UNIVERSITY OF CALIFORNIA SAN DIEGO

Attentive Representations for Objects Detection and Instance Segmentation

A thesis submitted in partial satisfaction of the
requirements for the degree Master of Science

in

Electrical Engineering
(Intelligent System, Robotics and Control)

by

Xudong Wang

Committee in charge:

Professor Nuno Vasconcelos, Chair
Professor Hao Su
Professor Mohan M. Trivedi
Professor Zhuowen Tu

2019

The thesis of Xudong Wang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

Chair

University of California San Diego

2019

DEDICATION

To my parents:

Xiurong Wang, Wei Wang.

EPIGRAPH

Don't judge each day by the harvest you reap but by the seeds that you plant

—Robert Louis Stevenson

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I hope that I can take this opportunity to express my gratitude and thanks to my family, friends, colleagues and advisors who helped me during my master's degree. Without their help, it is impossible for me to finish the graduate work and enjoy the life in San Diego, their support benefits me a lot.

First of all, I want to express my sincerely thanks to my supervisor, Professor Nuno Vasconcelos. During past research journy in statistical visual and computing lab, his serious research attitude and splendid research ideas led me into this whole new world for me, I feel so fortunate to have him as my first mentor. His deep understanding and mastery of fundamental problems of computer vision, strong passion in research, keen insight into emerging research areas and hard-working attitude benefits me a lot. I can still remember he modified my conference manuscript from late evening until the first light of rising sun touched the land. He is a prominent role model to me, always have, always will.

I also really appreciate the help from Dr. Dashan Gao. His great leadership, rigorous and serious research attitude and optimistic attitude towards life gave me a very deep impression. He has been working to promote the use of artificial intelligence and computer vision in industry and medical applications. The year I worked under his supervision has boosted my research a lot.

During the past several years in statistical visual and computing lab, many colleagues and friends in SVCL helped me a lot: Zhaowei Cai, Yi Li, Pei Wang, Bo Liu, Pedro Morgado, Yunsheng Li, John Ho and Zhihang Ren. I started work with PhD candidate, Bo Liu, I co-authored the first paper in SVCL with him. He taught me the fundamental knowledge and coding skills in computer vision field. Dr.Cai is the person who leads me into objects detection, I learned a lot from him. Many ideas and valuable suggestions from him boosted my research and inspired me, we co-authored another paper together. Pei Wang and Yi Li joined the lab one year later than me, we talked with each other quite a lot and enjoyed many beautiful lunch time together. I want to show thanks to Pedro Morgado, he took a lot of time on maintaining the GPU server and linux

server, which provide solid backup for all the members in SVCL. I would like to thank them all and give my best wishes to their future research and career.

I also want to say thank you to my close friends, Ziyao Tang and Haifeng Huang in San Diego, I will miss the weekly dinner with them. We three almost tried every Chinese restaurants in San Diego together, every Friday night, the delicious dinner with them will always dispel all the unhappiness away from me. I would like to thank the colleagues, Dr. Dashan Gao, Dr. Yunqiang Chen, Xin Zhong, Dr. Patrick Langechuan Liu, Dr. Darryl Lin, Yuanpeng Wu, Weiwei Liu, Dr. Jiao Wang, Dr. Shizhong Han, Dr. Yunxiang Mao, Nariaki Yamada and Jing Cai in 12 Sigma Technologies for their support, friendship and assistant during my internship period.

The last but very important, I want to appreciate the support from my family. No words can express, no act of gratitude can relay, no gift can represent what their encouragement and support have meant to me.

Chapter 2, in part, is a reprint of the material as it appears in In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. Xudong Wang, Zhaowei Cai, Dashan Gao, Nuno Vasconcelos., IEEE, 2019. The thesis author was the primary investigator and author of this paper.

Chapter 3, in part, has been submitted for publication of the material as it may appear in International Conference on Medical Image Computing and Computer Assisted Intervention(MICCAI), 2019, Xudong Wang, Zhaowei Cai, Dashan Gao, Nuno Vasconcelos., Springer, 2019. The thesis author was the primary investigator and author of this paper.

VITA

2016	B. S. in Engineering, Jilin University, Jilin, China
2019	Master of Science, Electrical Engineering, Intelligent System, Robotics and Control, University of California, San Diego

PUBLICATIONS

Xudong Wang, Shizhong Han, Dashan Gao, Nuno Vasconcelos. Volumetric Attention for 3D Medical Image Segmentation and Detection. *under review in International Conference on Medical Image Computing and Computer Assisted Intervention(MICCAI)*, 2019

Xudong Wang, Zhaowei Cai, Dashan Gao, Nuno Vasconcelos. Towards Universal Object Detection by Domain Attention.*In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2019.

Bo Liu, Xudong Wang, Mandar Dixit, Roland Kwitt, Nuno Vasconcelos. Feature Space Transfer for Data Augmentation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018

ABSTRACT OF THE THESIS

Attentive Representations for Objects Detection and Instance Segmentation

by

Xudong Wang

Master of Science in Electrical Engineering
(Intelligent System, Robotics and Control)

University of California San Diego, 2019

Professor Nuno Vasconcelos, Chair

In this thesis, we focused on investigating novelty modules integrated into popular detection network for assisting it to learn attentive representations for several practical applications in objects detection and instance segmentation tasks, including universal object detection and 3D medical image segmentation tasks. For universal object detection task, despite increasing efforts on universal representations for visual recognition, few have addressed object detection. In this thesis, we develop an effective and efficient universal object detection system that is capable of working on various image domains, from human faces and traffic signs to medical CT images. Unlike multi-domain models, this universal model does not require prior knowledge

of the domain of interest. This is achieved by the introduction of a new family of adaptation layers, based on the principles of squeeze and excitation, and a new domain-attention mechanism. In the proposed universal detector, all parameters and computations are shared across domains, and a single network processes all domains all the time. Experiments, on a newly established universal object detection benchmark of 11 diverse datasets, show that the proposed detector outperforms a bank of individual detectors, a multi-domain detector, and a baseline universal detector, with a $1.3\times$ parameter increase over a single-domain baseline detector. For 3D medical images segmentation tasks, although high resolution 3D medical images offer abundant detail information of human body parts and allow early detection of small lesions, due to the limitation of GPU memory, most methods either use down-sampled 3D volume as input, which significantly affects the detectability of small lesions, or use 2.5D networks to crop out neighboring image slices at original resolution, which loses context information along z direction. Both ways can significantly affect the performance of final model. In this paper, we propose a cross-slice spatial and channel attention module, which can maintain spatial resolution of input data, and effectively utilize context information along z direction of 3D volume. In order to get higher quality mask prediction, a cascade mask refinement module is designed to provide an objectiveness pixel-wise attention map for input feature maps. Furthermore, our scheme allows us to utilize the pretrained 2D detection models to achieve good results even with limited amount of training data, which is often met in medical applications and imposes big challenge to many deep learning methods. By utilizing the two novel modules, we achieve state-of-art performance 74.10 dice per case on Liver Tumor Segmentation Challenge(LiTS), which outperforms previous year challenge winner by 6.7 points and rank as 1st on leader board of LiTS benchmark upon submission of this paper.

Chapter 1

Introduction

1.1 2D Objects Detection

There has been significant progress in object detection in recent years [Gir15, RHGS15, CFFV16, LDG+17, HGDG17, CV18], powered by the introduction of convolution network(ConvNet) and the availability of challenging and diverse object detection datasets, e.g. PASCAL VOC [EEVG+15], COCO [LMB+14], KITTI [GLU12], WiderFace [YLLT16], etc. Generally, ConvNet based object detection network can be divided as two-stage based detector and single-stage based detector, which will be described as followings.

R-CNN[GDDM14] utilizes a selective search method to generate limited amount of proposals, instead of using sliding windows methods with huge number of regions, for regressing. ConvNet will be used for feature extrction followed by a SVM as regionwise proposals classifier to get per-proposal category prediction. However, due to the fact that selective search method is not trainable, proposal quality is not satisfactory. Also, the testing and training speed of R-CNN is very slow and is not able to be used as real-time detector. To bypass above limitations of R-CNN, Fast R-CNN[Gir15] is proposed. Fast R-CNN will identify region of interests from heat maps generated by feeding images to CNN layers, the trainable CNN layers boosts proposal qualities of Fast R-CNN compared with R-CNN. For speeding up inference time, SVM is replace by a fully connected layers and a softmax layer, which can process all ROIs extracted features together, therefore, inference time can decrease a lot. Faster R-CNN[RHGS17] further increased training and inference speed by using CNN layers called region proposal network(RPN) to provide proposal predictions, predicted regions proposals are used for extracting interested features regions with ROI pooling operator and the second stages will jointly do classification and bounding box offsets prediction. Many works have expanded Faster R-CNN base architecture. For example, MS-CNN [CFFV16] and FPN [LDG+17] built a feature pyramid to effectively detect objects of various scales; the R-FCN [DLHS16] proposed a position-sensitive pooling to achieve further speed-ups; and the Cascade R-CNN [CV18] introduced a multi-stage cascade for

high quality object detection.

In parallel, single-stage object detectors, such as YOLO [RDGF16] and SSD [LAE+16], became popular for their fairly good performance and high speed. [RDGF16] proposed much faster network with single stage which predicts bounding boxes and class probabilities directly from full images in one evaluation, however, the performance YOLO is still much worse than Faster R-CNN. SSD[LAE+16] will provide category scores and box offset of different scales from different feature scales, they also proposed new matching strategies, based on these method, SSD reached much higher performance than YOLO with competitive inference and training speed. Some anchor free detectors are also proposed for realizing comparable performance with complicated two-stage detector and meanwhile, maintaining the speed adavantages of one-stage detector. CornerNet[LD18] proposed a single-stage network which gives the bounding box prediction by predicting the bottom-right and top-left corners, instead of regressing manually designed anchor boxes as other single-stage detector, such as YOLO[RF18] and SSD[LAE+16]. CenterNet[ZWK19] will consider object detection problems as key point prediction problems, the main task will be predicting center point of objects bounding boxes and regressing object size(width and height of bounding box), dimension(2D or 3D), 3D context, orientation and some other related properties of target objects. Due to the simplicity of anchor free methods, they are able to reach pretty good speed-accuracy trade-off on objects detection tasks.

However, all these network need to be fine-tuned on target domain if they want to reach good performance on target datasets. Therefore, none of these detectors could reach high detection performance on more than one dataset/domain without finetuning. In the pre-deep learning era, [KZM+12] proposed a universal DPM [FGMR10] detector, by adding dataset specific biases to the DPM. But this solution is limited since DPM is not comparable to deep learning detectors.

1.2 3D Objects Segmentation

A natural solution to 3D medical image segmentation and detection problems is to rely on 3D convolutional networks, such as the 3D U-Net of [ÇAL+16] or the extended 2D U-Net of [RFB15]. However, current GPU memory limitations prevent the processing of 3D volumes with high resolution. This is problematic, because the use of low-resolution volumes leads to low precision or miss-detection of small lesions and tumors and blur in lesion mask predictions, especially on boundaries. Hence, there is a need to trade-off the spatial resolution of each 2D slice for the number of slices processed. This implies a trade-off between the precision with which segmentation or detection can be performed and the amount of contextual information, in the z direction, that can be leveraged. A popular solution is to a use a 2D network to segment or detect the structures of interest in 2D or 2.5D slices and then concatenate the results to build a 3D segmentation mask or bounding box.

Christ et al. proposed a 2D U-Net for liver and tumor segmentation, followed by a conditional random field for segmentation refinement [CEE+16]. Li et al. proposed a hybrid Dense 2D/3D UNet of three-stages [LCQ+18]. They found that a pre-trained 2D model can significantly boost performance of their network. Bi et al. proposed a two-stage cascaded deep residual network for liver lesion segmentation [Han17]. These approaches are limited by the lack of contextual information. Since even human experts need to inspect multiple slices to reach confident assessments of confusing lesions, this is likely to upper bound their performance. Ding et al. applied 2D networks to generate lesion candidates, then 3D CNN classifiers were trained for false positive reduction (FPR). To address this problem, Yan et al. [YBS18] proposed a 3D context enhanced region-based CNN. However, their method is based on a region proposal network (RPN) and cannot be implemented as a single-stage detector, such as SSD and YOLO, or a segmentation network, such as U-Net, without an RPN component. Furthermore, because only the feature map derived from a central image is processed by the RPN to generate proposals,

the proposal generation process has no access to 3D context. Given that missed proposals can not be recovered, this places an upper bound on detection performance.

1.3 Multi-Domain Learning/Adaptation

Multi-domain learning (MDL) addresses the learning of representations for multiple domains, known a priori [JCDR12, NH16a]. It uses a combination of parameters that are shared across domains and domain-specific parameters. The latter are adaptation parameters, inspired by works on domain adaptation [PGLC15, LCWJ15, RT18, MDL18], where a model learned from a source domain is adapted to a target domain. [THZ+14] proposed deep domain confusion(DDC) maximum module for learning domain invariant representations for both target and source domains, which is powered by MMD based domain confusion loss and cross-entropy loss. However, they will only add one domain adaptation layer, which may not be powerful enough to solve domain adaptation problem, [LCWJ15] proposed Deep Adaptation Network(DAN) architecture for learning transferable features with multiple kernels and layers, [YCBL14] shows that the deeper layers will tend to learn domain-specific bias and domain discrepancy will drops significantly along the transition from higher to lower layer of CNN network. DAN is composed of several MK-MDD(multi-kernel variant of maximum mean discrepancies) module which adopted mean embedding of different domain distributions for enhancing test power and decreasing test error[LCWJ15]. For the recent years, generative adversarial net(GAN)[GPAM+14] is proposed for estimating generative models via an adversarial process. [THSD17] utilized GAN for training a domain discriminator and a target encoder for domain adaptation. CycleGAN [ZPIE17] introduce the constraint of cycle-consistency to regularize the GAN model without necessarily having a one-to-one mapping between images from input to target domain in the training set. The Domain Transfer Network[TPW16] proposed an unsupervised image style transfer network trained with pixel-wise loss and semantic loss. CyCADA [HTP+17] applied feature loss and pixel loss for

adapting the learned representations at pixel and feature levels. In order to maintain structural consistency, similar to [THSD17], cycle-consistency is used for regularizing model and semantic loss is also applied for maintaining semantic information.

