

# Automatic Multiparametric Magnetic Resonance Imaging-Based Prostate Lesions Assessment with Unsupervised Domain Adaptation

Jing Dai, Xiaomei Wang, Yingqi Li, Zhiyu Liu, Yui-Lun Ng, Jiaren Xiao, Joe King Man Fan, James Lam, Qi Dou, Varut Vardhanabhuti,\* and Ka-Wai Kwok\*

Multiparametric magnetic resonance imaging (mpMRI) has emerged as a valuable diagnostic tool in prostate lesion assessment. However, training convolutional neural networks (CNNs) inevitably involves magnetic resonance (MR) images from multiple cohorts. There always exists variation in scanning protocol among cohorts, inducing significant changes in data distribution between source and target domains. This challenge has greatly limited clinical adoption on a large scale. Herein, a coarse mask-guided deep domain adaptation network (CMD<sup>2</sup>A-Net) is proposed to develop a fully automated framework for prostate lesion detection and classification (PLDC). No category or mask label is required from the target domain. A coarse segmentation module is trained to cover the possible lesion-related regions, so that attention maps can be generated to dedicate the local feature extraction of lesions within those regions. Experiments are performed on 512 mpMRI sets from datasets of PROSTATEx (330 sets) and two cohorts, A (74 sets) and B (108 sets). Using ensemble learning, CMD<sup>2</sup>A-Net accomplishes an AUC of 0.921 in cohort A and 0.913 in cohort B, demonstrating its transferability from a large-scale public dataset PROSTATEx to small-scale target domains. Results from an ablation study also support its effectiveness in classification between benign and malignant lesions, compared to the state-of-the-art models. An interactive preprint version of the article can be found here: <https://doi.org/10.22541/au.166081031.11420810/v1>.


## 1. Introduction

Prostate cancer (PCa) is the second most prevalent cancer among males.<sup>[1]</sup> The number of diagnoses is estimated to increase by  $\approx 1.7$  million worldwide by 2030.<sup>[2]</sup> Accurate prostate lesion assessment, particularly for classifying clinically significant PCa (csPCa; Gleason score [GS]  $\geq 7$ )<sup>[3]</sup> from indolent non-csPCa, can vastly improve the facilitation of tailored treatments.<sup>[4]</sup> The broad range of PCa's behavioral pathology makes assessment challenging.<sup>[5]</sup> Current clinical assessment relies on prostate-specific antigen (PSA) blood testing, which, if positive, requires a transrectal ultrasound (TRUS) biopsy. However, PSA in conjunction with blind TRUS biopsy has a high false-negative rate ( $\approx 20\%$ ), resulting in unnecessary biopsies.<sup>[6]</sup> It is also highly prone to causing underdetection of csPCa or overdetection of non-csPCa.<sup>[7]</sup>

Multiparametric magnetic resonance imaging (mpMRI) has become a gold standard for PCa diagnosis, even prior to biopsy.<sup>[8]</sup> It typically involves T2-weighted (T2), high diffusion-weighted imaging

J. Dai, X. Wang, Y. Li, Z. Liu, Y.-L. Ng, J. Xiao, J. Lam, K.-W. Kwok  
Department of Mechanical Engineering  
The University of Hong Kong  
Hong Kong 999077, China  
E-mail: kwokkw@hku.hk

X. Wang  
Multi-Scale Medical Robotics Center Ltd.  
Hong Kong 999077, China

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/aisy.202200246>.

© 2023 The Authors. Advanced Intelligent Systems published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/aisy.202200246

J. K. M. Fan  
Department of Surgery  
The University of Hong Kong-Shenzhen Hospital  
Shenzhen, Guangdong 518000, China

J. K. M. Fan  
Department of Surgery  
Li Ka Shing Faculty of Medicine  
The University of Hong Kong  
Hong Kong 999077, China

Q. Dou  
Department of Computer Science and Engineering  
The Chinese University of Hong Kong  
Hong Kong 999077, China

V. Vardhanabhuti  
Department of Diagnostic Radiology  
The University of Hong Kong  
Hong Kong 999077, China  
E-mail: varv@hku.hk

(hDWI) sequence, and its derivative apparent-diffusion coefficient (ADC) maps.<sup>[4,9]</sup> Although the magnetic resonance imaging (MRI) acquisition and interpretation have been standardized with the guidance of Prostate Imaging Reporting and Data System version 2.1 (PI-RADS v2.1),<sup>[10]</sup> image interpretation is still time-consuming for the readers,<sup>[5]</sup> and inevitably significant inter-reader variation still exists.<sup>[11]</sup> To this end, numerous learning-based methods have been proposed to facilitate efficient, accurate, and reliable prostate lesion assessment. In 2017, an international contest PROSTATEx Challenge<sup>[12]</sup> was organized. Twenty-one teams proposed their models with the area under receiver operating characteristics (ROC) curves (AUC) ranging from 0.80 to 0.87.<sup>[13]</sup> Unlike the traditional methods relying on inputs with handcrafted features,<sup>[14]</sup> all of them employed CNNs<sup>[15]</sup> to detect the complex semantic features automatically, demonstrating significant advantages of PLDC over traditional methods.

To enhance network training, prostate MR images have to be pre-cropped manually, so as to retain the prostate region that originally occupies a small portion of the entire image set. Few recent studies, e.g., refs. [2,16], proposed automated PLDC frameworks to reduce effort from repeated manual prostate segmentation. CNNs were also utilized to segment the target region, identifying the prostate profile. These studies, despite notable progress, still assumed the training/testing datasets have to be shared the same data distribution from the source and target domains. This would be an overly ideal assumption,<sup>[17]</sup> as in normal practice, prostate MR images from a single cohort could not avoid the nature of medical data scarcity, or they are typically publicly unavailable.<sup>[14]</sup> Most likely, it is necessary to collect and aggregate images from multiple cohorts to maintain sufficient samples for robust model training. Inevitably, these multisite images exhibit apparent discrepancies in terms of scanning protocols, in-plane resolutions, field of views (FoV), etc.<sup>[17,18]</sup> These inherent intersite discrepancies would cause “domain shift” while having the models trained in the source domain, but applied in the target domain. This can significantly degrade the overall model performance, biasing the PLDC results.

Several paradigms have been proposed to resolve the domain shift. An intuitive solution is directly mixing heterogeneous images from multiple cohorts to make the training data adequate. However, in this approach, the model’s prediction capability could not be explicitly improved, and in contrast, would be limited by overfitting when distribution heterogeneity is significant.<sup>[18,19]</sup> Another common practice is pretraining the model in the source domain and then fine-tuning it in the target domain. This generally requires sufficient labeled data from the target domain to manually tune massive network parameters, which can still be a labor-intensive process. Domain adaptation (DA) has emerged as a more promising method, allowing effective knowledge transfer<sup>[17,20]</sup> from the label-rich source domain to the target domain. Recently, unsupervised DA (UDA) methods have drawn increased attention, as they do not require target labels for training.<sup>[21]</sup> These can be generally categorized as image translation and feature alignment approaches. In the former, the models can align image appearance<sup>[17,22]</sup> by translating images from one domain to another using generative models, such as generative adversarial networks (GANs).<sup>[23]</sup> Difficulties mainly come from whole-slide image translation, and image synthesis due to insufficient image similarity. Additionally, these models usually

focus on low-level feature extraction, suffering from inconspicuous lesion texture and characteristics.<sup>[24]</sup> In contrast, the latter, feature alignment-based models could be more effective in resolving domain shift by extracting domain-invariant features, either minimizing correlation distance between domains,<sup>[25]</sup> or assimilating feature distributions through adversarial learning.<sup>[26]</sup> Yet, very few of them are dedicated to prostate lesion detection and/or classification, particularly using mpMRI. Therefore, an effective UDA model for fully automated mpMRI-based PLDC is highly desirable for use prior to any invasive biopsy.

In this work, we develop a CMD<sup>2</sup>A-Net for both coarse prostate lesion detection and lesion malignancy classification. We also extend the proposed network to an open-sourced system. This executable end-to-end system takes mpMRI sequences as input, and outputs coarse lesion contours as well as lesion malignancy. The system can also be downloaded online. Our work contributions can be summarized below: 1) Development of a deep-learning-based system for fully automated prostate lesion assessment. Our end-to-end system is dedicated to PLDC on multicohort mpMRI without the need for prior manual processing on mpMRI sequences. 2) Design of a UDA model (i.e., CMD<sup>2</sup>A-Net) capable of leveraging cross-site representation transfer to realize accurate PLDC without requiring target labels. Weakly supervised coarse lesion segmentation modules are incorporated to extract informative lesion features, thus facilitating feature alignment between domains. 3) Experimental evaluation of CMD<sup>2</sup>A-Net on one public dataset (i.e., PROSTATEx<sup>[12]</sup>) and three local cohort datasets, including lesion assessments with various mpMRI sequence inputs, comparisons with state-of-the-art models, as well as an ablation study. The capability of transferring knowledge from PROSTATEx to our small-scale local cohort datasets is demonstrated against the state-of-the-art models.

