Multi-Modal Multi-Stream UNET Model for Liver Segmentation

Hagar Louye Elghazy
computer science
Arab Academy for science ,technology and maritime transport,
college of computing and information technology
Cairo ,Egypt
hagargazy@aast.edu

Mohamed Waleed Fakhr
computer engineering
Arab Academy for science ,technology and maritime transport,
college of enginering and technology
Cairo ,Egypt
waleedf@aast.edu

***Abstract*— Computer segmentation of abdominal organs using CT and MRI images can benefit diagnosis, treatment, and workload management. In recent years, UNETs have been widely used in medical image segmentation for their precise accuracy. Most of the UNETs current solutions rely on the use of single data modality. Recently, it has been shown that learning from more than one modality at a time can significantly enhance the segmentation accuracy, however most of available multi-modal datasets are not large enough for training complex architectures. In this paper, we worked on a small dataset and proposed a multi-modal dual-stream UNET architecture that learns from unpaired MRI and CT image modalities to improve the segmentation accuracy on each individual one. We tested the practicality of the proposed architecture on Task 1 of the CHAOS segmentation challenge. Results showed that multi-modal/multi-stream learning improved accuracy over single modality learning and that using UNET in the dual stream was superior than using a standard FCN. A "Dice" score of 96.78 was achieved on CT images. To the best of our knowledge, this is one of the highest reported scores yet.**

Keywords—medical images, UNET, dual stream, segmentation.

I. Introduction

Semantic segmentation is a computer vision problem [1], where each pixel in an image is labeled. In recent years, semantic segmentation for medical images has become very popular due to providing more accurate and objective diagnoses [2], which in turn leads to more efficient treatments and better pre-operative planning. In addition, being an automated solution, segmentation could greatly impact the field of imaging-based screening through reducing the workload and screening time considerably [3]. Examples of widely used imaging modalities include ultrasonography (US), computed tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET) [4]. These give clear images of the various organs such as the liver, spleen, and kidneys, among others. However, since each of these devices uses a different imaging technique, the visual output produced from each modality is completely different. Acquiring sufficient datasets for machine learning from only one of these modalities is very challenging. Therefore, segmentation using multiple imaging modalities (two or more imaging modalities) was introduced with a potential to provide better results and more accurate segmentation compared to single modalities [5]. Two examples of multiple imaging modalities [6] are image fusion and dual stream architectures, which will be discussed in the current work.

After obtaining data from different medical imaging modalities, training through deep learning models is used to segment these data. Utilizing deep convolutional neural network (CNN) architectures [7] for semantic segmentation has shown superior segmentation performances compared to traditional techniques, especially in medical imaging. One of the most widely used techniques of CNN is fully convolutional networks (FCN) [8]; a basic deep semantic segmentation architecture which inspired most of the subsequent deep semantic segmentation techniques and later UNET [9]. Hence, UNET is a refined architecture based on FCN that provides the best semantic segmentation performance so far in many domains and applications, including single and multiple medical image segmentation.

Availability of imaging data with reasonable sizes is a challenge in itself; most of previously built FCN models need a large amount of data to be able to produce competing results in segmentation [10]. Most of medical imaging data are relatively small, but can include images from multiple modalities [11], for example both MRI and CT images, with each having different features. Therefore, there exists a need for developing computational tools that can provide accurate semantic segmentation for small data as well as being able to combine the various features produced from multiple imaging modalities.

978-1-6654-3568-0/21/$31.00 ©2021 IEEE

Our main aim in this paper is to capitalize on the superior performance of the UNET in the medical semantic segmentation. this paper proposes a multi-modal, multi-stream architecture which employs the UNET as the segmentation core in each stream. This will allow achieving more accurate segmentation through exploiting the advantages gained from multiple modalities and multi-stream architecture.

We also show that using dual stream has improved the segmentation of each modality compared to a separate UNET for each modality .

II. REALTED WORK

A main approach for semantic segmentation is using encoder decoder architectures such as FCN [8], UNET [9], SEGNET [12] and Mask-R-CNN [13].

A. Architechtures of medical image segmentation

It is definite that CNNs are doing great progress in this field, leading to outstanding performances in many medical problems [14]. Most available medical image segmentation architectures and applications are built by fully convolutional neural network FCN [8] or UNET [9] .

In FCN [8], a dense pixel segmentation is achieved. To recover the original resolution of the input image, the prediction is up-sampled. To improve prediction abilities, skip connections are incorporated to recover some of the lost spatial data [15].

