IMPROVEMENTS IN CARDIAC SEGMENTATION FOR CROSS-MODALITY DOMAIN ADAPTATION

by

RUSHIN HITESH GINDRA

A thesis submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Computer Science

Written under the direction of

Dimitris Metaxas

and approved by

_____

_____

_____

New Brunswick, New Jersey

May, 2020

ABSTRACT OF THE THESIS

Improvements in cardiac segmentation for cross-modality domain adaptation

By Rushin Hitesh Gindra

Thesis Director:

Dimitris Metaxas

In medical image computing, the problem of heterogeneous domain shift is quite common and severe, causing many deep convolutional networks to under-perform on various imaging modalities. Retraining the network is difficult since annotating the new domain data is prohibitively expensive, specifically in medical areas that require expertise. While recent works show approaches to tackle this problem using unsupervised domain adaptation, segmentation modules in such methods can be improved vastly. Our implementation provides a segmentation improvement on the current state-of-the-art framework, Synergistic Image and Feature Adaptation(SIFA). We revisit atrous spatial pyramid pooling while using convolutional features as well as image features for multi-scale object segmentation. We have validated the effectiveness of the improvement on the framework using the challenging application of cross-modality segmentation of cardiac structures. To demonstrate the robustness of the module, extensive experiments have been performed on Long-Axis(MMWHS) cross-modal cardiac segmentation tasks.

Acknowledgements

I would like to sincerely thank my advisor and committee chair, Dr. Dimitris Metaxas for continued guidance and support during this research.

I am grateful to Dr. Konstantinos Michmizos and Dr. Karl Stratos for being a part of my defense committee and sharing their valuable knowledge and insights on my work.

I also thank my colleagues Qiaoying Huang and Ligong Han for assisting me with the required literature and implementations throughout the research.

Lastly, I would like to thank my family and friends who have always been supportive of me and have made it possible for me to follow my passion.

Thanks for all your encouragement!

Dedication

Dedicated to my parents, Alpa and Hitesh Gindra, and my former mentor, Dr. Vivek Dave, L. V. Prasad Eye Institute, India.

Table of Contents

List of Figures

List of Tables

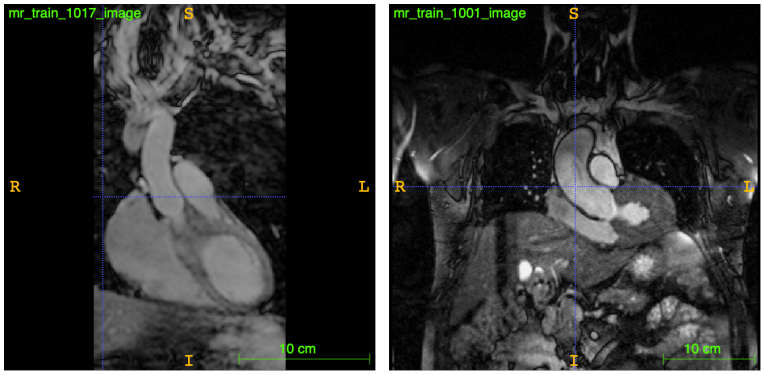
Chapter 1

Introduction

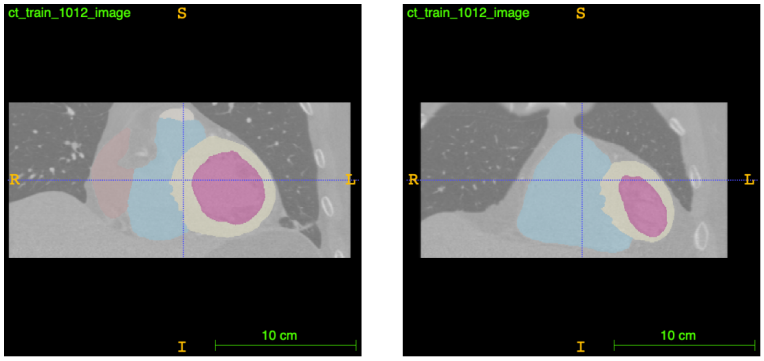
Deep Convolutional Neural Networks(DCNNs) are currently dominating numerous challenging tasks like semantic segmentation, object detection, motion analysis, etc, yielding outstanding performance in several medical imaging tasks [32]. These supervised algorithms frequently assume that the training and testing data are independent and identically (i.i.d) distributed. However this assumption rarely holds true in real life. A number of recent theoretical and empirical results have pointed out the problem of performance degradation when encountering a domain shift between the training data(source domain) and testing data(target domain) [4, 37, 50]. Such domain shifts are even more natural and severe in the case of medical image computing. Certain scenarios may include, training and testing images coming from different sites, different scanning protocols or even different imaging modalities [63, 62].