## 2. Related Work

CNNs have been proved effective and widely applied for mpMRI-based PCa classification with promising performance. Wang et al.<sup>[13a]</sup> explored optimal combinations of mpMRI sequences as input for the CNN, and their model achieved an AUC of 0.95, which was reported to outperform all models in the PROSTATEx Challenge. Instead of only PCa classification, Kiraly et al.<sup>[27]</sup> developed a model with an encoder–decoder architecture to detect prostate lesions and simultaneously classify lesion malignancy. However, these studies required manually cropped regions of the prostate, which would be time-consuming and expensive.<sup>[22a,28]</sup>

End-to-end PLDC frameworks have also been investigated, with the aim to avoid the need for manual prostate segmentation. Yang et al.<sup>[2]</sup> incorporated a CNN for automatic segmentation in advance to the PLDC. Insufficient prostate image features extracted by the shallow network (i.e., five layers) could deteriorate the overall segmentation performance. Later, Wang et al.<sup>[29]</sup> proposed a deeper prostate segmentation model capable of detecting more complex features. Apart from improving the segmentation performance, fusing spatial features using 3D CNNs is also another means to enhance the accuracy of PCa classification. Mehta et al.<sup>[30]</sup> employed a patient-level 3D model for binary classification using volumetric mpMRI, achieving an AUC of 0.79 and 0.86 on their local cohort dataset and PROSTATEx,

respectively. However, only single-cohort datasets were used to evaluate the model. Domain shift would occur when it is directly applied to an unseen cohort.<sup>[17,18]</sup> Provided with very few studies (e.g., Mehta et al.<sup>[30]</sup>) that use mpMRI sequences from multiple cohorts, they could just directly combine the heterogeneous images, giving rise to sufficient samples for model training, but inevitably ignore data source heterogeneity. This approach would be prone to suffering from severe domain shift, thus biasing predictions by particular cohorts.

Very recently, many studies have attempted to investigate DA approaches to alleviate intersite distributional variability, among which UDA methods demonstrated their advantages in exploiting unlabeled target samples.<sup>[20]</sup> Such UDA methods can be categorized into two groups: 1) image translation and 2) feature alignment approaches. The former performs image appearance alignment.<sup>[17,22]</sup> The resultant models translate images across domains using GAN-based networks.<sup>[23]</sup> However, texture similarity between the synthesized target image and the source image would be crucial for the PLDC problem. The DA process would fail with insufficient texture similarity, particularly found in the generated lesion area.<sup>[22c]</sup> Lesions could also be missed during the translation process due to varying transferability among image regions, thus worsening the DA process.<sup>[31]</sup> Moreover, the GAN models would distort the nonlesion region's appearance, further causing unreliable lesion assessment results.<sup>[24]</sup>

By using feature alignment approaches, domain-invariant features are extracted to reduce domain shift.<sup>[26]</sup> A common way is to minimize distribution similarity (e.g., second-order correlation<sup>[25]</sup>) between domains using Siamese network architecture. Adversarial learning<sup>[26a]</sup> can also align features by enforcing the cross-domain features indistinguishable using a domain classifier. For instance, Wang et al.<sup>[14]</sup> developed a GAN-based method to learn domain-invariant features from mammographic images acquired for breast cancer screening. However, these models were usually trained with the entire images, treating all voxels equally.<sup>[26b,28]</sup> Previous works<sup>[24,26b]</sup> revealed that not all image regions can facilitate knowledge transfer across domains. Roughly aligning the features in the whole image set would introduce irrelevant knowledge, resulting in ineffective DA. It is hypothesized that the background regions on mpMRI sequences, such as regions outside the prostate gland, would not significantly improve DA in our PLDC problem. To our knowledge, only few researchers have reported PCa classification using multisite ultrasound images,<sup>[32]</sup> histopathology images,<sup>[33]</sup> or T2 image slices only.<sup>[13b]</sup>

### 3. Results and Discussion

#### 3.1. Datasets

Five datasets were utilized in this study, i.e., Initiative for Collaborative Computer Vision Benchmarking (I2CVB),<sup>[34]</sup> PROSTATEx (P-x), and three datasets from Hong Kong hospital local cohorts, LC-A, LC-B, and LC-C. Note that, LC-A and LC-B were acquired from the same MR imaging center. **Table 1** shows the characteristics of these five datasets. Note that I2CVB is already available online (<https://i2cvb.github.io/>), which has been widely investigated for prostate zone segmentation.<sup>[8]</sup> It contains 646 T2 images acquired from 36 patients. Fifteen patients were scanned by 3.0-T Siemens scanners and 21 patients by 1.5-T General Electric scanners. Given the segmentation labels on the prostate, central gland, peripheral zone, and lesion, only image slices covering the prostate were selected as our samples. A Mask R-CNN model was employed for prostate segmentation using this dataset. P-x (<https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=23691656>), LC-A, LC-B, and LC-C are mpMRI-based datasets marked with point labels. The four datasets share the same set of category labels (i.e., csPCa and non-csPCa). Datasets, P-x, LC-A, and LC-B, were utilized to evaluate the PLDC performance of our CMD<sup>2</sup>A-Net, including 330 cases from P-x, 74 cases from LC-A, and 108 cases from LC-B. To avoid “over-fitting” caused by LC-C (29 cases) with its small size, it was only used for cross-site heterogeneity analysis.

The mpMRI samples from multiple domains exhibit apparent interdomain heterogeneity,<sup>[35]</sup> which was caused by differences in MRI scanners, diffusion *b*-values, in-plane resolutions, FoV, and subject cohorts/patient populations. As shown in **Figure 1**, the MRI<sup>[36]</sup> examples from P-x and LC-A present apparent interdomain heterogeneity, demonstrating visible discrepancies in lesion morphology, prostate gland appearance, and image intensity distribution. These inherent multidomain discrepancies are inevitable, and cause “domain shift”<sup>[37]</sup> that significantly degrades overall model performance in PLDC.

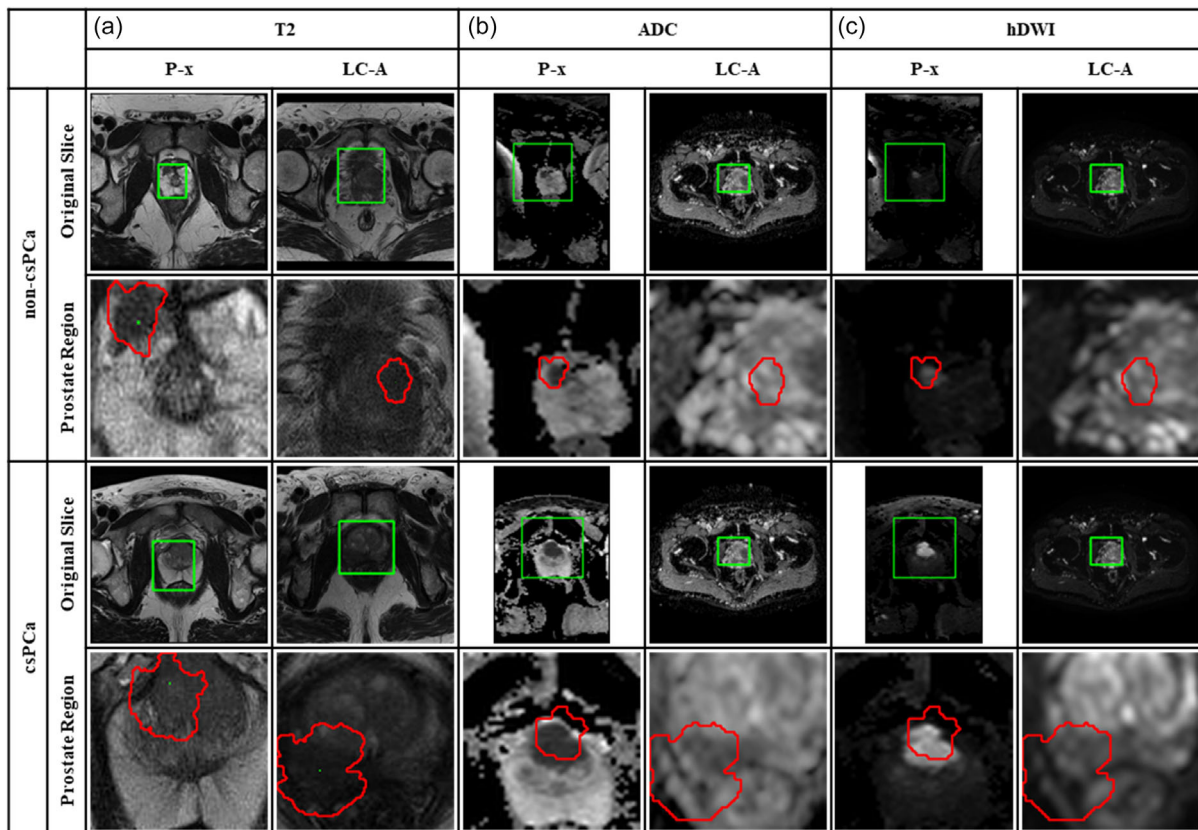
#### 3.2. Analysis of Cross-Site Heterogeneity

We first evaluated prostate segmentation performance using mean intersection over union (IoU), in order to ensure that the prostate regions can be predicted accurately. The IoU indicates the intersection between the predicted prostate contour and the ground truth mask label, which was measured on the

**Table 1.** Characteristics of the five MRI datasets for prostate segmentation and PLDC.

Datasets	Total cases	Positive cases	Negative cases	MRI scanner	Diffusion <i>b</i> -value [ <i>s</i> mm <sup>-2</sup> ]	In-plane resolution [mm]	Task		
							PSeg <sup>a)</sup>	CHA <sup>b)</sup>	PLDC
I2CVB	646	N/A	N/A	Siemens, and General Electric	N/A	0.68–0.79	✓	N/A	N/A
P-x	330	76	254	Siemens Trio and Skyra	50, 400, and 800	0.5	N/A	✓	✓
LC-A	74	51	23	Philips Achieva	0 and 1400	0.229	N/A	✓	✓
LC-B	108	14	94	Philips Achieva	1000 and 1400	0.315	N/A	✓	✓
LC-C	29	11	18	Siemens Skyra	N/A	0.625	N/A	✓	N/A

<sup>a)</sup>PSeg, prostate segmentation; <sup>b)</sup>CHA, cross-domain heterogeneity analysis.



