperDenseNet*, a 3D fully convolutional neural network that extends the definition of dense connectivity to multi-modal segmentation problems. Each imaging modality has a path, and dense connections occur not only between the pairs of layers within the same path, but also between those across different paths. This contrasts with the existing multi-modal CNN approaches, in which modeling several modalities relies entirely on a single joint layer (or level of abstraction) for fusion, typically either at the input or at the output of the network. Therefore, the proposed network has total freedom to learn more complex combinations between the modalities, *within and in-between all the levels of abstraction*, which increases significantly the learning representation. We report extensive evaluations over two different and highly competitive multi-modal brain tissue segmentation challenges, iSEG 2017 and MRBrainS 2013, with the former focusing on 6-month infant data and the latter on adult images. *HyperDenseNet* yielded significant improvements over many state-of-the-art segmentation networks, ranking at the top on both benchmarks. We further provide a comprehensive experimental analysis of features re-use, which confirms the importance of hyper-dense connections in multi-modal representation learning. Our code is publicly available.

Index Terms—Deep learning, brain MRI, segmentation, 3D CNN, multi-modal imaging

1 Introduction

Multi-modal imaging enables the combination of different anatomical and functional information. Therefore, the joint analysis of multi-modal data has emerged as a natural approach to medical image analysis. Its use is becoming increasingly common for the study of a breadth of diseases [1], being of primary importance in developing comprehensive models of pathologies, as well as in increasing the statistical power of current imaging biomarkers. In multi-modal studies, images of the same target structures are acquired with different techniques. This combination of complementary information enables accurate visualization and delineation of the structures of interest.

Integrating several modalities provides solutions that overcome the limitations of independent imaging techniques. For instance, magnetic resonance imaging (MRI) depicts different tissue contrasts by varying the pulse sequences. On the one hand, MR-T1 (T1) yields a good image contrast between the gray matter (GM) and white matter (WM) tissues. On the other hand, MR-T2 weighted ($T2_w$) and proton density (PD) pulses are powerful in visualizing tissue abnormalities, such as lesions. A special case of $T2_w$ is the fluid attenuated inversion recovery (FLAIR) pulse sequence. This modality enhances the image contrast of white matter lesions (WMLs), such as multiple sclerosis [2]. Considering multiple MRI modalities can recover low tissue contrast, for example, between brain tissues (Fig. 1). This span of imaging possibilities has, therefore, led to an outstanding progress in exploring and understanding brain anatomy and brain-related disorders.

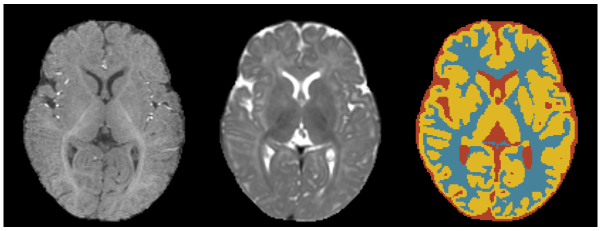


Fig. 1. Example of data from a training subject. Neonatal isointense brain images from a mid-axial T1 slice (*left*), the corresponding T2 slice (*middle*), and manual segmentation (*right*).

In general, the advances in multi-modal imaging have increased the quality of diagnosis, treatment and follow-up of various diseases. This comes, however, at the price of an inherently large amount of data produced by multi-modal studies, imposing a burden on disease assessments. Visual inspections of such an enormous amount of medical images are prohibitively time-consuming, prone to errors and unsuitable for large-scale studies. Therefore, automatic and reliable multi-modal segmentation algorithms are of high interest to the clinical community.

1.1 Prior work

The rapidly increasing use of multi-modal data prompted substantial research efforts in image segmentation algorithms that account for multi-modal scenarios, in various diseases or clinical studies. For example, PET-CT imaging has been widely used for joint tumor segmentation [3]–[12]. Deformable, level-set models, which combine individual

- *J. Dolz, K. Gopinath, H. Lombaert, C. Desrosiers and I. Ben Ayed are with the École de technologie Superieure, Montreal, Canada. email:jose.dolz@livia.etsmtl.ca*
- *J. Yuan is with the Xidian University, School of Mathematics and Statistics, Xi'an, China.*

Manuscript received XXX; revised XXX.

segmentations from PET and CT images, were used in some of these approaches [3]. In addition, textural features from CT images were extracted to distinguish cancerous tissue types, incorporating PET to this information [4], [7]. These methods have several limitations. In fact, independent segmentations do not fully account for all the available multimodal information. The underlying assumptions oversimplify the complex relationships that may exist between different imaging modalities. For instance, finding the tumor structure independently in PET and CT implicitly assumes that tumor volume is identical in both modalities, which is not the case. In fact, assuming simple correspondences between anatomical and functional images may not be realistic. Furthermore, sub-optimal solutions to the individual problems, due to optimization issues, along with heavy computational loads, impeded significantly the use of these approaches in practice.

The studies in [5], [8] stated the problem as a Markov Random field (MRF) on a graph, which encodes the information from both modalities. The ensuing PET-CT co-segmentation algorithms seek the tumor structures concurrently in both modalities. For these models, discrete graph-cut optimization can achieve globally optimal solutions in low-order polynomial time.

Multi-modal image segmentation in brain-related applications has also received a substantial research attention, for instance, brain tumors [13]–[16], brain tissues of both infant [17]–[27] and adult [28], [29], subcortical structures [30], among other problems [31]–[33]. Atlas-propagation approaches are commonly used in multi-modal scenarios [34], [35]. These methods rely on registering one or multiple atlases to the target image, followed by a propagation of manuals labels. When several atlases are considered, labels from individual atlases can be combined into a final segmentation via a label fusion strategy [18], [20], [23]. When relying solely on atlas fusion, the performance of such techniques might be limited and prone to registration errors. Parametric or deformable models [21] can be used to refine prior estimates of tissue probability [24]. For example, the study in [24] investigated a patch-driven method for neonatal brain tissue segmentation, integrating the probability maps of a subject-specific atlas into a level-set framework.

More recently, our community has witnessed a wide adoption of deep learning techniques, particularly, convolutional neural networks (CNNs), as an effective alternative to traditional segmentation approaches. CNN architectures are supervised models, trained end-to-end, to learn a hierarchy of image features representing different levels of abstraction. In contrast to conventional classifiers based on hand-crafted features, CNNs can learn both the features and classifier simultaneously, in a data-driven manner. They achieved state-of-the-art performances in a broad range of medical image segmentation problems [36], [37], including multi-modal tasks [14]–[16], [25]–[27], [29], [32], [33], [38], [39].

1.1.1 Fusion of multi-modal CNN feature representations

Most of the existing multi-modal CNN segmentation techniques followed an *early-fusion* strategy, which integrates the multi-modality information from the original space of low-level features [15], [25], [29], [33], [38], [39]. For instance, in [25], MR-T1, T2 and fractional anisotropy (FA) images are simply merged at the input of the network. However, as argued in [40] in the context of multi-modal learning, it is difficult to discover highly non-linear relationships between the low-level features of different modalities, more so when such modalities have significantly different statistical properties. In fact, early-fusion methods implicitly assumes that the relationship between different modalities are simple (e.g., linear). For instance, the early fusion in [25] learns complementary information from T1, T2 and FA images. However, the relationship between the original T1, T2 and FA image data may be much more complex than complementarity, due to significantly different image acquisition processes [26]. The work in [26] advocated *late fusion* of high-level features as a way that accounts better for the complex relationships between different modalities. They used an independent convolutional network for each modality, and fused the outputs of the different networks in higher-level layers, showing better performance than early fusion in the context infant brain segmentation. These results are in line with a recent study in the machine learning community [40], which investigated multimodal learning with deep Boltzmann machines in the context of fusing data from color images and text.

1.1.2 Dense connections in deep networks

Since the recent introduction of residual learning in [42], shortcut connections from early to late layers have become very popular in a breadth of computer vision problems [43], [44]. Unlike traditional networks, shortcut connections back-propagate gradients directly, thereby mitigating the gradient-vanishing problem and allowing deeper networks. Furthermore, they transform a whole network into a large ensemble of shallower networks, yielding competitive performances in various applications [29], [45]–[47]. DenseNet [48] extended the concept of shortcut connections, with the input of each layer corresponding to the outputs from all previous layers. Such a dense network facilitates the gradient flow and the learning of more complex patterns. DenseNet yielded significant improvements in accuracy and efficiency for natural image classification tasks [48]. However, the impact of dense connectivity in medical image segmentation, particularly in multi-modal problems, remains unexplored.