A typical situation in the medical field is the use of different indispensable imaging modalities like MRI and CT for cardiac imaging. For example, MRIs capture great contrast between soft tissues and provide high resolution in the temporal space. On the contrary, CT imaging is quick and provides great spatial resolution. One can observe, these different modalities play important complementary roles in clinical procedures for disease diagnosis and treatment.

In practice, often the same image analysis task is required for multiple related but different domains, like segmentation of cardiac structures from MRI and CT scans. As anticipated, DCNNs trained on MRI data only, perform poorly when tested on CT image scans, concluding that there exists a domain shift. To recover the model performance, one rudimentary method is to fine-tune the model [51], which requires



(a) Different MR Scans, Different scales



(b) Same CT Scan(multiple slices)

Figure 1.1: Existence of multiple scale samples of same cardiac structures

additional labeled data from the target domain. However, in many supervision dependent tasks like segmentation, labeling data for new domains in extremely expensive and cumbersome. Another option is to use purely synthetic data for model training. Unfortunately, models naively trained on synthetic data do not generalize well to real image samples.

This problem of domain shift has motivated several research works on unsupervised domain adaptation(UDA). It is a methodology that attempts to learn a model that performs well in target domain using solely unlabeled target domain data, and labelled data from the source domain. Recent works on UDA can be divided among two streams, Image Adaptation and Feature Adaptation. Briefly, image adaptation deals with the domain shift at the input level using pixel-level transformations, while feature adaptation learns a model that extracts domain invariant features.

Recent studies have pointed out, both streams address domain shift from complementary perspectives. Furthermore, several promising works have emerged that perform both the adaptations together [14, 24, 57]. The current state-of-the-art framework, Synergistic Image and Feature Adaptation (SIFA) [6] leverages the mutual interaction and benefits of both adaptations by enabling a synergistic fusion of the adaptations from both the feature and image perspectives. The network shares the feature encoder for simultaneously transforming image appearances and extracting domain invariant representations for segmentation tasks. The network is explained in detail in the Methods section.

To the best of our knowledge, all recent UDA for segmentation methods, including the state-of-the-art SIFA framework, use dilated residual networks (DRNs) [54] as their segmentation modules. While, DRNs are efficient and work well for segmenting out distinct structures, they do not capture the multi-scale context effectively (See figure 1.1). There remains extensive room for improvement in the segmentation module of the network. Semantic segmentation being a separate research problem, a fairly large number of methods [9, 33] have been developed to improve the performance in computer vision systems. Most recent works include use of atrous convolutions and spatial pyramid pooling [10] to capture multi-scale features.

In this thesis, noticing the growing inclination of researchers towards using PyTorch as their default deep learning library, we re-implement the SIFA framework in PyTorch. As far as we know, our implementation is the first reproduction of the TensorFlow implementation [5] of SIFA in PyTorch. Most importantly, we propose an additional component to the DRN module called atrous spatial pyramid pooling(ASPP)[10], and successfully validate the method on the challenging task of cross-modality cardiac structure segmentation [63]. The segmentation module is an independent entity, that can be appended to any domain adaptation modules effectively without interfering with the adaptation process. We encourage researchers to use our implementation, in whole or in part, for future improvements.

The major highlights of the thesis are as follows:

1. With growing inclination of researchers to use PyTorch, we re-implement the

SIFA network in PyTorch to encourage further improvements in UDA for medical imaging.

2. We propose an addition of atrous spatial pyramid pooling module to the segmentation network, to capture multi-context features and thus improve segmentation performance without interfering with the adaptation process.

3. We validate the robustness of the SIFA(PyTorch) module with the additional ASPP and compare the results with the TensorFlow implementation of SIFA.

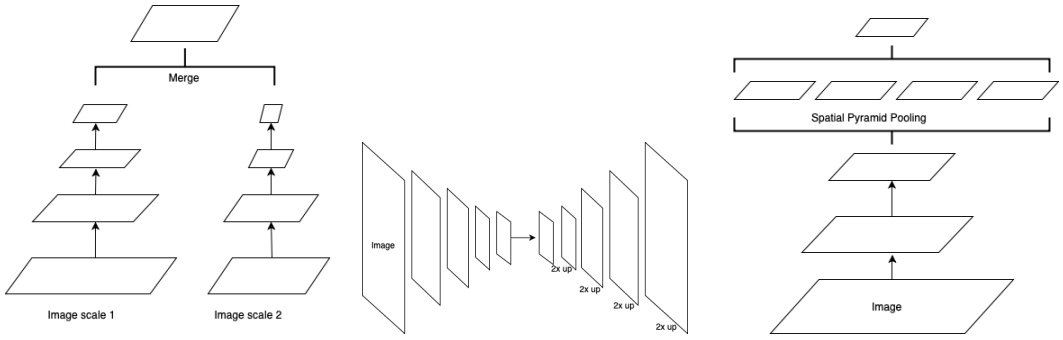
Chapter 2

Literature Review

Our implementation involves improving the segmentation networks that can be used in domain adaptation frameworks without affecting the adaptation process. As a result, in this chapter, we discuss two independent modules and the previous works that inspired this thesis. The modules discussed below, when integrated into one unified network, exploit information effectively to perform domain-invariant and accurate segmentation.