**Figure 1.** Non-csPCa and csPCa mpMRI examples (from P-x and LC-A) of a) T2, b) ADC, and c) hDWI. The prostate gland is contoured with rectangles (in green) on the original slices (first and third row). The coarse lesion is contoured in red within the cropped prostate regions (second and fourth row) using the level set method, showing lesion morphological discrepancy of the benign and malignant samples. Apparent intersite heterogeneity (e.g., FoV, image intensity distribution) of the samples demonstrates domain shift between P-x and LC-A.

test split of I2CVB. The mean IoU of the prostate region, central gland, and peripheral zone is 0.843, 0.781, and 0.516, respectively. These results are comparable with the work of Alkadi et al.<sup>[8]</sup> which attained an IoU of 0.673 and 0.599 for the central gland and peripheral zone, respectively. This implies that the training set, which contains MR images from 36 patients, is already sufficient for accurate prostate segmentation. Additionally, the segmentation results are found to be promising on the image obtained from either a 1.5-T or 3.0-T MRI machine, indicating that the IoU measuring is not sensitive to the scanner types (see details in Figure S1, Supporting Information).

Then, we analyzed cross-site heterogeneity on our multicohort datasets (P-x, LC-A, LC-B, and LC-C). We aim to verify whether the prior MR image intensity normalization (e.g., Liu et al.<sup>[38]</sup>) is effective to reduce domain shift, when domain knowledge is not considered. Coarse Mask-guided Network (i.e., CM-Net, in Figure 5) was utilized for cross-site heterogeneity analysis. Here, training a model on an individual dataset is defined as a “separate learning approach,” while training a model using a combined dataset from multiple cohorts is defined as a “joint learning approach.” As shown in Table 2, we trained the CM-Net using the individual and combined datasets from P-x, LC-A, and LC-B. The three separate models were individually trained in these three domains. They were set as the baselines for comparisons with the joint models. During the testing phase, each separate

model was tested on the four datasets. LC-C only acted as the hold-out testing set for domain shift analysis, as its small size (only 29) would cause overfitting in training and biased prediction in testing. Note that, owing to the limited sample size of local cohorts (74 and 108 cases on LC-A and LC-B, respectively), separate models of LC-A and LC-B were pretrained on the large-scale dataset P-x (330 cases) and then fine-tuned on the corresponding image domain. Such a transfer learning strategy would reduce overfitting caused by data scarcity. A common preprocessing method, scaled, was employed to normalize the image intensities within [0,1].

The results of separate models from P-x, LC-A, and LC-B are shown in Table 2. For the three sequences (i.e., T2, ADC, and hDWI), the AUCs of three separate models are relatively high when tested within their respective domains, but these AUCs sharply drop when directly tested in the unseen domains. Such results show the sensible cross-domain discrepancy (i.e., domain shift) among the four datasets. Note that, in terms of the T2 sequence, separate models of LC-A and LC-B accomplish the highest testing AUCs (0.66 and 0.67) in the unseen domain, LC-C, which is just marginally higher than those (0.61) within their corresponding domains. A potential reason for the biased predictions is the deficiency of testing samples (i.e., 29) on LC-C. When it comes to the joint models in the table, they cannot bring remarkable improvements in each sequence compared with the



**Table 2.** Comparisons of AUC using separate and joint learning approaches.

Datasets	T2				ADC				hDWI			
	P-x	LC-A	LC-B	LC-C	P-x	LC-A	LC-B	LC-C	P-x	LC-A	LC-B	LC-C
P-x only	<b>0.91</b>	0.35	0.55	0.65	<b>0.67</b>	0.53	0.42	0.51	<b>0.81</b>	0.41	0.68	0.50
LC-A only	N/A	<b>0.61</b>	0.55	0.66	N/A	<b>0.69</b>	0.38	0.53	N/A	<b>0.70</b>	0.57	0.49
LC-B only	N/A	0.39	<b>0.61</b>	0.67	N/A	0.52	<b>0.61</b>	0.48	N/A	0.47	<b>0.88</b>	0.59
Joint P-x, LC-A	0.89	0.67	N/A	N/A	0.73	0.54	N/A	N/A	0.73	0.54	N/A	N/A
Joint P-x, LC-B	0.88	N/A	0.59	N/A	0.66	N/A	0.55	N/A	0.76	N/A	0.87	N/A
Joint LC-A, LC-B	N/A	0.63	0.53	N/A	N/A	0.74	0.61	N/A	N/A	0.72	0.91	N/A

The bold/shading data highlight the baseline models performance in the three sequences (T2, ADC, hDWI).

separate models; instead, they may even lead to performance degradation due to cross-site heterogeneity.

With severe discrepancies among our datasets, we intend to validate whether the rigorous MR image preprocessing methods can contribute to classification performance of the joint models. Similar to scaled, whitening is another common preprocessing method, capable of normalizing the pixel values with a mean of zero and unit variance. We took the combined dataset, P-x and LC-A, as a representative for evaluation. In **Table 3**, scaled, whitening, and their combined function with bias field correction (BFC) or noise filtering (NF), six preprocessing methods in total, were adopted as in ref. [38]. The joint models using scaled and whitening acted as the two baselines for comparisons with the

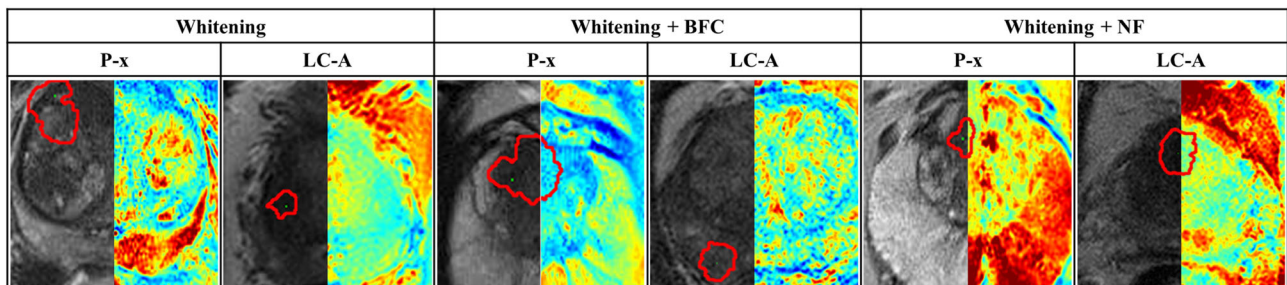
rigorous MR image preprocessing methods (i.e., BFC and NF). **Figure 2** depicts the image preprocessing examples of three methods (i.e., whitening, whitening + BFC, and whitening + NF). The left and right halves of each sample represent before and after preprocessing, respectively. Before preprocessing, we can observe noticeable intensity distribution discrepancies among the samples. The samples from LC-A are characterized by larger numbers of low-intensity grayscale pixels as compared with the images of P-x. Subsequently, the jet color maps were employed to visualize the intensity distribution between domains after preprocessing. All the color maps shared the same intensity color scale. Similar intensity distributions among the samples can be found after preprocessing, demonstrating the effectiveness of the methods in image distribution harmonization.

**Table 3.** Comparisons of AUC using six image preprocessing methods.

Preprocessing methods	T2		ADC		hDWI	
	P-x	LC-A	P-x	LC-A	P-x	LC-A
Scaled	0.89	0.67	<b>0.73</b>	0.54	0.73	0.54
Whitening	0.87	0.73	0.65	<b>0.72</b>	0.56	0.54
Scaled + BFC	0.90	0.71	0.67	0.65	0.76	<b>0.65</b>
Whitening + BFC	<b>0.91</b>	<b>0.80</b>	0.68	0.68	0.73	0.55
Scaled + NF	0.89	0.75	0.66	0.61	<b>0.80</b>	0.56
Whitening + NF	0.84	0.72	0.64	0.66	0.79	0.57

The bold/shading data indicate the maximum number in the columns.

