Generalizing Medical Image Segmentation In Multiple Domains
Deep Learning in Medical Imaging Course Final Project

David Sriker
Tel-Aviv University
David.Sriker@gmail.com

Abstract

Medical image segmentation has an essential role in computer-aided diagnosis systems in different applications. Accurate segmentation of medical images is a key step in contouring during radiotherapy planning. Computed topography (CT) and Magnetic resonance (MR) imaging are the most widely used radiographic techniques in diagnosis, clinical studies and treatment planning. This importance attract researchers to implement new medical image-processing algorithms. In this work we used both lungs CT and prostate MRI to train our models. We showed that although U-net is one of the best segmentation network currently available, we were able to improve the results be introducing GAN approach to the training procedure and by doing so we got smoother and more accurate results while keeping the processing time as constant using the U-net as our segmentation architecture.

1. Introduction

Medical image segmentation, identifying the pixels of organs or lesions from background medical images such as CT or MR images, is one of the most challenging tasks in medical image analysis that is to deliver critical information about the shapes and volumes of these organs. Many researchers have proposed various automated segmentation systems by applying available technologies. Earlier systems were built on traditional methods such as edge detection filters and mathematical methods. Then, machine learning approaches extracting hand-crafted features have became a dominant technique for a long period. Designing and extracting these features has always been the primary concern for developing such a system and the complexities of these approaches have been considered as a significant limitation for them to be deployed. In the 2000s, owing to hardware improvement, deep learning approaches came into the picture and started to demonstrate their considerable capabilities in image processing tasks. The promising ability of deep learning approaches has put them as a primary option for image segmentation, and in particular for medical image segmentation. Especially in the previous few years, image segmentation based on deep learning techniques has received vast attention. With increasing use of Computed topography (CT) and Magnetic resonance (MR) imaging for diagnosis, treatment planning and clinical studies, it has become almost compulsory to use computers to assist radiological experts in clinical diagnosis, treatment planning. Reliable algorithms are required for the delineation of anatomical structures and other regions of interest (ROI). The goals of computer-aided diagnosis (CAD) are:

- To automate the process so that large number of cases can be handled with the same accuracy i.e. the results are not affected as a result of fatigue, data overload or missing manual steps.
- To achieve fast and accurate results. Very high-speed computers are available at modest costs, speeding up computer-based processing in the medical field.
- To support faster communication, wherein patient care can be extended to remote areas using information technology.

The techniques available for segmentation of medical images are specific to application, imaging modality and type of body part to be studied. For example, requirements of prostate MRI segmentation are different from those of lungs CT segmentation. The artifacts, which affect the prostate image, are different - partial volume effect is more prominent in prostate while in the lungs it is motion artifact which is more prominent. Thus while selecting a segmentation algorithm one is required to consider all these aspects. The problems common to both CT and MR medical images are:

- Partial volume effect.
- Different artifacts: physiologic artifacts, inherent physical artifacts etc.
- Noise due to sensors and related electronic system.

There is no universal algorithm for segmentation of every medical image. Each imaging system has its own specific limitations. For example, in MRI one has to take care of bias field noise (intensity in-homogeneities in the RF field). Of course, some methods are more general as compared to specialized algorithms and can be applied to a wider range of data.
In this work we will suggest a general architecture for segmenting medical images that is applicable to multiple domains, while one outperform the common U-net architecture while conserving the interpolation time and other that is not due to insufficient amount of data.

2. Related Work

Recently, several deep learning-based pixel-wise classification methods have been proposed in computer vision area and some of them have been successfully applied in medical imaging. Early deep learning-based methods are based on bounding box [17]. The task is to predict the class label of the central pixels via a patch including its neighbors. Kallenberg et al. [5] designed a bounding box based deep learning method to perform breast density segmentation and scoring of mammographic texture. Shin et al. [16] compared several networks on the performance of computer-aided detection and proposed a transfer learning method by utilizing models trained in computer vision domain for medical imaging problem. Instead of running a pixel-wise classification with a bounding box, Long et al. [8] proposed a fully convolutional network (FCN) for semantic segmentation by replacing the fully connected layers with convolutional layers. An Auto-Encoder alike structure has been used by Noh et al. [12] to improve the quality of the segmented objects. Later, Ronneberger et al. [13] proposed a U-net model for segmentation, which consists of a contracting part as an encoder to analyze the whole image and an expanding part as a decoder to produce a full-resolution segmentation. The U-net architecture is different from [8] in that, at each level of the decoder, a concatenation is performed with the correspondingly cropped feature maps from the encoder. This design has been widely used and proved to be successful in many medical imaging applications such as Lumbar Surgery [1] and gland segmentation [10]. Most recently, Lalonde et al. [6] designed a convolutional-deconvolutional capsule network, called SegCaps, to perform lung segmentation, where they proposed the concept of deconvolutional capsules. After the emergence of Generative Adversarial Network (GAN) [3] based models, which have shown a better efficiency in leveraging the inconsistency of the generated image and ground truth in the task of image generation, Luc et al. [9] proposed a GAN based semantic segmentation model. The motivation is to apply GAN to detect and correct the high order inconsistencies between ground truth segmentation maps and the generated results. The model trains a segmentation network along with an adversarial network that discriminates segmentation maps coming either from the ground truth or from the segmentation network.

