Multi-Modal Brain Segmentation Using Hyper-Fused Convolutional Neural Network

Wenting Duan[1], Lei Zhang[1], Jordan Colman[1,2], Giosue Gulli[2] and Xujiong Ye[1]

[1] Department of Computer Science, University of Lincoln, UK
[2] Ashford and St Peter's Hospitals NHS Foundation Trust, Surrey, UK
wduan@lincoln.ac.uk

Abstract. Algorithms for fusing information acquired from different imaging modalities have shown to improve the segmentation results of various applications in the medical field. Motivated by recent successes achieved using densely connected fusion networks, we propose a new fusion architecture for the purpose of 3D segmentation in multi-modal brain MRI volumes. Based on a hyper-densely connected convolutional neural network, our network features in promoting a progressive information abstraction process, introducing a new module – ResFuse to merge and normalize features from different modalities and adopting combo loss for handing data imbalances. The proposed approach is evaluated on both an outsourced dataset for acute ischemic stroke lesion segmentation and a public dataset for infant brain segmentation (iSeg-17). The experiment results show our approach achieves superior performances for both datasets compared to the state-of-art fusion network.

Keywords: Multi-Modal Fusion, Dense Network, Brain Segmentation.

1 Introduction

In medical imaging, segmentation of lesions or organs using a multi-modal approach has become a growing trend strategy as more advanced systems and data becomes available. For example, magnetic resonance imaging (MRI) that is widely used for brain lesion or tumor detection and segmentation comes in several modalities including T1-weighted (T1), T2-weighted (T2), FLuid Attenuated Inversion Recovery (FLAIR) and Diffusion-weighted image (DWI), etc. Compared to single modality, the extraction of information from multi-modal images brings complementary information that contributes to reduced uncertainty and an improved discriminative power of the clinical diagnosis system [1]. Motivated by the success of deep learning, image fusion strategies have largely moved from probability theory [2] or fuzzy concept [3] based methods to deep convolutional neural network based approaches [1, 4].

Promising performance has been achieved by deep learning based methods for medical image segmentation from multi-modal images. The most widely applied strategy is simply concatenating images or image patches of different modalities to learn a unified image features set [5–7]. Such networks combine the data at the input level to form a multi-channel input. Another straightforward fusion strategy is for images of each modality to learn an independent feature map. Then these single-modality feature sets will, either learn their separate classifiers and use 'votes' to arrive at a final output, or learn a multi-modal classifier integrating high-level representations of different modalities [8-10]. In comparison to the strategies mentioned previously where fusion happens either at the input level or the output/classifier level, some recent works [11-14] have proved that performing fusion within the convolutional feature learning stage instead generally gives much better segmentation results. Tseng *et al.* [14] proposed a cross-modality convolution to aggregate data from different modalities within an encoder decoder network. The convolution LSTM is then used to model the correlations between slices. The method requires images of all modalities to be co-registered and the network parameters varies with the number of slices involved in the training dataset. For unpaired modalities such as CT and MRI, Dou *et* al. [15] developed a novel scheme involving separate feature normalization but shared convolution. Knowledge distillation-based loss is proposed to promote softer probability distribution over classes. However, the design so far is limited to two modalities. Another avenue of research on multi-modal fusion is based on DenseNet [16] where feature re-use is induced by connecting each layer with all previous layers. For example, Dolz *et al.* [13] extends the DenseNet so that the dense connections not only exist in the layers of same modality but also between the modalities. Their network (i.e. HyperDense-Net) made significant improvements over other state-of-art segmentation techniques and ranked first for two highly competitive multi-modal brain segmentation challenges. Dolz *et al.* [17] also explored the integration of DenseNet in U-Net, which involved a multi-path densely connected encoder and inception module-based convolution blocks with dilated convolution at different scales. However, the network input only accepts 2D slides and not 3D volumes.