1.2 Contributions

We propose *HyperDenseNet*, a 3D fully convolutional neural network that extends the definition of dense connectivity to multi-modal segmentation problems. Each imaging modality has a path, and dense connections occur not only between the pairs of layers within the same path, but also between those across different paths; see the illustration in Fig. 2. This contrasts with the existing multi-modal CNN approaches, in which modeling several modalities relies entirely on a single joint layer (or level of abstraction) for fusion, typically either at the input (early fusion) or at the output (late fusion) of the network. Therefore, the proposed network has total freedom to learn more complex combinations between the modalities, *within and in-between all the levels of abstractions*, which increases significantly the learning representation in comparison to early/late fusion.

TABLE 1
A brief summary of existing methods for multi-modal medical image segmentation for some applications.

Work	Modality	Target	Method
El Naqa et al., 2007 [3]	PET-CT	Lung and cervix cancer	Multi-Level sets 2D/3D
Yu et al.,2009 [4]	PET-CT	Head and neck cancer	Multi-Level sets 2D/3D
Han et al.,2011 [5]	PET-CT	Head and neck cancer	Markov Random Field (MRF)
Bagci et al.,2012 [6]	PET-CT	Lung disease abnormalities	Random Walker
Bagci et al.,2013 [9]	PET-CT, PET-MRI, PET-CT-MRI	Several lesions	Random Walker
Markel et al.,2013 [7]	PET-CT	Lung carcinoma	Decision tree with KNN
Song et al.,2013 [8]	PET-CT	Lung tumor	Markov Random Field (MRF)
Ju et al.,2015 [11]	PET-CT	Lung tumor	Random walker + Graph Cuts
Cui et al.,2016 [12]	PET-CT	Lung tumor	Random walker
Prastawa et al., 2005 [17]	T1,T2	Infant brain tissue	Multi-atlas
Weisenfeld et al., 2006 [18]	T1,T2	Infant brain tissue	Multi-atlas
Deoni et al., 2007 [30]	T1,T2	Thalamic nuclei	K-means clustering
Anbeek et al., 2008 [19]	T2,IR	Infant brain tissue	KNN
Weisenfeld and Warfield, 2009 [20]	T1,T2	Infant brain tissue	Multi-atlas
Wang et al., 2011 [21]	T1,T2,FA	Infant brain tissue	Multi-atlas + Level sets
Srhoj et al., 2012 [22]	T1,T2	Infant brain tissue	Multi-atlas + KNN
Wang et al., 2012 [23]	T1,T2	Infant brain tissue	Multi-atlas
Wang et al., 2014 [41]	T1,T2,FA	Infant brain tissue	Multi-atlas + Level sets
Kamnitsas et al., 2015 [38]	Flair, DWI, T1, T2	Brain lesion	3D FCNN + CRF
Zhang et al., 2015 [25]	T1,T2,FA	Infant brain tissue	2D CNN
Havaei et al., 2016 [14]	T1,T1c,T2,FLAIR	Multiple Sclerosis/Brain tumor	2D CNN
Nie et al., 2016 [26]	T1,T2,FA	Infant brain tissue	2D FCNN
Chen et al., 2017 [29]	T1,T1-IR,FLAIR	Brain tissue	3D FCNN
Dolz et al., 2017 [27]	T1,T2	Infant brain tissue	3D FCNN
Fidon et al., 2017 [16]	T1,T1c,T2,FLAIR	Brain tumor	CNN
Kamnitsas et al., 2017 [15]	T1,T1c,T2,FLAIR MPRAGE,FLAIR,T2,PD	Brain tumour/lesions	3D FCNN + CRF
Kamnitsas et al., 2017 [32]	MPRAGE,FLAIR,T2,PD	Traumatic brain injuries	3D FCNN(Adversarial Training)
Valverde et al., 2017 [33]	T1, T2,FLAIR	Multiple-sclerosis	3D FCNN

Furthermore, hyper-dense connections facilitate the learning as they improve gradient flow and impose implicit deep supervision. We report extensive evaluations over two different[1] and highly competitive multi-modal brain tissue segmentation challenges, iSEG 2017 and MRBrainS 2013. *HyperDenseNet* yielded significant improvements over many state-of-the-art segmentation networks, ranking at the top on both benchmarks. We further provide a comprehensive experimental analysis of features re-use, which confirms the importance of hyper-dense connections in multi-modal representation learning. Our code is publicly available[2].

A preliminary conference version of this work appeared at ISBI 2018 [49]. This journal version is a substantial extension, including (1) a much broader, more informative/rigorous treatment of the subject in the general context of multi-modal segmentation; and (2) comprehensive experiments with additional baselines and publicly available benchmarks, as well as a thorough investigation of the practical usefulness and impact of hyper-dense connections.

2 Methods and Materials

Convolutional neural networks (CNNs) are deep models that can learn feature representations automatically from the training data. They consist of multiple layers, each processing the imaging data at a different level of abstraction, enabling segmentation algorithms to learn from large datasets and discover complex patterns that can be further employed for predicting unseen samples. The first attempts to use CNNs in segmentation problems followed a sliding-window strategy, where the regions defined by the window are processed independently, which impedes segmentation accuracy and computational efficiency. To overcome these limitations, the network can be viewed as a single non-linear convolution, which is trained end-to-end, a process known as fully CNN (FCNN) [50]. The latter brings several advantages over standard CNNs. It can handle images of arbitrary sizes and avoid redundant convolution and pooling operations, enabling computationally efficient learning.

2.1 The proposed Hyper-Dense network

The concept of "*the deeper the better*" is considered as a key principle in deep learning [42]. Nevertheless, one obstacle when dealing with deep architectures is the problem of vanishing/exploding gradients, which hampers convergence during training. To address these limitations in very deep architectures, the study in [48] investigated densely connected

[1] iSEG 2017 focuses on 6-month infant data, whereas MRBrainS 2013 uses adult data. Therefore, there are significant differences between the two benchmarks in term of image data characteristics, e.g, the voxel spacing and number of available modalities.

[2] https://www.github.com/josedolz/HyperDenseNet

networks. DenseNets are built on the idea that adding direct connections from any layer to all the subsequent layers in a feed-forward manner makes training easier and more accurate. This is motivated by three observations. First, there is an implicit deep supervision thanks to the short paths to all feature maps in the architecture. Second, direct connections between all layers help improving the flow of information and gradients throughout the entire network. Third, dense connections have a regularizing effect, which reduces the risk of over-fitting on tasks with smaller training sets.

Inspired by the recent success of densely-connected networks in natural image classification tasks, as well as in a few recent medical image segmentation works [51]–[53], we propose *HyperDenseNet*, a hyper-dense architecture for multi-modal image segmentation. We extend the concept of dense connectivity to the multi-modal setting: Each imaging modality has a path, and dense connections occur not only between layers within the same path, but also between layers across different paths; see Fig. 2 for an illustration.

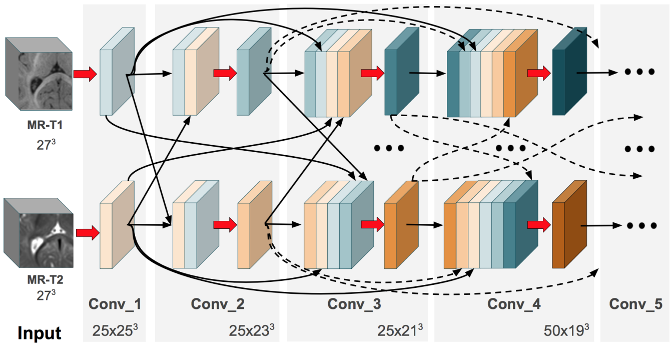


Fig. 2. A section of the proposed *HyperDenseNet* in the case of two image modalities. Each gray region represents a convolutional block. Red arrows correspond to convolutions and black arrows indicate dense connections between feature maps.

Let $\mathbf{x}_l$ be the output of the l^{th} layer. In CNNs, this vector is typically obtained from the output of the previous layer $\mathbf{x}_{l-1}$ by a mapping H_l composed of a convolution followed by a non-linear activation function:

$$\mathbf{x}_l \;=\; H_l(\mathbf{x}_{l-1}). \tag{1}$$

A densely-connected network concatenates all feature outputs in a feed-forward manner:

$$\mathbf{x}_l \;=\; H_l([\mathbf{x}_{l-1}, \mathbf{x}_{l-2}, \ldots, \mathbf{x}_0]), \tag{2}$$

where [...] denotes a concatenation operation.

