3DYNET: A DENSELY INTERCONNECTED NETWORK FOR VOLUMETRIC LIVER AND TUMORS SEGMENTATION

Gabriella d'Albenzio[1,*] *Yuliia Kamkova*[2,3,*] *Rabia Naseem*[4] *Mohib Ullah*[5]
Stefania Colonnese[6] *Rahul Prasanna Kumar*[1] *Faouzi Alaya Cheikh* [5]

[1] The Intervention Center, Oslo University Hospital, Norway
[2] Department of Research and Development, Oslo University Hospital, Oslo, Norway
[3]Department of Informatics, The University of Oslo, Oslo, Norway
[4] COMSATS University Islamabad, Pakistan
[5] Norwegian University of Science and Technology, Norway
[6] Sapienza University of Rome, Italy

ABSTRACT

Accurate segmentation of liver and tumours from CT volumes is crucial in hepatocellular carcinoma diagnosis and preoperative resection planning, as clinical decisions are made based on its output. Despite deep learning-based liver and tumor segmentation of abdomen CTs has become a robust tool in the overall resection plan, fully-automated segmentation of the liver and its lesion remains an open problem due to class imbalance and variation in their structure. In this work, we proposed an encoder-decoder architecture, 3DYNet, for liver and tumor segmentation. 3DYNet incorporates long skip-connections between each encoder branch and the decoder for effective use of low- and high-level features. We validated our newly proposed network by conducting experiments on LiTS dataset. The evaluation results demonstrate that 3DYNet improves the overall segmentation and outperforms the well-known 3DU-Net model.

Index Terms— CNN, CT, Multi-class Segmentation, HCC, Liver

1. INTRODUCTION

Liver cancer is the second leading cause of cancer death among men [1]. However, early-stage diagnosis and treatment can significantly reduce mortality rates. Segmentation is a prerequisite in a typical workflow of computer-assisted pre-operative liver surgical planning [2]. Despite several robust deep-learning methods have been proposed for medial image segmentation; however, fully-automated segmentation of the liver and its lesion remains an ill-posed problem [3], [4]. Some challenges associated with segmenting these structures include low contrast of CT images, diverse tumor shapes and high-class imbalance problem implying that the lesions are several times smaller than the organ. State-of-the-art segmentation approaches such as UNet [5] often fail in segmenting such imbalanced data. In this work, we propose an end-to-end trainable autoencoder network 3DYNet, which takes advantage of the contextual volumetric information in 3D medical imaging data combined with long skip connections to segment liver and lesions in CT scans. Our main contributions in this work can be summarized as follows:

1. We propose 3DYNet, an end-to-end autoencoder based architecture, with two encoder branches for automated liver and tumour segmentations in CT scans.
2. Long skip-connections between each encoders and the decoder are incorporated in this approach; we hypothesized that long-range skip-connections will further improve gradient flow and features reusability.
3. The proposed 3DYNet outperforms the well-known state-of-the art architecture 3D U-Net [6] in liver segmentation as validated by our extensive experimentations.

Several deep learning approaches have been presented to segment liver and lesions in CT volumes. Conventionally, Convolutional Neural Networks (CNN) and the Fully CNN (FCNs) show promising performance on segmentation tasks. U-Net, for instance remains one of the most well-known approaches applied to biomedical image segmentation. 2D CNNs/FCNs were presented initially for medical image segmentation followed by 3D CNNs that take advantage of contextual information present in the 2D images and replace 2D convolution operations with 3D convolutions. Residual skip connections were integrated with U-net to get rid of the vanishing gradient problem associated with deep neural networks. For instance, U-net3+ [7] used a full-scale skip connection; however, it sometimes fails in segmenting small objects in the presence of constrained training data. UNet# [8] aggregated dense scale and full-scale connections to learn precise object boundaries. Another end-to-end 3D medical image segmentation method 'V-net' was proposed which