In multi-domain learning field, [BV17] showed that multi-domain learning is feasible by simply adding domain-specific BN layers to an otherwise shared network. [RBV17] learned multiple visual domains with residual adapters, while [RBV18] empirically studied efficient parameterizations. However, they build on BN layers and are not suitable for detection, due to the batch constraints of detector training. Instead, we propose an alternative SE adapters, inspired by "Squeeze-and-Excitation" [HSS17], to solve this problem. Another similar topic is about Multi-task learning (MTL). MTL investigates how to jointly learn multiple tasks simultaneously, assuming a single input domain. Various multi-task networks [Kok17, ZSS+18, HGDG17, LSNK17, WSL+15, ZY17] have been proposed for joint solution of tasks such as object recognition, object detection, segmentation, edge detection, human pose, depth, action recognition, etc., by leveraging information sharing across tasks. However, the sharing is not always beneficial, sometimes hurting performance [EMP05, KKSA08]. To address this, [MSGH16] proposed a cross-stitch unit, which combines tasks of different types, eliminating the need to search through several architectures on a per task basis. [ZSS+18] studied the common structure and relationships of several different tasks.

1.4 Contributions of the Thesis

1.4.1 Universal representations for objects detection

There has been significant progress in object detection in recent years [Gir15, RHGS15, CFFV16, LDG+17, HGDG17, CV18], powered by the availability of challenging and diverse object detection datasets, e.g. PASCAL VOC [EEVG+15], COCO [LMB+14], KITTI [GLU12], WiderFace [YLLT16], etc. However, existing detectors are usually *domain-specific*, e.g. trained

and tested on a single dataset. This is partly due to the fact that object detection datasets are diverse and there is a nontrivial domain shift between them. As shown in Figure 1.1, detection tasks can vary in terms of categories (human face, horse, medical lesion, etc.), camera viewpoints (images taken from aircrafts, autonomous vehicles, etc.), image styles (comic, clipart, watercolor, medical), etc. In general, high detection performance requires a detector specialized on the target dataset.

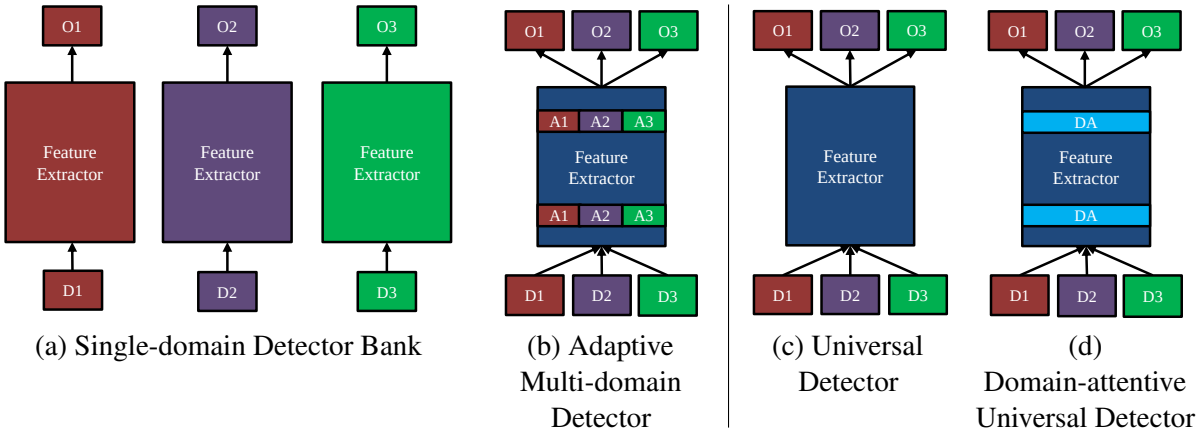
This poses a significant problem for practical applications, which are not usually restricted to any one of the domains of Figure 1.1. Hence, there is a need for systems capable of detecting objects regardless of the domain in which images are collected. A simple solution is to design a *specialized* detector for each domain of interest, e.g. use D detectors trained on D datasets, and load the detector specialized to the domain of interest at each point in time. This, however, may be impractical, for two reasons. First, in most applications involving autonomous systems the domain of interest can change frequently and is not necessarily known a priori. Second, the overall model size increases linearly with the number of domains D. A recent trend, known as general AI, is to request that a single universal model solves multiple tasks [KGS+17, Kok17, ZSS+18], or the same task over multiple domains [RBV17, BV17]. However, existing efforts in this area mostly address image classification, rarely targeting the problem of object detection. The fact that modern object detectors are complex systems, composed of a backbone network, proposal generator, bounding box regressor, classifier, etc., makes the design of a universal object detector much more challenging than a universal image classifier.

In this paper, we address the design of an object detector capable of detecting objects from multiple domains, e.g. the 11 datasets of Figure 1.1. Figure 1.2 summarizes a number of architectures that can be used to address the problem. In Figure 1.2, "D" is the domain, "O" the output, "A" domain-specific adapter, and "DA" the proposed domain attention module. The blue color and the DA are domain-universal, but the other colors domain-specific. The left two are multi-domain detectors, requiring prior knowledge of the domain of interest. The right two

Figure 1.1: Samples of our universal object detection benchmark.

are universal detectors, with no need for such knowledge. When operating on an unknown domain, the multi-domain detector have to repeat the inference process with different sets of domain-specific parameters, while the universal detector performs inference only once.

The detector of Figure 1.2 (a) is a bank of domain-specific detectors, with no sharing of parameters/computations. Multi-domain learning (MDL) [JCDR12, NH16b, KCK+17, YH14, JZCL08, DKC10] improves on this, by sharing parameters across various domains, and adding small domain-specific layers. In [RBV17, BV17], expensive convolutional layers are shared and complemented with light-weight domain-specific *adaptation layers*, implemented with a combination of batch normalization (BN) [IS15] and ResNet-style residual layers [HZRS16]. These are not practical for object detection, due to the constraints on BN in this setting. Instead, we propose a new class of light adapters, based on the squeeze and excitation (SE) mechanism of [HSS17], and denoted *SE adapters*. This leads to the *multi-domain detector* of Figure 1.2 (b), where domain-specific SE adaptation layers are introduced throughout the network to compensate



(a) Single-domain Detector Bank (b) Adaptive Multi-domain Detector (c) Universal Detector (d) Domain-attentive Universal Detector

Figure 1.2: Multi-domain and universal object detectors for three domains.

for domain shift. On 11 datasets, this detector outperforms Figure 1.2 (a) with ∼5 times less parameters.

In contrast, the *universal detector* of Figure 1.2 (c) shares all parameters/computations[1] across domains. It consists of a *single* network, which is always active. This is the most efficient solution in terms of parameter sharing, but it is difficult for a single model to cover many domains with nontrivial domain shifts. Hence, this solution underperforms the multi-domain detector of Figure 1.2 (b). To overcome this problem, we propose the *domain-attentive universal detector* of Figure 1.2 (d). This leverage a novel domain attention (DA) module based on the now proposed SE adapters. A bank of universal SE adapters, which are used at all times, is first added to the network. A feature-based attention mechanism is then introduced to achieve domain sensitivity. This mechanism learns to assign network activations to different domains automatically, soft-routing their responses through the different SE adapters. This enables the adapters to specialize on individual domains. Since the process is data-driven, the number of domains does not have to match the number of datasets and datasets can span multiple domains. This allows the network to leverage shared knowledge across domains, which is not available in the common single-domain detectors. Our experiments show that this data-driven form of parameter/computation sharing enables substantially better multi-domain detection performance than the remaining architectures

[1] other than output layers

of Figure 1.2. This allows the network to leverage the fact that some domains are close, e.g. the everyday objects of VOC [EEVG+15] and COCO [LMB+14], but some others are very distinct, e.g. the aerial objects of DOTA [XBD+18] and the medical images of DeepLesion [YWL+18].

1.4.2 3D context enhanced representations for 3D image segmentation

There has been significant progress in 2D object detection and segmentation in recent years [Gir15, RHGS15, LAE+16]. Mask-RCNN [HGDG17] tried to accomplish both object detection and 2D object instance segmentation tasks using one network by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. U-Net[RFB15] built upon fully convolutional network for 2D medical image segmentation. With 3D medical imaging, healthcare professionals can now access new angles, resolutions and details that offer an all-around better understanding of the body part in question, all while cutting the dosage of radiation for patients. Therefore developing a high accuracy network for segmentation and detection on 3D CT volume will be very necessary.

For 3D medical image segmentation and detection, one solution is using 2D network to do segmentation and detection on 2.5D or 2D slices, then concatenate all results together to build a 3D segmentation mask or give bounding box prediction on key slices. Another one is using 3D network directly, such as 3D U-Net[ÇAL+16] extended 2D U-Net[RFB15] to learn dense volumetric segmentation. Due to the limitation of GPU memory. If we want to use 3D network for 3D CT images detection and segmentation, low resolution CT volume will be used as input for 3D network to fit GPU memory, which will lead to low precision or miss-detection of small lesions and tumors and blur lesion mask prediction, especially on boundaries. If we want to keep the spatial resolution, we need to sacrifice the slices number, the advantage of utilizing contextual information in z direction will be damaged. Therefore, most researchers still choose to use the 2D architecture with 2D convolution for the 3D medical CT images segmentation. Christ et al. utilized a 2D U-Net for liver and tumor segmentation followed by a conditional random filed

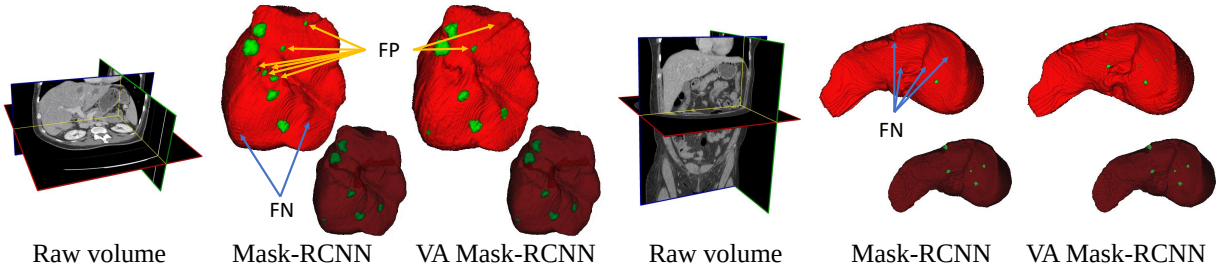


Figure 1.3: Examples of 3D segmentation results by Mask-RCNN and our proposed methods on the LiTS `val` set.

for segmentation refinement [CEE+16]. Vorontsov et al. proposed to jointly segment the liver and lesion together with two 2D U-Net [VTPK18]. Li et al. proposed a hybrid Dense 2D and 3D UNet with three-stage [LCQ+18]. They found that the 2D pretrained model can significantly boost the performance. Bi et al. proposed a two-stage cascaded deep residual network for liver lesion segmentation [Han17]. However even human experts need to look through multiple slices before making accurate analysis on confusing lesions. Yan et al. [YBS18] proposed 3D context enhanced region-base CNN to incorporate 3D context information and make bounding box prediction on key slices. However, their method is based on region proposal network(RPN), it is impossible to extend this method to one-stage detector, such as SSD and YOLO, or some segmentation network, such as U-Net, which do not contain RPN part. Also, only the feature map derived from the central image is sent to the RPN to generate lesion proposals and crop feature maps across multiple images, so RPN will not be able to benefit from 3D context, missed proposals can not be traced back.

In order to learn long-range dependencies between video frames and exploit spatial information within 3D video data, [WGGH18] added a non-local network to 3D convolutional network(C3D/I3D) for video classification, based on a spacetime dependency/attention mechanism. Inspired by [WGGH18], we proposed a flexible and computation efficient **C**ross-**S**lices **C**hannel and **S**patial attention(**CSCS**) module, which sequentially infers 3D enhanced attention maps along two separate dimensions, channel and spatial, then the attention maps are multiplied to the input feature map for adaptive feature refinement, based on 2D network. Similar to

[HSS17] and [WPLK18], global spatial pooling and global channel pooling are used for reducing computation cost. Our proposed module has several advantages: 1) It will be able to afford high spatial resolution images, meanwhile utilize 3D context information for segmentation and detection on 3D CT volume. 2) CSCS can be combined with any existing architectures, both one-stage/two-stage detectors and segmentation network can benefit from it. 3) Due to spatial and channel pooling, our proposed module is computation cost efficient. 4) This module can be added on customized positions, RPN can also benefit from our proposed block. 5) Because our module can be combined with 2D network, we can leverage pre-trained 2D CNN weights for transfer learning. As we can see from Fig.1.3, with our proposed module, the improved Mask-RCNN will not only reduce false positive segmentation, but also retrieve missed tiny lesions in original Mask-RCNN model. In Fig.1.3, the red area denote the segmented liver while the green ones denote the segmented lesions. Zoomed out 3D mask on the bottom right are ground truth mask, the dark red area denote the liver area while the dark green ones denote the lesions area.As we can see from left example, our method can detect all the lesions with only one false positive, Mask-RCNN not only misses one lesion but also has five false positive instance segmentation. For the right example, our method can find all the 5 tiny lesions, Mask-RCNN will miss 4 of them. These examples proved that our module will not only enhance small lesions prediction ability but also benefit the network to remove false positive cases.