As reviewed in [3], dense connection-based layer-level fusion improves the effectiveness and efficiency of multi-modal segmentation network through better information propagation, implicit deep supervision and reduced risk of over-fitting on small datasets. While recognising the advantages provided by densely connected networks for multi-modal fusion, HyperDense-Net architecture has some limitations which we address in this paper. The first lies in the variation of filter depth. Compared to many other segmentation networks such as U-Net, HyperDense-Net contains no pooling layer between convolutional layers and is overall not so deep (i.e. contains nine convolution blocks and four fully-convolutional layers). However, it retained the conventional way of increasing the number of filters (just like the networks with pooling layers) by doubling or multiplying 1.5 after every three consecutive convolution blocks, resulting in a drastic change in feature abstraction in the 4th and 7th layers and moderate learning in other layers. The other lacking aspect we identified is the way multi-modal feature maps concatenate. In HyperDense-Net, the feature maps from all modalities as well as previous layers are simply fused using concatenation along the channel dimension. We speculate this approach fails to consider the discrepancy in visual features under different modalities and the importance of modal-specific learning, resulting in ineffective multi-modal feature merging and propagation.

Given the challenges and limitations described above, we propose a new densely connected fusion architecture, which we refer to as HyperFusionNet, for multi-modal brain tissue segmentation. The proposed network is trained in an end-to-end fashion, where a progressive feature abstraction process is ensured, and a better feature fusion strategy is integrated to alleviate the interference and incompatibility of feature maps generated from different modality paths. We compare the proposed architecture to the state-of-art method using both a private dataset on acute ischemic stroke lesions and data from the iSeg-2017 MICCAI Grand challenge [18] on 6-month infant brain MRI Segmentation.

2 Method

2.1 Baseline Architecture

The pipeline of the baseline architecture – HyperDense-Net [13] is shown in Fig. 1, but without the added ResFuse modules. Taking the fusion of three modalities as an example, each imaging modality has its own stream for the propagation of the features until it reaches the fully convolutional layer. Every convolutional block includes batch normalization, PReLU activation and convolution with no spatial pooling. For a convolutional block in a conventional CNN, the output of the current layer, denoted as x_l, is obtained by applying a mapping function $F_l(\cdot)$ to the output x_{l-1} of the previous layer, i.e.

$$x_l = F_l(x_{l\text{-}1}) \tag{1}$$

However, in the HyperDense-Net, feature maps generated from different modalities as well as the feature outputs from previous layers are concatenated in a feed-forward manner to be input to the convolution block. Let M represents the number of modalities involved in the multi-modal network, the output of the l^{th} layer along a stream $m = 1,2,\dots,M$ in the baseline architecture is then defined as

$$x_l^m = F_l([x_{l\text{-}1}^1, x_{l\text{-}1}^2, ..., x_{l\text{-}1}^M, x_{l\text{-}2}^1, x_{l\text{-}2}^2, ..., x_{l\text{-}2}^M, ..., x_0^M]) \tag{2}$$

All streams are then concatenated together before entering the fully convolutional layers. The output of the network is fed into a softmax function to generate the probabilistic map. The final segmentation result is computed based on the highest probability value. The baseline network is optimised using Adam optimiser and cross-entropy loss function.

2.2 Proposed Architecture

To avoid drastic changes of feature abstraction, we first modified the number of filter sizes in the baseline network. Instead of having equal number of filters for every three consecutive convolutional blocks, we gradually increase the number of filters in the successive blocks. Let w denotes the increased value in filter number in the original network, we add ${}^{W}/_{3}$ filters to the successive convolutional layers in the proposed network. The effectiveness of such design is demonstrated previously in [19].

To improve the fusion of the multi-modal features along each modality path, we propose to merge the feature maps via a ‘ResFuse Module’. Inspired by [20], the module (illustrated in Fig. 2) contains a residual connection where the main information belonging to that specific modality path is traversed directly. A 1×1 convolutional layer is also introduced to allow some comprehension of channel correspondence between the features of the specific path and the merging information from other modalities. For the concatenated feature maps, we apply non-linear PReLU activation before summation in order to promote better mapping and information flow in the fused propagation. Equation 2 is then updated to

$$x_l^m = H_l(x_{l-1}^m) + G_l([x_{l-1}^1, x_{l-1}^2, x_{l-1}^M, x_{l-2}^1, x_{l-2}^2, \dots, x_{l-2}^M, \dots, x_0^M]) \tag{3}$$

where H_l applies the dimension expansion of x_{l-1}^m via the 1×1 convolution and G_l performs the concatenation of features from all modalities and the activation.