2.1 Semantic Image Segmentation

Semantic segmentation was an extremely challenging task in the previous decade, when deep learning was still on the rise. Most of the successful segmentation algorithms involved using hand-crafted features with shallow classifiers like Random Forests [48] and Support Vector Machines [17]. The performance of these systems has always been compromised by the limited representational power of the hand-crafted features. With the rise in DCNNs deployed in a fully convolutional manner (i.e Fully Convolutional Networks or FCNs [47, 38, 36]), the performance of these models on several semantic



(a) Image Pyramid (b) Encoder-Decoder (c) Spatial Pyramid Pooling

Figure 2.1: Segmentation model designs

segmentation benchmarks has been phenomenal.

We know that basic DCNNs normally are invariant to local image transforms [55], allowing them to learn abstract representations. While this is desirable for high-level vision tasks like image classification or object detection, it can hamper the performance of dense prediction tasks like semantic segmentation. Another seemingly trivial but difficult problem usually faced for segmentation tasks is the existence of objects at multiple scales [63].

Current state-of-the-art(SoA) DCNN systems for low-level prediction tasks revolve around two essential principles to solve the above problems. The first one is using optimal neural network design and the second is ability to capture multi-scale context. It has been known that context information is crucial for pixel labeling [22, 13, 39, 7, 31, 41, 46, 52]. Several model designs have been proposed for the same. (See figure 2.1)

2.1.1 Image Pyramid

The model is used repeatedly on multiple scaled inputs to capture multi-scale context. Small scale inputs encode long-range context in features. On the other hand, large scale inputs encode details relevant to small objects. For example, Farabet et al. [16], uses a laplacian pyramid to obtain images at multiple scales. These images are then given as an input to the model. The feature maps generated from the model, for all scales, are then merged for final pixel-level labeling. [31, 11, 9] involved directly resizing the inputs to various scales and fusing all the intermediate features from all the scales. The image pyramid design of networks is not a favorable method, and doesn't scale well for deeper DCNNs due to its heavy computations and limited GPU memory. Such models are usually applied during the Inference stage [12].

2.1.2 Encoder-decoder

This type of model comprises of two distinct components. The Encoder captures long-range information easily in the deeper layers, using consecutive pooling or convolutional striding operations. The Decoder inputs the captured information from the Encoder to recover the object details and spatial resolution that is necessary for dense-labeling

tasks.

One popular research example that uses the encoder-decoder design is U-Nets [44]. Comprising of several variations, U-Nets skip connections from encoder features to decoder activations, thus regaining all spatial resolution lost due to pooling and striding. [19] employs Laplacian pyramid reconstruction for decoders. SegNet [1] uses the pooling indices of the encoder for upsampling the convolutions in the decoder. Many such variations [30, 43, 42, 26] have performed effectively in several segmentation benchmarks.

2.1.3 Context Modules

In this design, models optimized for high-level tasks are appended with additional modules in cascade to encode long range context. Quite frequently, researchers have explored CRFs as context modules. One effective method is to append DenseCRFs [28] as an independent module over the DCNNs [8]. Furthermore, [60, 31, 46] have proposed to jointly train CRFs & DCNN components in a unified network.

Models based on atrous convolutions are also explored as context modules for segmentation. For example, [35, 53, 54] experiment with modifying the atrous rates in consecutive layers, to capture long range information effectively.

2.1.4 Spatial Pyramid Pooling

As the name suggests, this type of model [21, 29] incorporates pooling of feature maps obtained from several multiple-scaled inputs. Few works [25, 8, 53, 9] have explored atrous convolutions as context modules for spatial pyramid pooling(SPP) such that they can be applied to any network. Specifically, they duplicate several copies of the last block of the network (like ResNet [23]) and arrange them in a cascade before adding the SPP module. Atrous Spatial Pyramid Pooling(ASPP) [10] is a popular method where a layer with parallel atrous convolutions captures multi-scale information. In [34, 59], ASPP module is augmented with image-level features as well to capture global context.

While the above methods represent semantic segmentation in traditional vision systems, medical imaging studies frequently borrow them to effectively perform complex segmentations.

2.2 Domain Adaptation

Numerous investigations of deep learning have focused on reducing performance degradation of DCNNs under domain shift. As mentioned earlier, several research proposals tackle this problem from the perspective of image adaptation, feature adaptation or a mixture of both. This section gives an overview of the SoA approaches and focuses on unsupervised domain adaptation. (Assuming annotations are available only for the source domain). Studies on both traditional vision systems and medical imaging systems are covered.