Based on our proposed modules, we got state-of-art performance, 74.1 dice per case on LiTS liver tumor segmentation challenge `test` set, which outperforms $\sim$5 points than previous year challenge winner and got the 1st place on leaderboard upon submission of this paper. In order to prove the generalization ability of our methods on 3D CT volume, we further did extension experiments on DeepLesion dataset, which provide 3D CT volume to do 2D bounding box prediction on key slices, our proposed method got 77.22 sensitivity at 1 false positives(FPs) per image, which outperforms best published results by $\sim$ 4 points.

1.5 Organization of the Thesis

The rest of this thesis will be organized as follows, in Chapter 2 we will introduce the data-drive domain attentive representations for universal objects detection. We will also discuss the difference of single domain objects detector bank, adaptive multi-domain detector, universal detector and domain-attentive universal detector. The proposed universal objects detection benchmark(UODB) will also be discussed in this chapter. In chapter 3, we will focus on 3D medical image segmentation problems and discussed the proposed volumetric attention module which leverages 3D context enhanced representations with 3D medical CT volume as training set. To further prove the effectiveness of our volumetric attention module, we also did extension experiments on objects detection datasets, deeplesion, which provides neighbor slices of key slices with bounding boxes annotations. The final chapter 4 will draw conclusions on these algorithms and talk about the future works we are going to work on and the possible applications.

Chapter 2

Domain sensitive representations for universal objects detection

2.1 Multi-domain Object Detection

2.1.1 Multi-domain Datasets

To train and evalaute multi-domain object detection systems, we proposed to use 11 datasets: Pascal VOC [EEVG+15], WiderFace [YLLT16], KITTI [GLU12], LISA [MTM12], DOTA [XBD+18], COCO [LMB+14], Watercolor [IFYA18], Clipart [IFYA18], Comic [IFYA18], Kitchen [GRM+16] and DeepLesions [YWL+18].

This set includes the popular VOC [EEVG+15] and COCO [LMB+14], composed of images of everyday objects, e.g. bikes, humans, animals, etc. The 20 VOC categories are replicated on CrossDomain [IFYA18] with three subsets of Watercolor, Clipart and Comic, with objects depicted in watercolor, clipart and comic styles, respectively. Kitchen [GRM+16] consists of common kitchen objects, collected with an hand-held Kinect, while WiderFace [YLLT16] contains human faces, collected on the web. Both KITTI [GLU12] and LISA [MTM12] depict traffic scenes, collected with cameras mounted on moving vehicles. KITTI covers the categories of vehicle, pedestrian and cyclist, while LISA is composed of traffic signs. DOTA [XBD+18] is a surveillance-style dataset, containing objects such as vehicles, planes, ships, harbors, etc. imaged from aerial cameras. Finally DeepLesion [YWL+18] is a dataset of lesions on medical CT images. A representative example of each dataset is shown in Figure 1.1. Some more details are summarized in Table 2.1. Altogether, the 11 datasets cover a wide range of variations in category, camera view, image style, etc. They thus establish a good suite for the evaluation of multi-domain object detection.

2.1.2 Single-domain Detector Bank

The Faster R-CNN [RHGS15] is used as the backbone architecture of all detectors proposed in this work. As a single-domain object detector, the Faster R-CNN is implemented in two stages. First, a region proposal network (RPN) produces preliminary class-agnostic detection

Table 2.1: The dataset details, the domain-specific hyperparameters and the performance of the single-domain detectors. "T/V/T" means train/val/test, "size" the shortest side of inputs, *BS* RPN batch size, and S/R anchor "scales/aspect ratios".

dataset	dataset details			hyperparameters				mAP
	class	T/V/T	domain	size	*BS*	RoIs	S/R	
KITTI	3	7k/-/7k	traffic	576	256	128	12/3	64.3
WiderFace	1	13k/3k/16k	face	800	256	256	12/1	48.9
VOC	20	8k/8k/5k	natural	600	256	256	4/3	78.5
LISA	4	8k/-/2k	traffic	800	64	32	4/3	88.3
DOTA	15	14k/5k/10k	aerial	600	128	128	12/3	57.5
COCO	80	35k/5k/-	natural	800	256	256	4/3	47.3
Watercolor	6	1k/-/1k	watercolor	600	256	256	4/3	52.4
Clipart	6	0.5k/-/0.5k	clipart	600	256	256	4/3	32.1
Comic	20	1k/-/1k	comic	600	256	256	4/3	45.8
Kitchen	11	5k/-/2k	indoor	800	256	256	12/3	87.7
DeepLesion	1	23k/5k/5k	medical	512	128	64	12/3	51.3
Average	-	-	-	-	-	-	-	59.4

hypotheses. The second stage processes these with a region-of-interest detection network to output the final detections.

As illustrated in Figure 1.2 (a), the simplest solution to multi-domain detection is to use an independent detector per dataset. We use this detector bank as a multi-domain detection baseline. This solution is the most expensive, since it implies replicating all parameters of all detectors. Figure 2.1 shows the statistics (mean and variance) of the convolutional activations of the 11 detectors on the corresponding dataset. Some observations can be made. First, these statistics vary non-trivially across datasets. While the activation distributions of VOC and COCO are similar, DOTA, DeepLesion and CrossDomain have relatively different distributions. Second, the statistics vary across network layers. Early layers, which are more responsible for correcting domain shift, have more evident differences than latter layers. This tends to hold up to the output layers. These are responsible for the assignment of images to different categories and naturally differ. Interestingly, this behavior also holds for RPN layers, even though they are category-independent. Third, many layers have similar statistics across datasets. This is especially true for intermediate layers, suggesting that they can be shared by at least some domains.

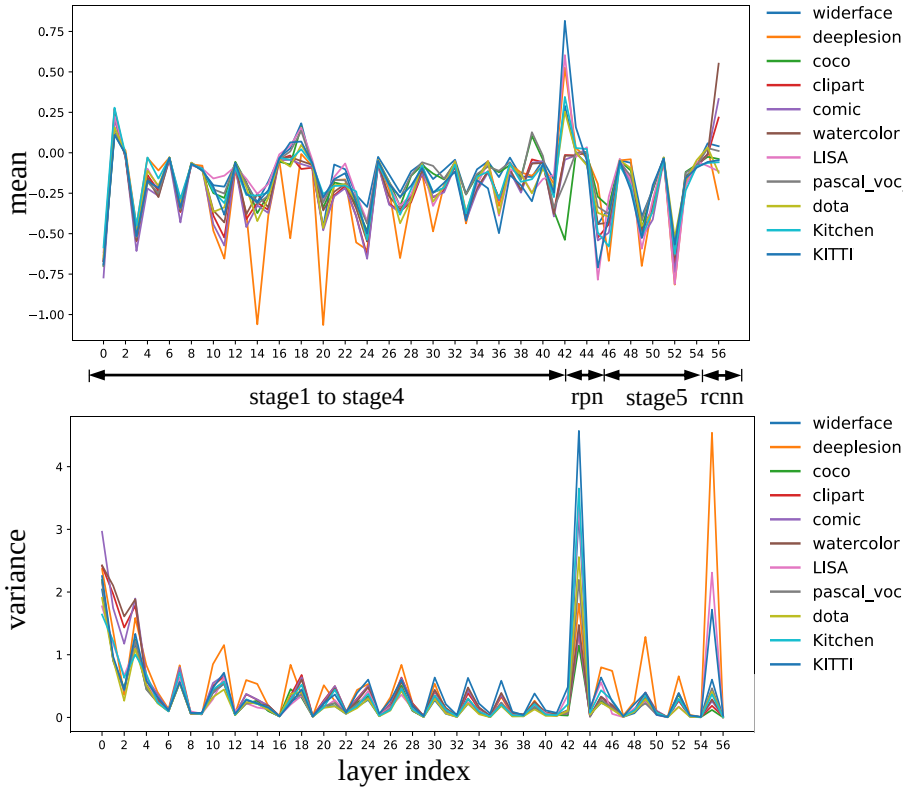


Figure 2.1: The statistics of each convolutional activation of all single-domain detectors.

2.1.3 Adaptive Multi-domain Detector

Inspired by Figure 2.1, we propose an adaptive multi-domain detector, shown in Figure 1.2 (b). In this model, the output and RPN layers are domain-specific. The remainder of the newtork, e.g. all convolutional layers, is shared. However, to allow adaptation to new domains, we introduce some additional domain-specific layers, as is commonly done in MDL [RBV17, BV17]. These extra layers should be 1) sufficiently powerful to compensate for domain shift; 2) as light as possible to minimize parameters/computation. The adaptation layers of [RBV17, BV17] rely extensively on batch normalization. This is unfeasible for detection, due to the small batch sizes allowable for detector training. For detection, BN layers have to be frozen.

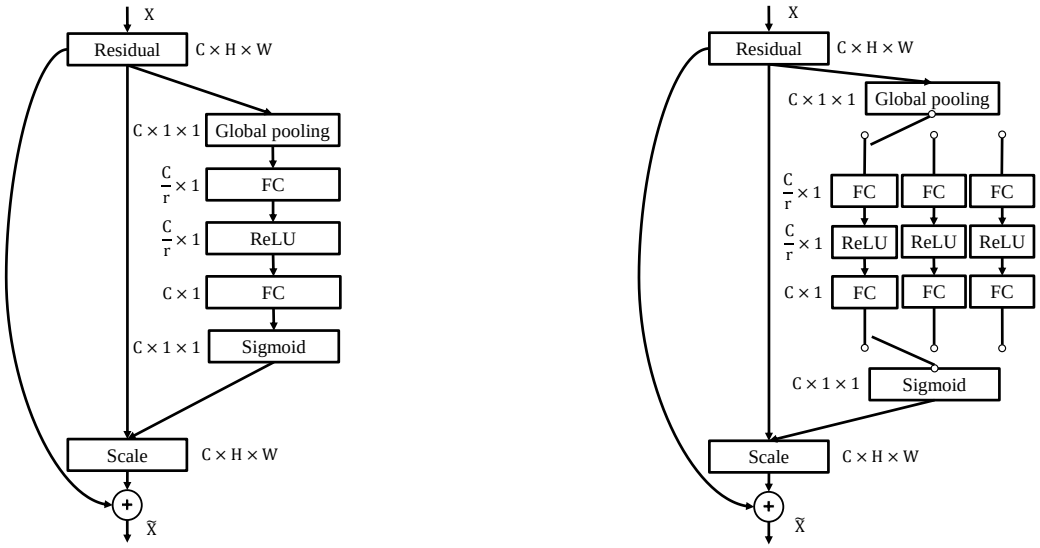
As suggested by the universal classification [RBV17, BV17] and our observations in Figure 2.1, we propose a semi-shared multi-domain detector, with the architecture of Figure

1.2 (c). In this model, beyond the domain-specific output and RPN layers, some light-weight intermediate layers are domain-specific as well, but the heavy-weight layers are shared, e.g. the convolutional layers. Since the domain-specific layers will add extra parameters/computations, they should 1) effectively account for domain shift throughout the entire network; 2) be light such that the memory/computation budget will be under control when more than a dozen domains to address. In [RBV17, BV17], the batch normalization (BN) layers are domain-specific, because BN layers meet these two requirements. However, in object detection, BN layers have to be frozen due to the small batch size during training. Thus, the domain-specific BN is not available for multi-domain detection.

Instead, we have experimented with the squeeze-and-excitation (SE) module [HSS17] of Figure 2.2 (a). There are a few reasons for this. First, feature-based attention is well known to be used in mammalian vision as a mechanism to adapt perception to different tasks and environments [Yar67, Pal99, Wol00, IB05, Yan98]. Hence, it seems natural to consider feature-based attention mechanisms for domain adaptation. Second, the SE is a module that accounts for interdependencies among channels to modulate channel responses. This can be seen as a a feature-based attention mechanism. Third the SE module has enabled the SENet to achieve state-of-the-art classification on ImageNet. Finally, it is a light-weight module. Even when added to each residual block of the ResNet [HZRS16] it increases the total parameter count by only $\sim$10%. This is close to what was reported by [RBV17] for BN-based adapters. For all these reasons, we adopt the SE module as the atomic adaptation unit, used to build all domain adaptive detectors proposed in this work, and denote it by the *SE adapter*.

2.1.4 SE Adapters

Following [HSS17], the *SE adapter* consists of the sequence of operations of Figure 2.2 (a): a global pooling layer, a fully conected (FC) layer, a ReLU layer, and a second FC layer,



(a) SE adapter (b) SE adapter bank

Figure 2.2: (a) block diagram of the SE adapter and (b) SE adapter bank.

implementing the computation

$$\mathbf{X}_{SE} = \mathbf{F}_{SE}(\mathbf{F}_{avg}(\mathbf{X})), \tag{2.1}$$

where $\mathbf{F}_{avg}$ is a global average pooling operator, and $\mathbf{F}_{SE}$ the combination of FC+ReLU+FC layers. The channel dimension reduction factor r, in Figure 2.2, is set as 16 in our experiments. To enable multi-domain object detection, the SE adapter is generalized to the architecture of Figure 2.2 (b), which is denoted as the *SE adapter bank.* This consists of adding a SE adapter branch per domain and a domain-switch, which allows the selection of the SE adapter associated with the domain of interest. Note that this architecture assumes this domain to be known a priori. It leads to the *multi-domain detector* of Figure 1.2 (b). Compared to Figure 1.2 (a), this model is up to 5 times smaller, while achieving better overall performance across the 11 datasets.