[*] Authors with equal contribution, names are listed in alphabetical order

was an FCN embedded with residual connections [9]. V-net also introduced a loss function to handle the class imbalance problem better than classic weighted cross-entropy loss. Similarly, 3DUNet embedded with a skip connection was introduced by Cicek et al.[6]. 3DUNet extends the original U-net by replacing 2D operations with the corresponding 3D operations and it is a semi-supervised approach segmenting 3D volume from sparse annotation. Besides, it bypasses bottlenecks in the architecture and uses batch normalization to achieve rapid convergence. A hybrid encoder-decoder network for localizing polyps in colonoscopic images was proposed by Ahmed et al. [10]. This method fuses pretrained and untrained encoder and a decoder trained from scratch with sum-skip-concatenation connections to tackle the limited amount of labelled training data. A recently proposed end-to-end autoencoder-based architecture Y-Net [11] combines frequency and time domain features to segment retinal optical coherence tomography images. Regular convolution layers cannot efficiently extract the global patterns in the images unlike fast Fourier convolutions. The method performs very well in fluid segmentation, exhibiting 13% dice score improvement compared to U-net.

2. PROPOSED METHOD

In this section, an overview of the core components of our network is presented. Our 3DYNet network predicts the segmentation map $\hat{y}$ given an input CT volume $x \in \mathbb{R}^{\mathbb{D}\times\mathbb{H}\times\mathbb{W}\times\nVdash}$, and its corresponding multi-class segmentation label $y \in \mathbb{Z}^{\mathbb{D}\times\mathbb{H}\times\mathbb{W}\times\mathbb{C}}$, where D, H, W are the dimensions of the voxels and C is the number of classes (0-background, 1-liver, 2-tumour respectively) The architecture of our 3DYNet model is shown in Fig (). The network consists of two contractions (encoder) and an expanding (decoder) path. The decoder receives the combined features from the two encoders and generates the segmentation map $\hat{Y}$.

Encoders: The two encoders share the same architecture, based on a modified 3D version of VGG19 [12], without the three fully connected layers. The first two layers consist of two $3 \times 3 \times 3$ convolutions with stride and pad of 1, each followed by a rectified linear unit (ReLu) and a $2 \times 2 \times 2$ max pooling operation with strides of two in each dimension. The last 3 layers follow the same hierarchy as the previous two, but 4 convolutions are applied this time. Moreover, the designed skip connections enable flexible feature fusion between the two VGG19 decoders, after each max pooling operation, allowing the network to encode more features.

Decoder: Our network's decoder is similar to the original 3DU-Net [6] with four convolutional blocks. The feature maps are upsampled by using a $3 \times 3 \times 3$ transposed convolution operation with strides of 2 in each dimension, followed by three $3 \times 3 \times 3$ convolutions, each followed by an LReLU. The skip connections encoders-decoder from layers of equal resolution allow the network to recover the spatial information lost by pooling operations. Finally, a $1\times1\times1$ convolution operation generates the final segmentation map.

Long-range skip connections: To improve the interconnectivity between the two encoders and the decoder, we designed long-range skip connections similar to UNet. Our skip connections allow the network to encode additional features simultaneously. This was done by concatenating the feature maps of the two encoders before each max pooling operation onto the upsampling function of expanding path. Moreover, the two feature maps obtained for each resolution layer in the decoder are again concatenated together to enable flexible feature fusion, as illustrated in Fig.1.

2.1. Multi-class Loss Function

Our networks are trained with a combination of Soft Dice Loss and Cross-Entropy loss. In the following, we merely report the definition. However, a detailed description of both loss functions can be found here [13]. Let $\hat{Y}$ be the reference foreground segmentation (ground truth) with voxel values $\{\hat{y}\}_n = 1, ..., N$, and P the predicted probabilistic map for the foreground label over N image elements $\{p_n\}_n = 1, ..., N$, with the background class probability being 1-P. The ϵ provides numerical stability to prevent division by zero, and $\{c_n\}_n = 1, ..., C$ indicates the class label. The Soft Dice Loss is defined as:

$$\mathcal{L}_{DL} = 1 - \frac{\sum_{n=1}^{N}\sum_{c=1}^{C} p_n^c \hat{y}_n^c + \epsilon}{\sum_{n=1}^{N}\sum_{c=1}^{C} p_n^c + \hat{y}_n^c + \epsilon} \tag{1}$$

and the Cross-Entropy Loss is as follow:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} r_n \lg(p_n^c) + (1 - \hat{y}_n^c) log(1 - p_n) \tag{2}$$