2.2.1 Image Adaptation

Image adaptation is the process of aligning the image appearance between domains with pixel-to-pixel transformations. With the introduction of GANs [20] and several of its variations, image adaptation addresses the domain shift addressed at input level. To preserve the semantic information of the input images, the entire process is usually guided by a cycle-consistency constraint [61]. One stream of solutions is to test the transformed source like images on the trained model [56, 45]. Alternatively, generated target-like images can be used to train the model in the target domain(i.e using synthetic data for model training). However, these methods is usually not effective as they don't generalize to real images [4, 58]. CycleGANs [61] have gained much recognition in image-to-image transformations. Many natural [45] as well as medical imaging [24] studies have heavily borrowed the idea of CycleGANs for the task of segmentation. Bateson et al. [2] proposes a general constrained image adaptation approach for spine segmentation. It encodes simple domain-invariant prior knowledge about the segmentation regions, like region size, or region shape.

2.2.2 Feature Adaptation

Meanwhile, several feature-level adaptation methods have also been researched, with the end goal being to reduce domain shift deeper in the network. The focus is on aligning the feature distributions by minimizing a certain chosen measure of distances like Maximum

Mean Discrepancy(MMD) [37] between features from the source and target domain. Instead of using the distance measurements between distributions, methods like DANN [18] and ADDA [50] have effectively used discriminators to differentiate the feature space across domains. Several medical studies [27, 15] have adopted this framework for segmenting tasks. A Recent studies [49], propose to project the high-dimensional feature space to more compact spaces like the semantic prediction space, and use it as the discriminator input to derive the adversarial loss.

Several methods have emerged that combine image and feature adaptation to achieve a stronger process. For example CyCada [24], PnP-AdaNet [14] and FCAN [57] achieve significant improvement in adaptation between synthetic and real world scenarios, by sequentially performing image and feature adaptation.

While all the above methods assume that abundant unlabeled target data is available, which is not always realistic in medical image computing, OUyang et al. [40] uses a different direction for 3D cross modality cardiac segmentation. The method heavily borrows the idea of one shot domain translation [3].

It is worth noting that, to the best of our knowledge, most adaptation methods for semantic segmentation above utilize dilated residual networks(DRNs) [54] as their segmentation module.

Considering that cardiac MRI or CT can be taken across different sites or even different angles depending on the purpose of diagnosis, it is highly possible that the cardiac structures present are of multiple scales [63]. As a result, simple segmentation modules that don't effectively capture multi-scale context may not be sufficient in this challenging task of cardiac segmentation. To tackle this task, we propose to merge the current SoA SIFA framework [6] with the current SoA segmentation module to fully exploit their mutual benefits toward whole-heart segmentation without interfering with the unsupervised domain adaptation process. However there are certain constraints or restrictions of using the segmentation research for domain adaptation. For example U-Nets [44] cannot be utilized in adaptation frameworks because of the skip connections since, along with regaining the resolution maps, the decoder also gains domain specific features through the skip connections which is not desired for adaptation frameworks.

As a result, we need to focus on segmentation modules [10] that can perform without interfering with the adaptation process. To smoothen the integration process, we implement the two modules independent of each other such that the segmentation module can be used as an individual entity for use in future improved domain adaptation frameworks.

Chapter 3

Dataset

To compare the effectiveness of our method with it's predecessor, SIFA [6], we test it on the MICCAI Multi-modality Whole Heart Segmentation(MMWHS) Challenge [63] dataset (Figure 3.1). It contains 20 MRI and 20 CT cardiac images with manual segmentation annotations. The images in either domains are unpaired and are obtained from different anonymous patients and sites. Similar to SIFA data preparation and usage, we evaluate the adaptation process on segmentation of four structures: ascending aorta(AA), left atrium blood cavity(LA-blood/LAC), left ventricle blood cavity(LV-blood/LVC), and the myocardium of the left ventricle(LV-myo/MYO).

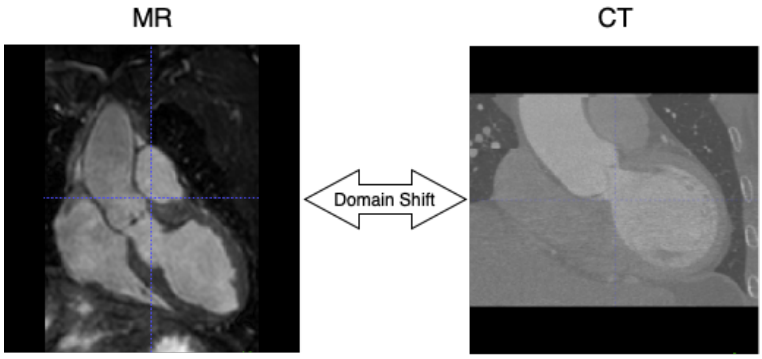


Figure 3.1: Severe cross-modality domain shift in cardiac imaging

To simplify the procedure and perform an equivalent evaluation like in SIFA, we use the released pre-processed data from PnP-AdaNet [14]. The pre-processing of samples includes normalizing all the data with zero mean and unit variance. The images were cropped to the size 256×256 and additionally augmentated with rotation, scaling and affine transforms. For the learning process, each modality was randomly split with 80% cases for training, and 20% cases for testing. MR scans are considered as the source domain, and CT scans are considered as the target domain. Please note, the

ground-truth of the target domain(i.e CT image scans) were used only for validation and testing phase.