2.2 Universal Object detection

The detectors of the previous section require prior knowledge of the domain of interest. This is undesirable for autonomous systems, like robots or self-driving cars, where determining

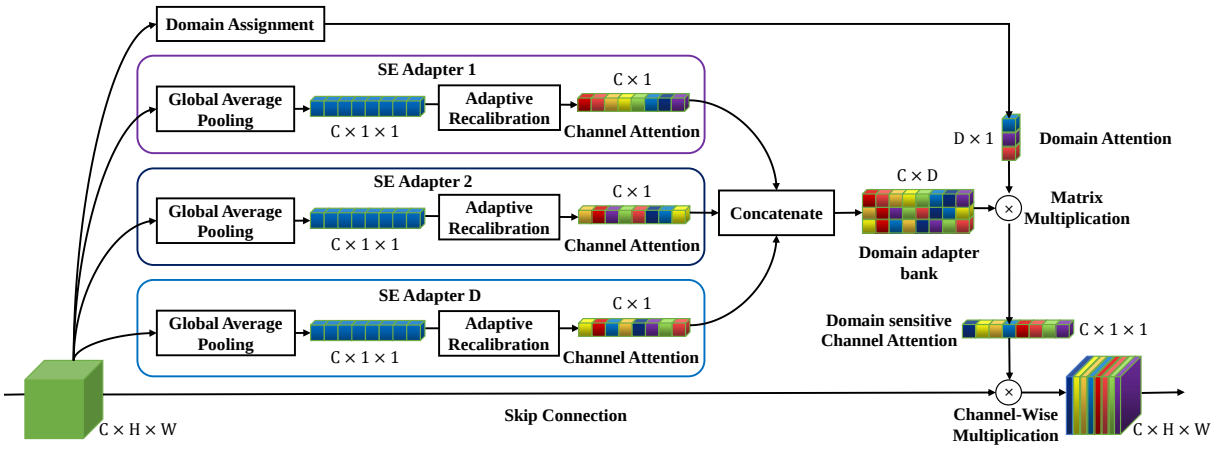


Figure 2.3: Block diagram of the proposed domain adaptation module.

the domain is part of the problem to solve. In this section, we consider the design of *universal detectors,* which eliminate this problem.

2.2.1 Universal Detector

The simplest solution to universal detection, shown in Figure 1.2 (c), is to share a single detector by all tasks. Note that, even for this detector, the output layer has to be task-specific, by definition of the detection problem. We have found that there is also a benefit in using task-specific RPN layers, due to the observations of Figure 2.1. This is not a problem because the task, namely what classes the system is trying to detect, is always known. Universality refers to the domain of input images that the detector processes, which does not have to be known in the case of Figure 1.2 (c). Beyond universal, the fully shared detector is the most efficient of all detectors considered in this work, as it has no domain-specific parameters. On the other hand, by forcing the same set of parameters/representations on all domains, it has little flexibility to deal with the statistical variations of Figure 2.1. In our experiments, this detector usually underperforms the multi-domain detectors of Figure 1.2 (a) and (b).

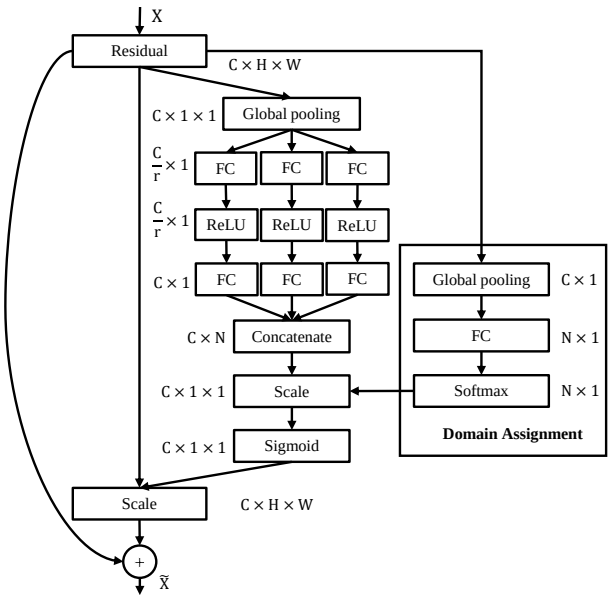


Figure 2.4: Detailed view of the universal adapter

2.2.2 Domain-attentive Universal Detector

Ideally, a universal detector should have some domain sensitivity, and be able to adapt to different domains. While this has a lot in common with multi-domain detection, there are two main differences. First, the domain must be inferred automatically. Second, there is no need to tie domains and tasks. For example, the traffic tasks of Figure 1.1 operate on a common visual domain, "traffic scenes", which can have many sub-domains, e.g. due to weather conditions (sunny vs. rainy), environment (city vs. rural), etc. Depending on the specific operating conditions, any of the tasks may have to be solved in any of the domains. In fact, the domains may not even have clear semantics, i.e. they can be data-driven. In this case, there is no need to request that each detector operates on a single domain, and a soft domain-assignment makes more sense. Given all of this, while domain adaptation can still be implemented with the SE adapter of Figure 2.2 (a), the hard attention mechanism of Figure 2.2 (b), which forces the network to fully attend to a single domain, can be suboptimal. To address this limitations, we propose the domain adaptation (DA) module of Figure 2.4. This has two components, a *universal SE adapter bank* and a *domain attention* mechanism, which are discussed next.

2.2.3 Universal SE Adapter Bank

The universal SE (USE) Adapter Bank, shown in Figure 2.4, is an SE adapter bank similar to that of Figure 2.2 (b). The main difference is that there is no domain switching, i.e. the adapter bank is *universal.* This is implemented by concatenating the outputs of the individual domain adapters to form a universal representation space

$$\mathbf{X}_{USE} = [\mathbf{X}_{SE}^{1}, \mathbf{X}_{SE}^{2}, ..., \mathbf{X}_{SE}^{N}] \in \mathbb{R}^{C \times N}, \tag{2.2}$$

where N is the number of adapters and $\mathbf{X}_{SE}^{i}$ the output of each adapter, given by (2.1). Note that N is not necessarily identical to the number of detection tasks. The USE adapter bank can be seen as a non-linear generalization of the filter banks commonly used in signal processing [Vai93]. Each branch (non-linearly) projects the input along a subspace matched to the statistics of a particular domain. The attention component then produces a domain-sensitive set of weights that are used to combine these projections in a data-driven way. In this case, there is no need to know the operating domain in advance. In fact there may not even be a single domain, since an input image can excite multiple SE adapter branches.

2.2.4 Domain Attention

The attention component, of Figure 2.4, produces a domain-sensitive set of weights that are used to combine the SE bank projections. Motivated by the SE module, the domain attention component first applies a global pooling to the input feature map, to remove spatial dimensions, and then a softmax layer (linear layer plus softmax function)

$$\mathbf{S}_{DA} = \mathbf{F}_{DA}(\mathbf{X}) = \text{softmax}(\mathbf{W}_{DA}\mathbf{F}_{avg}(\mathbf{X})), \tag{2.3}$$

where $\mathbf{W}_{DA} \in \mathbb{R}^{N \times C}$ is the matrix of softmax layer weights. The vector $\mathbf{S}_{DA}$ is then used to weigh the USE bank output $\mathbf{X}_{USE}$, to produce a vector of domain adaptive responses

$$\mathbf{X}_{DA} = \mathbf{X}_{USE}\mathbf{S}_{DA} \in \mathbb{R}^{C \times 1}. \tag{2.4}$$

As in the SE module of [HSS17], $\mathbf{X}_{DA}$ is finally used to channel-wise rescale the activations $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ being adapted,

$$\widetilde{\mathbf{X}} = \mathbf{F}_{scale}(\mathbf{X}, \sigma(\mathbf{X}_{DA})) \tag{2.5}$$

where $\mathbf{F}_{scale}(\cdot)$ implements a channel-wise multiplication, and σ is the sigmoid function.

In this way, the USE bank captures the feature subspaces of the domains spanned by all datasets, and the DA mechanism soft-routes the USE projections. Both operations are data-driven, and operate with no prior knowledge of the domain. Unlike the hard attention mechanism of Figure 2.2 (b), this DA module enables information sharing across domains, leading to a more effective representation. In our experiments, the domain-attentive universal detector outperforms the other detectors of Figure 1.2.

2.3 Experiments

In all experiments, we used a PyTorch implementation [YLBP17] of the Faster R-CNN with the SE-ResNet-50 [HSS17] pretrained on ImageNet, as the backbone for all detectors. Training started with a learning rate of 0.01 for 10 epochs and 0.001 for another 2 epochs on 8 synchronized GPUs, each holding 2 images per iteration. All samples of a batch are from a single (randomly sampled) dataset, and in each epoch, all samples of each dataset are processed only once. As is common for detection, the first convolutional layer, the first residual block and all BN layers are frozen, during training. These settings were used in all experiments, unless otherwise noted. Both multi-domain and universal detectors were trained on all domains of

interest simultaneously.

The Faster R-CNN has many hyperparameters. In the literature, where detectors are tested on a single domain, these are tuned to the target dataset, for best performance. This is difficult, and very tedious, to do over the 11 datasets now considered. We use the same hyperparameters across datasets, except when this is critical for performance and relatively easy to do, e.g. the choice of anchors. The main dataset-specific hyperparameters are shown in Table 2.1.

2.3.1 Datasets and Evaluation

Our experiments used the new UODB benchmark introduced in Section 2.1.1. For Watercolor [IFYA18], Clipart [IFYA18], Comic [IFYA18], Kitchen [GRM+16] and DeepLesion [YWL+18], we trained on the official `trainval` sets and tested on the `test` set. For Pascal VOC [EEVG+15], we trained on VOC2007 and VOC2012 `trainval` set and tested on VOC2007 `test` set. For WiderFace [YLLT16], we trained on the `train` set and tested on the `val` set. For KITTI [GLU12], we followed the train/val splitting of [CFFV16] for development and trained on the `trainval` set for the final results on `test` set. For LISA [MTM12], we trained on the `train` set and tested on the `val` set. For DOTA [XBD+18], we followed the pre-processing of [XBD+18], trained on `train` set and tested on `val` set. For MS-COCO [LMB+14], we trained on COCO 2014 `valminusminival` and tested on `minival`, to shorten the experimental period.

All detectors were evaluated on each dataset individually. The Pascal VOC mean average precision (mAP) was used for evaluation in all cases. The average mAPs was used as the overall measure of universal/multi-domain detection performance. The domain attentive universal detector was also evaluated using the official evaluation tool of each dataset, for comparison with the literature.

Table 2.2: The comparison on multi-domain detection. † denotes fixed assignment. "time" is the relatively run-times on the five datasets when the domain is unknown.

	Params	time	KITTI	VOC	WiderFace	LISA	Kitchen	Avg
single-domain	31.06M×5	5x	64.3	78.5	48.8	88.3	87.7	73.5
adaptive	42.37M	6x	67.8	78.9	49.9	**88.5**	86.0	74.2
BNA [BV17]	31.72M	5x	64.0	71.9	44.0	66.8	84.3	66.2
RA [RBV17]	82.72M	6x	64.3	70.5	46.9	69.1	84.6	67.1
universal	31.64M	1x	66.3	76.7	45.5	88.4	85.4	72.5
universal+DA†	42.37M	1.3x	67.5	79.0	49.8	88.2	88.0	74.6
universal+DA	42.44M	1.33x	**67.9**	**79.2**	**52.2**	87.5	**88.5**	**75.1**

Table 2.3: Overall results on the full universal object detection benchmark (11 datasets).

	# adapters	Params	DA index	KITTI	VOC	WiderFace	LISA	Kitchen	COCO	DOTA	Lesion	Comic	Clipart	Watercolor	Avg
single-domain	-	31.06M×11	-	64.3	78.5	48.8	88.3	87.7	**47.3**	**57.5**	51.2	45.8	32.1	52.6	59.4
universal	-	32.60M	-	67.5	80.9	45.5	87.1	88.5	45.5	54.7	45.3	51.1	43.1	47.0	59.7
adaptive	11	58.13M	-	68.0	82.1	50.6	**88.5**	87.2	45.7	54.1	53.0	50.0	56.1	57.8	63.0
universal+DA	11	58.29M	all	**68.1**	82.0	51.6	88.3	**90.1**	46.5	57.0	**57.3**	50.7	53.1	**58.4**	63.8
universal+DA*	6	41.74M	first+middle	67.6	**82.7**	**51.8**	87.9	88.7	46.8	57.0	54.8	**52.6**	**54.6**	58.2	**63.9**

2.3.2 Single-domain Detection

Table 2.1 shows the results of the single-domain detector bank of Figure 1.2 (a) on all datasets. Our VOC baseline with the SE-ResNet-50 is 78.5, and better than the Faster R-CNN performance of [RHGS17, HZRS16] (76.4 mAP for ResNet-101). The other entries in the table are incomparable to the literature, where different evaluation metrics/tools are used for different datasets. The detector bank is a fairly strong baseline for multi-domain detection (average mAP of 59.4).