Pushing this idea further, HyperDenseNet introduces a more general connectivity definition, in which we link the outputs from layers in different streams, each associated with a different image modality. In the multi-modal setting, our hyper-dense connectivity yields a much more powerful feature representation than early/late fusion as the network learns the complex relationships between the modalities within and in-between all the levels of abstractions. For simplicity, let us consider the scenario of two image modalities, although extension to N modalities is straightforward. Let $\mathbf{x}_l^1$ and $\mathbf{x}_l^2$ denotes the outputs of the l^{th} layer in streams 1 and 2, respectively. The output of the l^{th} layer in a stream s can then be defined as follows:

$$\mathbf{x}_l^s \;=\; H_l([\mathbf{x}_{l-1}^1, \mathbf{x}_{l-1}^2, \mathbf{x}_{l-2}^1, \mathbf{x}_{l-2}^2, \ldots, \mathbf{x}_0^1, \mathbf{x}_0^2]). \tag{3}$$

Figure 2 depicts a section of the proposed architecture, where each gray region represents a convolutional block. For simplicity, we assume that the red arrows indicate convolution operations only, whereas the black arrows represent the direct connections between feature maps from different layers, within and in-between the different streams. Thus, the input of each convolutional block (maps before the red arrow) is the concatenation of the outputs (maps after the red arrow) of all the preceding layers from both paths.

2.2 Baselines

To investigate thoroughly the impact of hyper-dense connections between different streams in multi-modal image segmentation, we considered several baselines. First, we extend the semi-dense architecture proposed in [27] to a fully-dense one, by connecting the output of each convolutional layer to all subsequent layers. In this network, we follow an early-fusion strategy, in which MR-T1 and T2 are integrated at the input of the CNN and processed jointly along a single path (Fig. 8, *left*). The connectivity setting followed by this model is explained in eq. 2. Second, instead of merging both modalities at the input of the network, we considered a late-fusion strategy, where each modality is processed independently in different streams and learned features are fused before the first fully connected layer (Fig. 1 *right* in supplemental materials). In this model, the dense connections are included within each path, assuming the connectivity definition described in Eq. 2 for each of the streams.

2.3 Network architecture

To have a large receptive field, FCNNs typically use full images as input. The number of parameters is then limited via pooling/unpooling layers. A problem with this approach is the loss of resolution from repeated down-sampling operations. In the proposed method, we follow the strategy in [15], where sub-volumes are used as input, avoiding pooling layers. While sub-volumes of size $27 \times 27 \times 27$ are considered for training, we used $35 \times 35 \times 35$ sub-volumes during inference, as in [15], [36]. This strategy offers two considerable benefits. First, it reduces the memory requirements of our network, thereby removing the need for spatial pooling. More importantly, it substantially increases the number of training examples and, therefore, does not need data augmentation.

The network parameters are optimized via the RMSprop optimizer, using cross-entropy as cost function. Let $\boldsymbol{\theta}$ denotes the network parameters (i.e., convolution weights, biases and a_i from the parametric rectifier units), and y_s^v the label of voxel v in the s-th image segment. We optimize the following:

$$J(\boldsymbol{\theta}) \;=\; -\frac{1}{S \cdot V}\sum_{s=1}^{S}\sum_{v=1}^{V}\sum_{c=1}^{C}\delta(y_s^v = c)\cdot \log\, p_c^v(\mathbf{x}_s), \tag{4}$$

TABLE 2
The layers used in our architectures and the corresponding values with an input of size $27\times27\times27$. In the case of multi-modal images, the convolutional layers (conv_x) are present in any network path. All the convolutional layers have a stride of one pixel.

	Conv. kernel	# kernels	Output Size	Dropout
conv_1	3×3×3	25	25×25×25	No
conv_2	3×3×3	25	23×23×23	No
conv_3	3×3×3	25	21×21×21	No
conv_4	3×3×3	50	19×19×19	No
conv_5	3×3×3	50	17×17×17	No
conv_6	3×3×3	50	15×15×15	No
conv_7	3×3×3	75	13×13×13	No
conv_8	3×3×3	75	11×11×11	No
conv_9	3×3×3	75	9×9×9	No
fully_conv_1	1×1×1	400	9×9×9	Yes
fully_conv_2	1×1×1	200	9×9×9	Yes
fully_conv_3	1×1×1	150	9×9×9	Yes
Classification	1×1×1	4	9×9×9	No

where $p_c^v(\mathbf{x}_s)$ is the softmax output of the network for voxel v and class c, when the input segment is $\mathbf{x}_s$.

To initialize the weights of the network, we adopted the strategy proposed in [54], which yields fast convergence for very deep architectures. In this strategy, a zero-mean Gaussian distribution of standard deviation $\sqrt{2/n_l}$ is used to initialize the weights in layer l, where n_l denotes the number of connections to the units in that layer. Momentum was set to 0.6 and the initial learning rate to 0.001, being reduced by a factor of 2 after every 5 epochs (starting from epoch 10). The network was trained for 30 epochs, each composed of 20 subepochs. At each subepoch, a total of 1000 samples were randomly selected from the training images and processed in batches of size 5.

3 Experiments and Results

The proposed *HyperDenseNet* architecture is evaluated on two challenging multi-modal image segmentation tasks, using publicly available data provided by two MICCAI challenges: infant brain tissue segmentation, iSEG[3], and adult brain tissue segmentation, MRBrainS[4]. Quantitative evaluations and comparisons with the state-of-the-art methods are reported for each of these applications. First, to evaluate the impact of dense connectivity on performance, we compared the proposed *HyperDenseNet* to the baselines described in section 2.2 on infant brain tissue segmentation. Then, our results, compiled by the iSEG challenge organizers on testing data, are compared to those from the other competing teams. Second, to juxtapose the performance of *HyperDenseNet* to other segmentation networks under the same conditions, we provide a quantitative analysis of the results of current state-of-the-art segmentation networks for adult brain tissue segmentation. This includes comparison to the participants the MRBrainS challenge. Finally, in section 3.3, we report a comprehensive analysis of feature re-use.

3.1 iSEG Challenge

The focus of this challenge was to compare (semi-) automatic stat-of-the-art algorithms for the segmentation of 6-month infant brain tissues in T1- and T2-weighted brain MRI scans. This challenge was carried out in conjunction with MICCAI 2017, with a total of 21 international teams participating in the first round.

3.1.1 Evaluation

The MICCAI iSEG-2017 organizers used three metrics to evaluate the accuracy of the competing methods: Dice Similarity Coefficient (DSC) [55], Modified Hausdorff distance (MHD), where the 95-*th* percentile of all Euclidean distances is employed, and Average Surface Distance (ASD). The first measures the degree of overlap between the segmentation region and ground truth, whereas the other two evaluate boundary distances.

3.1.2 Results

We report infant brain tissue segmentation results using MR-T1w and MR-T2w images. Table 3 compares the performance achieved by *HyperDenseNet* to those of the baselines introduced in Section 2.2, for CSF, GM and WM brain tissues. The first observation that we can make from these results is that the late fusion of the high-layer features of independent paths did not provide a clear improvement over the single-path version. While the dual-path baseline reported better results for DC and ASD, processing both modalities in a single path achieved a better performance for MHD. *HyperDenseNet* outperformed both baselines, yielding better DC and ASD accuracy values for all cases, and better MHD values in two out of the three tissues. It achieved a lower MHD for the GM and WM tissues. Considering standard deviations, *HyperDenseNet* showed lower variances than the baselines, in the GM and WM regions. The difference in performance might be explained by the fact that our hyper-dense connectivity allows the network to access any feature map from almost any point in the architecture, both in depth and in width, thanks to the use of multiple inter-connected streams. This enable the network to freely learn features from several image modalities, at any level of abstraction, thereby identifying more complex patterns.

Figure 3 depicts a comparison of the training and validation accuracy between the baselines and *HyperDenseNet*. In these figures, the mean DC for the three brain tissue is evaluated on training (*Top*) and validation (*Bottom*) data after each sub-epoch. One can see that, in both cases, *HyperDenseNet* outperforms the baselines, achieving better results in lesser epochs. This might be attributed to two factors. The first is the high number of direct connections between different layers, which facilitates back-propagation of the gradient to shallow layers. The second is the freedom of the network to explore more complex patterns thanks to the combination of several image modalities at any level of abstraction.

Figure 4 depicts visual results for the subject used in validation. It can be observed that *HyperDenseNet* recovers thin regions better than the baselines, which can explain the improvements observed in the distance-based metrics. As confirmed in Table 3, this effect is most prominent in the boundaries between the gray and white matter. Furthermore, *HyperDenseNet* produces fewer false positives for

[3] http://iseg2017.web.unc.edu
[4] http://mrbrains13.isi.uu.nl

TABLE 3
Performance measures provided by the iSEG Challenge organizers for the analyzed methods. The best performance for each metric is highlighted in bold.