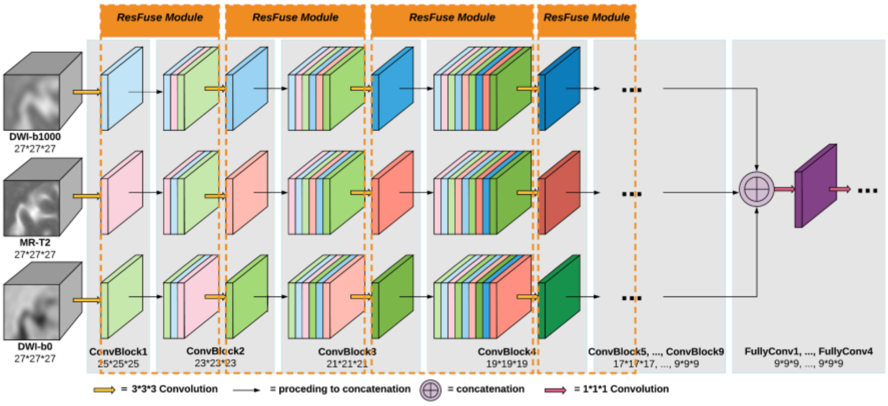

Fig. 1. The proposed HyperFusionNet architecture in the case of three imaging modalities. The feature map generated by each convolutional block is colour coded; the deeper the colour the deeper the layer. The stacked feature maps show how the dense connection and layer shuffling happens originally along each path. The ResFuse Module is added to replace the original concatenation.

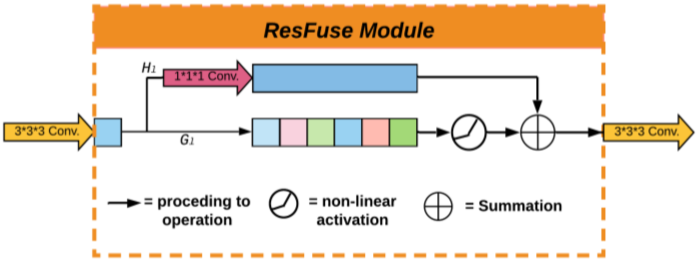

Fig. 2. Proposed residual fusion module for the multi-modal feature merging.

Table 1. The HyperFusionNet architecture detail. Notations: CB - convolutional block; RFM – residual fusion module; FC - fully convolutional layer.

Network components	No. filters	Output size	Network components	No. filters	Output size
CB1	25	25^3	RM6	819	13^3
			CB7	75	
RFM1	75	23^3	RM7	1044	11^3
CB2	33		CB8	83	
RFM2	174	21^3	RM8	1293	9^3
CB3	41		CB9	91	
RM3	297	19^3	RM9	1566	9^3
CB4	50		FC1	600	
RM4	447	17^3	FC2	300	9^3
CB5	58		FC3	150	9^3
RM5	621	15^3	FC4	No. classes	9^3
CB6	66				

The layer details are presented in Table 1, which shows the layer parameters involved in the proposed network. The overall architecture layout is presented in Figure 1, which we term HyperFusionNet.

2.3 Learning Process and Implementation Details

Another change we made to the baseline network was to the loss function. Instead of using cross entropy loss, we propose to use Combo Loss, which is the combined function of Dice Loss (DL) and Cross-Entropy (CE). The Combo Loss function allows us to benefit from DL for better handling the lightly imbalanced class and the same time leverage the advantage of CE for curve smoothing. It is defined as

$$L = \alpha\left(-\frac{1}{N}\sum_{i=1}^{N}\beta(g_i \log s_i) + (1-\beta)\left[(1-g_i)\log(1-s_i)\right]\right) - (1-\alpha)\left(\frac{2\sum_{i=1}^{N} s_i g_i + \varepsilon}{\sum_{i=1}^{N} s_i + \sum_{i=1}^{N} g_i + \varepsilon}\right) \tag{4}$$

where g_i is the ground truth for pixel i, and s_i is the corresponding predicted probability.

The model is implemented in PyTorch and trained on a single NVIDIA GTX 1080Ti GPU. Images from each modality are skull stripped and normalized by subtracting the mean value and dividing by the standard deviation. 3D image patches of size 27×27×27 are randomly extracted and only ones with lesion voxels are used for training. The Adam optimization algorithm used for optimization is set with default parameter values. The network was trained for 600 epochs. For model inference, the testing images are first normalised and non-overlapping 3D patches are extracted. The output, which is the 9×9×9 voxel-wise classification obtained from the prediction at the centre of the patch, is used to reconstruct the full image volume by reversing the extraction process. The source code for the implemented model is available on GitHub[1].