2.3.3 Multi-domain Detection

Table 2.2 compares the multi-domain object detection performance of all architectures of Figure 1.2. For simplicity, only five datasets (VOC, KITTI, WiderFace, LISA and Kitchen) were used in this section. The table confirms that the adaptive multi-domain detector of Section 2.1.3 ("adaptive") is light-weight, only adding $\sim$11M parameters to the Faster R-CNN over the five datasets. Nevertheless, it outperforms the much more expensive single-domain detector bank by 0.7 points. Note that the latter is a strong baseline, showing the multi-domain detector can

Table 2.4: The effect of SE adapters number.

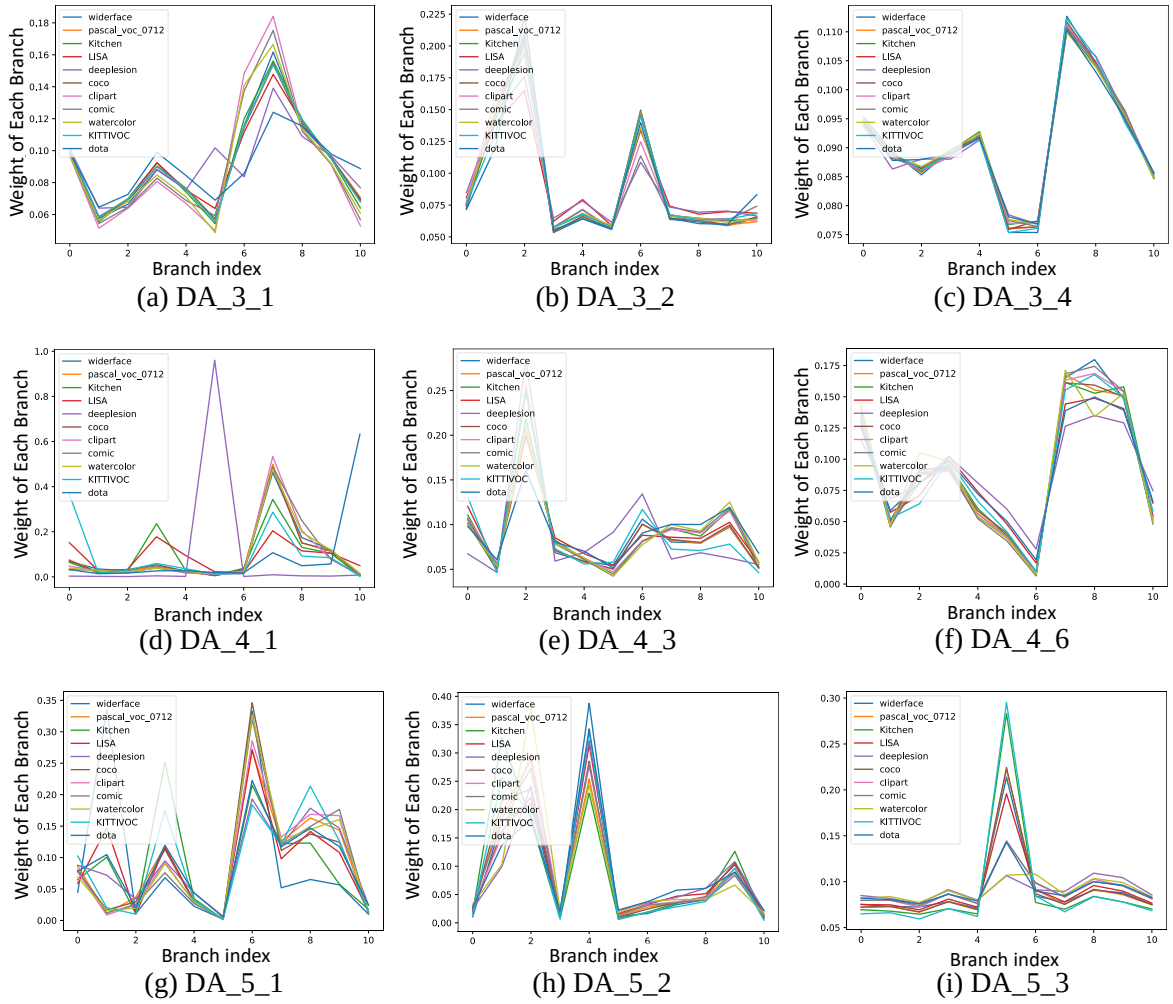
# adapters	Params	KITTI	VOC	WiderFace	LISA	Kitchen	Avg
single	31.06M×5	64.3	78.5	48.8	88.3	87.7	73.5
1	32.32M	66.3	74.9	43.5	87.4	85.4	71.3
3	37.38M	67.8	78.4	47.1	87.7	**89.0**	74.1
5	42.44M	**67.9**	79.2	**52.2**	87.5	88.5	75.1
7	47.50M	**67.9**	**79.6**	**52.2**	**89.5**	88.7	**75.6**

beat individually trained models with a fraction of the computation. Table 2.2 also shows that the proposed SE adapter significantly outperforms the BN adapter (BNA) of [BV17] and the residual adapter (RA) or [RBV17], previously proposed for classification. This is not surprising, given the above discussed inadequacy of BN as an adaptation mechanism for object detection.

The universal detector of Figure 1.2 (c) is even more efficient, adding only 0.5M parameters to the Faster R-CNN, accounting for domain-specific RPN and output layers. However, its performance ("universal" in Table 2.2) is much weaker than that of the adaptive multi-domain detector (1.7 points). Finally, the domain-attentive universal detector ("universal+DA") has the best performance. With a $\sim$7% parameter increase per domain, i.e. comparable to the multi-domain detector, it outperforms the single-domain bank baseline by 1.6 points. To assess the importance of data-driven domain attention mechanism of Figure 2.4 (b), we fixed the soft domain assignments, simply averaging the SE adapter responses, during both training and inference. This (denoted "universal+DA†") caused a performance drop of 0.5 point. Finally, Table 2.2 shows the relative run-times of all methods on the five datasets, when the domain is unknown. It can be seen that "universal+DA" is about 4$\times$ faster than the multi-domain detectors ("single-domain" and "adaptive") and only 1.33$\times$ slower than "universal".

2.3.4 Effect of the number of SE adapters

For the USE bank of Figure 2.4 (b), the number N of SE adapters does not have to match the number of detection tasks. Table 2.4 summarizes how the performance of the domain attentive universal detector depends on N. For simplicity, we again use 5 datasets in this experiment. For



(a) DA_3_1 (b) DA_3_2 (c) DA_3_4

(d) DA_4_1 (e) DA_4_3 (f) DA_4_6

(g) DA_5_1 (h) DA_5_2 (i) DA_5_3

Figure 2.5: Soft assignments across SE units for all datasets.

a single adapter, the DA module reduces to the standard SE module, and the domain attentive universal detector to the universal detector. This has the worst performance. Performance improves with the number of adapters. On the other hand, the number of parameters increases linearly with the number of adapters. In these experiments, the best trade-off between performance and parameters is around 5 adapters. This suggests that, while a good rule of thumb is to use "as many adapters as domains", fewer adapters can be used when complexity is at a premium.

2.3.5 Results on the full benchmark

Table 2.3 presents results on the full benchmark. The settings are as above, but we used 10 epochs with learning rate 0.1, and then 4 epochs with 0.01 on 8 GPUs, each holding 2 images. The universal detector performs comparably to the single-domain detector bank, with 10 times fewer parameters. The domain-attentive universal detector ("universal+DA") improves baseline performance by 4.4 points with a 5-fold parameter decrease. It has large performance gains (>5 points) on DeepLesion, Comic, and Clipart. This is because Comic/Clipart contain underpopulated classes, greatly benefiting from information leveraged from other domains. The large gain of DeepLesion is quite interesting, given the nontrivial domain shift between its medical CT images and the RGB images of the other datasets. The gains are mild for VOC, KITTI, Kitchen, WiderFace and WaterColor (1$\sim$5 points), and none for COCO, LISA and DOTA. In contrast, for the universal detector, joint training is not always beneficial. This shows the importance of domain sensitivity for universal detection.

To investigate what was learned by the domain attention module of Figure 2.4 (b), we show the soft assignments of each dataset, averaged over its validation set, in Figure 2.5. Only the first and last blocks of the 4th and 5th residual stages are shown. The fact that some datasets, e.g. VOC and COCO, have very similar assignment distributions, suggests a substantial domain overlap. On the other hand, DOTA and DeepLesion have distributions quite distinct from the remaining. For example, on block "DA_4_1", DeepLesion fully occupies a single domain. These observations are consistent with Figure 2.1, indicating that the proposed DA module is able to learn domain-specific knowledge.

A comparison of the first and the last blocks of each residual stage, e.g. "DA_4_1" v.s. "DA_4_6", shows that the latter are much less domain sensitive than the former, suggesting that they could be made universal. To test this hypothesis, we trained a model with only 6 SE adapters for the 11 datasets, and only in the first and middle blocks, e.g. "DA_4_1" and "DA_4_3". This model, "universal+DA*", achieved the best performance with much less parameters than the

Table 2.5: The comparison with official evaluation on Pascal VOC, KITTI, DeepLesion, Clipart, Watercorlor, Comic and WiderFace.

(a) The comparison on VOC 2007 `test`.

	Backbone	mAP
Faster-RCNN [RHGS15]	ResNet-101	76.4
R-FCN [DLHS16]	ResNet-50	77.0
Faster-RCNN‡ [RHGS17]	VGG16	78.8
RefineDet512 [ZZW+17]	VGG-16	81.8
Faster-RCNN (ours)	SE-ResNet-50	78.5
Faster-RCNN+DA	SE-ResNet-50	79.6
Faster-RCNN+DA†	SE-ResNet-50	82.7

(b) The comparison on WiderFace `Val`.

	Backbone	Easy	Medium	Hard
CascadeCNN [ZZLQ16]	VGG-16	0.851	0.820	0.607
Faster-RCNN [RHGS15]	VGG-16	0.907	0.850	0.492
MS-CNN [CFFV16]	VGG-16	0.916	0.903	0.802
HR [HR17]	ResNet-101	0.925	0.910	0.806
SSH [NSCD17]	VGG-16	0.931	0.921	0.845
Faster-RCNN (ours)	SE-ResNet-50	0.910	0.872	0.556
Faster-RCNN+DA	SE-ResNet-50	0.914	0.882	0.587

(c) Sensitivity on DeepLesion `test`.

	Backbone	Sensitivity
Faster-RCNN [RHGS15]	VGG-16	81.62
R-FCN [DLHS16]	VGG-16	82.21
R-FCN† [DLHS16]	VGG-16	82.76
3-DCE, 9 Slices [YBS18]	VGG-16	84.34
3-DCE, 27 Slices [YBS18]	VGG-16	85.65
Faster-RCNN (ours)	SE-ResNet-50	82.44
Faster-RCNN+DA	SE-ResNet-50	87.29

(d) The comparison on KITTI `test` set of car.

	Backbone	Moderate	Easy	Hard
Faster-RCNN [RHGS15]	VGG-16	81.84	86.71	71.12
SDP+CRC [YCL16]	VGG-16	83.53	90.33	71.13
YOLOv3 [RF18]	Darknet-53	84.13	84.30	76.34
MS-CNN [CFFV16]	VGG-16	88.83	90.46	74.76
F-PointNet [QLW+18]	PointNet	90.00	90.78	80.80
Faster-RCNN (ours)	SE-ResNet-50	81.83	90.34	71.23
Faster-RCNN+DA	SE-ResNet-50	88.23	90.45	74.21

(e) The comparison on Clipart, Watercolor and Comic `test` set.

	Backbone	Clipart	Watercolor	Comic
ADDA [THSD17]	VGG-16	27.4	49.8	49.8
Faster-RCNN[RHGS15]	VGG-16	26.2	-	-
SSD300 [LAE+16]	VGG-16	26.8	49.6	24.9
Faster-RCNN+DT+PL[IFYA18]	VGG-16	34.9	-	-
SSD300+DT+PL[IFYA18]	VGG-16	46.0	54.3	37.2
Faster-RCNN (ours)	SE-ResNet-50	32.1	52.6	45.8
Faster-RCNN+DA	SE-ResNet-50	54.6	58.2	52.6

"universal+DA" detector of 11 adapters. It outperformed the single domain baseline by 4.5 points.

2.3.6 Official evaluation

Since, to the best of our knowledge, this is the first work to explore universal/multi-domain object detection on 11 datasets, there is no literature for a direct comparison. Instead, we compared the "universal+DA*" detector of Table 2.3 to the literature using the official evaluation for each dataset. This is an unfair comparison, since the universal detector has to remember 11 tasks. On VOC, we trained two models, with/without COCO. Results are shown in Table 2.5a, where all methods were trained on Pascal VOC 07+12 `trainval`, ‡/† denotes with COCO `trainval`/`val`. Note that our Faster R-CNN baseline (SE-ResNet-50 backbone) is stronger than

that of [HZRS16] (ResNet-101). Adding universal domain adapters improved on the baseline by more than 1.1 points. Adding COCO enabled another 3.1 points. Note that, 1) this universal training is different from the training scheme of [RHGS17] (the network trained on COCO then finetuned on VOC), where the final model is only optimized for VOC; and 2) only the 35k images of COCO2014 `valminusminival` were used.