3. Data-Sets

In order to check the robustness of our model we tested two data-sets which are different in the acquisition process. Both data-sets were obtain from Kaggle and has only one channel in the image domain and a binary classification mask. We found two publicly available dataset that fit our demands.

- Lung CT scans and their corresponding binary masks [4, 2], the data-set consist of 700 pairs and can directly downloaded from Lung Data-Set. Sample of the data can be seen in Fig. 1.

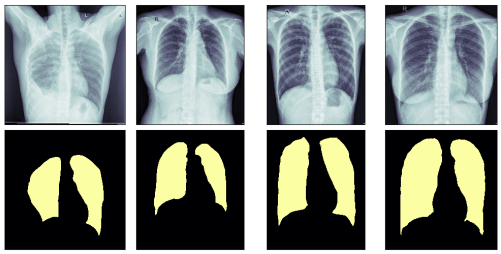

Figure 1. Lung CT and their corresponding segmentation

- Prostate MRI scans and their corresponding binary masks [7], the data-set consist of 900 pairs and can directly downloaded from Prostate Data-Set Sample of the data can be seen in Fig. 2.

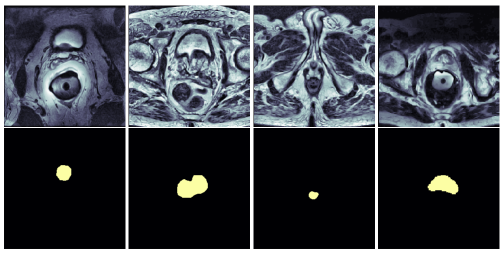

Figure 2. Prostate MRI and their corresponding segmentation

Although both data-sets is a simple binary segmentation tasks, the different acquisition and anatomic rise challenges and we will able to demonstrate the generalization ability of our model and can be easily extended to multi-class segmentation task.

4. Implementations

The supplementary code is available at the following GitHub repositories which all was implemented in *Python* using *PyTorch* API and was trained on a NVIDIA GeForce RTX 2080:

- U-net
- LGAN
- CycleGan

4.1. U-Net

Unet [14] architecture is illustrated in Fig. 3 and was implemented from scratch. It consists of a contracting path (left side) and an expansive path (right side). The contracting path follows the typical architecture of a convolutional network. It consists of the repeated application of two 3x3 convolutions (unpadded convolutions), each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. At each downsampling step we double the number of feature channels. Every step in the expansive path consists of an upsampling of the feature map followed by a 2x2 convolution ("up-convolution") that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3x3 convolutions, each followed by a ReLU. The cropping is necessary due to the loss of border pixels in every convolution. At the final layer a 1x1 convolution is used to map each 64 component feature vector to the desired number of classes.

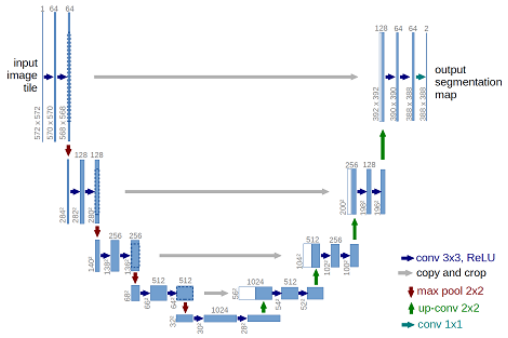

Figure 3. U-net architecture

4.2. LGAN

LGAN [18] is a Generative Adversarial Network (GAN) which designed for lung segmentation the schema is designed to force the generated lung segmentation mask to be more consistent and close to the ground truth and its architecture is illustrated in Fig. 4. LGAN consists of two networks: the generator network and the discriminator network, and both of them are convolution neural networks. The generator network is to predict the lung segmentation masks based on the gray scale input images while the discriminator computes the EM distance between the predicted masks and the ground truth masks. We build the generator different from the original paper, the generator network is the same U-net as described in section 4.1 while the discriminator is a simple contracting convolutional neural network also implemented entirely by us, there is no existing code published yet so we were based solely on the paper.