3 Experiments and Results

3.1 Datasets

The proposed HyperFusionNet is evaluated both on a hospital-collected multi-modal dataset of acute stroke lesion segmentation and on the public iSeg-17 MICCAI Grand Challenge dataset. The hospital-collected dataset was divided into 90 training cases and 30 testing cases, with three modalities in each case, i.e., T2, DWI-b1000 and DWI-b0. All images are of size 256 × 256 × 32. The ground truth for the acute stroke lesion in the dataset is annotated by experienced physicians and there are two classes involved: lesion and non-lesion. Comparably, iSEG17 is a much smaller dataset containing 10 available volumes with two modalities, i.e., T1- and T2- weighted. To be consistent with the experiment carried out in the original baseline paper [13], we also split the dataset into training, validation and testing sets, each having 6, 1, 3 subjects, respectively. There are four classes involved in iSeg-17 dataset, i.e., background, cerebrospinal fluid (CSF), grey matter (GM) and white matter (WM).

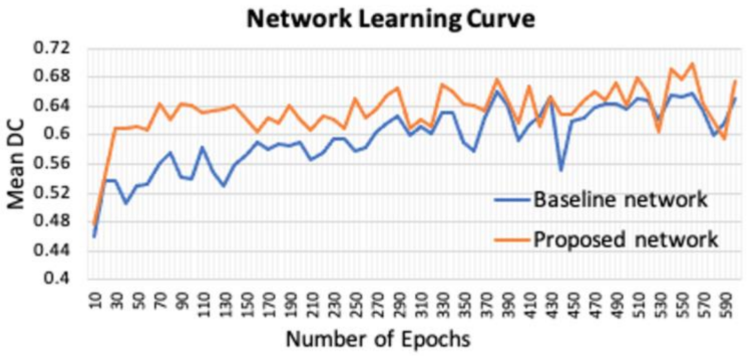

Fig. 3. Validation accuracy measured using mean DC during proposed model training on the stroke lesion dataset.

3.2 Results and Discussion

The proposed network is first evaluated by assessing its performance at segmenting acute stroke lesions in the hospital-collected dataset. In this experiment, the batch size was set to 10 and learning rate was set to 0.0002. Fig. 3 shows the comparison of the validation accuracy between the baseline and HyperFusionNet. The mean Dice score

[1] https://github.com/Norika2020/HyperFusionNet

of the validation set is calculated after every ten epochs. We can see from the learning curve that HyperFusionNet is not only more accurate compared to the baseline but also converges faster. This can be attributed to the synergy between the residual connections and the feature activation after concatenation. Table 2 shows the segmentation results on the testing volumes in metrices Dice coefficient (DC) and Hausdorff distance (HD). Both measurements suggest that the proposed network provides more effective fusion of multi-modal features than the original approach. Fig. 4 shows some examples of qualitative results on three kinds of stroke lesion conditions: a big lesion, multiple lesions and a small lesion. Overall, we observe that the proposed network is better at discarding outliers and predict stroke lesion regions of higher quality.

To better understand how the proposed modifications to the baseline contribute to the network performance, we also did an ablation study. In this experiment, the 3D networks were changed to 2D (i.e. slice-by-slice input with patch size 27×27) to save training and computation time. As shown in Table 3, the accuracy is immediately decreased when the network is changed to 2D. This is expected and it also emphasises the importance of exploiting the slice dimension information for such networks. The results show the clear improvements made by each modification to the 2D baseline network, with ResFuse module making the biggest contribution. We also tested other loss functions – Dice Loss, Focal Loss and Tversky Loss. Comparably, Combo Loss has shown to be more advantageous in our proposed network.

Table 2. The testing results on stroke lesion segmentation measured in DC (%) and HD with their associated standard deviation for the experimented networks.