In **Table 3**, for the T2 sequence, BFC with either scaled or whitening outperforms the baselines. BFC with whitening also achieves the best AUCs of 0.91 and 0.80 on P-x and LC-A, respectively. However, these findings are not consistent with the results in ADC and hDWI. In terms of ADC, the models preprocessed with BFC or NF underperform the baselines. Instead, the baseline models receive the highest AUCs, where scaled alone and whitening alone accomplish 0.73 and 0.72 on P-x and LC-A, respectively. When it comes to the sequence of hDWI, both BFC and NF demonstrate limited improvement over the baselines. On P-x, the AUC increases marginally from 0.73 (scaled only) to 0.80 (scaled with NF); on LC-A, only an AUC of 0.65 is achieved using scaled with BFC. The above results of the three sequences show that these preprocessing approaches could improve CM-Net's



**Figure 2.** Image preprocessing examples (from P-x and LC-A) in quantitative analysis on intersite heterogeneity. Among the six methods in **Table 3**, whitening, whitening + BFC, and whitening + NF act as representatives. Coarse lesion region is contoured (in red) on the randomly selected pre-cropped T2 images. Prior to the preprocessing (left half), the heterogeneity of intensity distribution can be observed obviously in the original samples, while the distributions are harmonized after the preprocessing (right half). All the jet color maps share the same scale.

classification performance when combining our two datasets. However, none of the methods is capable of boosting the joint models' generalization considerably, as compared with the separate models of P-x and LC-A (in Table 2). This indicates that the preprocessing methods are probably insufficient to solve domain shift fundamentally. A possible reason is that the severe discrepancies may also come from the intersite discrepancies (in Table 1), rather than only from the intensity distribution of the heterogeneous mpMRI sequences (see details in Figure 1).

### 3.3. Cross-Domain Malignancy Classification and Lesion Detection

We emphasized the importance of knowledge transfer from a large-scale public dataset to a small-scale target domain. The malignancy estimation performance of CMD<sup>2</sup>A-Net (the architecture is shown in Figure 5) was evaluated. The dataset, P-x, was only regarded as the source domain. Either LC-A or LC-B was also set as the source domain for knowledge transfer between local cohorts. The scaled method was employed for image preprocessing. In general, available types of MR sequences may vary in healthcare institutions. Thus, we employed ensemble learning to handle multiple sequences, allowing the use of single and multiple sequence(s) in our framework. Three common metrics were adopted for classification performance evaluation, i.e., AUC, sensitivity (SEN), and specificity (SPE).

Table 4 illustrates the classification results (i.e., csPCa or non-csPCa). Seven sequence combinations were involved for comparisons. The former and the later domains in the table are denoted as the source and target domains, respectively. We define such pairs of domains/cohort as DA settings. First, we compared CMD<sup>2</sup>A-Net with the separate and joint models (in Table 2) in terms of AUC. Take the first DA setting (P-x → LC-A) as an example. In the T2 sequence, CMD<sup>2</sup>A achieves an AUC of 0.87 in the target domain (i.e., LC-A), outperforming both the separate model (AUC: 0.61) and the joint model (AUC: 0.67). Consistent findings can be observed in ADC and hDWI. When it comes to the other three DA settings, CMD<sup>2</sup>A-Net also demonstrates its advantage in resolving domain shift between two of our datasets. This validates our hypothesis that incorporating prostate lesion information in prior to the DA process can facilitate PCa classification.

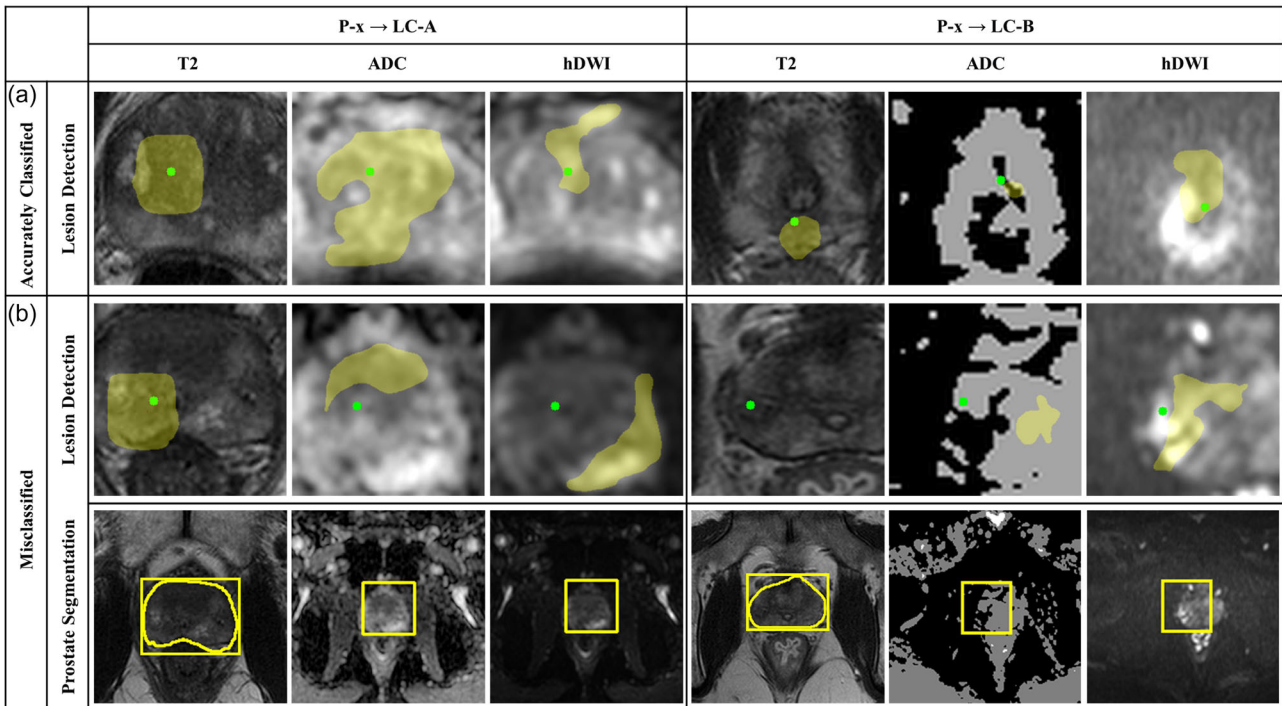
Second, we analyzed our model's PCa classification performance using a single sequence, i.e., T2, ADC, or hDWI. In most source-target DA settings, T2 is the most effective, while ADC receives the lowest AUC. The sequence, hDWI, shows unstable performance in the four DA settings. For example, it accomplishes the most superior performance (w.r.t. AUC, SEN, and SPE) in "P-x → LC-B," but underperforms T2 and ADC in "LC-A → LC-B." This could be caused by heterogeneous *b*-values among the domains. As shown in Table 1, *b*-values of 50, 400, and 800 s mm<sup>-2</sup> were employed on P-x, while 0 and 1400 s mm<sup>-2</sup> were used in LC-A, and 1000 and 1400 s mm<sup>-2</sup> were used in LC-B. Thus, we can conclude that the significant discrepancies in the acquisition parameters would result in the inconsistent performance of hDWI. Note that there were no widely accepted guidelines regarding *b*-value until the release of PI-RADS in 2019, which recommended a minimum value of 1200 s mm<sup>-2</sup>.

We also investigated the effect of ensemble learning using multiple sequences, which could provide references to choose appropriate sequences for PLDC. In each DA setting, the models using multiple sequences are always more effective than those relying on a single sequence. Besides, although ADC or hDWI always leads to the worst classification results, T2 ensemble with one or both can explicitly enhance the model's performance. This finding is consistent with the clinical practice of using mpMRI for PCa diagnosis. Sequences ADC and hDWI are usually regarded as secondary references by radiologists. It should be noted that the all-sequence-ensembled (i.e., ensemble of T2, ADC, and hDWI) models show significant predictions in most DA settings. Although an ensemble of the three sequences could not yield the best performance in the second DA setting (i.e., P-x → LC-A), the model still achieved a remarkable AUC of 0.91, which is only about 1% lower than the highest AUC (0.92). It can be concluded that using more sequences would help multicohort MRI harmonization, thus boosting the final classification performance. Moreover, with the same target domain (i.e., either LC-A or LC-B), the CMD<sup>2</sup>A-Net transferred from P-x attains a higher AUC than transferred from a local cohort domain in each sequence combination. This implies more source samples could enhance the model's cross-domain knowledge transferability, thus improving the model's generalization in the target domain. The superior performance also demonstrates CMD<sup>2</sup>A-Net's capability of transferring the knowledge from a public dataset to our local cohort domains.

**Table 4.** Malignancy classification results in the target domains in four combinations of source–target domains.

Sequences combinations	P-x → LC-A			P-x → LC-B			LC-A → LC-B			LC-B → LC-A		
	AUC	SEN	SPE	AUC	SEN	SPE	AUC	SEN	SPE	AUC	SEN	SPE
T2	0.87	0.79	0.80	0.84	0.75	0.79	0.65	0.60	0.59	0.70	0.60	0.62
ADC	0.79	0.76	0.66	0.75	0.74	0.73	0.64	0.62	0.59	0.67	0.62	0.66
hDWI	0.79	0.74	0.74	0.90	0.80	0.80	0.61	0.57	0.55	0.70	0.69	0.68
T2 + ADC	0.91	0.76	0.86	0.86	0.76	0.75	0.69	0.60	0.64	0.74	0.67	0.64
T2 + hDWI	0.92	0.80	0.84	<b>0.92</b>	<b>0.94</b>	0.68	0.68	0.61	0.64	0.73	0.62	0.68
ADC + hDWI	0.84	0.79	0.74	0.90	0.83	0.73	0.68	0.62	0.68	0.74	0.64	0.62
T2 + ADC + hDWI	<b>0.92</b>	<b>0.83</b>	<b>0.90</b>	0.91	0.79	<b>0.82</b>	<b>0.74</b>	<b>0.73</b>	<b>0.70</b>	<b>0.77</b>	<b>0.74</b>	<b>0.74</b>

The bold/shading data indicate the maximum number in the columns



**Figure 3.** Coarse lesion detection results of a) accurately classified and b) misclassified examples in target domains, LC-A and LC-B, relative to the source, P-x. All-sequence-enssembled (T2, ADC, hDWI) approach is employed. In lesion detection results (first and second rows), the lesions (ground-truth) are pointed in green. The predicted coarse lesion regions are colored in yellow. Promising prediction of lesion region, i.e., containing the ground-truth in all sequences, can yield the higher correctness of classification as in (a). Moreover, undersegmented prostate regions marked with yellow boxes/contours (i.e., the example of LC-B in the third row) would also worsen the classification outcome.