	DC	MHD	ASD
		CSF	
Single-Path Baseline	0.953 (0.007)	**9.296** (0.942)	0.128 (0.016)
Dual-Path Baseline	0.953 (0.008)	9.354 (1.242)	0.128 (0.018)
HyperDenseNet	**0.957 (0.007)**	9.421 (1.392)	**0.119 (0.017)**
		Gray Matter	
Single-Path Baseline	0.916 (0.009)	7.131 (1.729)	0.346 (0.041)
Dual-Path Baseline	0.918 (0.008)	7.643 (1.698)	0.343 (0.041)
HyperDenseNet	**0.920 (0.008)**	**5.752 (1.078)**	**0.329 (0.041)**
		White Matter	
Single-Path Baseline	0.895 (0.015)	6.903 (1.140)	0.406 (0.051)
Dual-Path Baseline	0.896 (0.013)	7.434 (1.571)	0.397 (0.045)
HyperDenseNet	**0.901 (0.014)**	**6.659 (0.932)**	**0.382 (0.047)**

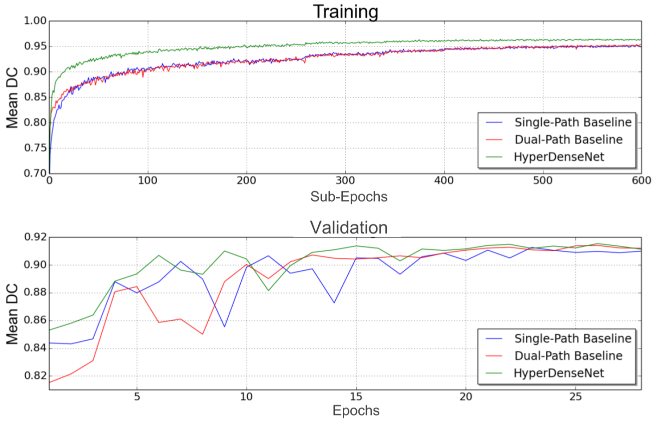
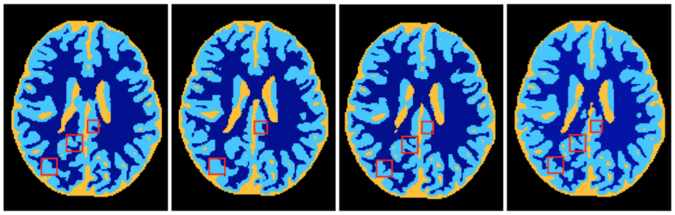


Fig. 3. Training (*top*) and validation (*bottom*) accuracy plots for the proposed architecture and the baselines.

WM than the baseline, which tend to over-estimate the segmentation in this region.


Single-Path Baseline

Dual-Path Baseline

HyperDenseNet

Ground Truth

Fig. 4. Qualitative results of segmentation achieved by the baselines and *HyperDenseNet* on the validation subject. The red squares indicate some spots, where *HyperDenseNet* successfully reproduced the ground-truth whereas the baselines failed.

Comparing these results to those of all the methods in the first round of the iSEG Challenge (Table 4), one can see that *HyperDenseNet* achieved a state-of-the-art accuracy for this task. It obtained the best performance in 7 out of the 9 metrics. A noteworthy point is the general decrease in performance, among all the methods, for the segmentation of GM and WM, with lower DC and larger ASD values. This suggests that the segmentation of these tissues is challenging due to the unclear boundaries between them.

3.2 MRBrainS Challenge

This MRBrainS challenge was carried out in conjunction with MICCAI 2013, and a total of 47 international teams have participated up to date. The focus is on adult brain tissue segmentation in the context of aging, and three modalities have been used for this purpose, MR-T1, MR-T1 Inversion Recovery (IR) and MR-FLAIR.

3.2.1 Evaluation

The organizers used three types of evaluation measures: a spatial overlap measure (DC), a boundary distance measure (MHD) and a volumetric measure (the percentage of absolute volume difference).

3.2.2 Architectures for comparison

We start by comparing *HyperDenseNet* to the state-of-the-art networks in medical image segmentation. The first architecture is a 3D fully convolutional neural network with residual connections [56], which we denote *FCN_Res3D*. Then, U-Net [57] with residual connections in the encoder and 3D volumes as input, referred to as *UNet3D*, is evaluated. Finally, *DeepMedic* [15], which has shown an outstanding performance in brain lesion segmentation, is included in the comparison. The implementation details are described in the supplemental materials.

3.2.3 Results

We performed a leave-one-out-cross-validation (LOOCV) on the training set to compare the proposed network to the state-of-the-art architectures listed above, using four subjects for training and one for validation. This process was repeated for 3 different subjects, and an average was computed. For this set of comparisons, we used all the three modalities, MR-T1, MR-T1 IR and FLAIR, for all the competing methods. In a second set of experiments, we assessed the impact of integrating multiple imaging modalities on the results of *HyperDenseNet* using all the possible combinations of two modalities as input.

Table 5 reports the mean DSC and standard-deviation values, with *FCN_Res3D* exhibiting the lowest mean DSC, a performance that might be explained by the transpose convolutions in *FCN_Res3D*, which may cause voxel misclassification within small regions. Furthermore, the downsampling and upsampling operations in *FCN_Res3D* make the feature maps in hidden layers sparser than the original inputs, causing a loss of the image details. This limitation is overcome by skip connections, which propagate information at different levels of abstraction between the encoding and decoding paths, as in *UNet3D*. This is reflected in the results, where the latter clearly outperformed *FCN_Res3D* in all the metrics. *DeepMedic* obtained better results than its competitors, yielding a performance close to the different two-modality configurations of *HyperDenseNet*. The dual multiscale path is an important feature of *DeepMedic*, which gives the network a larger receptive field via two paths, one for the input image and the other processing a low-resolution version of the input. This, in addition to the

TABLE 4
Results on the iSEG-2017 data for *HyperDenseNet* and the methods ranked in the top-5 at the first round of submissions (in alphabetical order). The bold fonts highlight the best performances. For additional details, we refer the reader to the challenge's website.

Method	**CSF**			**GM**			**WM**		
	DSC	**MHD**	**ASD**	**DSC**	**MHD**	**ASD**	**DSC**	**MHD**	**ASD**
Bern_IPMI	**0.96**	9.62	0.13	**0.92**	6.46	0.34	**0.90**	6.78	0.40
LIVIA(Ensemble)	**0.96**	9.13	**0.12**	**0.92**	6.06	0.34	**0.90**	7.45	0.41
MSL_SKKU	**0.96**	**9.07**	**0.12**	**0.92**	5.98	**0.33**	**0.90**	**6.44**	0.39
nic_vicorob	0.95	9.18	0.14	0.91	7.65	0.37	0.89	7.15	0.43
TU/e IMAG/e	0.95	9.43	0.15	0.90	6.86	0.38	0.89	6.91	0.43
HyperDenseNet (Ours)	**0.96**	9.42	**0.12**	**0.92**	**5.75**	**0.33**	**0.90**	6.66	**0.38**

TABLE 5
Comparison to several state-of-the-art 3D networks on the MRBrainS challenge.

Method	**Mean DSC (std dev)**		
	CSF	**GM**	**WM**
FCN_Res3D [58] (3-Modalities)	0.7685 (0.0161)	0.8163 (0.0222)	0.8607 (0.0178)
UNet3D [57] (3-Modalities)	0.8218 (0.0159)	0.8432 (0.0241)	0.8841 (0.0123)
DeepMedic [15] (3-Modalities)	0.8292 (0.0094)	0.8522 (0.0193)	0.8884 (0.0137)
HyperDenseNet (T1-FLAIR)	0.8259 (0.0133)	0.8620 (0.0260)	0.8982 (0.0138)
HyperDenseNet (T1_IR-FLAIR)	0.7991 (0.0181)	0.8226 (0.0255)	0.8654 (0.0087)
HyperDenseNet (T1-T1_IR)	0.8191 (0.0297)	0.8498 (0.0173)	0.8913 (0.0082)
HyperDenseNet (3-Modalities)	**0.8485** (0.0078)	**0.8663** (0.0247)	**0.9016** (0.0109)

removal of pooling operations in *DeepMedic*, could explain the increase in performance with respect to *FCN_Res3D* and *UNet3D*.

The two-modality versions of *HyperDenseNet* yielded competitive performances, although there is a significant variability between the three configurations. Notice that using MR-T1 and FLAIR already places *HyperDenseNet* first for two DSC measures (GM and WM), and second for the remaining measure (CSF), even though the competing methods used all three modalities. *HyperDenseNet* with three modalities yielded significantly better segmentations, obtaining the highest mean DSC values for all three tissues. These results confirm the importance of handling image modalities in separate paths with dense connections within and in-between the paths, facilitating the flow of information.