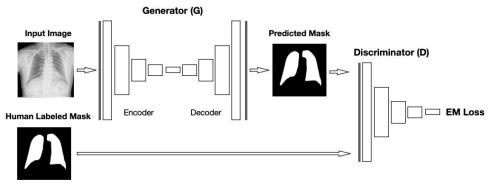

Figure 4. LGAN architecture

4.3. Cycle-GAN

The proposed architecture for semi-supervised segmentation, illustrated in Fig 5, is based on the cycle consistent GAN (Cycle-GAN) model [19] which has shown outstanding performance for unpaired image-to image translation. This architecture is composed of four inter-connected networks, two conditional generators and two discriminators, which are trained simultaneously. In the original Cycle-GAN model, the generators are employed to learn a bidirectional mapping from an image domain to the other. On the other hand, discriminators try to determine whether an image from the corresponding domain is real or generated. By fooling the discriminators through adversarial learning, the model thus learns to generate images from the true distribution without requiring paired images. A cycle-consistency loss is also added to ensure that generators are consistent, i.e. that we recover the same image when going through both generators sequentially. The generators are the a U-net generators as described in section 4.1 and the discriminator is a simple patch discriminator, there is multiple ways to choose the generator and discriminator architecture, we chose the U-net as the generator in order to be able to compare the results to our base line U-net and the discriminator was chosen as the one that mostly used in GAN segmentation networks. Traditionally in a cycle consistency approach, two generators ($G_{A\to B}$ and $G_{B\to A}$) and two discriminators (D_A and D_B) are trained in order to achieve unsupervised image-to-image translation and an objective loss can be devised by comparing the pairs of samples $\{X_A, X_{A\to B\to A}\}$ and $\{X_B, X_{B\to A\to B}\}$.

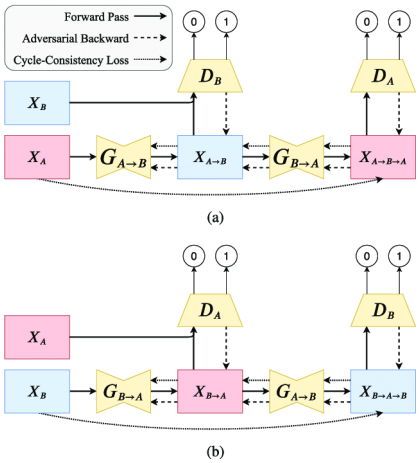

Figure 5. GAN architecture based on cycle consistency

5. Evaluation Metrics

The metrics used in this study are widely used for the evaluation of segmentation algorithms:

* Intersection over Union (IoU)
* Dice coefficient (Dice)
* Hausdorff distance (HD)

5.1. Intersection over Union

The Intersection-Over-Union (IoU), also known as the Jaccard Index, is one of the most commonly used metrics in semantic segmentation. The IoU is a very straightforward metric that's extremely effective. It is the ratio of the overlapping area of ground truth and predicted area to the total area.

$$IoU(X,Y) = \frac{X \cap Y}{X \cup Y} \tag{1}$$

A visual explanation of the metric can be seen in Fig. 6.

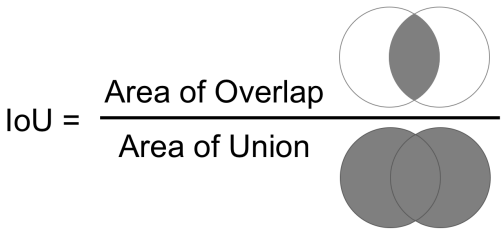

Figure 6. A visual explanation of IoU

5.2. Dice coefficient

The Dice similarity coefficient, also known as the Sørensen–Dice index or simply Dice coefficient, is a statistical tool which measures the similarity between two sets of data. This index has become arguably the most broadly used tool in the validation of image segmentation algorithms created with AI, but it is a much more general concept which can be applied sets of data for a variety of applications including NLP. Simply put, the Dice Coefficient is two times the area of overlap divided by the total number of pixels in both images.

$$Dice(X,Y) = \frac{2 \cdot (X \cap Y)}{|X| + |Y|} \tag{2}$$

A visual explanation of the metric can be seen in Fig. 7.