The baseline was the default Faster R-CNN that initially worked on VOC, with minimum dataset-specific changes, e.g. in Table 2.1. Table 2.5d shows that this performed weakly on KITTI. However, the addition of adapters, enabled a gain of 6.4 points (`Moderate` setting). This is comparable to detectors optimized explicitly on KITTI, e.g. MS-CNN [CFFV16] and F-PointNet [QLW$^+$18]. For WiderFace, which has enough training face instances, the gains of shared knowledge are smaller (see Table 2.5b). On the other hand, on DeepLesion and CrossDomain (Clipart, Comic and Watercolor), see Table 2.5c and 2.5e respectively, the domain attentive universal detector significantly outperformed the state-of-the-art. Overall, these results show that a single detector, which operates on 11 datasets, is competitive with single-domain detectors in highly researched datasets, such as VOC or KITTI, and substantially better than the state-of-the-art in less explored domains. This is achieved with a relatively minor increase in parameters, vastly smaller than that needed to deploy 11 single task detectors.

2.4 Acknowledgment

This work was partially funded by NSF awards IIS-1546305 and IIS-1637941, a gift from 12 Sigma Technologies, and NVIDIA GPU donations.This chapter, in part, is a reprint of the material as it appears in In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. Xudong Wang, Zhaowei Cai, Dashan Gao, Nuno Vasconcelos., IEEE, 2019. The thesis author was the primary investigator and author of this paper.

Chapter 3

3D context enhanced representations for 3D image segmentation

3.1 Introduction

There has been significant progress in 2D object detection and instance segmentation in recent years. Many two-stage object detectors, such as Fast-RCNN[Gir15] and Faster-RCNN[RHGS15], and one-stage detector, such as SSD[LAE+16] and YOLO[RF18], reached pretty good results on object detection tasks. Mask-RCNN [HGDG17] tried to accomplish both object detection and 2D object instance segmentation tasks using one network by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. U-Net[RFB15] built upon fully convolutional network for 2D medical image segmentation.

For 3D medical image segmentation, 3D U-Net extended 2D U-Net to learn dense volumetric segmentation and became one of the most popular network for 3D medical image segmentation. However, due to the limitation of GPU memory, we need to make a trade-off between slice resolution and feature bags size. If we want to keep the spatial resolution, we need to sacrifice the feature bags size, which can damage the ability of utilizing contextual information in z direction for more accurate lesion detection and mask generation. Another choice is to use low resolution 3D volume as input of 3D U-Net, which usually leads to miss-detection of small lesions and low precision of mask.

Our network is built based on 2D Mask-RCNN for 2D instance segmentation. Although high resolution images will be able to fit into 2D network, slice dependencies will not be able to be utilized for 2D network. In order to overcome this problem and provide another direction to solve detection and segmentation problems with 3D contextual information, we introduced a cross-slice spatial and channel attention module to learn and utilize the relationship between different slices to help improve 3D object/lesion detection and mask prediction tasks.

As shown in Figure 1, we introduced two new modules for high precision segmentation, they are cross-slice spatial and channel attention module and mask refinement module. Cross-slice spatial and channel attention module is used for combining information from multiple slices in

the volume and learn the dependencies between them. The reasons why multi-slices are useful is that, for many lesions, it is very easy even for human doctors to be confused between lesions and blood vessel without looking through nearby slices above and below key slice. Only rely on one single slice to get high accuracy lesion and organ segmentation is neither reliable nor practical, as human doctors also need to get access to multiple slices for final disease prediction and lesion detection on many cases, which is also the reason why a volume of CT scan is necessary not only several slices. Also, getting access to nearby slices will provide neural network extra information for getting more precise prediction on the key slices lesion detection and mask prediction.

We also built anther module named mask refinement module for higher performance mask prediction. For human visual system, saliency is used for processing visual information and images. Human brains will automatically blur the unfocused part of input image and salient the focused part. Inspired by this, we add a cascade mask refinement module which will be able to provide an objectiveness attention map for input feature maps and help mask head to get higher quality mask prediction.

Based on above two novel modules, we got state-of-art performance, 74.1 dice per case on lesion segmentation, on one of the biggest 3D medical image segmentation challenges-LiTS liver tumor segmentation challenge, which outperforms 5 points than previous year challenge winner and got the 1st place on leaderboard upon submission of this paper. Overall architecture can be seen in Figure 3.1 and details will be discussed in following sections.

3.2 Related Work

Attention Module: [VSP+17] proposed a self-attention module for machine translation, obtaining the response at a position by attending to all positions in a sequence. Similarly, [WGGH18] proposed a non-local network for video classification, based on a spacetime dependency/attention mechanism. [HSS17] focused on channel relationships, introducing the SE module to adapta-

tively recalibrate channel-wise feature responses. The resulting SENet achieved good results on ImageNet recognition. [WPLK18] proposed Convolutional Block Attention Module(CBAM), which sequentially infers attention maps along two separate dimensions, channel and spatial, then the attention maps are multiplied to the input feature map for adaptive feature refinement. Compared with previous work, which will do self-attention operation, our work will build a bag of features by concatenating the features extracted from a list of neighbor slices and focused on learning the dependencies in feature space between key slice and a bag of neighbor slices for boosting the performance on key slice.

2D Objects Detection and Instance Segmentation: The two stage detection framework of the R-CNN [GDDM14], Fast R-CNN [Gir15] and Faster R-CNN [RHGS15] detectors has achieved great success in recent years. Many works have expanded this base architecture. For example, MS-CNN [CFFV16] and FPN [LDG$^+$17] built a feature pyramid to effectively detect objects of various scales; the R-FCN [DLHS16] proposed a position-sensitive pooling to achieve further speed-ups; and the cascade R-CNN [CV18] introduced a multi-stage cascade for high quality object detection. In parallel, single-stage object detectors, such as YOLO [RDGF16] and SSD [LAE$^+$16], became popular for their fairly good performance and high speed. However, all of these detectors will only deal with 2D images/samples for objects detection without considering merging information from other images/samples.

3D Medical Image Segmentation: A natural solution to 3D medical image segmentation and detection problems is to rely on 3D convolutional networks, such as the 3D U-Net of [ÇAL$^+$16] or the extended 2D U-Net of [RFB15]. However, current GPU memory limitations prevent the processing of 3D volumes with high resolution. This is problematic, because the use of low-resolution volumes leads to low precision or miss-detection of small lesions and tumors and blur in lesion mask predictions, especially on boundaries. Hence, there is a need to trade-off the spatial resolution of each 2D slice for the number of slices processed. This implies a trade-off between the precision with which segmentation or detection can be performed and the amount

of contextual information, in the z direction, that can be leveraged. A popular solution is to a use a 2D network to segment or detect the structures of interest in 2D or 2.5D slices and then concatenate the results to build a 3D segmentation mask or bounding box.

Christ et al. proposed a 2D U-Net for liver and tumor segmentation, followed by a conditional random field for segmentation refinement [CEE+16]. Li et al. proposed a hybrid Dense 2D/3D UNet of three-stages [LCQ+18]. They found that a pre-trained 2D model can significantly boost performance of their network. Bi et al. proposed a two-stage cascaded deep residual network for liver lesion segmentation [Han17]. These approaches are limited by the lack of contextual information. Since even human experts need to inspect multiple slices to reach confident assessments of confusing lesions, this is likely to upper bound their performance. Ding et al. applied 2D networks to generate lesion candidates, then 3D CNN classifiers were trained for false positive reduction (FPR). To address this problem, Yan et al. [YBS18] proposed a 3D context enhanced region-based CNN. However, their method is based on a region proposal network (RPN) and cannot be implemented as a single-stage detector, such as SSD and YOLO, or a segmentation network, such as U-Net, without an RPN component. Furthermore, because only the feature map derived from a central image is processed by the RPN to generate proposals, the proposal generation process has no access to 3D context. Given that missed proposals can not be recovered, this places an upper bound on detection performance.

3.3 Volumetric Attention

The overall architecture of the VA Mask R-CNN is shown in Fig.3.1. The VA attention module operates on the Mask R-CNN feature pyramids extracted from a *target* 2.5D image, where detection takes place, and neighboring *contextual* 2.5D images. The 2.5D images are each composed of 3 adjacent slices. The attention module has three components: bag of long-range features, volumetric channel attention, and volumetric spatial attention. Unlike the self-attentive

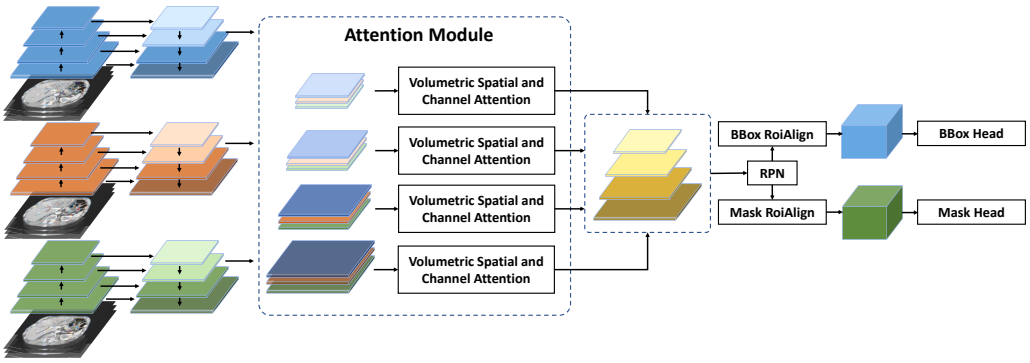


Figure 3.1: Architecture of the Volumetric Attention(VA) Mask-RCNN. Three continuous 2.5D images, each composed of 3 adjacent slices, are shown as example.

feature map of [WGGH18], VA uses long-range features from neighboring slices, which are combined with the feature map of the target slice to generate spatial and channel attention responses. A detailed scheme of the attention module is given in Fig. 3.2. We next discuss the three components combined with Mask-RCNN in detail.

3.3.1 Bag of Long-range Features

To account for dependencies along the z direction of the 3D CT volume, the VA Mask R-CNN complements the target 2.5D image, with neighboring images, both above and below the target image. These are denoted as contextual images. The features extracted from these images are concatenated for each level of the spatial pyramid, according to

$$\mathbf{X}^i_{long} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N] \in \mathbb{R}^{N \times C^i \times H^i \times W^i}, \tag{3.1}$$

where i is the pyramid level, $C^i \times H^i \times W^i$ its dimensions (chanel, height, and width, respectively), $\mathbf{X}^i_{long}$ the corresponding bag of long-range features, and N the number of contextual images. The features $\mathbf{X}_k$ are sorted by the order of the corresponding images along the z direction of the 3D volume.

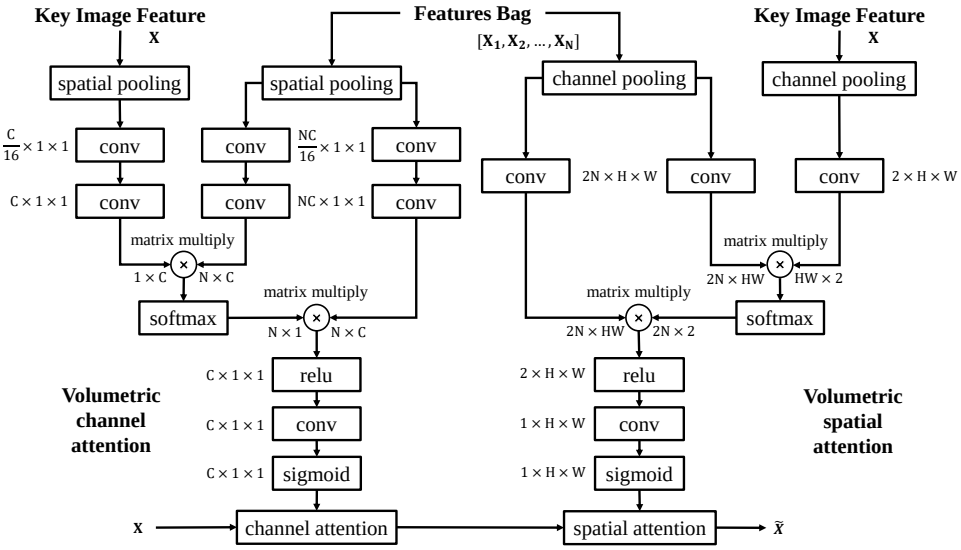


Figure 3.2: Volumetric Spatial and Channel Attention Module. N is the bag size, C, H, W the feature map channel size, height and width, respectively. Spatial and channel pooling are used to reduce computation.