The MRBrainS challenge organizers compiled the results and a ranking of 47 international teams[5]. In Table 6, we report the results of the top-10 methods, with *HyperDenseNet* ranked first. This performance confirms, again, the importance of hyper-dense connections for multi-modal segmentation

A typical example of the obtained segmentation result is depicted in Fig. 5. In these images, red arrows indicate regions where the two-modality versions of *HyperDenseNet* fail in comparison to the three-modality version. Most of the errors of these networks occur at the boundaries between the GM and WM, which makes sense given the very weak contrast between these tissues; see the images in Fig. 1, for example. We can observe how *HyperDenseNet* with three modalities can handle thin regions better than its two-modality versions. This example demonstrates how the integration of correlated information can overcome the limitations or weaknesses of single modalities.

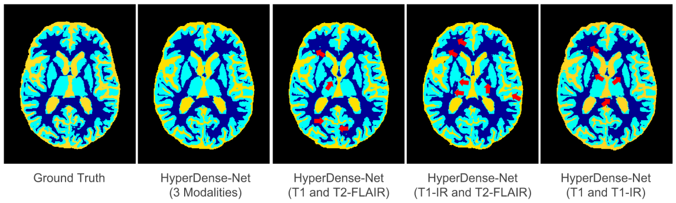

Ground Truth
HyperDense-Net (3 Modalities)
HyperDense-Net (T1 and T2-FLAIR)
HyperDense-Net (T1-IR and T2-FLAIR)
HyperDense-Net (T1 and T1-IR)

Fig. 5. A typical example of the segmentations achieved by the proposed *HyperDenseNet* in a validation subject (Subject 1 in the training set) for 2 and 3 modalities. The red arrows indicate some of the differences between the segmentations. For instance, one can see here that *HyperDenseNet* with three modalities can handle thin regions better than its two-modality versions.

3.3 Analysis of features re-use

Dense connectivity enables each network layer to access feature maps from all its preceding layers, strengthening feature propagation and encouraging feature re-use. To investigate the degree of usability of features in a trained network, we performed additional experiments. For each convolutional layer, we computed the average L_1-norm of its filter weights assigned to the connections to the previous layers from any stream. This serves as a surrogate for the dependency of a given layer on its preceding layers. We normalized the values between 0 and 1 to facilitate visualization. Fig. 6 depicts the weights of *HyperDense-Net* trained with two modalities, for both iSEG and MRBrainS challenges. As the MRBrainS data set contains three modalities, we have three different two-modality configurations. Fig 7

[5] http://mrbrains13.isi.uu.nl/results.php

TABLE 6
Results of the MICCAI MRBrainS challenge of different methods (DC, HD (mm) and AVD). Only the top-10 methods are included in this table. More details on the results of 47 international teams can be found in the MRBrainS Challenge Website.

Method	GM			WM			CSF			Sum
	DSC	HD	AVD	DSC	HD	AVD	DSC	HD	AVD	
HyperDenseNet (ours)	**0.8633**	**1.34**	6.19	**0.8946**	**1.78**	6.03	0.8342	2.26	7.31	**48**
VoxResNet [29] + Auto-context	0.8615	1.44	6.60	0.8946	1.93	6.05	**0.8425**	2.19	7.69	54
VoxResNet [29]	0.8612	1.47	6.42	0.8939	1.93	5.84	0.8396	2.28	7.44	56
MSL-SKKU	0.8606	1.52	6.60	0.8900	2.11	**5.54**	0.8376	2.32	6.77	61
LRDE	0.8603	1.44	6.05	0.8929	1.86	5.83	0.8244	2.28	9.03	61
MDGRU	0.8540	1.54	6.09	0.8898	2.02	7.69	0.8413	2.17	7.44	80
PyraMiD-LSTM2	0.8489	1.67	6.35	0.8853	2.07	5.93	0.8305	2.30	7.17	83
3D-UNet [57]	0.8544	1.58	6.60	0.8886	1.95	6.47	0.8347	2.22	8.63	84
IDSIA [59]	0.8482	1.70	6.77	0.8833	2.08	7.05	0.8372	**2.14**	7.09	100
STH [60]	0.8477	1.71	**6.02**	0.8845	2.34	7.67	0.8277	2.31	**6.73**	112

depicts the average weights for the case of three modalities. A dark square in these plots indicates that the target layer (*on x-axis*) makes a strong use of the features produced by the source layer (*on y-axis*).

An important observation that one can make from both figures is that, in most cases, all layers spread the importance of the connections over many previous layers, not only within the same path, but also from the other streams. This indicates that features extracted by shallower layers are directly used by deeper layers from both paths, which confirms the usefulness of hyper-dense connections in facilitating information flow and in learning complex relationships between the modalities within different levels of abstractions.

Looking into the details of each particular application, we can observe that, for *HyperDenseNet* trained on iSEG (top row of Fig 6), immediate previous layers have typically higher impact on the connections from both paths. Furthermore, the connections having access to the MR-T2 features typically have the strongest values, which may indicate that MR-T2 is more discriminative than T1 in this particular situation. We can also observe some regions with high (>0.5) feature re-usability patterns from shallow to deep layers. The same behaviour is observed for *HyperDenseNet* trained on two modalities from the MRBrainS challenge, where immediate previous layers have a high impact on the connections within and in-between the paths. The re-use of low-level features by deeper layers is more evident than in the previous case. For example, in *HyperDenseNet* trained with T1-IR and FLAIR, the deep layers in the T1-IR path make a strong use of the features extracted in shallower layers in the same path, as well as in the path that processed the FLAIR modality. This strong re-use of the features extracted early in the network from both paths occurred across all the configurations. The same pattern is observed when using three modalities (Fig 7), with a strong re-use of shallow features from the last layers in the network. This reflects the importance of allowing the deepest layers in the network to access to the early-extracted features. Additionally, it demonstrates that learning how and where to fuse information from multiple sources is a better strategy than combining image modalities in early or late stages.

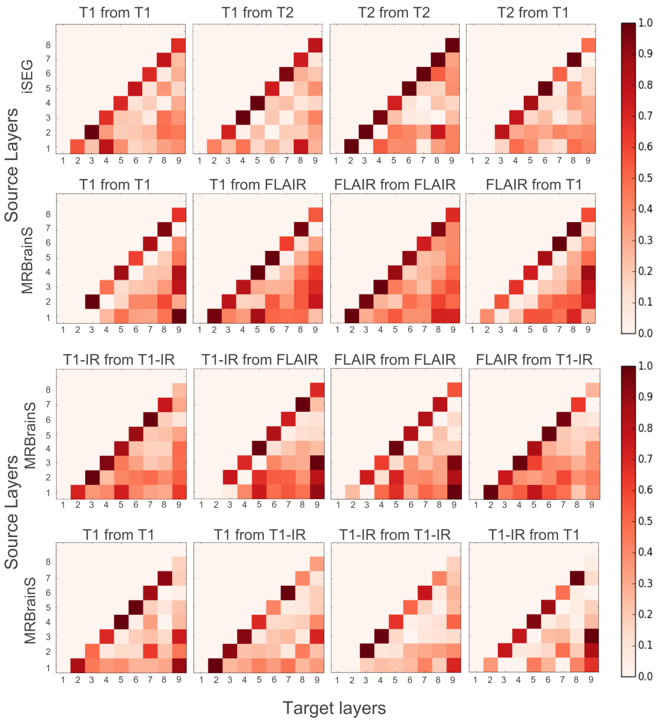


Fig. 6. The average of the L1-norm of the filters from *HyperDenseNet* trained on the iSEG (*top*) and MRBrainS (*from 2nd to 4th rows*) challenges with two modalities. The color at each location encodes the average L1 norm of the weights connecting a convolutional-layer source to a convolutional-layer target. These values were normalized between 0 and 1 by accounting for all the values within each layer.

4 Discussion

We presented a novel densely connected network, *HyperDenseNet*, which makes efficient use of multiple imaging modalities in the context of segmentation. In our model, each imaging modality has a path, and dense connections occur not only between the pairs of layers within the same path, but also between those across different paths.

We presented extensive experiments on two challenging public brain segmentation benchmarks, one focusing on 6-month infant data and the other on adult images, with significant differences in the image characteristics between the two benchmarks. *HyperDenseNet* yielded state-of-the-art performances on both. Considering several baselines,

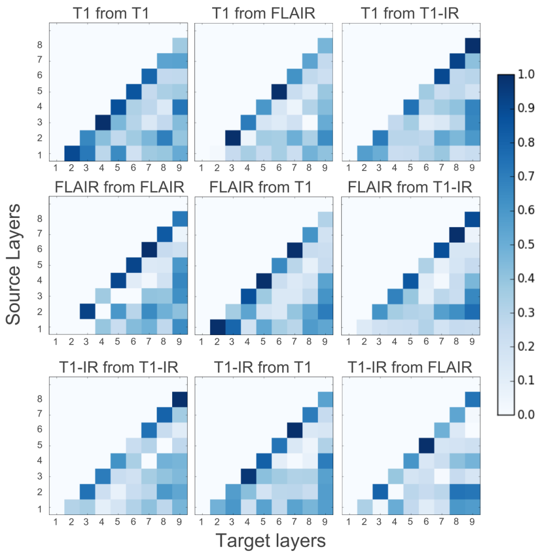


Fig. 7. The average of the L1-norm of the filters from *HyperDenseNet* trained on the MRBrainS challenge with three modalities (MR-T1, FLAIR and MR-T1 IR). The color at each location encodes the average L1 norm of the weights connecting a convolutional-layer source to a convolutional-layer target. These values were normalized between 0 and 1 by accounting for all the values within each layer.

we showed experimentally the positive impact of hyper-dense connections between different streams in multi-modal segmentation. It is worth mentioning that the architecture and hyper-parameters were unchanged through the experiments of both challenges, showing the flexibility of the proposed model in brain related problems. We anticipate that our model has a much larger applicability scope, going beyond brain tissue regions, in a breadth of other multi-modal volumetric medical image segmentation problems.