**Figure 3** shows coarse lesion detection results of the accurately classified and misclassified examples. Two DA settings (i.e., P-x to LC-A, and P-x to LC-B) were selected as representatives for lesion detection evaluation. Results of the all-sequence-enssembled method were selected as a representative for analysis. In the correctly classified examples, coarse lesion contours could encircle the lesion ground-truth point in all sequences (as shown in Figure 3a). However, in the unclassified examples, the coarse lesion position could not be precisely detected in most sequences as shown in the third row. In the example of LC-A, the lesion on the T2 image was correctly detected, but the lesion contours on ADC and hDWI maps were falsely identified. The possible reason is that the coarse lesion masks applied as the training ground truth could not depict the actual lesion contours accurately. Therefore, we can observe that accurate detection on ADC and hDWI also play a role in enhancing the ensemble classification, although lesion detection generally heavily relies on T2 images. In the future, robust weak label processing methods (e.g., deep extreme level set evolution method<sup>[39]</sup>) will be employed. For the example from LC-B, undersegmentation of the prostate region can be found on the T2 image, which could lead to failed lesion detection. As the prostate regions on ADC and hDWI were transformed using T2, under/oversegmentation of the prostate gland on T2 would deteriorate the lesion detection in the other two sequences. Despite the inaccurate lesion detection on ADC and hDWI, it should be noted that the models with multisequences input

still outperform the models using T2 alone in lesion classification, accredited to the reuse of prostate features from ADC and hDWI.

### 3.4. Comparisons with the State-of-the-Art Methods

We compared our model with three state-of-the-art models using AUC, i.e., Resnet50,<sup>[40]</sup> DANN,<sup>[41]</sup> and Deep Coral.<sup>[25]</sup> The dataset, P-x, was used as the source domain. Our local cohort datasets, LC-A and LC-B, acted as the target domains. The individual (i.e., T2, ADC, and hDWI) and the ensemble (i.e., T2 + ADC + hDWI) sequences were involved. The other ensemble sequences, T2 + ADC, T2 + hDWI, and ADC + hDWI, were not involved here due to their inferior performance as discussed in Section 4.2. Detailed comparison results are summarized in **Table 5**.

Resnet50 is a common classification model. It was pretrained from the source domain, and then tuned and tested in the target domain; therefore, no DA was utilized. In **Table 5**, Resnet50 underperforms all other methods in all sequences. A possible reason may be the weak cross-domain knowledge transferability of the fine-tune strategy. This shows the advantage of domain adaptation methods over the fine-tune strategy. Another model, DANN, is GAN-based and has been widely employed in lesion assessment. It can extract low-level features from the entire image. Deep Coral was also introduced, which can leverage domain knowledge transfer by aligning the second-order



**Table 5.** AUC comparisons on malignancy classification (i.e., csPCa or non-csPCa) with the three existing models.

Methods	P-x → LC-A				P-x → LC-B			
	T2	ADC	hDWI	Ensemble	T2	ADC	hDWI	Ensemble
Resnet50	0.65	0.70	0.55	0.70	0.66	0.71	0.68	0.71
DANN	0.70	0.72	0.66	0.79	0.68	0.66	0.75	0.80
Deep Coral	0.71	0.75	0.67	0.75	0.74	0.73	0.80	0.85
<b>CMD<sup>2</sup>A-Net (Ours)</b>	<b>0.87</b>	<b>0.79</b>	<b>0.79</b>	<b>0.92</b>	<b>0.84</b>	<b>0.75</b>	<b>0.90</b>	<b>0.91</b>

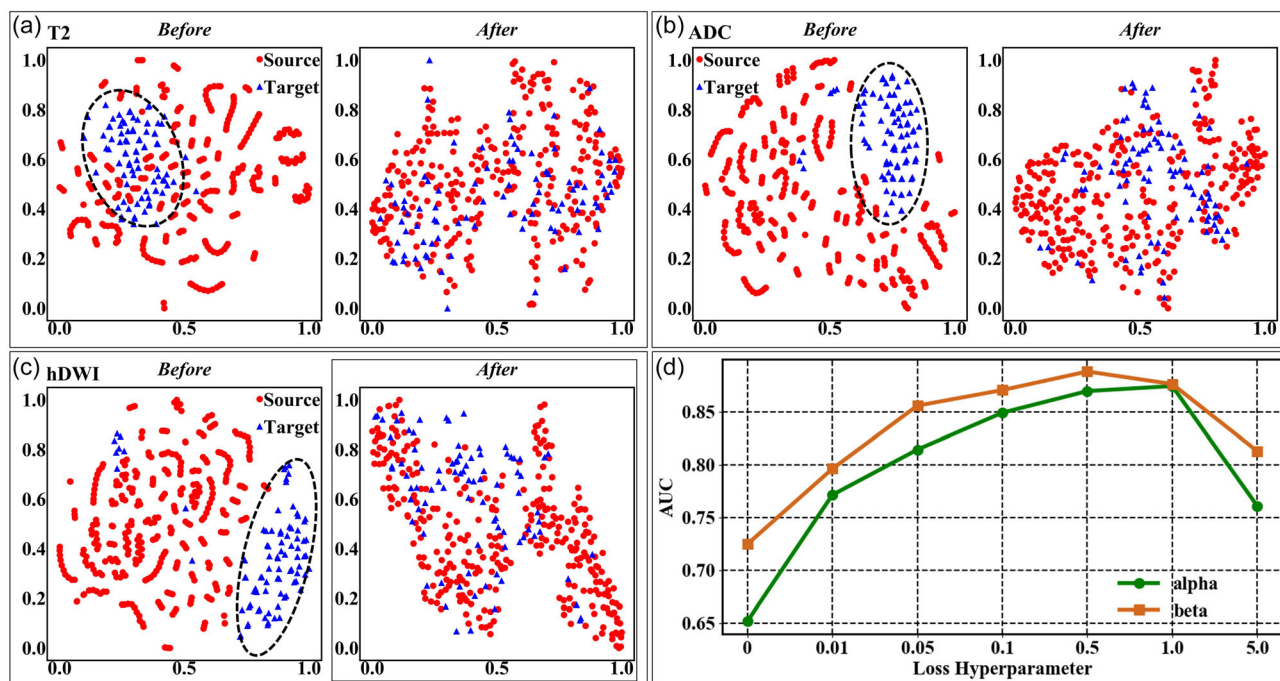
The bold/shading data indicate the maximum number in the columns

statistics. Similar to DANN, it also adopts a common encoder for feature extraction from the input of a whole image slice. Comparatively, our model could fuse both lesion features and prostate features for effective DA, instead of extracting the prostate features. We “strengthened” the point labels to be coarse mask labels, such that features, particularly lesion features, can be robustly aligned for DA using the mask labels. In Table 5, CMD<sup>2</sup>A-Net outperforms the two UDA models in all the sequences in terms of AUC, indicating the effectiveness of our model in cross-domain feature harmonization and its advantage in prostate lesion classification. It is worth noting that all four models accomplish their highest AUCs using the ensembled sequence. A consistent conclusion can be found in Section 4.2, showing the benefits of the all-sequence-ensbled method again.