UNET [9], on the other hand, contains two paths. The first is the contraction path (also called the encoder), which is used to capture the context in the image. The encoder is a stack of convolutional layers and max pooling layers. The second path is the symmetric expanding path (also called the decoder) which is used to allow exact localization using transposed convolution layers. Skip connections or transfer layers are used to concatenate the features from contraction and expansion path layers. This allows for retrieval of lost features during the encoder path thus maintaining localization [16].

SEGNET [12] contains both encoder and decoder networks, but with no fully connected layers; only convolution layers. The transferred pool indices output from each layer in the encoder network is inputted to the corresponding one in the decoder network.

Mask-R-CNN [13] consists of two stages, stage 1 applies a suggestion to where an object in a picture might be, then in stage 2, it predicts the classes of objects defined in stage 1, refines the boundaries of the boxes around predicted objects and finally generates a pixel level mask of the object predicted in the first stage.

Mali proposed a single stream 3D UNET model [17], using the Combined Healthy Abdominal Organ Segmentation (CHAOS) challenge dataset [18]. His approach involved two phases, first, pre-training the model using unsupervised data and second, training/fine tuning using supervised data. Both pre-training and training were applied on three models: CT data only, MRI data only and finally using combined CT and MRI data. Rather than using CT and/or MRI slices, he used patched 3D images to create the unsupervised dataset for the pretraining.

B. Dual streams architectures

Dual streams architectures are used when one or more modalities are represented in the medical dataset. They are particularly useful if the datasets available are not large. They are applied on all single stream architectures like those previously mentioned, and they are mostly created by duplicating the architecture and adding some connecting layers in the middle for feature sharing between two streams. Therefore, dual streams benefit from different modalities in order to improve segmentation accuracy [19].

Another previous study by Valindria et al used an encoder-decoder FCN architecture with residual layers for multi-organ image segmentation [20]. To effectively merge the multi-modal features from CT and MRI, they proposed a dual-stream network architecture and used individual streams for each modality. However, FCN loses a lot of information in the decoding phase, which is a drawback in case of medical imaging segmentation where every piece of information is crucial. Four versions (v1, v2, v3, v4) for the segmentation process were used by this study [20].

- V1: Two encoders, one for each modality and one decoder.
- V2: both modalities share same encoder and decoder, however, a separate stream layer (one for each modality) is applied before the encoder.
- V3: One encoder where both modalities are inputted in the same stream, and two decoders.
- V4: separate encoder and decoder for each modality, but they were connected in the middle to share weight between encoders and decoders.

According to the results reported by Valindria et al [20], the 2-encoder and 2-decoder architecture V4 has outperformed the other versions. Thus, this architecture was adopted in our study, however, UNET was used instead of FCN to capitalize on its superior performance and efficient utilization of limited training data.

III. METHODOLGY

In the current study, the proposed model is a dual stream UNET that accepts inputs from different modalities , in this case MRI and CT scans. The model consists of two encoders and two decoders, merged with a convolution layer before the decoder section as shown in Fig. 1.We built our model based on UNET architecture because of its ability to keep the localization information through the transfer layers, as opposed to FCN which lacks transfer layers. Since the dataset was in DICOM form, the Pydicom [21] library was used to convert the input images to PNG.

In our architecture there were two input layers that take RGB images (256*256*3), Keras [22] layers were used to build the model, it contains two identical streams, each comprising an encoder and a decoder path. The two encoders were merged by a convolutional layer.

Each encoder path consisted of 5 convolutional layers, each layer is followed by a max pooling layer and a merging layer. The decoder constitutes five convolutional layers, each layer followed by a convolutional transpose layer. Each convolutional – convolutional transpose pair was concatenated with correspondent layers from the encoder path. All activation functions used were Relu (rectified linear unit).

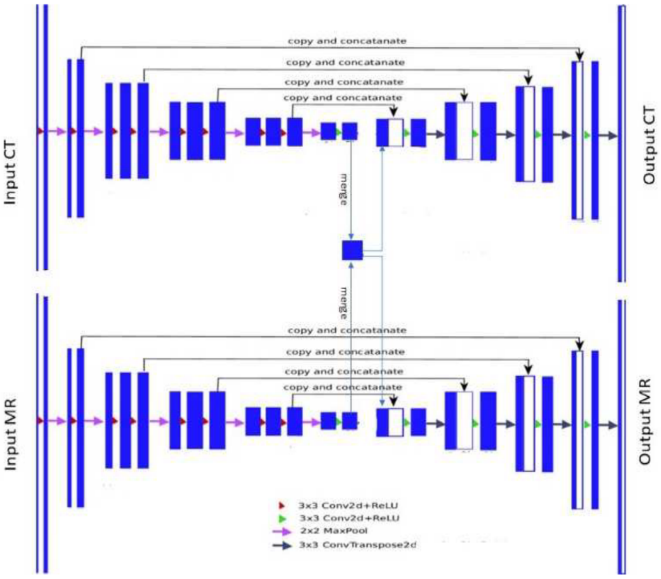


Fig. 1. Proposed dual stream UNET architecture

IV. Experimental setup

In this section, the dataset used and the proposed model will be discussed in details.