The accuracy improvements that *HyperDenseNet* brings might be explained by several factors. First, the network has total freedom to learn complex combinations between the modalities, within and in-between all the levels of abstraction, which increases significantly the learning representation. Second, it implicitly imposes deep supervision from the loss function that individual layers receive via the dense connections. The benefits of such deep supervision have been observed previously in the different context of deeply-supervised nets [61], which impose explicit layer-wise supervision by associating a classifier to each of the hidden layers. Finally, hyper-dense connections facilitate gradient flow, allowing feature re-use throughout the network, as evidenced by our experimental analysis. In fact, shallow layers typically operate at a fine-grained scale, to extract low-level features, while deeper layers extract coarse-scale information so as to infer the global context. Both scales are important, but occur at different levels in the network. Including dense connections alleviates this problem by facilitating the information flow between the different scales.

We evaluated the impact of the number of modalities on the performance and found that, even with less modalities, *HyperDenseNet* yields competitive performances in comparison to several state-of-the-art segmentation networks under the same conditions. An important finding in this experiment was that *DeepMedic* outperformed two versions of *HyperDenseNet*, when the latter is trained with only two modalities instead of three: T1-T1_IR and T1_IR-FLAIR. In fact, *DeepMedic* explicitly uses multi-scale information, incorporating larger receptive fields via two paths, one for the input image and the other processing a low-resolution version of the input. This suggests that integrating hyper-dense connections with explicit multi-scale information and larger receptive fields might further improve the network performance. Indeed, Huang et al. [62] investigated the use of dense connections in a multi-scale branch network, *Multi-Scale Dense Net* (MSDN), for image classification, showing that the multi-scale version outperforms its predecessor *DenseNet*.

Our experimental analysis of feature re-use revealed a strong information flow between the deep and shallow layers, particularly at the high levels of abstraction, an observation aligned with the findings in [48]. This indicates that the early-layer features are directly used by the deep layers. Our feature re-use study has an important benefit: it may allow us to remove useless connections, keeping only those having a strong information flow. This will ideally generate a computationally efficient model, without affecting performance.

5 Conclusion

This study investigated a hyper-densely connected 3D fully CNN, *HyperDenseNet*, with applications to brain tissue segmentation in multi-modal MRI. Our model leverages dense connectivity beyond recent works, exploiting the concept in multi-modal problems. Dense connections occur not only within each single-modality stream, but also across the streams, which increases significantly the learning representation in multi-modal problems: The network has total freedom to explore complex combinations between the different modalities, within and in-between all the levels of abstraction. We reported comprehensive evaluations and comparisons using the benchmarks of two highly competitive challenges, iSEG-2017 for 6-month infant brain segmentation and MRBrainS for adult data, showing state-of-the-art performances of *HyperDenseNet* on both. The experiments presented in this work provided new insights on the inclusion of short-cut connections in deep neural networks for segmentating medical images, particularly in multi-modal scenarios. *HyperDenseNet* demonstrated its potential to tackle multi-modal volumetric medical image segmentation problems.

Acknowledgments

This work is supported by the National Science and Engineering Research Council of Canada (NSERC), discovery grant program, and by the ETS Research Chair on Artificial Intelligence in Medical Imaging. The authors would like to thank both iSEG and MRBrainS organizers for providing data benchmarks and evaluations.

References

[1] D. Delbeke, H. Schöder, W. H. Martin, and R. L. Wahl, “Hybrid imaging (SPECT/CT and PET/CT): improving therapeutic decisions,” in *Seminars in nuclear medicine*, vol. 39, no. 5. Elsevier, 2009, pp. 308–340.

[2] X. Lladó, A. Oliver, M. Cabezas, J. Freixenet, J. C. Vilanova, A. Quiles, L. Valls, L. Ramió-Torrentà, and À. Rovira, "Segmentation of multiple sclerosis lesions in brain MRI: a review of automated approaches," *Information Sciences*, vol. 186, no. 1, pp. 164–185, 2012.

[3] I. El Naqa, D. Yang, A. Apte, D. Khullar, S. Mutic, J. Zheng, J. D. Bradley, P. Grigsby, and J. O. Deasy, "Concurrent multimodality image segmentation by active contours for radiotherapy treatment planning," *Medical physics*, vol. 34, no. 12, pp. 4738–4749, 2007.

[4] H. Yu, C. Caldwell, K. Mah, and D. Mozeg, "Coregistered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning," *IEEE Transactions on Medical Imaging*, vol. 28, no. 3, pp. 374–383, 2009.

[5] D. Han, J. Bayouth, Q. Song, A. Taurani, M. Sonka, J. Buatti, and X. Wu, "Globally optimal tumor segmentation in PET-CT images: a graph-based co-segmentation method," in *Biennial International Conference on Information Processing in Medical Imaging*. Springer, 2011, pp. 245–256.

[6] U. Bagci, J. K. Udupa, J. Yao, and D. J. Mollura, "Co-segmentation of functional and anatomical images," in *International Conference on MICCAI*. Springer, 2012, pp. 459–467.

[7] D. Markel, C. Caldwell, H. Alasti, H. Soliman, Y. Ung, J. Lee, and A. Sun, "Automatic segmentation of lung carcinoma using 3D texture features in 18-FDG PET/CT," *International journal of molecular imaging*, vol. 2013, 2013.

[8] Q. Song, J. Bai, D. Han, S. Bhatia, W. Sun, W. Rockey, J. E. Bayouth, J. M. Buatti, and X. Wu, "Optimal co-segmentation of tumor in PET-CT images with context information," *IEEE Transactions on Medical Imaging*, vol. 32, no. 9, pp. 1685–1697, 2013.

[9] U. Bagci, J. K. Udupa, N. Mendhiratta, B. Foster, Z. Xu, J. Yao, X. Chen, and D. J. Mollura, "Joint segmentation of anatomical and functional images: Applications in quantification of lesions from PET, PET-CT, MRI-PET, and MRI-PET-CT images," *Medical image analysis*, vol. 17, no. 8, pp. 929–945, 2013.

[10] X. Wang, C. Ballangan, H. Cui, M. Fulham, S. Eberl, Y. Yin, and D. Feng, "Lung tumor delineation based on novel tumor-background likelihood models in PET-CT images," *IEEE Transactions on Nuclear Science*, vol. 61, no. 1, pp. 218–224, 2014.

[11] W. Ju, D. Xiang, B. Zhang, L. Wang, I. Kopriva, and X. Chen, "Random walk and graph cut for co-segmentation of lung tumor on PET-CT images," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5854–5867, 2015.

[12] H. Cui, X. Wang, W. Lin, J. Zhou, S. Eberl, D. Feng, and M. Fulham, "Primary lung tumor segmentation from PET–CT volumes with spatial–topological constraint," *International journal of computer assisted radiology and surgery*, vol. 11, no. 1, pp. 19–29, 2016.

[13] 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*, vol. 34, no. 10, pp. 1993–2024, 2015.

[14] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "HeMIS: Hetero-modal image segmentation," in *International Conference on MICCAI*. Springer, 2016, pp. 469–477.

[15] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61–78, 2017.

[16] L. Fidon, W. Li, L. C. Garcia-Peraza-Herrera, J. Ekanayake, N. Kitchen, S. Ourselin, and T. Vercauteren, "Scalable multimodal convolutional networks for brain tumour segmentation," in *International Conference on MICCAI*. Springer, 2017, pp. 285–293.

[17] M. Prastawa, J. H. Gilmore, W. Lin, and G. Gerig, "Automatic segmentation of MR images of the developing newborn brain," *Medical image analysis*, vol. 9, no. 5, pp. 457–466, 2005.

[18] N. I. Weisenfeld, A. Mewes, and S. K. Warfield, "Segmentation of newborn brain MRI," in *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*. IEEE, 2006, pp. 766–769.