### 3.5. Visualization of Sample Distribution and Ablation Study

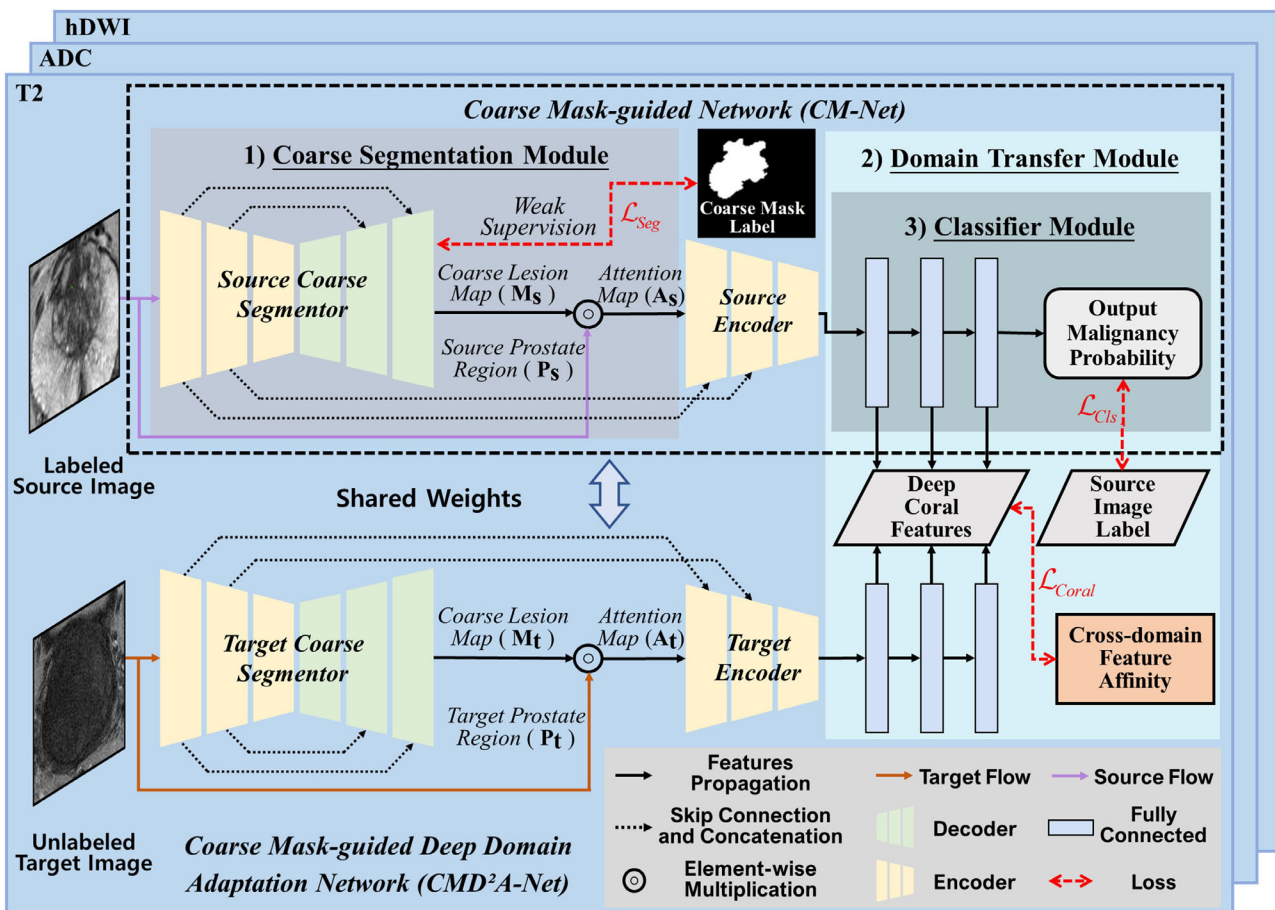
Apart from AUC, we also visualized the sample distribution of source and target domains, in support to any improved performance of handling domain shift intuitively. Datasets, P-x and LC-A, were adopted to visualize the data distribution before and after the DA. Algorithm, t-SNE,<sup>[42]</sup> was employed to visualize the data distributions of all sequences, i.e., T2, ADC, and hDWI. Fifty mpMRI cases from each dataset were randomly chosen. As shown in Figure 4a–c, obvious clustering can be observed before DA in each sequence, indicating severe domain shift between the two domains. After CMD<sup>2</sup>A-Net training (i.e., DA), domain-invariant features were extracted by the well-trained model. After the DA, samples from the two cohorts for each sequence are evenly distributed, proving that CMD<sup>2</sup>A-Net could assure feature alignment on the heterogenous mpMRI sequences.

To carry out the ablation study, we selected two key components, i.e., the coarse segmentation module and the domain transfer module, to analyze their contribution to lesion malignancy classification using T2 images. We compared our CMD<sup>2</sup>A-Net with its two variants using AUC, i.e., 1) CMD<sup>2</sup>A-Net excluding the domain transfer module (i.e., CM-Net, shown in the black dashed box in Figure 5) and 2) CMD<sup>2</sup>A-Net excluding the coarse segmentation modules (D<sup>2</sup>A-Net). As the CM-Net does not contain DA modules, it was trained in the source domain, and then fine-tuned and tested in the target domain. Datasets, P-x and LC-A, were selected as the source and target domain, respectively. D<sup>2</sup>A-Net obtains a lower AUC (0.65) compared with CMD<sup>2</sup>A-Net (0.87). This suggests that the coarse segmentation module is essential for domain-invariant feature



**Figure 4.** Sample distribution before and after DA for sequences a) T2, b) ADC, and c) hDWI using t-SNE. Similar change in distributions can be observed in all the sequences. Before DA, sample distributions of the source (red dots) and target (blue triangles) are dispersed in separate clusters, indicating severe domain shift. The mixed and even distribution after DA demonstrates the effectiveness of CMD<sup>2</sup>A-Net in feature alignment. d) Indicates the impact of hyperparameters (i.e.,  $\alpha$  and  $\beta$ ) in the loss sensitivity analysis.





**Figure 5.** Overview of the proposed CMD<sup>2</sup>A-Net using T2, ADC, and hDWI image inputs. Each image sequence network features two parallel branches with respect to the source and target domains. Three main modules in the source one: 1) coarse segmentation module for coarse lesion detection and feature alignment enhancement; 2) domain transfer module for knowledge transfer between domains; and 3) classifier module for malignancy classification.

extraction between domains. This also supports our hypothesis that the coarse lesion maps would enhance the malignancy classification accuracy. CM-Net obtains an AUC of 0.67, also less than CMD<sup>2</sup>A-Net. This indicates that the domain transfer module can substantially mitigate domain shift, thus enhancing CMD<sup>2</sup>A-Net's PCa classification performance.

The loss parameters sensitivity was also analyzed. CMD<sup>2</sup>A-Net was trained using P-x (source domain) and LC-A (target domain). Hyperparameters  $\alpha$  and  $\beta$  (i.e., weighting parameters of the total loss) in Equation (6) would influence the model's generalization ability essentially. The two hyperparameters could not be learned, which were preset prior to the model training. They were used to balance the contributions of the three network modules, such that joint optimization on all modules can be realized, thus facilitating the training process to reach equilibrium. Therefore, we manually tuned the hyperparameters in {0, 0.1, 0.05, 0.1, 0.5, 1.0, 5.0} to analyze the modules' contribution to lesion malignancy classification. As shown in Figure 4d, we could see a group of hyperparameters were preset to yield an optimal model. We can observe that our model demonstrates superior classification performance with  $\alpha$  within [0.1, 1.0] and  $\beta$  within [0.05, 1.0]. It should be noted that our model receives the lowest AUC when

either  $\alpha$  or  $\beta$  is set to 0, showing that the coarse segmentation module and the domain transfer module could enhance cross-domain knowledge transferability positively, thus improving lesion classification accuracy.

#### 4. Conclusion

In this article, we address the issue of performance heterogeneity in target domains arising from real-world usage across multiple sites/cohorts. We present a fully automated framework for mpMRI-based prostate lesion assessment. The framework involves a Mask R-CNN network to pre-crop the prostate region, and a novel UDA network (i.e., CMD<sup>2</sup>A-Net) for coarse lesion detection and malignancy classification (i.e., csPCa or non-csPCa). By introducing weakly supervised coarse segmentation modules, CMD<sup>2</sup>A-Net can incorporate both the prior lesion features and prostate features into the domain knowledge transfer process, yielding robust feature alignment between heterogeneous datasets. No labeling is necessary for the target domain. CMD<sup>2</sup>A-Net serves as a general UDA model primarily designed for PLDC, which could also be applied in other lesion assessment

tasks (e.g., liver tumors). Its PLDC performance has been evaluated on datasets of P-x, LC-A, and LC-B. The models with multi-sequence input accomplish higher AUC than any model using a single sequence only. The all-sequence-ensembled (T2, ADC, and hDWI) model demonstrates the most superior PCa classification performance w.r.t. AUC, SEN, and SPE. Additionally, when P-x acts as the source domain, the model ensemble with all the three sequences accomplishes an AUC of 0.921 in LC-A and 0.913 in LC-B, demonstrating its transferability from a large-scale public dataset P-x (330 cases) to our small-scale local cohorts (LC-A with 74 cases and LC-B with 108 cases). Experimental results also show that our model accomplishes higher AUC in PCa malignancy classification, compared to the state-of-the-art models, Resnet50, DANN, and Deep Coral. Other experimental results, including an ablation study and visualization of data distribution, further support the effectiveness of CMD<sup>2</sup>A-Net in domain adaptation. It is worth noting that our open-sourced system can be downloaded from GitHub (<https://github.com/jdai019/domain-adaptation-lesion-assessment.git>), capable to streamline the PLDC in an end-to-end manner without requiring manual prostate segmentation and annotation. We would be the first to develop a PLDC executable system available online for open usage, which is also deep-learning-based and trained by multicohort mpMRI sequences. In our future work, we will resolve few limitations: currently, lesions are distributed in different prostate zones (e.g., transition zone and peripheral zone). We will incorporate the prostate zones as input parameters to our model, in order to attain a higher AUC for prostate lesion assessment. Deep learning will also be used properly to facilitate effective feature extraction for the prostate zones.

## 5. Experimental Section

**Weakly Supervised Coarse Lesion Detection:** We employed Mask R-CNN<sup>[43]</sup> to crop the prostate region accurately for PLDC. When a sample is fed in, the prostate region and the remaining areas can be separated, respectively, as foreground (as known as image mask) and background. In general, the T2 sequence is necessary for the model input, while other ADC and hDWI are optional. A circumscribed rectangle (as the bounding box) of the detected prostate contour marks out the regions of interest (ROIs). The prostate mask on T2 images can also be applied to other accompanied input images (e.g., ADC, hDWI) through coordinate transformation, to obtain their corresponding ROIs. In comparison with Yang et al.<sup>[2]</sup> using a five-layer shallow segmentor, we chose a deeper feature extractor, i.e., Resnet50,<sup>[40]</sup> such that more complex features can be learned for more accurate prostate detection. Besides, a multiscale deep spatial feature extraction module, i.e., Feature Pyramid Networks,<sup>[44]</sup> was utilized to deal with FoV difference and varying prostate cross-sectional size.