Similar to the SIFA framework, for segmentation evaluation, Dice coefficient(%) and average surface distance (ASD) metrics were used. Higher Dice coefficient and Lower ASD indicated better segmentation performance.

Chapter 4

Methods

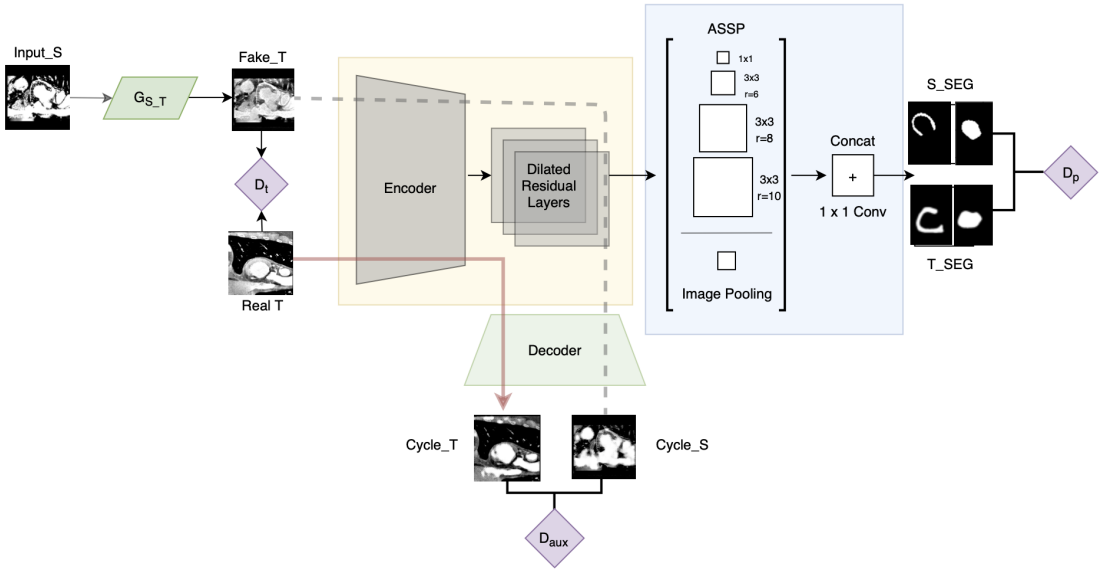


Figure 4.1: Improved segmentation network: SIFA + ASPP

Generator G_{S_T}: source $\rightarrow$ target transformation.
Encoder E + Decoder U: target $\rightarrow$ source transformation.
Encoder E + ASPP segmentor S: Segmentation mask predictions.

4.1 Segmentation module

As mentioned previously, to the best of our knowledge, all recent UDA for segmentation frameworks use DRNs as their segmentation module. Several segmentation improvements have emerged since DRNs, that capture multi-scale context from data easily.

A variation of atrous spatial pyramid pooling proposed in [10] is one such improvement which uses the design of DRNs as a context module and tool for adding multi-scale context modules. The ASPP module captures multi-scale information by resampling features at different scales using multiple atrous convolution rates as well as adopting

image-level features. We refer the reader to head to [10] for specific configurations and a detailed explanation of the design.

In the end, the module consists of the following parallel layer branches on top of the base DRN design (Figure 4.1):

1. One 1×1 convolution
2. Two 3×3 convolutions with $rates = (4, 6, 8)$ (Experimented with $rates = (6, 8, 10)$ as well. Got comparable results)
3. Image-level features i.e upsampled feature space obtained from global average pooling of the input feature map to module.

The resulting features from all the parallel branches are then concatenated and passed through one more convolution before the final classification layer.

4.2 Synergistic Image and Feature Adaptation

As mentioned before, in order to enable cross-modality cardiac segmentation, SIFA [6] proposes a synergistic integration of both the perspectives of adaptation, i.e. image adaptation and feature adaptation, in a single unified network, such that both aspects can mutually benefit each other during training (See figure 4.1).

4.2.1 Overview

Given source input domain X^s, set of labeled samples $\{x_i^s, y_i^s\}_{i=1}^N$, and target domain X^t, set of unlabeled samples $\{x_j^t\}_{j=1}^M$, the aim is to reduce the domain shift such that, the same segmentation network can be used to segment structures from either domain inputs.

4.2.2 Image Adaptation

The image adaptation process narrows the domain shift between the source and target domain by aligning the image appearances. This is achieved using GANs [20] for pixel-to-pixel image transformations of source images x^s to target-like images $x^{s \to t}$.