[19] P. Anbeek, K. L. Vincken, F. Groenendaal, A. Koeman, M. J. Van Osch, and J. Van der Grond, "Probabilistic brain tissue segmentation in neonatal magnetic resonance imaging," *Pediatric research*, vol. 63, no. 2, pp. 158–163, 2008.

[20] N. I. Weisenfeld and S. K. Warfield, "Automatic segmentation of newborn brain MRI," *Neuroimage*, vol. 47, no. 2, pp. 564–572, 2009.

[21] L. Wang, F. Shi, W. Lin, J. H. Gilmore, and D. Shen, "Automatic segmentation of neonatal images using convex optimization and coupled level sets," *NeuroImage*, vol. 58, no. 3, pp. 805–817, 2011.

[22] V. Srhoj-Egekher, M. Benders, K. J. Kersbergen, M. A. Viergever, and I. Isgum, "Automatic segmentation of neonatal brain MRI using atlas based segmentation and machine learning approach," *MICCAI Grand Challenge: Neonatal Brain Segmentation*, vol. 2012, 2012.

[23] S. Wang, M. Kuklisova-Murgasova, and J. A. Schnabel, "An atlas-based method for neonatal MR brain tissue segmentation," *Proceedings of the MICCAI Grand Challenge: Neonatal Brain Segmentation*, pp. 28–35, 2012.

[24] L. Wang, F. Shi, G. Li, Y. Gao, W. Lin, J. H. Gilmore, and D. Shen, "Segmentation of neonatal brain MR images using patch-driven level sets," *NeuroImage*, vol. 84, pp. 141–158, 2014.

[25] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, 2015.

[26] D. Nie, L. Wang, Y. Gao, and D. Sken, "Fully convolutional networks for multi-modality isointense infant brain image segmentation," in *13th International Symposium on Biomedical Imaging (ISBI), 2016*. IEEE, 2016, pp. 1342–1345.

[27] J. Dolz, C. Desrosiers, L. Wang, J. Yuan, D. Shen, and I. Ben Ayed, "Deep CNN ensembles and suggestive annotations for infant brain MRI segmentation," *arXiv preprint arXiv:1712.05319*, 2017.

[28] A. M. Mendrik, K. L. Vincken, H. J. Kuijf, M. Breeuwer, W. H. Bouvy, J. De Bresser, A. Alansary, M. De Bruijne, A. Carass, A. El-Baz *et al.*, "MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans," *Computational intelligence and neuroscience*, vol. 2015, p. 1, 2015.

[29] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images," *NeuroImage*, 2017.

[30] S. C. Deoni, B. K. Rutt, A. G. Parrent, and T. M. Peters, "Segmentation of thalamic nuclei using a modified k-means clustering algorithm and high-resolution quantitative magnetic resonance imaging at 1.5T," *Neuroimage*, vol. 34, no. 1, pp. 117–126, 2007.

[31] O. Commowick, F. Cervenansky, and R. Ameli, "MSSEG Challenge proceedings: Multiple Sclerosis Lesions Segmentation Challenge using a data management and processing infrastructure," in *MICCAI*, 2016.

[32] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert *et al.*, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *International Conference on IPMI*. Springer, 2017, pp. 597–609.

[33] S. Valverde, M. Cabezas, E. Roura, S. González-Villà, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, A. Oliver, and X. Lladó, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach," *NeuroImage*, vol. 155, pp. 159–168, 2017.

[34] S. González-Villà, A. Oliver, S. Valverde, L. Wang, R. Zwiggelaar, and X. Lladó, "A review on brain structures segmentation in magnetic resonance imaging," *Artificial intelligence in medicine*, vol. 73, pp. 45–69, 2016.

[35] A. Makropoulos, S. J. Counsell, and D. Rueckert, "A review on automatic fetal and neonatal brain MRI segmentation," *NeuroImage*, 2017.

[36] J. Dolz, C. Desrosiers, and I. Ben Ayed, "3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study," *NeuroImage*, 2017.

[37] T. Fechter, S. Adebahr, D. Baltas, I. Ben Ayed, C. Desrosiers, and J. Dolz, "Esophagus segmentation in CT via 3D fully convolutional neural network and random walk," *Medical Physics*, 2017.

[38] K. Kamnitsas, L. Chen, C. Ledig, D. Rueckert, and B. Glocker, "Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI," *Ischemic Stroke Lesion Segmentation*, vol. 13, 2015.

[39] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. Benders, and I. Išgum, "Automatic segmentation of MR brain images with a convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1252–1261, 2016.

[40] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *Journal of Machine Learning Research*, vol. 15, pp. 2949–2980, 2014.

[41] L. Wang, F. Shi, Y. Gao, G. Li, J. H. Gilmore, W. Lin, and D. Shen, "Integration of sparse multi-modality representation and anatomical constraint for isointense infant brain MR image segmentation," *NeuroImage*, vol. 89, pp. 152–164, 2014.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on CVPR*, 2016, pp. 770–778.

[43] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *ECCV*. Springer, 2016, pp. 646–661.

[44] G. Larsson, M. Maire, and G. Shakhnarovich, "Fractalnet: Ultra-deep neural networks without residuals," *arXiv preprint arXiv:1605.07648*, 2016.

[45] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[46] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[47] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in *AAAI*, 2017, pp. 4278–4284.

[48] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE CVPR*, 2017.

[49] J. Dolz, I. Ben Ayed, J. Yuan, and C. Desrosiers, "Isointense infant brain segmentation with a Hyper-dense connected convolutional neural network," in *Biomedical Imaging (ISBI), 2018 IEEE 15th International Symposium on*. IEEE, 2018, pp. 616–620.

[50] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE CVPR*, 2015, pp. 3431–3440.

[51] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P. A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and liver tumor segmentation from CT volumes," *arXiv:1709.07330*, 2017.

[52] L. Yu, J.-Z. Cheng, Q. Dou, X. Yang, H. Chen, J. Qin, and P.-A. Heng, "Automatic 3D cardiovascular MR segmentation with densely-connected volumetric convnets," in *International Conference on MICCAI*. Springer, 2017, pp. 287–295.

[53] L. Chen, Y. Wu, A. M. DSouza, A. Z. Abidin, A. Wismuller, and C. Xu, "MRI tumor segmentation with densely connected 3D CNN," *arXiv preprint arXiv:1802.02427*, 2018.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE ICCV*, 2015, pp. 1026–1034.

[55] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *ECCV*. Springer, 2016, pp. 630–645.

[57] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on MICCAI*. Springer, 2016, pp. 424–432.

[58] N. Pawlowski, S. I. Ktena, M. C. Lee, B. Kainz, D. Rueckert, B. Glocker, and M. Rajchl, "DLTK: State of the art reference implementations for deep learning on medical images," *arXiv preprint arXiv:1711.06853*, 2017.

[59] M. F. Stollenga, W. Byeon, M. Liwicki, and J. Schmidhuber, "Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation," in *NIPS*, 2015, pp. 2998–3006.

[60] A. Mahbod, M. Chowdhury, Ö. Smedby, and C. Wang, "Automatic brain segmentation using artificial neural networks with shape context," *Pattern Recognition Letters*, vol. 101, pp. 74–79, 2018.

[61] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial Intelligence and Statistics*, 2015, pp. 562–570.

[62] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger, "Multi-scale dense convolutional networks for efficient prediction," *arXiv preprint arXiv:1703.09844*, 2017.

[63] Y. D. Reijmer, A. Leemans, M. Brundel, L. J. Kappelle, G. J. Biessels *et al.*, "Disruption of the cerebral white matter network is related to slowing of information processing speed in patients with type 2 diabetes," *Diabetes*, vol. 62, no. 6, pp. 2112–2115, 2013.

[64] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: a toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

[65] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical parametric mapping: the analysis of functional brain images*. Academic press, 2011.

Supplemental Materials

Dataset iSEG:

The images were acquired at the UNC-Chapel Hill on a Siemens head-only 3T scanner with a circular polarized head coil, and were randomly chosen from the pilot study of the Baby Connectome Project (BCP)[6]. During scan, infants were asleep, unsedated and fitted with ear protection, with the head secured in a vacuum-fixation device. T1-weighted images were acquired with 144 sagittal slices using the following parameters: TR/TE = 1900/4.38 ms, flip angle = 7° and resolution = $1{\times}1{\times}1$ mm^3. Likewise, T2-weighted images were obtained with 64 axial slices, TR/TE = 7380/119 ms, flip angle = 150° and resolution =$1.25{\times}1.25{\times}1.95$ mm^3. T2 images were linearly aligned onto their corresponding T1 images. All the images were resampled into an isotropic $1{\times}1{\times}1$ mm^3 resolution. Standard image pre-processing steps were then applied using in-house tools, including skull stripping, intensity inhomogeneity correction, and removal of the cerebellum and brain stem. For this application, 9 subjects were employed for training and 1 for validation.