Since fine delineation of the lesion region (e.g., the pixel-level label) is time-consuming, and demanding to even professionals, prostate lesions are commonly marked with a typical weak label,<sup>[45]</sup> i.e., point label. In general, weak labels are widespread in real-world applications. Recent work has explored various forms of weakly supervised labeling to alleviate the annotation effort, including point annotations, scribbles, and bounding boxes.<sup>[46]</sup> To accomplish prostate lesion detection, pixel-level labels are required prior to model training. However, the point label is insufficient to represent the prostate lesion area for training, as the lesion area not marked or pinpointed with such a point label would be probably miscategorized as healthy tissue.

We attempted to “strengthen” the existing point-level labels to coarse lesion areas by aggregating their neighbor pixels into a region through preprocessing. Such preprocessed areas would be comparable to “strong”

labels (i.e., manually labeled pixel-level contours), providing promising cues for lesion detection model training. Recently, Kiraly et al.<sup>[27]</sup> expanded the single marked pixel to a small-diameter circle using Gaussian kernels, but such a processing method focuses on lesion localization rather than contour approximation. Here, we applied a more sophisticated weak label processing method, i.e., distance regularized level set evolution,<sup>[47]</sup> to automatically generate a pixel-level weak label, i.e., a coarse mask label (in Figure 5). This level set method is an edge-based active contour approach. The labels can be produced in three steps: 1) initialize a level set function to represent the lesion contour originated from a manually marked point; 2) expand the lesion contour outward and update the level set function; and 3) terminate the expansion and finalize the function once exceeding the predefined iteration steps. As shown in Figure 1, several examples of the coarse mask labels were automatically generated using such a level set method. The coarse mask contours were annotated (in red) on the cropped prostate regions (2nd and 4th row). Therefore, the weak labels were “strengthened” from points to coarse lesion areas through preprocessing. The weak supervision would significantly reduce the time needed for accurate pixel-level annotation by experts, so as to enable coarse lesion detection and enhance malignancy classification.

Figure 5 illustrates the network architecture of the proposed CMD<sup>2</sup>A-Net. The coarse segmentation module outputs coarse lesion contour and also enables local feature extraction on lesion regions. Provided with more lesion features, the domain transfer module is introduced to facilitate feature alignment. A classifier module is incorporated for malignancy prediction. CMD<sup>2</sup>A-Net is trained on the three sequences (i.e., T2, ADC, and hDWI) individually. Based on the model output (i.e., lesions malignancy probability) of the three sequences, we can obtain the final malignancy predictions using ensemble learning. CMD<sup>2</sup>A-Net has two parallel branches with respect to (w.r.t.) the source and target domain, where two encoders extract features of prostate MR images separately in the two domains. The segmentors from the two domains share the same weights. The source segmentor is optimized by a supervised loss function (i.e., coarse lesion segmentation loss). Samples and coarse mask labels from the source domain are required for training. The segmentation loss  $\mathcal{L}_{\text{Seg}}$  can be defined as

$$\mathcal{L}_{\text{Seg}} = 1 - \frac{2 \sum_i^w \sum_j^h [m_{ij} s_{ij}] + \epsilon}{\sum_i^w \sum_j^h m_{ij} + \sum_i^w \sum_j^h s_{ij} + \epsilon} \quad (1)$$

where  $s_{ij}$  and  $m_{ij}$  indicate the pixel element values of mask label  $S$  and predicted lesion map  $M$ , respectively. Indices  $i$  and  $j$  denote the  $i$ th column and  $j$ th row of the image matrix in a dimension of  $w \times h$ . Constant value,  $\epsilon$  (set to  $10^{-5}$ ), is applied to avoid the zero-denominator case, as well as to guarantee numerical stability.

**Attention-Based Malignancy Estimation:** In recent studies of prostate lesion classification (e.g., Guan et al.<sup>[28]</sup>), lesion identification was suggested to be highly associated with disease-related regions in MR images. Instead of treating all pixels in the entire MR slice equally, an attention mechanism can be introduced to specifically extract lesion features. With these insights, we hypothesized that incorporating the prior knowledge of lesion regions into the DA process could enhance the model's classification performance. As illustrated in Figure 5, the two branches follow the same pipeline to generate attention feature maps. In each branch, the attention map can be produced using the prostate region and the coarse lesion mask, enabling our model to focus on the lesion region and also extract more lesion representations. The prostate region and the coarse lesion mask are denoted as  $P$  and  $M$ , respectively. Note that the subscripts “s” and “t” of variables (e.g.,  $M_s$  and  $M_t$ ) in Figure 5 represent the source and target domains, respectively. The attention maps of source and target domains,  $A_s$  and  $A_t$ , respectively, can be calculated by

$$A_i = \sigma(P_i \circ M_i), \quad i = s, t \quad (2)$$

where the operation  $\circ$  means the element-wise product, and the sigmoid function is denoted by  $\sigma$  which is adopted as the nonlinear activation to generate attention maps. Such a simple but effective function can

constrain each element of the feature maps in  $[0,1]$ , thus weighting the importance of regions. As a result, guided by coarse mask labels, the lesion areas would be assigned higher weights than the noninformative background (i.e., healthy tissue) in the feature maps.

To achieve accurate lesion classification, features from the lesion attention maps can be extracted by an encoder, such that high-level lesion features can be captured for the classifier module. Thus, in each branch, an encoder is incorporated after the segmentor to extract each domain's specific features. Besides, we proposed to fuse the lesion features and the prostate features to boost the classification accuracy. Skip connection and concatenation operations are introduced to reuse prostate features from the segmentors.

We designed a domain transfer module (in Figure 5) without requiring target labels in the training process. The semantics features from both the prostate region and attention map are fused, such that deep coral features from fully connected (FC) layers can be captured for feature affinity. Deep Coral loss<sup>[25]</sup> is employed to minimize cross-domain feature distribution discrepancy, owing to its generality, transferability, and ease of implementation. It is defined as the difference of second-order covariances between domains. Our domain transfer loss  $\mathcal{L}_{Coral}$  is defined as

$$\mathcal{L}_{Coral} = \sum_{i=1}^l \lambda_i \frac{1}{4d_i^2} \|C_s - C_t\|_F^2 \quad (3)$$

where  $l$  indicates the number of FC layers. Constants  $\lambda_i$ ,  $i = 1, 2, \dots, l$  are the weights that balance the contribution of FC layers, which are set to 1 here. The squared matrix Frobenius norm is denoted as  $\|\cdot\|_F^2$ . The dimension of the  $i$ th FC layer is indicated by  $d_i$ . The feature covariance matrices of source and target domains,  $C_s$  and  $C_t$ , respectively, can be calculated by

$$C_i = \frac{1}{n_i - 1} \left( D_i^T D_i - \frac{1}{n_i} (1^T D_i)^T (1^T D_i) \right), \quad i = s, t \quad (4)$$

where  $n_i$  denotes the number of images in the corresponding domain, and  $D_i$  indicates the feature matrices of the corresponding FC layer, and  $1$  is a column vector with all elements as 1.

To accomplish malignancy prediction using mpMRI, an ensemble learning approach is employed to fuse the predictions of the three separated models (w.r.t T2, ADC, and hDWI). We trained the classifier module, as in Figure 5, using labeled source data. The FC layers in the source domain are employed, not only for cross-domain feature affinity, but also for malignancy classification. The cross-entropy loss is utilized to optimize the classifier module. Our classification loss  $\mathcal{L}_{Cls}$  can be defined as

$$\mathcal{L}_{Cls} = \frac{1}{n_s} \sum_{i=1}^{n_s} -\hat{y}_i^s \log r - (1 - \hat{y}_i^s) \log(1 - r) \quad (5)$$

where variables  $\hat{y}_i^s$  and  $r$  denote the ground truth and the malignancy prediction w.r.t. each source sample, respectively.

The ultimate purpose of CMD<sup>2</sup>A-Net is to accomplish accurate PLDC. To this end, we simultaneously trained the coarse segmentation module, domain transfer module, and classifier module. Note that, minimizing segmentation loss alone would cause overfitting to the source domain, and only optimizing domain transfer loss would lead to generalization degradation in the target domain. Therefore, joint optimization on the total loss could facilitate the training process to reach equilibrium, such that the domain-invariant features could be extracted to achieve accurate classification. The total loss  $\mathcal{L}_{Total}$  can be defined to

$$\mathcal{L}_{Total} = \alpha \mathcal{L}_{Seg} + \beta \mathcal{L}_{Coral} + \mathcal{L}_{Cls} \quad (6)$$

where  $\alpha$  and  $\beta$  are weighting hyperparameters of the total loss. Both of them were set to 0.5 in our experiments.