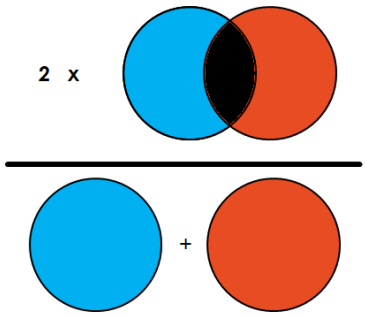

Figure 7. A visual explanation of Dice

5.3. Hausdorff distance

The Hausdorff distance measures how far two subsets of a metric space are from each other. Informally, two sets are close in the Hausdorff distance if every point of either set is close to some point of the other set. The Hausdorff distance is the longest distance you can be forced to travel by an adversary who chooses a point in one of the two sets, from where you then must travel to the other set. In other words, it is the greatest of all the Euclidean distances from a point in one set to the closest point in the other set.

$$HD(X,Y) = \underset{x \in X}{Max} \left(\underset{y \in Y}{Min}\, \mathbf{D}(x,y) \right) \tag{3}$$

A visual explanation of the metric can be seen in Fig. 8.

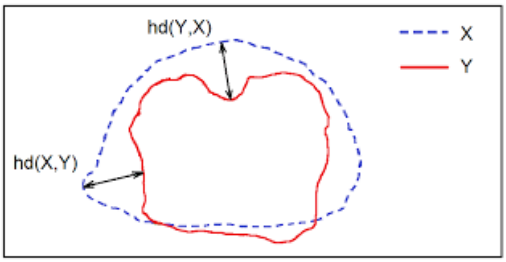

Figure 8. A visual explanation of Hausdorff distance

6. Results

Before comparing the models, we tested each trained model on each data-set to validate the success of the training procedure with the given data-sets. The results per individual model is presented in the following sub-sections. In order to properly compare the results we used the same U-net architecture as our baseline as well as the generators. This was in order to not rely on results from different sources that was probably optimized to the end.

6.1. U-net

First, we trained the model for a total of 75 epochs. We did not increased the number of epochs since we observed that the model stopped improving. In Fig. 9 we can see the averaged loss of the training procedure for the Lung CT data-set and in Fig. 10 for the Prostate MRI data-set. It is easy to observe that we indeed reached plateau around 65 epochs. While training we watched two performance parameters, the IoU and Dice and we saved the weights that correspond to the highest on both measures, as was introduced in section 5, and in the case there was no improvement we decreased the learning rate so to make the training more delicate. In Fig. 11 we can see the score on the validation set for both measure for the Lung CT data-set and in Fig. 12 for the Prostate MRI data-set. We can see again that the results did not improve after 65 epochs, validating that indeed training for 75 epochs is sufficient. In the table 1 the results on the test-set are summarized, we averaged each measure for in order to estimate the overall test-set performance.

The results shown in table 1 will serve as our baseline to evaluate the performance of the other models this is due to the fact the in both other models we used the same U-net as the generator architecture.

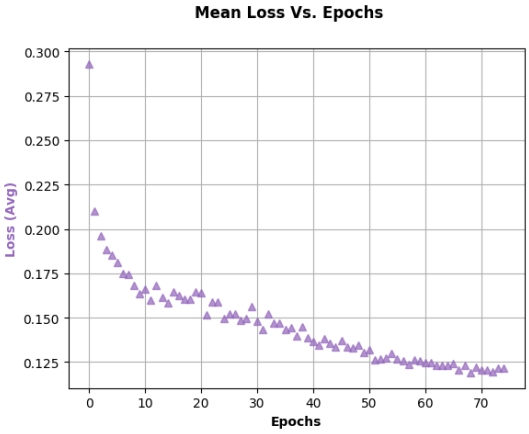

Figure 9. Mean Loss Vs. Epochs (Lung CT data-set)

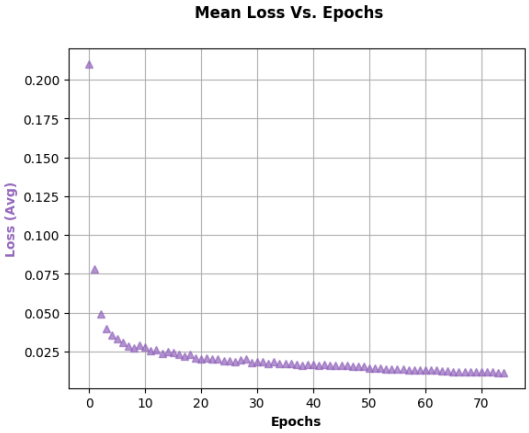

Figure 10. Mean Loss Vs. Epochs (Prostate MRI data-set)