Then, the combined loss function is equivalent to:

$$\mathcal{L}_T = \mathcal{L}_{DL} + \mathcal{L}_{CE}; \tag{3}$$

3. EXPERIMENTS AND ANALYSIS

3.1. Dataset

We tested our network on the competitive LiTS dataset of MICCAI 2017 Liver Tumor Segmentation Challenge [14]. The LiTS dataset consists of 200 contrast-enhanced 3D abdominal CT scans (131 provided with labels for training and 70 for the test set) from several clinical sites with different scanners and protocols, leading to extensively varying spatial resolution and field-of-view. For our experiments, all images are resized to 256x256x64 due to the GPU limitation. Moreover, since the focus of this task is the liver and lesions, the Hounsfield unit (HU) values were windowed in the range [-175], [250] to exclude artifacts and irrelevant organs and tissues. For intensifying the network generalization, data augmentation is done "on-the-fly" during training, including a

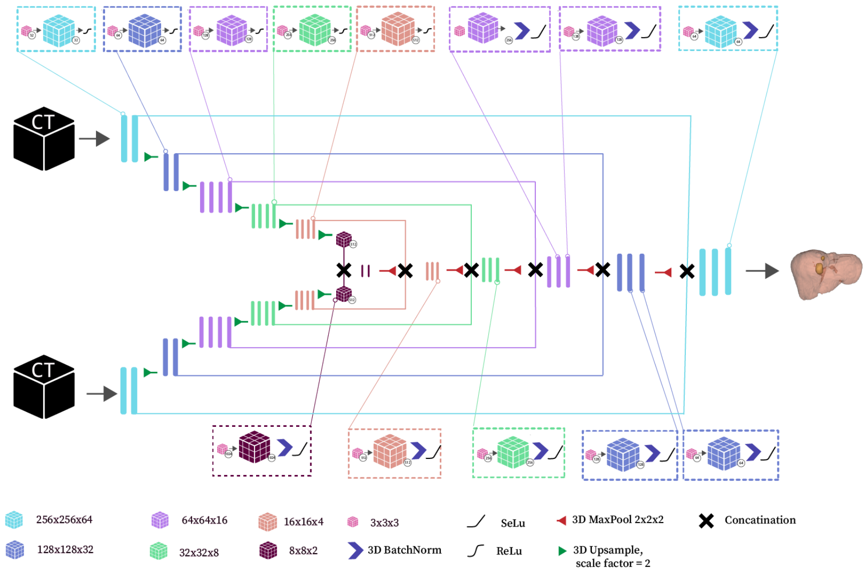

Fig. 1. The proposed by the authors 3DYNet architecture. Our backbone consists of a two encoders, whose structure is based on a modified versionof VGG19 [12] and one decoder similare to 3DU-Net [5]. The lines shows the skip connections in our backbone, where each color represent a different resolution layer.

series of geometric transformations, such as random flipping, shifting, scaling, and resampling.

3.2. Implementation Details

The implementation of our network is based on PyTorch 1.12.0 [15]. The network is trained for 500 epochs on an NVIDIA GeForce RTX 3090 graphic memory with a batch size of 1. The best validation accuracy for all models was used to determine the number of training epochs. The LiTS dataset was split into 100 volumes for training, 15 for the validation phase, and 15 for testing. The normal distribution initializer [16] is employed for initializing the weights since its robustness in considering the rectifier nonlinearities. For training our deep neural network, the learning rate adaptive optimizer ADAM [17], was used. ADAM optimizer dominates the field of deep learning due to its fast convergence. We set the initial learning rate to 0.0001, decaying the learning rate with a cosine annealing for each batch as proposed in [18]. For evaluating the liver and tumor segmentation performance, we used 4 metrics: Sensitivity, Specificity, Precision and Dice Score (DS).