Specifically, a target generator G_{S_T} and a target discriminator D_t is built, forming a minimax two player component in the target domain $[x^{s\to t} = G_{S\_T}(x^s)]$.

The discriminator opposes the generator, trying to differentiate the fake image $x^{s\to t}$ from the real target image x^t. The adverserial loss L^t_{adv} is defined as :

$$\begin{aligned} L^t_{adv}(G_{S\_T}, D_t) &= E_{x^t\sim X^t}[\log D_t(x^t)] \\ &\quad + E_{x^s\sim X^s}[\log(1 - D_t(G_{S\_T}(x^s)))] \end{aligned}$$

The training of this network is guided by a cyclic consistency constraint using a reverse generator $G_s = E \circ U$ and source discriminator D_s. (E : encoder and U : upsampling decoder). This is necessary to preserve the semantic context of the input during adversarial training. The pair (G_s, D_s) is trained in the same way as (G_{S_T}, D_t) using adversarial loss L^s_{adv}:

$$\begin{aligned} L^s_{adv}(G_s, D_s) &= E_{x^s\sim X^s}[\log D_s(x^s)] \\ &\quad + E_{x^{s\,!\,t}\sim X^{s\,!\,t}}[\log(1 - D_s(G_s(x^{s\to t})))] \end{aligned}$$

The cyclic consistency loss is defined as follows:

$$\begin{aligned} L_{cyc}(G_{S\_T}, E, U) &= E_{x^s\sim X^s} \,||\, U(E(G_{S\_T}(x^s))) - x^s \,|| \\ &\quad + E_{x^t\sim X^t} \,||\, G_{S\_T}(U(E(x^t))) - x^t \,|| \end{aligned}$$

In short, the aim is to obtain : $U(E(G_{S\_T}(x^s))) \simeq x^s$; $G_{S\_T}(U(E(x^t))) \simeq x^t$.

Ideally this should bring $x^{s\to t}$ closer to the data distribution of target domain. With that assumption the segmentation network is trained with synthetic target-like images. To elaborate, the feature maps extracted from $E(x^{s\to t})$ are input to the pixel-level classifier (in our case, ASPP module), S for predicting segmentation masks $\hat{y}^{s\to t} = S(E(x^{s\to t}))$.

The segmentation loss to be optimized is:

$$L_{seg}(E, S) = H(y^s, \hat{y}^{s\to t}) + \alpha \cdot Dice(y^s, \hat{y}^{s\to t})$$

where, H is the weighted cross-entropy loss from ground-truth, predicted segmentation mask, $Dice$ is the Dice Loss, and α is the weighting hyper-parameter.

4.2.3 Feature Adaptation

When domain shift is severe, like in the case of MRI scans & CT scans, only image adaptation can be insufficient to achieve the desired performance. One needs to address the remaining domain gap between the synthesized target images and real target images using feature adaptation i.e. enabling the network to extract domain invariant features such that it becomes indistinguishable or difficult to interpret, which domain it came from.

It is worth noting that the feature space extracted by the DCNNs will be of very high dimension. It can be difficult to learn the mappings from the specific domain distribution X to the underlying domain-invariant feature space Z. To overcome this problem, two independent discriminators, $D_s and D_p$ are used to distinguish the domain invariance in two lower dimensional spaces: Semantic Segmentation space and Auxiliary Feature space.

Discriminator D_p distinguishes whether the predicted segmentation mask comes from $x^{s \to t}$ or x^t. Since segmentations are supposedly simple anatomical structures, if the segmenter receives features that are invariant, the discriminator would fail to tell apart the domain of the predicted masks.

The adversarial loss for the semantic segmentation space is:

$$\begin{aligned} L^p_{adv}(E, S, D_p) &= E_{x^{s\,!\,t} \sim X^{s\,!\,t}}[\log\left(D_p(S(E(x^{s \to t})))\right)] \\ &\quad + E_{x^t \sim X^t}[\log\left(1 - D_p(S(E(x^t)))\right)] \end{aligned}$$

While the cyclic consistency constraint helps preserve the semantic information during pixel to pixel transformation, it can also be used to remove any remaining domain characteristics in the latent feature space Z. Specifically, an auxiliary task is added to the source discriminator D_s to distinguish whether the generated source-like images $x^{t \to s}$ came from real target images x^t or generated target-like images $x^{s \to t}$.

The domain invariance training in the auxiliary feature space is guided by the following adversarial loss:

$$
\begin{aligned}
L_{adv}^{aux}(E, D_s) &= E_{x^{s\to t}\sim X^{s\to t}}[\log(D_s(U(E(x^{s\to t}))))] \\
&\quad + E_{x^t\sim X^t}[\log(1 - D_s(U(E(x^t))))]
\end{aligned}
$$

Overall, we enforce the encoder to extract domain invariant features by training it from two perspectives: segmentation prediction (High-level semantics); generated image space(low-level appearance).