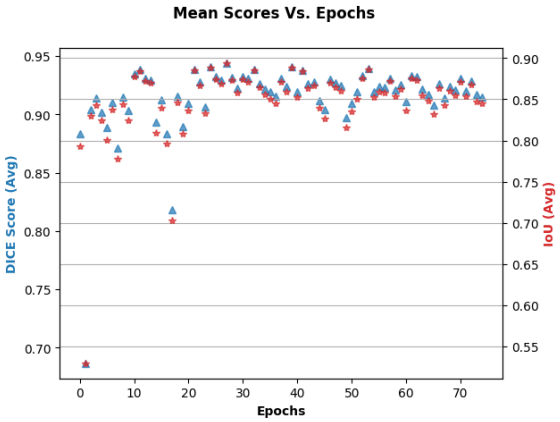

Figure 11. Score (IoU, Dice) Vs. Epochs (Lung CT data-set)

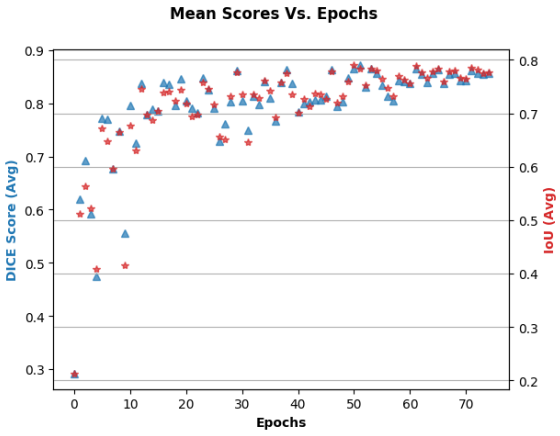

Figure 12. Score (IoU, Dice) Vs. Epochs (Prostate MRI data-set)

	Mean IoU	Mean Dice	Mean HD
Lung CT	0.8974	0.9453	4.9348
Prostate	0.8141	0.8946	3.9642

Table 1. Test Performance

6.2. LGAN

We repeated the same training procedure as described in subsection 6.1. The reason to do so is because we used the same U-net architecture as our generator thus we can be quite sure that they will be very similar. In Fig. 13 we can see the averaged losses of the training procedure for the Lung CT data-set and in Fig. 14 for the Prostate MRI data-set. It is easy to observe that we indeed reached plateau around 65 epochs.

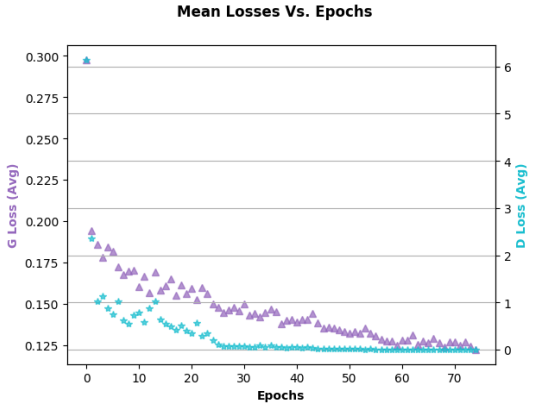

Figure 13. Mean Loss Vs. Epochs (Lung CT data-set)

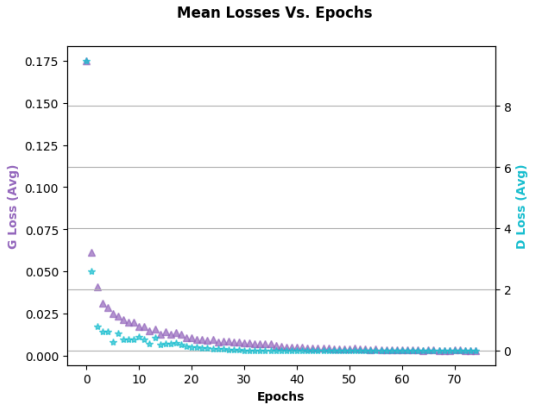

Figure 14. Mean Loss Vs. Epochs (Prostate MRI data-set)

As in the section 6.1 during training we watched two performance parameters, the IoU and Dice and we saved the weights that correspond to the highest on both measures, as was introduced in section 5, and in the case there was no improvement we decreased the learning rate so to make the training more delicate. In Fig. 15 we can see the score on the validation set for both measure for the Lung CT data-set and in Fig. 16 for the Prostate MRI data-set.