Network	Mean DC	DC Std	Mean HD	HD Std
Baseline	65.6	18.0	87.756	20.386
HyperFusionNet	67.7	16.5	85.462	14.496

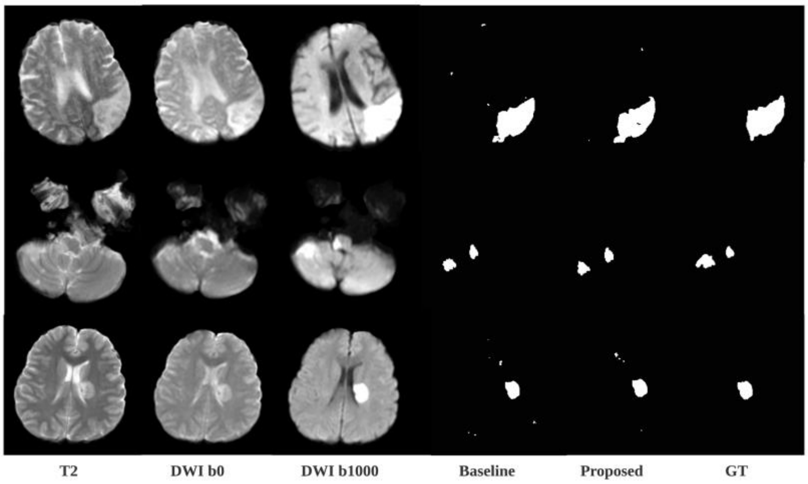

Fig. 4. Qualitative results obtained for the stroke dataset using the baseline and the proposed networks.

Table 3. The testing results of the proposed modification to the baseline on stroke lesion segmentation measured in DC (%).

Network modification	DC	Other loss function	DC
Baseline 2D (CE loss)	41.9	HyperFusionNet (CE loss)	46.6
+ Incremental filters	43.0	+ Dice loss	45.9
+ ResFuse module	46.6	+ Focal loss	39.0
+ Combo loss	**47.1**	+ Tversky loss	43.6

Table 4. The performance comparison on the testing set of the iSeg17 brain segmentation measured in DC (%).

Architecture	CSF	WM	GM
Baseline	93.4 ± 2.9	89.6 ± 3.5	87.4 ± 2.7
HyperFusionNet	93.6 ± 2.5	90.2 ± 2.2	87.8 ± 2.3

We also tested the HyperFusionNet on the iSeg-17 dataset to investigate its performance on a smaller dataset with more classes involved. To allow a fair comparison, the parameters such as batch size (=5) and learning rate (initially =0.001 and reducing by a factor of 2 every 100 epochs) are set to match the baseline paper. The results for the baseline are reproduced using their published code written in PyTorch[2] in order to compare results under the same experimental setting. Results in Table 4 shows the proposed network yields better segmentation results than the baseline. Although there is not a significant improvement in the averaged Dice score, we observed that it worked well for challenging cases of segmenting GM and WM. Fig. 5 depicts such a challenging example where the proposed HyperFusionNet shows a better contour recovery than that obtained by the baseline.

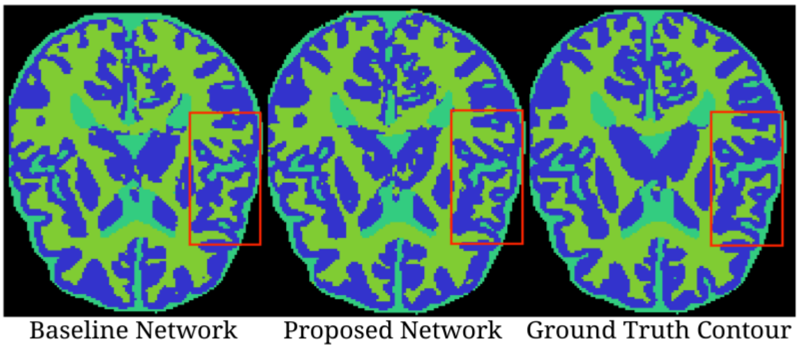

Fig. 5. Qualitative results achieved by the baseline and proposed network compared to the ground truth contour.

[2] https://github.com/josedolz/HyperDenseNet_pytorch

4 Conclusion

In this work, we propose a novel method HyperFusionNet for brain segmentation using 3D images captured with multiple modalities. The proposed network presents a new way to fuse features from different modalities in a densely connected architecture. A progressive feature abstraction process is promoted and a ResFuse module is introduced to replace the simple concatenated fusion used in the baseline network. The network is improved further with a Combo loss function. We evaluate the proposed network in both ischemic acute lesion segmentation and infant brain segmentation and compare it to a state-of-art multi modal fusion network. The experimental results demonstrate the effectiveness of HyperFusionNet and its capability to tackle challenging multi-modal segmentation tasks with different applications and dataset sizes. Our research largely focused on the fusion network itself, and little data augmentation and post processing was included. For future work, we will improve the network further by implementing pre and post enhancements. The influence of each modality on different applications will also be investigated.