3.3. Results

We compared our model with two widely recognized networks for medical segmentation, which are 3DU-Net and V-Net, on the LiTS datasets. In addition, we also implemented another variant of our 3DYNet, which we called 3DYNet-EE. In this version, the feature maps of each encoder are first summed and then concatenated into the decoder layers of equal resolution, similar to [10]. The intuition behind this is that the features learned by each block of the first encoder can complement the second encoder. The four contending methods were trained under identical settings for a fair comparison. Quantitative comparison in Table 1 shows that both 3DYNet and 3DYNet-EE outperform 3DU-Net and V-Net in liver and tumour segmentation. The Dice score and Sensitivity achieved by 3DYNet for the liver class were 0.889 and 0.933, respectively. For the tumor class, the precision score improved from 0.370 to 0.483. Compared with the widely used 3D U-Net, our network improved the Dice score by around 14% for liver class and 28% for tumor class. Due to the high-class imbalance of the LiTS dataset, the performance of all four models in segmenting the lesions is less accurate compared to the liver class. Moreover, we also report that 3D U-Net does not perform well in tumor segmentation tasks, where it fails to segment the lesions in the test set. In Fig. 2, we show some qualitative results to compare segmentation performance of our model to 3DU-Net, V-Net, 3DYNet-EE.

4. DISCUSSION & CONCLUSION

Automatic liver and tumor segmentation plays a critical role in clinical planning. It can reduce the time clinicians spend on this task while assisting them in the diagnosis process. In this work, we present an end-to-end encoder-decoder architecture for segmenting the liver and its lesions in CT images. Our proposed network, 3DYNet extracts spatial features using two encoder branches, propagating them onto one decoder, which

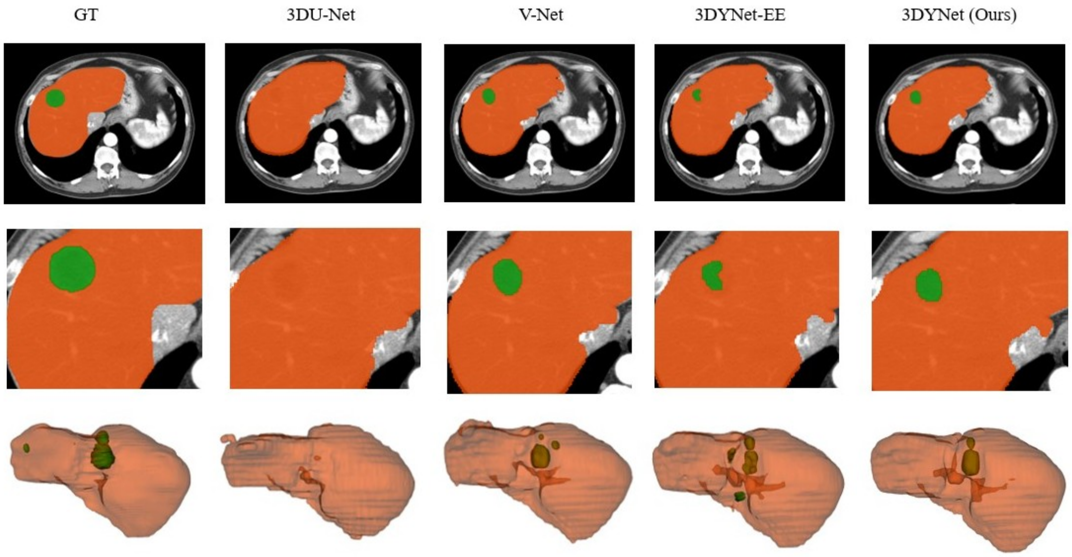

Fig. 2. Visual comparison of the competing networks for liver and tumors segmentation on LiTS dataset: the firt two rows represents the axial view of CT images of a patient from LiTS training set. The third row shows the 3D model generated from the obtained liver and tumors segmentation. GT: ground truth. Orange: liver segmentation. Green: tumors.