Dataset MRBrainS:

20 subjects with a mean age of 71 ± 4 years (10 male, 10 female) were selected from an ongoing cohort study of older ($65 - 80$ years of age), functionally-independent individuals without a history of invalidating stroke or other brain diseases [63]. To test the robustness of the segmentation algorithms in the context of aging-related pathology, the subjects were selected to have varying degrees of atrophy and white-matter lesions, and the scans with major artifacts were excluded. The following sequences were acquired and used for the evaluation framework: 3D T1 (TR: 7.9 ms, TE: 4.5 ms), T1-IR (TR: 4416 ms, TE: 15 ms, and TI: 400 ms) and T2- FLAIR (TR: 11000 ms, TE: 125 ms, and TI: 2800 ms). The sequences were aligned by rigid registration using Elastix [64], along with a bias correction performed using SPM8 [65]. After the registration, the voxel size within all the provided sequences (T1, T1 IR, and T2 FLAIR) was $0.96{\times}0.96{\times}3.00$ mm^3. Five subjects that were representative for the overall data (2 male, 3 female and varying degrees of atrophy and white-matter lesions) were selected for training. The remaining fifteen subjects were provided as testing data. While ground truth was provided for the 5 training subjects, manual segmentations were unknown for the testing data set. The following structures were segmented and were available for training: (a) cortical gray matter, (b) basal ganglia, (c) white matter, (d) white matter lesions, (e) peripheral cerebrospinal fluid, (f) lateral ventricles, (g) cerebellum and (h) brainstem. These structures can be merged into gray matter (a-b), white matter (c-d), and cerebrospinal fluid (e-f). The cerebellum and brainstem were excluded from the evaluation.

Dice similarity coefficient (DSC):

Let V_{ref} and V_{auto} be, respectively, the reference and automatic segmentations of a given tissue class and for a given subject. The DSC for this subject can be defined as:

$$\mathrm{DSC}\big(V_{\mathrm{ref}}, V_{\mathrm{auto}}\big) \;=\; \frac{2 \mid V_{\mathrm{ref}} \cap V_{\mathrm{auto}} \mid}{\mid V_{\mathrm{ref}} \mid + \mid V_{\mathrm{auto}} \mid} \tag{5}$$

DSC values are within a $[0,1]$ range, 1 indicating perfect overlap and 0 corresponding to a total mismatch.

Modified Hausdorff distance (MHD):

Let P_{ref} and P_{auto} denote the sets of voxels within the reference and automatic segmentation boundary, respectively. MHD is given by:

$$\mathrm{MHD}\big(P_{\mathrm{ref}}, P_{\mathrm{auto}}\big) \;=\; \max\Big\{ \max_{q\in P_{\mathrm{ref}}} d(q, P_{\mathrm{auto}}),\; \max_{q\in P_{\mathrm{auto}}} d(q, P_{\mathrm{ref}}) \Big\}, \tag{6}$$

where $d(q,P)$ is the point-to-set distance defined by: $d(q,P) = \min_{p\in P} \|q-p\|$, with $\|.\|$ denoting the Euclidean distance. Low MHD values indicate high boundary similarity.

Average surface distance (ASD):

Using the same notation as the Hausdorff distance above, the ASD corresponds to:

$$\mathrm{ASD}\big(P_{\mathrm{ref}}, P_{\mathrm{auto}}\big) \;=\; \frac{1}{|P_{\mathrm{ref}}|} \sum_{p\,\in\,P_{\mathrm{ref}}} d(p, P_{\mathrm{auto}}), \tag{7}$$

where $|.|$ denotes the cardinality of a set. In distance-based metrics, smaller values indicate higher proximity between two point sets and, thus, a better segmentation.

5.0.0.1 Absolute Volume Differences:

$$\mathrm{AVD}\big(V_{\mathrm{ref}}, V_{\mathrm{auto}}\big) \;=\; \frac{\mid V_{\mathrm{ref}} - V_{\mathrm{auto}} \mid}{V_{\mathrm{ref}}} \cdot 100 \tag{8}$$

where V_{auto} and V_{ref} are the volume of the segmentation result and reference, respectively. These measures were used to evaluate the following brain structures in each of the fifteen test datasets: GM, WM, CSF, brain (GM + WM), and intracranial volume (GM + WM + CSF). The brainstem and cerebellum are excluded from the evaluation.

Implementation:

We extended our 3D FCNN architecture proposed in [36], which is based on Theano. The source code of this architecture is publicly available[7]. Training and testing was performed on a server equipped with a NVIDIA Tesla P100 GPU with 16 GB of RAM memory. Training *HyperDenseNet* took around 70 min per epoch, and around 35 hours in total for the two-modality version. With three image modalities, training each epoch took nearly 3 hours. Inference on a whole 3D MR scan took on average from 70-80 to 250-270 seconds, for the two- and three-modality versions, respectively.

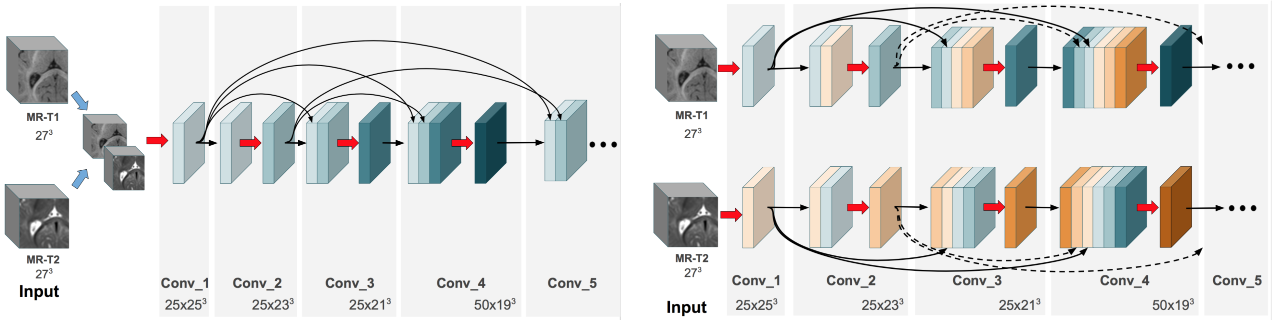
FCN_Res3D: The architecture of *FCN_Res3D* consists on 5 convolutional blocks with residual units on the encoder path, with 16, 64, 128, 256 and 512 kernels. The decoding path contains 4 convolutional upsampling blocks, each composed of 4 kernels, one per class. At each residual block, batch normalization and a Leaky ReLU with a leakage value of 0.1 are employed before the convolution. Instead of including max-pooling operations to re-size the images, stride values of $2 \times 2 \times 2$ are used in layers 2, 3 and 4. Volume size at the input of the network is $64 \times 64 \times 24$. The implementation of this network is provided in [58] [8].

UNet3D: Although quite similar to *FCN_Res3D*, *UNet3D* presents some differences, particularly in the decoding path. It contains 9 convolutional blocks in total, 4 in the encoding and 5 in the decoding path. The number of kernels in the

[6] http://babyconnectomeproject.org

[7] https://github.com/josedolz/SemiDenseNet

[8] https://github.com/DLTK/DLTK



Single Dense Path

Dual Dense Path

Fig. 8. Section of baseline architectures: single-path dense (*left*) and dual-path dense (*right*). While in the first case both modalities are concatenated at the input of the network, each modality is analyzed independently in the second architecture, and features fused at the end of the streams. Each gray region represents a convolutional block. Red arrows correspond to convolutions and black arrows indicate dense connections between feature maps. Dense connections are propagated through the entire network.

encoding path are 32, 64, 128 and 256, with strides of 2 × 2 × 2 at layers 2, 3 and 4. In the decoding path, the number of kernels are 256, 128, 64, 32 and 4, from the first to the last layer. Furthermore, skip connections are added at the convolutional blocks of the same scale between the encoding and decoding paths. As in *FCN_Res3D*, batch normalization and a Leaky ReLU with a leakage value of 0.1 are employed before the convolution at each block. Volume size at the input of the network is also 64 × 64 × 24. The implementation is provided in [58].

DeepMedic**:** We used the default architecture of *DeepMedic* in our experiments. This architecture includes two paths with 8 convolutional blocks: 30, 30, 40, 40, 40, 40, 50, 50 kernels of size 3×3×3. At the end of both paths, two fully connected convolutional layers with 150 1×1×1 filters each are added, before the last classification layer. The second path is used with a low-resolution version of the input at the first path, for a larger receptive field. The input patch size is 27×27×27 and 35×35×35 for training and segmentation, respectively. The official code [9] is employed to evaluate this architecture.

[9] https://github.com/Kamnitsask/deepmedic