A. Dataset and pre-processing

We used the CHAOS challenge dataset [18] [14] [15]. The aim of the challenge is to perform segmentations on abdominal organs (liver, kidney, and spleen) form CT and MRI images. The challenge contains five tasks from which participants can choose.

- Task1: Liver Segmentation (CT & MRI)
- Task2: Liver Segmentation (CT only)
- Task3: Liver Segmentation (MRI only)
- Task4: Segmentation of abdominal organs (CT & MRI)
- Task5: Segmentation of abdominal organs (MRI only)

In this project, Task1 was selected; segmentation of liver using two modalities CT and MRI. The training dataset consisted of CT DICOM images of 40 patients and MRI DICOM images of 120 patients, while the MRI images are divided into two different sequences, T1-DUAL (in phase: 40 datasets and out phase: 40 datasets) and T2-SPIR (40 datasets).

The testing dataset provided contained the same amount of DICOM images as the training data set but without ground truth, therefore we were not able to utilize it in this project. Instead, the training dataset was divided into two folders, 70% for training and 30% for testing. The dividing process was carried out randomly and repeated three times, creating three training folders and three testing folders.

After conversion of the images to PNG format, the annotations for the ground truths were removed from all organs except the liver (coloured as white).

B. Training

The model was implemented using the basic UNET architecture, but with two encoders and two decoders joined with a convolution layer at the middle.

The model was trained using Adam optimizer with mini batch of size =1. The number of filters for each layer of the encoder was 32, 64, 128 and 256, while for the decoder, it was 256, 128, 64 and 32. Loss was calculated [24] using binary cross entropy, dice coefficient [25] and accuracy were used as metrics for evaluation, and finally the model was trained for 30 epochs to achieve best results. Each train batch had a validation split of 20% of unseen data.

For evaluating the segmented data output, dice coefficient was used, which represents the overlap between two masks where score 1 is perfect match or exact overlap (masks identical) and score 0 means no overlap whatsoever [25].

$$DICE = \frac{2 \times TP}{(TP + FP) + (TP + FN)}$$

V. Results and discussion

Using Multimodal learning can use the information from both modalities in one pass, unlike normal single modal/stream models that learn sequentially from each modality. In the current study, we demonstrate that multi stream/modality architecture gives accurate results than traditional architectures.

As mentioned before, our data was randomized into 3 training folders and 3 corresponding test folders, each training folder comprised 70% of the whole data and each test folder was the remaining 30%. Therefore, the final dataset included 3 training folders (train1, train2, train3) and 3 test folders (test1, test2, test3).

VI. TABLES AND FIGUERS

Table 1 training results on 3 training batches

Point of comparison	Train 1		Train 2		Train3	
Modality type	CT	MR	CT	MR	CT	MR
Dice	0.967	0.795	0.967	0.799	0.956	0.769
Accuracy	0.997	0.996	0.997	0.996	0.997	0.997
loss	0.005	0.008	0.005	0.009	0.005	0.009

Table 1 shows the training on the 3 different batches of training data, all training parameters were the same for the 3 runs as previously mentioned. CT was found to perform significantly better than MR.

Table 2 dice results on 3 testing batches

Point of comparison	Test1		Test2		Test3	
Modality type	CT	MR	CT	MR	CT	MR
Dice	0.952	0.687	0.961	0.701	0.976	0.721

Table 2 demonstrates the performance of the 3 testing unseen images when attempting segmentation. CT achieved >0.95 in dice while MR didn't perform as well, due to using different modalities of MR on the same stream. However, this helped the CT to get higher results because of the dual architecture used and the merging between the layers of both streams. Table 3 presents a clear comparison between the same UNET layers in dual stream vs single stream with CT and MR as inputs.