3.3.2 Volumetric Channel Attention

This attention mechanism is inspired by that of [HSS17, WGGH18]. The bag of features $\mathbf{X}_{long} \in \mathbb{R}^{N\times C\times H\times W}$ and corresponding target image feature map $\mathbf{X}_{tgt} \in \mathbb{R}^{C\times H\times W}$ are each subject to a global average pooling operator $\mathbf{F}^c_{avg}$. Following [HSS17], computation is reduced by replacing the linear embedding layer of the original non-local blocks of [WGGH18] by two 1×1 convolutional layers with reduction ratio of 16. This is implemented as $\mathbf{F}^c_{emb}(\mathbf{X}) = W_2\delta(W_1\mathbf{F}^c_{avg}(\mathbf{X}))$, where $W_1 \in \mathbb{R}^{\frac{C}{16}\times C}$, $W_2 \in \mathbb{R}^{C\times\frac{C}{16}}$ and δ is the RELU function. The slices attention signal is finally computed with a softmax

$$\mathbf{S}^c_{att} = \text{softmax}(\mathbf{F}^c_{emb}(\mathbf{X}_{tgt}) \cdot \mathbf{F}^c_{emb}(\mathbf{X}_{long})) \in \mathbb{R}^{1\times N} \tag{3.2}$$

along dimension N, where $\mathbf{F}^c_{emb}(\mathbf{X}_{tgt}) \in \mathbb{R}^{1\times C}$, $\mathbf{F}^c_{emb}(\mathbf{X}_{long}) \in \mathbb{R}^{C\times N}$ and $\cdot$ refers to matrix multiplication. The slice attention signal $\mathbf{S}^c_{att}$ is then applied to $\mathbf{F}^c_{emb}(\mathbf{X}_{long}) \in \mathbb{R}^{N\times C}$ according to $\mathbf{S}^c_{att} \cdot \mathbf{F}^c_{emb}(\mathbf{X}_{long})$ and this is followed by a relu layer, a 1×1 conv layer and a sigmoid layer, to learn a nonlinear interaction $\mathbf{S}_c \in \mathbb{R}^{C\times 1\times 1}$ between channels. Then channel-wise multiplication is

applied on $\mathbf{X}_{tgt} \in \mathbb{R}^{C\times H\times W}$.

3.3.3 Volumetric Spatial Attention

The volumetric spatial attention module uses max and average pooling to shrink feature maps along the channel dimension, concatenating them into two channel feature maps $\mathbf{F}^s_{pool}(\mathbf{X}) = [\mathbf{F}^s_{max}(\mathbf{X}), \mathbf{F}^s_{avg}(\mathbf{X})] \in \mathbb{R}^{2\times H\times W}$. An embedding function is then implemented as $\mathbf{F}^s_{emb}(\mathbf{X}) = W\mathbf{F}^s_{pool}(\mathbf{X})$, where W is a learned convolutional weight layer. The slice attention signal is finally computed with a softmax

$$\mathbf{S}^s_{att} = \text{softmax}(\mathbf{F}^s_{emb}(\mathbf{X}_{tgt}) \cdot \mathbf{F}^s_{emb}(\mathbf{X}_{long})) \in \mathbb{R}^{1\times N} \tag{3.3}$$

along dimension N. A spatial attention map $S_s \in \mathbb{R}^{1\times H\times W}$ is then generated with an architecture similar to that of Section 3.3.2 and element-wise multiplied with $\mathbf{X}_{tgt} \in \mathbb{R}^{C\times H\times W}$.

3.4 Experiments

The volumetric attention was evaluated on two public datasets, Liver Tumor Segmentation (LiTS) [BCV+19] and DeepLesion[YWLS18]. All experiments used a PyTorch implementation [CPW+18] of the Mask-RCNN and Faster R-CNN. Unless otherwise noted, all hyperparameters are as in [LDG+17] for the Faster-RCNN and [HGDG17] for the Mask-RCNN.

3.4.1 Datasets and Evaluation

LiTS is a dataset of liver lesions, including 131 training and 70 test CT scans, acquired in six different clinical sites using different protocols and scanners. Lesion segmentation performance is evaluated and ranked by the Dice coefficient per volume, averaged over all test cases. For additional insight on the quality of the segmentation, we also break down the average

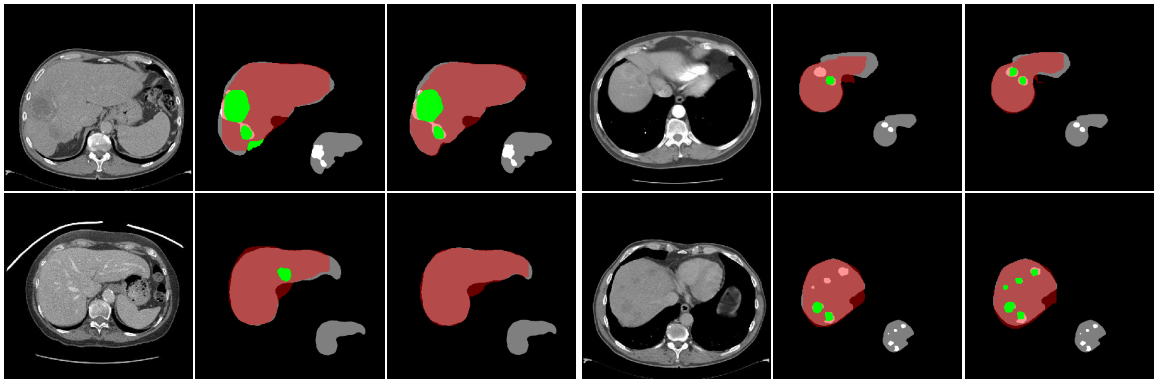


Figure 3.3: 2D visualization of segmentations by Mask-RCNN and VA Mask R-CNN on LiTS `val` set.

Dice/lesion per lesion size: the coefficients measured for small (diameter $<$ 15 mm), medium (diameter between [15mm, 30mm] and large (diameter $>$ 30mm) legions are denoted as Dice_s, Dice_m and Dice_l respectively. DeepLesion is a dataset with a larger variety of lesions, including 33,688 bookmarked radiology images from 10,825 studies of 4,477 unique patients. For each bookmarked image, a bounding box is generated to indicate the location of each lesion. We use the official split (70% training, 15% validation and 15% test) at the patient level, for training and testing. For consistency with prior art, detection results are evaluated with the False Positives (FPs) per Image metric.

3.4.2 Pre-processing for LiTS and DeepLesion Datasets

For the pre-processing of LiTS data. The volume intensity values are truncated to the HU range of [-200, 300] for removing the irrelevant information and then normalized to [0, 1]. The slice spacing is resampled to 1.5 mm. The original in-plane spacing is kept same because of the small variation. The adjacent three axial slices are concatenated as a 2D image for the Mask-RCNN training and inference. The ground truth lesions are removed in the training stage, if the number of pixels is smaller than 5. We add another body mask for the Mask-RCNN learning, because the Mask-RCNN can not learn anything from the negative sample without any ground

truth target. The body mask was generated by a binary threshold with -200 HU value and followed by hole filling in each slice. Then the Mask-RCNN model is trained with multi tasks supervised by lesion, liver, and body mask together.

For DeepLesion datasets, the 12-bit CT intensity range was rescaled to floating-point numbers in [0,255] using a single windowing (1024 to 3071 HU) that covers the intensity ranges of lung, soft tissue,and bone. Every image slice was resized to512512. To encode 3-D information, we used three axial slices to compose a three-channel image and input it to the network. The slices were the center slice that contains the bookmark and its neigh-boring slices interpolated at 2-mm slice intervals.

3.4.3 Post-processing

In the inference stage of LiTS dataset, the 2D live and lesion segmentation can be predicted from the trained Mask-RCNN model. The 3D liver and lesion segmentation can be get by stacking the 2D segmentation.

3.4.4 LiTS Experiments

Pre-processing: For 3D liver/lesion detection and segmentation, we stack three adjacent axial slices into a 3-channel image and apply the Mask-RCNN to detect and segment the liver/lesion for the center slice. 3D segmentation results are then obtained by stacking the masks generated for all slices. The Mask-RCNN is trained to detect both liver and lesions, to enable the removal of false lesions outside the liver by simply computing the logical AND of the predicted liver and lesion masks. Since the focus of this task is on the liver and lesions, the CT scan's Hounsfield unit (HU) is clamped between [-200, 300] and normalized to a floating point between [0, 1]. Each slice is scaled to 1024×1024 pixels and the slice-thickness resampled to 1.5mm.

Benchmark results: To evaluate performance on LiTS, the feature bag size of (3.1) was set to 9,

Table 3.1: Comparison with LiTS Challenge leaderboard, as of April 2, 2019

Team	Model	Dice per case
3D U-Net(Ours) [ÇAL+16]	3D U-Net	55.0
X. Li et al. [LCQ+18]	3D DenseUNet	59.4
G. Chlebus [CMMS17]	2D U-Net	65.0
E. Vorontsov et al. [VTPK18]	2D + 3D FCN	65.0
Y. Yuan [Yua17]	Deconv-Conv Net	65.7
X. Han [Han17]	2D U-Net	67.0
Mask-RCNN(Ours)[HGDG17]	Mask-RCNN	70.3
X. Li et al.[LCQ+18]	H-DenseUNet	72.2
VolumetricAttention	VA Mask-RCNN	**74.1**

the weights of the feature extractor and RPN copied from detectors pre-trained on the MS-COCO and DeepLesion datasets, and the smallest image scale set to 1024. Table 3.1 presents a copy of the LiTS leaderboard, at the time of submission of this paper. All algorithms are evaluated on the LiTS `test` set. The VA Mask R-CNN achieves state-of-the-art performance, with 74.10 dice per case. This outperforms the previous year challenge winner by 6.8 points and the best published results by 1.6 points.

Ablation study and evaluation: To better understand the proposed architecture, the LiTS dataset was split, using 75% of the `train` data to create a training set and the remaining 25% as a `val` set for a local ablation study. Table 3.2 summarizes the resulting dice per volume, averaged over all cases, and dice per cases, averaged over small, medium and large lesions. All these are control experiments, all hyper-parameters and settings remaining the same as in the benchmark experiments, unless otherwise noted.

Benefits of VA attention: Three conclusions can be drawn from Table 3.2a. First, the 2D approaches outperform the 3D U-Net, even before addition of the VA attention module. This shows that 2D networks are at least competitive for 3D mask segmentation. Since the Mask-RCNN achieved the best performance on these experiments, we use it as base model in the remainder of the paper. It should, however, be pointed out that VA could equally be combined with the 2D U-Net. Second, the addition of the VA module further increases performance, increasing the Dice

Table 3.2: Evaluation on LiTS `val` set, in terms of dice per volume, averaged over all cases, and dice per lesions, averaged over small, medium and large lesions.

(a) Dice comparison.

	Dice	Dice_s	Dice_m	Dice_l
3D U-Net	35.3	17.0	39.2	61.3
2D U-Net	48.8	39.7	58.2	68.3
Mask-RCNN	56.1	44.3	65.1	77.9
Ours	**63.3**	**54.3**	**73.7**	**80.3**

(b) Influence of image scales.

Scale	Dice	Dice_s	Dice_m	Dice_l
512	50.2	35.8	65.1	77.9
800	61.1	52.1	71.6	79.3
1024	**63.3**	**54.3**	**73.7**	**80.3**
1333	**63.5**	**54.8**	**73.5**	**80.4**

(c) Pre-training dataset.

	Pre-training dataset		
+ImageNet	✓	✓	✓
+MS-COCO		✓	✓
+DeepLesion			✓
Dice per case	59.3	61.2	**63.3**

(d) Influence of VA modules.

	Dice	Dice_s	Dice_m	Dice_l
Baseline	56.1	44.3	65.1	77.9
+channel att	61.5	52.2	72.7	78.7
+spatial att	**63.3**	**54.3**	**73.7**	**80.3**

(e) Influence of VA location.

	Dice	Dice_s	Dice_m	Dice_l
Baseline	56.1	44.3	65.1	77.9
RPN	**63.3**	**54.3**	**73.7**	**80.3**
RCNN	60.6	51.0	70.4	78.2

(f) Influence of number of slices.

# Slices	Dice	Dice_s	Dice_m	Dice_l
$9(3\times3)$	60.8	52.2	69.4	77.7
$21(3\times7)$	62.5	52.6	72.2	79.3
$27(3\times9)$	**63.3**	**54.3**	**73.7**	80.3
$33(3\times11)$	63.1	53.6	73.4	**80.6**

coefficient per case by 5.24 points. Third, this gain is especially large for small and medium lesions, e.g. 10 points (a $\sim$23% relative improvement) for small lesions. Note how the lack of contextual information along the z direction severely compromises the small lesion performance of the mask R-CNN. Fig.3.3 illustrates how VA attention enables the Mask R-CNN to reject confusing FP lesions and produce smoother segment boundaries. Segmented liver is shown in red and lesions in green. Zoomed out ground truth masks are shown on bottom right, with liver in gray and lesions in white. The VA Mask-RCNN produces smoother segmentation boundaries and lower FP and miss rates. In the top left, the gallbladder area is easily confused with the lesion area. VA Mask-RCNN leverages contextual slices to remove this FP.

Image scales. Table 3.2b shows that larger image scales lead to better performance, especially for small lesions. However, performance saturates at a scale of 1333 pixels. This is only marginally better than a scale of 1024 but requires substantially more memory. For this reason, a scale of 1024 is adopted in the remainder of the paper.

Influence of pre-training: [HGD18] claims that ImageNet pre-training does not improve accu-

racy of networks trained with as few as 10k COCO images. As shown in Table 3.2c, this does not hold for medical imaging where, due to the difficulties of collecting and labeling datasets, few datasets have 10k examples. Furthermore, while MS-COCO has $\sim$5 objects/image, this number is much smaller for medical image datasets. For LiTS the number is smaller than 1, especially when the 3D volume is split into 2D slices and these are considered different examples. Table 3.2c shows that, in this case, ImageNet pre-training still has an important role in combating overfitting. Adding MS-COCO to the pre-training dataset further improves performance by 1.9 points. This is mostly because the COCO tasks encourage the network to more accurately localize objects. Finally, due to the non-trivial domain shift between MS-COCO and LiTS, further pre-training on DeepLesion improves performance by an additional 2.1 points.

Spatial vs. Channel Attention: to compare the relative importance of the two attention mechanisms, the two modules were incrementally added to the 2D Mask-RCNN, with the results of Table 3.2d. These experiments use 9 slices. The addition of channel attention enhances performance by more than 4 points, and the subsequent addition of spatial attention increases performance by another 2 points. In summary, both attention mechanisms are important.