References

1. Tongxue, Z., Su, R., Stéphane, C.: A review: Deep learning for medical image segmentation using multi-modality fusion. Array 3–4, (2019).
2. Lapuyade-Lahorgue, J., Xue, J.H., Ruan, S.: Segmenting multi-source images using hidden markov fields with copula-based multivariate statistical distributions. IEEE Transaction on Image Processing 26(7), 3187–3195 (2017).
3. Balasubramaniam, P., Ananthi, N.: Image fusion using intuitionistic fuzzy sets. Information Fusion 20(1), 21-30 (2014).
4. Guo, Z., Li, X., Huang, H., Guo, N., Li, Q.: Deep learning-based image segmentation on multimodal medical imaging. IEEE Transactions on Radiation and Plasma Medical Sciences 3(2), 162-169 (2019).
5. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, D., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural networks. Medical Image Analysis 35, 18–31 (2017).
6. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3d CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis 36, 61– 78 (2017).
7. Lavdas, I., Glocker, B., Kamnitsas, K., Rueckert, D., Mair, H., Sandhu, A., Taylor, S.A., Aboagye, E.O., Rockall, A.G.: Fully automatic, multiorgan segmentation in normal whole body magnetic resonance imaging (MRI), using classification forests (CFs), convolutional neural networks (CNNs), and a multi-atlas (MA) approach. Medical Physics 44(10),5210-5220 (2017).
8. Cai, H., Verma, R., Ou, Y., Lee, S., Melhem, E.R., Davatzikos, C.: Probabilistic segmentation of brain tumors based on multi-modality magnetic resonance images. In 4th IEEE International Symposium on Biomedical Imaging, pp. 600–603, (2007).
9. Klein, S., van der Heide, U.A., Lips, I.M., van Vulpen, M., Staring, M., Pluim, J.P.: Automatic segmentation of the prostate in 3d mr images by atlas matching using localized mutual information. Medical Physics 35(4), 1407–1417 (2008).

10. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark. IEEE Transactions on Medical Imaging 34(10), 1993–2024 (2015).
11. Aygun, M., Sahin, Y.H., Unal, G.: Multimodal convolutional neural networks for brain tumor segmentation. arXiv preprint:1809.06191 (2018).
12. Chen, Y., Chen, J., Wei, D., Li, Y., Zheng, Y.: Octopusnet: a deep learning segmentation network for multi-modal medical images. In Multiscale Multimodal Medical Imaging (MMMI 2019), Lecture Notes in Computer Science, vol. 11977, (2019).
13. Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ben Ayed, I.: HyperDense-Net: A hyper-densely connected cnn for multi-modal image segmentation. IEEE Transactions on Medical Imaging 38(5), 1116–1126 (2019).
14. Tseng, K.L., Lin, Y.L., Hsu, W., Huang, C.Y.: Joint sequence learning and cross-modality convolution for 3d biomedical segmentation. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2017), pp. 3739–3746, (2017).
15. Dou, Q., Liu, Q., Heng, P.A., Glocker, B.: Unpaired Multi-Modal Segmentation via Knowledge Distillation. IEEE Transaction on Medical Imaging 39(7), 2415-2425 (2020).
16. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2017), pp. 2261–2269, (2017).
17. Dolz, J., Ben Ayed, I., Desrosiers, C.: Dense multi-path u-net for ischemic stroke lesion segmentation in multiple image modalities. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2018. Lecture Notes in Computer Science, vol. 11383, (2019).
18. Wang, L., et al.: Benchmark on automatic 6-month-old infant brain segmentation algorithms: the iSeg-2017 challenge. IEEE Transactions on Medical Imaging 38(9), 2219-2230 (2019).
19. Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C.: The importance of skip connections in biomedical image segmentation. In Deep Learning and Data Labeling for Medical Applications, pp. 179–187 (2016).
20. Ibtehaz, N., Sohel Rahman, M.: MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Networks 121, 74-87 (2020).