Table 3 comparison between dual stream vs. single stream UNET

Point of comparison Dice	CT segmented images	MR segmented images
Our Proposed multi stream/modal UNET	0.963	0.703
Single stream UNET	0.893	0.344

Table 4 dice results comparison between proposed UNET and related work

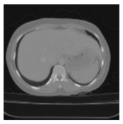
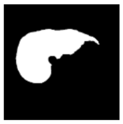
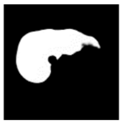
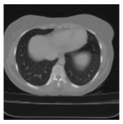
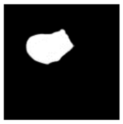
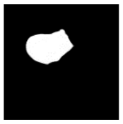
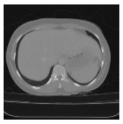
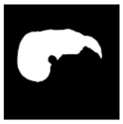
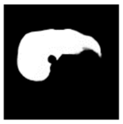
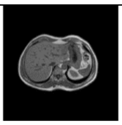
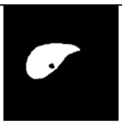
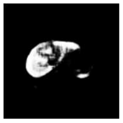
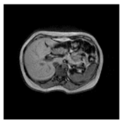
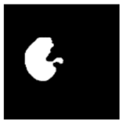
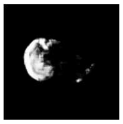
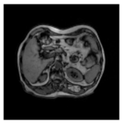
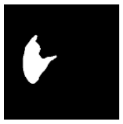
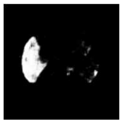
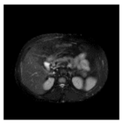
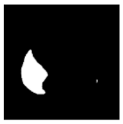
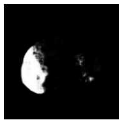
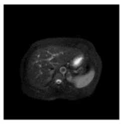
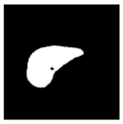
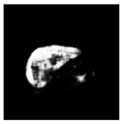
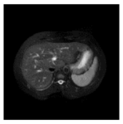
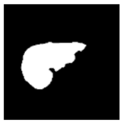
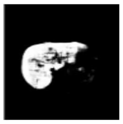
Point of comparison on Average Dice	CT segmented images	MR segmented images
Our Proposed multi stream/modal UNET	**0.963**	**0.703**
Shruti Atul's 3D UNET CT [17]	0.946	0.478
Shruti Atul's 3D UNET MRI [17]	0.947	0.493
Shruti Atul's 3D UNET COMBO [17]	0.946	0.510
Vanya's multi stream/modal FCN [20]	0.919	**0.914**

Table 4 shows average results of dice for all CT and MRI segmented images in comparison to related work. The above-mentioned study [20] used the same idea on a dual stream FCN but with single MR modality and CT unlike the current study where two MR modalities were used in the same stream. Therefore, our proposed method outperformed the previous study in terms of CT modality, but not MRI. Since both provided MRI types (T1-DUAL, T2-SPIR) were used, our approach succeeded to benefit the CT modalities to achieve better results (0.963 DICE) but deteriorated the MRI output due to segmentation-associated noise.

Our model has also outperformed the work done by Mali et al. in CT and MRI segmentation on the three models they applied. It is showing in our model outperformed theirs in CT and MRI segmentation on the three models they applied. It should be noted that both our study and theirs used the same CHOS dataset and that their model is also UNET-based.

Examples of segmented images are given in Table 5, where it can be seen with the naked eye that the segmentation is so close to the ground truth with some noise resulting from using two different MR modalities. This noise causes significant reduction of DICE.

Table 5 segmented images examples

Modality type	Input image	Ground truth	Output images
CT			
			
			
MR T1 DUAL			
			
			
MR T2 SPIR			
			
			

VII. Conclusion

This work demonstrates the value of multi-modal learning on unpaired multi-modal multi-stream CT and MRI segmentation. A dual-stream network architecture was presented. By multi-modal learning, shared representation on both modalities can help the network to segment at high efficiency with limited training data. Experimental results on both MR and CT demonstrate that the CT was improved on the liver segmentation task. The power of learning on multi-streams from different datasets appears to be encouraging to examine further more in future work as its results are quiet promising, finally it is also shown that using the UNET in the proposed architecture is superior than using the FCN.

VIII. References

[1] "A review of semantic segmentation using deep neural networks | SpringerLink." https://link.springer.com/article/10.1007/s13735-017-0141-z (accessed Apr. 06, 2021).

[2] J. Yanase and E. Triantaphyllou, "A Systematic Survey of Computer-Aided Diagnosis in Medicine: Past and Present Developments," *Expert Systems with Applications*, vol. 138, p. 112821, Jul. 2019, doi: 10.1016/j.eswa.2019.112821.