Notice that the encoder E is shared for both image adaptation and feature adaptation. This allows us to use the encoder simultaneously to extract features and transform image appearances in a multi-task scenario - a synergistic merge indeed.

Since both the adaptation processes are independent of each other, they form independent training graphs allowing us to train the entire network in an end-to-end manner. The final goal of the framework is to reduce the loss function:

$$
\begin{aligned}
L &= L_{adv}^{t}(G_{S\_T}, D_t) + \lambda_{adv}^{s} L_{adv}^{s}(E, U, D_s) + \lambda_{cyc} L_{cyc}(G_{S\_T}, E, U) \\
&\quad + \lambda_{seg} L_{seg}(E, S) + \lambda_{adv}^{p} L_{adv}^{p}(E, S, D_p) + \lambda_{adv}^{aux} L_{aux}^{aux}(E, D_s)
\end{aligned}
$$

where all the λs are the weighting hyper-parameters. We refer the reader to head to [10] for specific network configurations and a detailed explanation of the design.

4.2.4 Implementation Details

The entire network including the SIFA framework + ASPP segmentation module is implemented in PyTorch 1.2 (*Python 3.6*). Both independent modules are configured as proposed in [6] and [10], respectively.

The learning protocols of SIFA framework are used for training the network. This makes it easy for us to compare the results between the TensorFlow implementation & PyTorch implementation. It also enables us to correctly evaluate the additional segmentation improvements. Overall, the training parameters are as follows:

1. Adaptation module: we use the Adam optimizer with a learning rate of 2×10^{-4} with no decay.

2. Segmentation module: we use another Adam optimizer with a learning rate of 1×10^{-3} with a step decay of 0.9 every 2 epochs.

3. During testing, the image scan from the target domain is given directly as input to the domain invariant encoder E, followed by the segmenter S, to obtain the segmentation prediction.

Chapter 5

Results

For the purpose of this thesis, we compare our PyTorch implementation with the TensorFlow implementation of SIFA [5] framework without ASPP. Table 5.1 reports the comparison results. We can see that the PyTorch implementation of SIFA gives a slightly reduced performance. Since completely reproducible results can't be guaranteed across deep-learning libraries and across CPU/GPU platforms, our best guess is that we need better hyper-parameter tuning.

Nevertheless, when we integrate our SIFA implementation with the ASPP module, there is a significant boost in the Dice score for the Left Atrium blood cavity (LAC) and the Left Ventricle blood cavity (LVC). This demonstrates the effectiveness of multi-scale context modules for segmentation.

Evaluation										
Methods	Dice					ASD				
	AA	LAC	LVC	MYO	Average	AA	LAC	LVC	MYO	Average
SIFA	81.1	76.4	75.7	58.7	73.0	10.6	7.4	6.7	7.8	8.1
SIFA(PyTorch)	79.6	70.0	74.8	54.0	69.6	10.3	6.2	5.6	7.3	7.3
Proposed	79.6	76	77.2	55.3	72.02	10.8	7.1	5.6	7.3	7.7

Table 5.1: Evaluation Results

Performance comparison between our method(SIFA + ASPP) and SIFA for the task of cardiac cross-modality segmentation.

Chapter 6

Conclusion

The thesis proposes adding the atrous spatial pyramid pooling to the segmentation submodule of the domain adaptation network, enabling it to capture multi-scale context effectively and thus improving the segmentation performance.

We also provide a PyTorch implementation of SIFA, the current SoA domain adaptation framework for cross-modality cardiac segmentation. This is with the hope of encouraging further research among scientists who prefer PyTorch's dynamic ecosystem.

The method is validated on unpaired MRI to CT image adaptation for cardiac segmentation and compared with the base SIFA module. Our method is general and can be extended to other segmentation appplications in part (i.e SIFA and ASPP) or in whole for unsupervised domain adaptation. The code is publicly available at *https://github.com/rushin682/SIFA-PyTorch.git*.

Chapter 7

Discussion

When thinking of solutions for improvements in individual components of domain adaptation, a good start is to think of approaches that can simplify the framework without affecting the adaptation process. From our own experimental experiences, for effective model learning, it is necessary to have a refined dataset that doesn't include many outliers. Pre-processing can improve the training significantly.

It is also important to validate the robustness of the proposed method on multiple datasets. Several publicly available datasets exist for domain adaptation like the Multi-sequence Cardiac MR Segmentation Challenge 2019 [MS-CMR 2019] dataset [62]. We plan to explore optimal pre-processing techniques for medical image data as well as validate the robustness of our method on MS-CMR 2019 dataset in the near future.

With emerging research in Neural Architecture Search [33] beyond image classification, building a domain adaptation framework should be plausible. With this note, we encourage researchers to move ahead toward developing a general domain adaptation search space that can reduce and possibly eliminate the multi-domain difference.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation, 2015, 1511.00561.