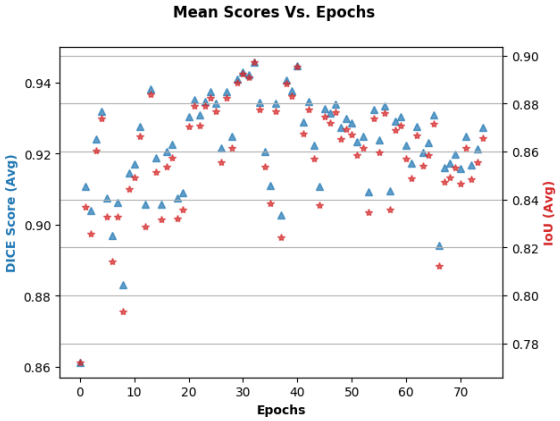

Figure 15. Score (IoU, Dice) Vs. Epochs (Lung CT data-set)

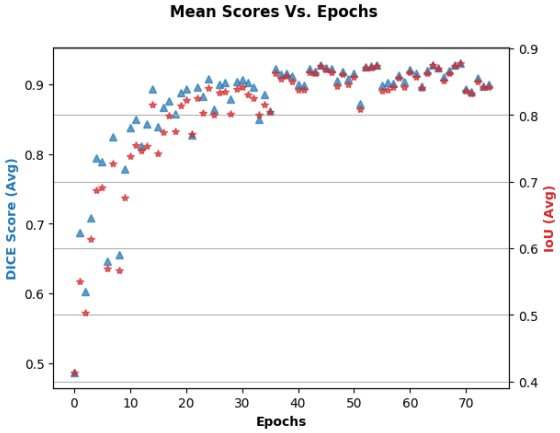

Figure 16. Score (IoU, Dice) Vs. Epochs (Prostate MRI data-set)

We can see that the scores in the Lung CT data-set is more fluctuating and that we received the peak at around 35 epochs. while in the Prostate MRI data-set we received the peak in the last epochs.
The results on the test-set is summarised in table 2 and we see that although that for the Lung CT data-set the best weights were obtained at around 35 epochs we still superb results.

	Mean IoU	Mean Dice	Mean HD
Lung CT	0.9028	0.9483	4.8363
Prostate	0.8842	0.9364	3.3754

Table 2. Test Performance

6.3. Cycle-GAN

We Trained the model for a total of 200 epochs, In Fig. 17 we can see the averaged losses of the training procedure for the Lung CT data-set and in Fig. 18 for the Prostate MRI data-set. It seems that the model did not improve after 125 epochs. But we can see from the following plots, Fig.

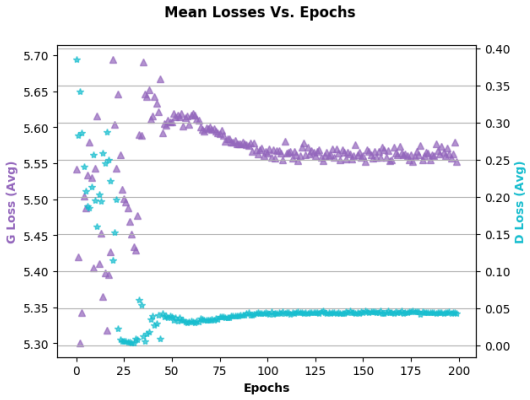

Figure 17. Mean Loss Vs. Epochs (Lung CT data-set)

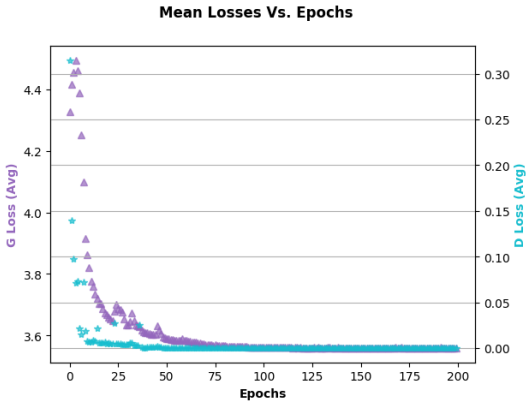

Figure 18. Mean Loss Vs. Epochs (Prostate MRI data-set)

19 is the score on the validation set for both measure for the Lung CT data-set and in Fig. 20 for the Prostate MRI data-set, although the losses did not continue to decrease instead of generalizing the data we over-fitted it. We did try and test other optimizer and learning rate decay procedure in order to improve the results but the problem is probably in the data itself. For both data-sets we have 700 and 900 low res images respectively while the amount of parameters of the model is huge (for each generator $\sim$ 42M and for each discriminator $\sim$ 10M) so over-fitting the data was inevitable. The results are summarised in table 3.