Location of attention module: the VA module can be added as shown in Fig.3.1, i.e. to the last stage of feature extraction, before the RPN, or after the bounding box ROI align and mask ROI align steps, i.e. before the RCNN. Table 3.2e shows that attention is more effective if introduced before the RPN. While this improves performance by 5.1 Dice points per case, the gain is only 1.7 points when attention is introduced after the RCNN. This shows that 3D context is important for high quality proposal generation. Since only RPN detected ROIs are used to crop feature maps, addition of attention after the RPN only improves the ability to reject FPs. In this case, attention cannot improve the retrieval of lesions that are otherwise missed.

Feature bags size: Table 3.2f compares the network performance as the feature bag size varies between 3, 7, 9 and 11 images. While dice per case increases with feature bag size, the small and medium lesion performance starts to worsen beyond a bag size of 9. We thus adopt this size in

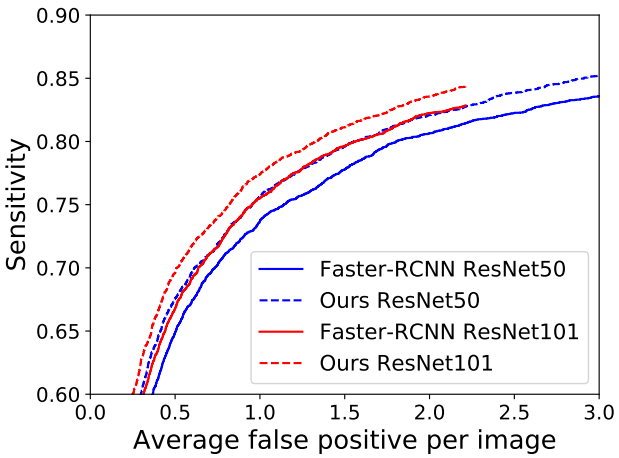


Figure 3.4: FROC curves on the DeepLesion `test` set.

Table 3.3: Sensitivity (%) at 1 and 2 FPs/image on the official split `test` set of DeepLesion.

Model	Backbone	1 FPs	2 FPs
No 3D context	VGG-16	60.57	71.19
Faster-RCNN[YWLS18]	VGG-16	67.26	75.57
R-FCN[DLHS16]	VGG-16	67.26	75.37
Improved R-FCN [DLHS16]	VGG-16	67.65	76.89
Data-level fusion, 11 slices	VGG-16	70.03	77.89
3-DCE,9 Slices[YBS18]	VGG-16	70.68	79.09
3-DCE,27 Slices[YBS18]	VGG-16	73.37	80.70
VA Faster-RCNN, 9 Slices	ResNet50	**75.63**	**82.45**
VA Faster-RCNN, 9 Slices	ResNet101	**77.42**	**83.67**

the remaining experiments. We note, however, that for applications sensitive to inference time, smaller bag size may be preferable.

3.4.5 Extension Experiments on DeepLesion

To test the effectiveness of volumetric attention for the processing of 3D CT volume datasets, we performed some extension experiments on DeepLesion. This dataset enables the use of part of the 3D CT volume as context for 2D bounding box prediction on target slices. Since DeepLesion does not provide mask groundtruth, the VA module was implemented on two Faster-RCNN-FPN detectors, with ResNet50 and ResNet101 backbones. As usual for DeepLesion, performance is evaluated with FPs per image. All experiments in this section are based on

training with the DeepLesion `train` and `val` sets, and testing on `test` set. Each 2.5D image is formed by concatenating 3 contiguous slices and scaled to 512$\times$ 512 pixels as in [YWLS18], the Faster-RCNN-FPN backbone is pretrained on ImageNet.

Table 3.3 and Fig. 3.4, compare the proposed networks to several methods from the literature. The proposed networks achieve state of the art results, increasing sensitivity by more than 4 points at 1 Fp/image and 2.97 at 2 FPs/image. Fig.3.4, shows that the proposed network with the ResNet50 backbone is almost equivalent to the much heavier Faster-RCNN with ResNet101 backbone. This confirms that VA attention is effective for both 3D segmentation and detection.

3.5 Acknowledgment

This Chapter, in part, has been submitted for publication of the material as it may appear in International Conference on Medical Image Computing and Computer Assisted Intervention(MICCAI), 2019, Xudong Wang, Zhaowei Cai, Dashan Gao, Nuno Vasconcelos., Springer, 2019. The thesis author was the primary investigator and author of this paper.

Chapter 4

Conclusions

In this thesis, we mainly focused on domain attentive representations for universal objects detection and volumetric attentive representations for 3D medical images segmentation. For the first one, We have investigated the unexplored and challenging problem of universal/multi-domain object detection. We proposed a universal detector that requires no prior domain knowledge, consisting of a *single* network that is active for all tasks. The proposed detector achieves domain sensitivity through a novel data-driven domain adaptation module and was shown to outperform multiple universal/multi-domain detectors on a newly established benchmark, and even individual detectors optimized for a single task.

We also proposed a volumetric attention module that enables 2.5D methods to leverage contextual information along the z direction and the use of pretrained 2D detection models when training data is limited, as is often the case for medical applications. VA can be combined with any CNN architecture, including one-stage and two-stage detectors and segmentation networks. It was shown that 2.5D networks with VA achieve state of the art results for *both* lesion segmentation and detection.

Bibliography

[BCV+19] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jrgen Hesser, Samuel Kadoury, Tomasz Konopczynski, Miao Le, Chunming Li, Xiaomeng Li, Jana Lipkov, John Lowengrub, Hans Meine, Jan Hendrik Moltz, Chris Pal, Marie Piraud, Xiaojuan Qi, Jin Qi, Markus Rempfler, Karsten Roth, Andrea Schenk, Anjany Sekuboyina, Eugene Vorontsov, Ping Zhou, Christian Hlsemeyer, Marcel Beetz, Florian Ettlinger, Felix Gruen, Georgios Kaissis, Fabian Lohfer, Rickmer Braren, Julian Holch, Felix Hofmann, Wieland Sommer, Volker Heinemann, Colin Jacobs, Gabriel Efrain Humpire Mamani, Bram van Ginneken, Gabriel Chartrand, An Tang, Michal Drozdzal, Avi Ben-Cohen, Eyal Klang, Marianne M. Amitai, Eli Konen, Hayit Greenspan, Johan Moreau, Alexandre Hostettler, Luc Soler, Refael Vivanti, Adi Szeskin, Naama Lev-Cohain, Jacob Sosna, Leo Joskowicz, and Bjoern H. Menze. The liver tumor segmentation benchmark (lits), 2019.

[BV17] Hakan Bilen and Andrea Vedaldi. Universal representations: The missing link between faces, text, planktons, and cat breeds. *arXiv preprint arXiv:1701.07275*, 2017.

[ÇAL+16] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[CEE+16] Patrick Ferdinand Christ, Mohamed Ezzeldin A Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin DAnastasi, et al. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 415–423. Springer, 2016.

[CFFV16] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *ECCV*, pages 354–370, 2016.

[CMMS17] Grzegorz Chlebus, Hans Meine, Jan Hendrik Moltz, and Andrea Schenk. Neural network-based automatic liver tumor segmentation with random forest-based candidate filtering. *arXiv preprint arXiv:1706.00842*, 2017.

[CPW+18] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. mmdetection. https://github.com/open-mmlab/mmdetection, 2018.

[CV18] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018.

[DKC10] Mark Dredze, Alex Kulesza, and Koby Crammer. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning*, 79(1-2):123–149, 2010.

[DLHS16] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[EEVG+15] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[EMP05] Theodoros Evgeniou, Charles A Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(Apr):615–637, 2005.

[FGMR10] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[GDDM14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[Gir15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[GLU12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[GRM+16] Georgios Georgakis, Md Alimoor Reza, Arsalan Mousavian, Phi-Hung Le, and Jana Kosecka. Multiview rgb-d dataset for object instance detection. *arXiv preprint arXiv:1609.07826*, 2016.

[Han17] Xiao Han. Automatic liver lesion segmentation using a deep convolutional neural network method. *arXiv preprint arXiv:1704.07239*, 2017.

[HGD18] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. *arXiv preprint arXiv:1811.08883*, 2018.

[HGDG17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[HR17] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *CVPR*, pages 1522–1530, 2017.

[HSS17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 7, 2017.

[HTP+17] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[IB05] Laurent Itti and Pierre Baldi. A principled approach to detecting surprising events in video. In *CVPR*, volume 1, pages 631–637, 2005.

[IFYA18] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, pages 5001–5009, 2018.

[IS15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[JCDR12] Mahesh Joshi, William W Cohen, Mark Dredze, and Carolyn P Rosé. Multi-domain learning: when do domains matter? In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1302–1312. Association for Computational Linguistics, 2012.

[JZCL08] Wei Jiang, Eric Zavesky, Shih-Fu Chang, and Alex Loui. Cross-domain learning methods for high-level visual concept classification. In *ICIP*, pages 161–164. IEEE, 2008.

[KCK+17] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

[KGS+17] Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017.

[KKSA08] Tsuyoshi Kato, Hisashi Kashima, Masashi Sugiyama, and Kiyoshi Asai. Multi-task learning via conic programming. In *NeurIPS*, pages 737–744, 2008.

[Kok17] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, page 8, 2017.

[KZM+12] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *ECCV*, pages 158–171. Springer, 2012.

[LAE+16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37, 2016.

[LCQ+18] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[LCWJ15] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.

[LD18] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[LDG+17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[LSNK17] Anan Liu, Yuting Su, Weizhi Nie, and Mohan S Kankanhalli. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(1):102–114, 2017.

[MDL18] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *ECCV*, pages 67–82, 2018.

[MSGH16] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, pages 3994–4003, 2016.

[MTM12] Andreas Møgelmose, Mohan M Trivedi, and Thomas B Moeslund. Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. *IEEE Trans. Intelligent Transportation Systems*, 13(4):1484–1497, 2012.

[NH16a] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*, pages 4293–4302, 2016.

[NH16b] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*, June 2016.

[NSCD17] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S Davis. Ssh: Single stage headless face detector. In *ICCV*, pages 4885–4894, 2017.

[Pal99] Stephen E Palmer. *Vision science: Photons to phenomenology*. MIT press, 1999.

[PGLC15] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

[QLW+18] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. 2018.

[RBV17] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *NeurIPS*, pages 506–516, 2017.

[RBV18] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *CVPR*, pages 8119–8127, 2018.

[RDGF16] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[RF18] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[RHGS15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015.

[RHGS17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):1137–1149, 2017.

[RT18] Amir Rosenfeld and John K Tsotsos. Incremental learning through deep adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[THSD17] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, page 4, 2017.

[THZ+14] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[TPW16] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

[Vai93] Parishwad P Vaidyanathan. *Multirate systems and filter banks*. Pearson Education India, 1993.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[VTPK18] Eugene Vorontsov, An Tang, Chris Pal, and Samuel Kadoury. Liver lesion segmentation informed by joint liver segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1332–1335. IEEE, 2018.

[WGGH18] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[Wol00] Jeremy M Wolfe. Visual attention. In *Seeing*, pages 335–386. Elsevier, 2000.

[WPLK18] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.

[WSL+15] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Towards unified depth and semantic prediction from a single image. In *CVPR*, pages 2800–2809, 2015.

[XBD+18] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proc. CVPR*, 2018.

[Yan98] Steven Yantis. Control of visual attention. *attention*, 1(1):223–256, 1998.

[Yar67] Alfred L Yarbus. Eye movements during perception of complex objects. In *Eye movements and vision*, pages 171–211. Springer, 1967.

[YBS18] Ke Yan, Mohammadhadi Bagheri, and Ronald M Summers. 3d context enhanced region-based convolutional neural network for end-to-end lesion detection. In *ICCV*, pages 511–519, 2018.

[YCBL14] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NeurIPS*, pages 3320–3328, 2014.

[YCL16] Fan Yang, Wongun Choi, and Yuanqing Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *CVPR*, pages 2129–2137, 2016.

[YH14] Yongxin Yang and Timothy M Hospedales. A unified perspective on multi-domain and multi-task learning. *arXiv preprint arXiv:1412.7489*, 2014.

[YLBP17] Jianwei Yang, Jiasen Lu, Dhruv Batra, and Devi Parikh. A faster pytorch implementation of faster r-cnn. *https://github.com/jwyang/faster-rcnn.pytorch*, 2017.

[YLLT16] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *CVPR*, pages 5525–5533, 2016.

[Yua17] Yading Yuan. Hierarchical convolutional-deconvolutional neural networks for automatic liver and tumor segmentation. *arXiv:1710.04540*, 2017.

[YWL+18] Ke Yan, Xiaosong Wang, Le Lu, Ling Zhang, Adam Harrison, Mohammadhadi Bagheri, and Ronald M Summers. Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. In *IEEE CVPR*, 2018.

[YWLS18] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5(3):036501, 2018.

[ZPIE17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[ZSS+18] Amir R Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, pages 3712–3722, 2018.

[ZWK19] Xingyi Zhou, Dequan Wang, and Philipp Krhenbhl. Objects as points, 2019.

[ZY17] Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2017.

[ZZLQ16] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[ZZW+17] Yousong Zhu, Chaoyang Zhao, Jinqiao Wang, Xu Zhao, Yi Wu, and Hanqing Lu. Couplenet: Coupling global structure with local parts for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4126–4134, 2017.