[2] M. Bateson, J. Dolz, H. Kervadec, H. Lombaert, and I. B. Ayed. Constrained domain adaptation for segmentation, 2019, 1908.02996.

[3] S. Benaim and L. Wolf. One-shot unsupervised cross domain translation, 2018, 1806.06029.

[4] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks, 2016, 1612.05424.

[5] C. Chen. Sifa-tensorflow: https://github.com/cchen-cc/sifa.git.

[6] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation, 2019, 1901.08211.

[7] L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform, 2015, 1511.03328.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs, 2014, 1412.7062.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2016, 1606.00915.

[10] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation, 2017, 1706.05587.

[11] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation, 2015, 1511.03339.

[12] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation, 2015, 1503.01640.

[13] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015.

[14] Q. Dou, C. Ouyang, C. Chen, H. Chen, B. Glocker, X. Zhuang, and P.-A. Heng. Pnp-adanet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation, 2018, 1812.07907.

[15] Q. Dou, C. Ouyang, C. Chen, H. Chen, and P.-A. Heng. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss, 2018, 1804.10916.

[16] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013.

[17] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *2009 IEEE 12th International Conference on Computer Vision*, pages 670–677, 2009.

[18] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks, 2015, 1505.07818.

[19] G. Ghiasi and C. C. Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation, 2016, 1605.02264.

[20] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014, 1406.2661.

[21] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1458–1465 Vol. 2, 2005.

[22] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization, 2014, 1411.5752.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015, 1512.03385.

[24] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation, 2017, 1711.03213.

[25] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian. A real-time algorithm for signal analysis with the help of the wavelet transform., 1989.

[26] M. A. Islam, M. Rochan, S. Naha, N. D. B. Bruce, and Y. Wang. Gated feedback refinement network for coarse-to-fine dense semantic image labeling, 2018, 1806.11266.

[27] K. Kamnitsas, C. Baumgartner, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks, 2016, 1612.08894.

[28] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials, 2012, 1210.5644.

[29] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2169–2178, 2006.

[30] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, 2016, 1611.06612.

[31] G. Lin, C. Shen, A. van dan Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation, 2015, 1504.01013.

[32] G. J. S. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *CoRR*, abs/1702.05747, 2017, 1702.05747.

[33] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. Yuille, and L. Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation, 2019, 1901.02985.

[34] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better, 2015, 1506.04579.

[35] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. Semantic image segmentation via deep parsing network, 2015, 1509.02634.

[36] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation, 2014, 1411.4038.

[37] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 97–105. PMLR, 2015.

[38] P. Luo, G. Wang, L. Lin, and X. Wang. Deep dual learning for semantic image segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2737–2745, 2017.

[39] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features, 2014, 1412.0774.

[40] C. Ouyang, K. Kamnitsas, C. Biffi, J. Duan, and D. Rueckert. Data efficient unsupervised domain adaptation for cross-modality image segmentation, 2019, 1907.02766.

[41] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly- and semi-supervised learning of a dcnn for semantic image segmentation, 2015, 1502.02734.

[42] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. Large kernel matters – improve semantic segmentation by global convolutional network, 2017, 1703.02719.

[43] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe. Full-resolution residual networks for semantic segmentation in street scenes, 2016, 1611.08323.

[44] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015, 1505.04597.

[45] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo. From source to target and back: symmetric bi-directional adaptive gan, 2017, 1705.08824.

[46] A. G. Schwing and R. Urtasun. Fully connected deep structured networks, 2015, 1503.02351.

[47] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks, 2013, 1312.6229.

[48] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[49] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation, 2018, 1802.10349.

[50] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation, 2017, 1702.05464.

[51] A. van Opbroek, M. A. Ikram, M. W. Vernooij, and M. de Bruijne. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE Transactions on Medical Imaging*, 34(5):1018–1030, 2015.

[52] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net, 2015, 1511.06881.

[53] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions, 2015, 1511.07122.

[54] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks, 2017, 1705.09914.

[55] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks, 2013, 1311.2901.

[56] Y. Zhang, S. Miao, T. Mansi, and R. Liao. Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation, 2018, 1806.07201.

[57] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei. Fully convolutional adaptation networks for semantic segmentation, 2018, 1804.08286.

[58] H. Zhao, H. Li, S. Maurer-Stroh, Y. Guo, Q. Deng, and L. Cheng. Supervised segmentation of un-annotated retinal fundus images by synthesis. *IEEE Transactions on Medical Imaging*, 38(1):46–56, 2019.

[59] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network, 2016, 1612.01105.

[60] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015.

[61] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

[62] X. Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):2933–2946, 2019.

[63] X. Zhuang and J. Shen. Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. *Medical Image Analysis*, 31:77 – 87, 2016.