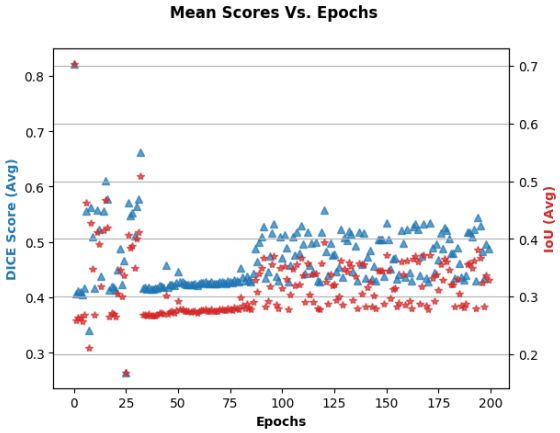

Figure 19. Score (IoU, Dice) Vs. Epochs (Lung CT data-set)

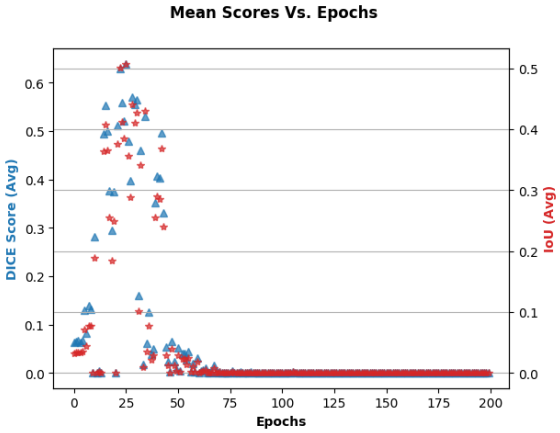

Figure 20. Score (IoU, Dice) Vs. Epochs (Prostate MRI data-set)

	Mean IoU	Mean Dice	Mean HD
Lung CT	0.5289	0.6803	7.8441
Prostate	0.4099	0.5302	4.8439

Table 3. Test Performance

6.4. Results Summary

Fig 21 present the results of the different models for both the Lung CT data-set and the Prostate MRI data-set for all the three models together we can see that Cycle-GAN gave us the worst results and by probably add a Conditional Random Fields (CRF) as a post-processing we will get a better and smoother results. We can also see that the LGAN gave us the closest to the ground truth which is impressive given the fact that it trained in a GAN manner which usually need more input sample to properly learn and generalize.

Summarizing the results of the evaluation metrics in table 4 we can validate that the LGAN model was able to surpass the other models in all criteria.