Network	Class	Sensitivity	Specificity	Precision	Dice
3D V-Net	Liver	0.920∓0.071	0.995∓0.001	0.855∓0.046	0.884∓0.035
	Tumors	0.200∓0.304	0.999∓0.001	0.370∓0.354	0.207∓0.259
3D U-Net	Liver	0.699∓0.266	0.997∓0.003	0.893∓0.039	0.749∓0.194
	Tumors	0	0	0	0
3DYNet-EE	Liver	0.921∓0.093	0.995∓0.002	0.848∓0.056	0.878∓0.043
	Tumors	0.262∓0.297	0.999∓0.001	0.394∓0.356	0.240∓0.193
3DYNet (Ours)	Liver	**0.933∓0.075**	0.996∓0.001	0.856∓0.052	**0.889∓0.037**
	Tumors	0.276∓0.309	**0.999∓0.009**	0.483∓0.383	0.289∓0.283

Table 1. Quantitative results: mean and standard deviation of sensitivity, specificity, precision and dice score per each class.

generates a 3D multi-class segmentation map. Moreover, our 3DYNet, takes advantage of accurately designed skip connections between the encoder and the decoder for effective use of low- and high-level features. Our results demonstrate that these long-range skip-connections via concatenation improve the interconnectivity between the final feature maps for each resolution level and allow efficient feature reusability, which can tackle class imbalance better than encoder-encoder connections when compared with 3DYNet-EE LiTS dataset is indeed affected by high-class imbalance, meaning that some of those classes we seek to identify and label appear significantly less frequently than other classes represented in the dataset. Rare classes could end up being bypassed because they are underrepresented during training, as reported in Table 1 for 3DU-Net. The proposed model yielded the highest Dice score compared to previous state-of-the-art approaches. However, 3DYNet suffers from high computational cost and GPU memory constraints due to employing VGG19-based architecture in the contracting phase. In the future, it is of interest to explore different network backbones with a reduced number of hyper-parameters and test our network on more medical image segmentation datasets.

5. ACKNOWLEDGMENTS

This work was supported by the grants from the Research Grants Council of ...Other types of acknowledgements can also be listed in this section.

6. REFERENCES

[1] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal, “Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] Tobias Blum, Hubertus Feußner, and Nassir Navab, “Modeling and segmentation of surgical workflow from laparoscopic video,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2010, pp. 400–407.

[3] Patrick Ferdinand Christ, Mohamed Ezzeldin A Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D’Anastasi, et al., “Automatic liver

and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 415–423.

[4] Fang Lu, Fa Wu, Peijun Hu, Zhiyi Peng, and Dexing Kong, "Automatic 3d liver location and segmentation via convolutional neural network and graph cut," *International journal of computer assisted radiology and surgery*, vol. 12, no. 2, pp. 171–182, 2017.

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

[6] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[7] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1055–1059.

[8] Ledan Qian, Xiao Zhou, Yi Li, and Zhongyi Hu, "Unet#: A unet-like redesigning skip connections for medical image segmentation," *arXiv preprint arXiv:2205.11759*, 2022.

[9] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[10] Ahmed Mohammed, Sule Yildirim, Ivar Farup, Marius Pedersen, and Øistein Hovde, "Y-net: A deep convolutional neural network for polyp detection," *arXiv preprint arXiv:1806.01907*, 2018.

[11] Azade Farshad, Yousef Yeganeh, Peter Gehlbach, and Nassir Navab, "Y-net: A spatiospectral dual-encoder network for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 582–592.

[12] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L Martel, "Loss odyssey in medical image segmentation," *Medical Image Analysis*, vol. 71, pp. 102035, 2021.

[14] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al., "The liver tumor segmentation benchmark (lits)," *arXiv preprint arXiv:1901.04056*, 2019.

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[17] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] Ilya Loshchilov and Frank Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.