[3] A. Rodriguez-Ruiz *et al.*, "Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study," *Eur Radiol*, vol. 29, no. 9, pp. 4825–4832, Sep. 2019, doi: 10.1007/s00330-019-06186-9.

[4] "WHO | Imaging Modalities," *WHO*. http://www.who.int/diagnostic_imaging/imaging_modalities/en/ (accessed Apr. 06, 2021).

[5] P. Moeskops *et al.*, "Deep Learning for Multi-task Medical Image Segmentation in Multiple Modalities," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Cham, 2016, pp. 478–486, doi: 10.1007/978-3-319-46723-8_55.

[6] "Multimodality imaging techniques - Martí-Bonmatí - 2010 - Contrast Media &amp; Molecular Imaging - Wiley Online Library." https://onlinelibrary.wiley.com/doi/full/10.1002/cmmi.393 (accessed Apr. 06, 2021).

[7] H. Shin *et al.*, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, May 2016, doi: 10.1109/TMI.2016.2528162.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation,"

arXiv:1411.4038 [cs], Mar. 2015, Accessed: Mar. 31, 2021. [Online]. Available: http://arxiv.org/abs/1411.4038.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv:1505.04597 [cs]*, May 2015, Accessed: Mar. 31, 2021. [Online]. Available: http://arxiv.org/abs/1505.04597.

[10] L. Bi, J. Kim, A. Kumar, M. Fulham, and D. Feng, "Stacked fully convolutional networks with multi-channel learning: application to medical image segmentation," *Vis Comput*, vol. 33, no. 6–8, pp. 1061–1071, Jun. 2017, doi: 10.1007/s00371-017-1379-4.

[11] M. D. Kohli, R. M. Summers, and J. R. Geis, "Medical Image Data and Datasets in the Era of Machine Learning—Whitepaper from the 2016 C-MIMI Meeting Dataset Session," *J Digit Imaging*, vol. 30, no. 4, pp. 392–399, Aug. 2017, doi: 10.1007/s10278-017-9976-3.

[12] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *arXiv:1511.00561 [cs]*, Oct. 2016, Accessed: Mar. 31, 2021. [Online]. Available: http://arxiv.org/abs/1511.00561.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *arXiv:1703.06870 [cs]*, Jan. 2018, Accessed: Mar. 31, 2021. [Online]. Available: http://arxiv.org/abs/1703.06870.

[14] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, May 2019, doi: 10.1016/j.zemedi.2018.11.002.

[15] H. R. Roth *et al.*, "An application of cascaded 3D fully convolutional networks for medical image segmentation," *Computerized Medical Imaging and Graphics*, vol. 66, pp. 90–99, Jun. 2018, doi: 10.1016/j.compmedimag.2018.03.001.

[16] W. Chen, B. Liu, S. Peng, J. Sun, and X. Qiao, "S3D-UNet: Separable 3D U-Net for Brain Tumor Segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Cham, 2019, pp. 358–368, doi: 10.1007/978-3-030-11726-9_32.

[17] S. A. Mali, "Multi-modal learning for Abdominal Organ Segmentation," p. 68.

[18] Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer, "CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data." Zenodo, Apr. 11, 2019, doi: 10.5281/zenodo.3431873.

[19] Y. Shu, J. Zhang, B. Xiao, and W. Li, "Medical image segmentation based on active fusion-transduction of multi-stream features," *Knowledge-Based Systems*, vol. 220, p. 106950, 2021, doi: https://doi.org/10.1016/j.knosys.2021.106950.

[20] V. V. Valindria *et al.*, "Multi-modal Learning from Unpaired Images: Application to Multi-organ Segmentation in CT and MRI," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, Mar. 2018, pp. 547–556, doi: 10.1109/WACV.2018.00066.

[21] "Pydicom |." https://pydicom.github.io/ (accessed Mar. 31, 2021).

[22] *keras-team/keras*. Keras, 2021.

[23] A. E. Kavur *et al.*, "CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, p. 101950, Apr. 2021, doi: 10.1016/j.media.2020.101950.

[24] S. Jadon, "A survey of loss functions for semantic segmentation," *arXiv:2006.14822 [cs, eess]*, Sep. 2020, doi: 10.1109/CIBCB48159.2020.9277638.

[25] Z. Wang, E. Wang, and Y. Zhu, "Image segmentation evaluation: a survey of methods," *Artif Intell Rev*, vol. 53, no. 8, pp. 5637–5674, Dec. 2020, doi: 10.1007/s10462-020-09830-9.