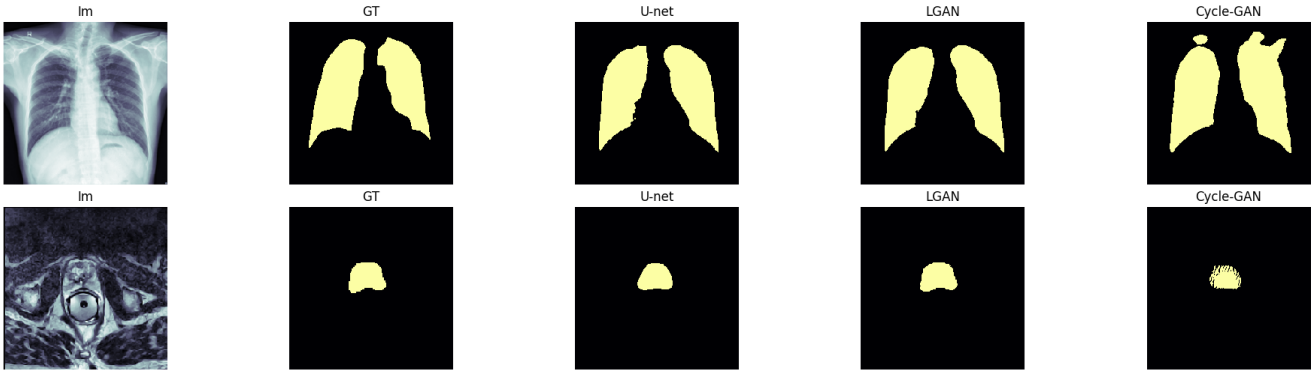

Figure 21. Results of the models for the two data-sets

	Lung CT			*Prostate MRI*		
	Mean IoU	**Mean Dice**	**Mean HD**	**Mean IoU**	**Mean Dice**	**Mean HD**
U-Net	0.8974	0.9453	4.9348	0.8141	0.8946	3.9642
LGAN	**0.9028**	**0.9483**	**4.8363**	**0.8842**	**0.9364**	**3.3754**
Cycle-GAN	0.5289	0.6803	7.8441	0.4099	0.5302	4.8439

Table 4. Summary of Performance

7. Conclusions

We could see that although the Cycle-GAN is very promising in domain-domain translation, unless the amount of data is sufficiently high it can not outperform state-of-the-art segmentation networks. Regarding the LGAN approach, we noticed that this approach not only got us better results in terms of the evaluation metrics but also a more smooth segmentation mask. This is a very promising route to go with, as we see that GAN training procedure can help in the process of segmenting anatomical parts in an efficient way.

8. Future Work

There is a lot of directions to go from this work. As we saw that we could improve the results from the U-net that served as our baseline using the LGAN model, this approach can be easily extended to multi-class segmentation and lesion detection. Other direction is to apply the same logic to a 3D object segmentation and replace the U-net with V-net [11] and test if we indeed get better performance. Another possibility is checking other types of discriminators such as Patch-GAN discriminator [15] etc.
In the case of Cycle-GAN we could try to test it with other and larger data-sets. We could also try and train it in a manner to create a synthetic and realistic data to thicken our data-set to be fed to other architectures.

References

[1] N. Baka, S. Leenstra, and T. van Walsum. Ultrasound aided vertebral level localization for lumbar surgery. *IEEE Transactions on Medical Imaging*, 36(10):2138–2147, Oct. 2017.

[2] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, 33(2):577–590, Feb. 2014.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[4] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh, S. Antani, G. Thoma, Y.-X. Wang, P.-X. Lu, and C. J. McDonald. Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging*, 33(2):233–245, Feb. 2014.

[5] M. Kallenberg, K. Petersen, M. Nielsen, A. Y. Ng, P. Diao, C. Igel, C. M. Vachon, K. Holland, R. R. Winkel, N. Karssemeijer, and M. Lillholm. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE Transactions on Medical Imaging*, 35(5):1322–1331, May 2016.

[6] R. LaLonde and U. Bagci. Capsules for object segmentation. arxiv 2018. *arXiv preprint arXiv:1804.04241*.

[7] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, R. Strand, F. Malmberg, Y. Ou, C. Davatzikos, M. Kirschner, F. Jung, J. Yuan, W. Qiu, Q. Gao,

P. “. Edwards, B. Maan, F. van der Heijden, S. Ghose, J. Mitra, J. Dowling, D. Barratt, H. Huisman, and A. Madabhushi. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Medical Image Analysis*, 18(2):359–373, Feb. 2014.

[8] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015.

[9] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. *CoRR*, abs/1611.08408, 2016.

[10] S. Manivannan, W. Li, J. Zhang, E. Trucco, and S. J. McKenna. Structure prediction for gland segmentation with hand-crafted and deep convolutional features. *IEEE Transactions on Medical Imaging*, 37(1):210–221, Jan. 2018.

[11] F. Milletari, N. Navab, and S. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *CoRR*, abs/1606.04797, 2016.

[12] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation, 2015.

[13] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[14] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[15] A. B. Saad, Y. Tamaazousti, J. Kherroubi, and A. He. Where is the fake? patch-wise supervised gans for texture inpainting, 2019.

[16] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, May 2016.

[17] J. Tan, Y. Huo, Z. Liang, and L. Li. Apply convolutional neural network to lung nodule detection: Recent progress and challenges. In *Smart Health*, pages 214–222. Springer International Publishing, 2017.

[18] J. Tan, L. Jing, Y. Huo, Y. Tian, and O. Akin. LGAN: lung segmentation in CT scans using generative adversarial network. *CoRR*, abs/1901.03473, 2019.

[19] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017.