To leverage the benefits of multiple sequences, we utilized the weighted average ensemble learning-based method. The outputs of the three

separated models were incorporated, thus contributing to the final ensemble prediction  $r_{ens}$  as follows

$$r_{ens} = \frac{r_T + \omega_A r_A + \omega_B r_B}{1 + \omega_A + \omega_B} \quad (7)$$

where  $r_T$ ,  $r_A$ , and  $r_B$  are the malignancy probability predictions of T2, ADC, and hDWI, for which the weights are 1,  $\omega_A$ , and  $\omega_B$ , respectively. Binary variables  $\omega_A$ ,  $\omega_B \in \{0, 1\}$  are assigned based on the availability of ADC and hDWI. For example, if the samples include ADC but without hDWI,  $\omega_A = 1$  and  $\omega_B = 0$ .

**Implementation Details:** Our models (i.e., Mask-RCNN model, CM-Net, and CMD<sup>2</sup>A-Net) were trained using a GeForce GTX 1080 Ti GPU (Nvidia, California, USA) with API Keras.<sup>[48]</sup> For the Mask-RCNN model training, data augmentation with random rotation was applied on the 646 T2 image slices on I2CVB. All the slices were split into training, validation, and testing sets in the ratio of 7:2:1. The input shape of Mask R-CNN was set to  $512 \times 512$  pixels. Adam optimizer was applied with a learning rate of  $10^{-3}$ . The batch size was set to 4 and the total epoch was 200. During the training process, the model with the highest dice coefficient score on the validation set was retained. For CM-Net and CMD<sup>2</sup>A-Net training, the prostate regions from P-x, LC-A, and LC-B were scaled to  $224 \times 224$  pixels. Random rotation of  $\{\pm 3^\circ, \pm 6^\circ, \pm 9^\circ, \pm 12^\circ, \pm 15^\circ\}$  was applied for data augmentation. Adam optimizer was chosen, and its learning rate was set to  $10^{-5}$ . The batch size was set as 2. In the training process of CM-Net, due to the limited sample size, all the slices were split into training and testing sets in the ratio of 4:1 using the hold-out method. The segmentation loss was optimized first to accelerate model convergence, and CM-Net with the pretrained coarse segmentation module was further trained. In terms of CMD<sup>2</sup>A-Net, we initialized both of its branches first using the weight of pretrained CM-Net, in order to facilitate its convergence. To be specific, we trained both the coarse segmentation module and classifier of CM-Net first, with the combined samples from both domains. Then, we optimized the total loss of CMD<sup>2</sup>A-Net with labeled source samples and unlabeled target samples. By cotraining all the modules, the model with the highest accuracy was saved for malignancy evaluation in the target domain.

We also offered our executable codes and files online available via GitHub, so as to allow any work extension or application by others. This open-sourced deep-learning-based model acts as an end-to-end system, with input from prostate mpMRI sequences (i.e., T2, ADC, and hDWI), and output to prediction results (i.e., prostate segmentation, coarse lesion detection, and malignancy estimation). The system supports multifomat inputs, including DICOM, jpeg, png, and jpg files. It is emphasized that no manual prostate segmentation or annotation is required.

## Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

## Acknowledgements

This work was partially supported by the Research Grants Council (RGC) of Hong Kong (Ref. Nos.: 17209021, 17207020, 17205919, and T42-409/18-R), Innovation and Technology Commission (ITC) (Project No.: MRP/029/20X), and Multi-scale Medical Robotics Center Limited. The authors thank J.D.L. Ho for assistance with polishing the language and proofreading the manuscript. K.W.K. and V.V. are co-corresponding authors.

## Conflict of Interest

The authors declare no conflict of interest.



## Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

## Keywords

convolutional neural networks, domain adaptation, multiparametric magnetic resonance imaging (mpMRI), prostate lesion detection and classification

Received: July 30, 2022

Revised: April 29, 2023

Published online: June 29, 2023

- [1] R. L. Siegel, K. D. Miller, A. Jemal, *Ca-Cancer J. Clin.* **2015**, *65*, 5.
- [2] X. Yang, C. Liu, Z. Wang, J. Yang, H. Le Min, L. Wang, K.-T. T. Cheng, *Med. Image Anal.* **2017**, *42*, 212.
- [3] A. C. Vidal, L. E. Howard, D. M. Moreira, R. Castro-Santamaria, G. L. Andriole, S. J. Freedland, P. Biomarkers, *Cancer Epidemiol.* **2014**, *23*, 2936.
- [4] Z. Wang, Y. Lin, K.-T. T. Cheng, X. Yang, *Med. Image Anal.* **2020**, *59*, 101565.
- [5] A. Saha, M. Hosseinzadeh, H. Huisman, *Med. Image Anal.* **2021**, *73*, 102155.
- [6] J. S. Quon, B. Moosavi, M. Khanna, T. A. Flood, C. S. Lim, N. Schieda, *Insights into Imaging* **2015**, *6*, 449.
- [7] F. H. Schröder, J. Hugosson, M. J. Roobol, T. L. Tammela, S. Ciatto, V. Nelen, M. Kwiatkowski, M. Lujan, H. Lilja, M. Zappa, *N. Engl. J. Med.* **2009**, *360*, 1320.
- [8] R. Alkadi, F. Taher, A. El-Baz, N. Werghi, *J. Digital Imaging* **2019**, *32*, 793.
- [9] D. Fehr, H. Veeraraghavan, A. Wibmer, T. Gondo, K. Matsumoto, H. A. Vargas, E. Sala, H. Hricak, J. O. Deasy, *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, E6265.
- [10] B. Turkbey, A. B. Rosenkrantz, M. A. Haider, A. R. Padhani, G. Villeirs, K. J. Macura, C. M. Tempny, P. L. Choyke, F. Cornud, D. J. Margolis, *Eur. Urol.* **2019**, *76*, 340.
- [11] T. T. Stolk, I. J. de Jong, T. C. Kwee, H. B. Luiting, S. V. Mahesh, B. H. Doornweerd, P.-P. M. Willemsse, D. Yakar, *Abdom. Radiol.* **2019**, *44*, 1044.
- [12] S. G. Armato, H. Huisman, K. Drukker, L. Hadjiiski, J. S. Kirby, N. Petrick, G. Redmond, M. L. Giger, K. Cha, A. Mamonov, *J. Med. Imaging* **2018**, *5*, 044501.
- [13] a) Y. Wang, M. Wang, *Phys. Med.* **2020**, *80*, 92; b) A. Grebenisan, A. Sedghi, J. Iazard, R. Siemens, A. Menard, P. Mousavi, in *IEEE 18th Int. Symp. Biomed. Imaging (ISBI)*, Nice, France **2021**, pp. 1218–1222
- [14] Y. Wang, Y. Feng, L. Zhang, Z. Wang, Q. Lv, Z. Yi, *Med. Image Anal.* **2021**, *73*, 102147.
- [15] S. Albawi, T. A. Mohammed, S. Al-Zawi, in *Int. Conf. Eng. Technol. (ICET)*, Antalya, Turkey **2017**, pp. 1–6.
- [16] S. Wang, L. Yu, K. Li, X. Yang, C.-W. Fu, P.-A. Heng, *IEEE Trans. Med. Imaging* **2020**, *39*, 4237.
- [17] H. Guan, M. Liu, *IEEE Trans. Biomed. Eng.* **2021**, *69*, 1173.
- [18] G. Mårtensson, D. Ferreira, T. Granberg, L. Cavallin, K. Oppedal, A. Padovani, I. Rektorova, L. Bonanni, M. Pardini, M. G. Kramerberger, *Med. Image Anal.* **2020**, *66*, 101714.
- [19] a) E. Gibson, Y. Hu, N. Ghavami, H. U. Ahmed, C. Moore, M. Emberton, H. J. Huisman, D. C. Barratt, in *Med. Image Comput. Assist. Interv.*, Granada, Spain **2018**, pp. 506–514; b) L. Rundo, C. Han, J. Zhang, R. Hataya, Y. Nagano, C. Militello, C. Ferretti, M. S. Nobile, A. Tangherloni, M. C. Gilardi, in *Neural Approaches to Dynamics of Signal Exchanges*, Vol. 151 (Eds: M. F.-Z. Anna Esposito, F. C. Morabito, E. Pasero) Springer, Singapore **2020**, Ch. 25; c) J. A. Nielsen, B. A. Zielinski, P. T. Fletcher, A. L. Alexander, N. Lange, E. D. Bigler, J. E. Lainhart, J. S. Anderson, *Front. Hum. Neurosci.* **2013**, *7*, 599.
- [20] C. Pei, F. Wu, L. Huang, X. Zhuang, *Med. Image Anal.* **2021**, *71*, 102078.
- [21] C. Bian, C. Yuan, J. Wang, M. Li, X. Yang, S. Yu, K. Ma, J. Yuan, Y. Zheng, *Med. Image Anal.* **2020**, *64*, 101732.
- [22] a) C. Chen, Q. Dou, H. Chen, J. Qin, P. A. Heng, *IEEE Trans. Med. Imaging* **2020**, *39*, 2494; b) V. Kearney, B. P. Ziemer, A. Perry, T. Wang, J. W. Chan, L. Ma, O. Morin, S. S. Yom, T. D. Solberg, *Radiol.: Art. Int.* **2020**, *2*, e190027; c) G. Zeng, F. Schmaranzer, T. D. Lerch, A. Boschung, G. Zheng, J. Burger, K. Gerber, M. Tannast, K. Siebenrock, Y.-J. Kim, in *Med. Image Comput. Assist. Interv.*, Lima, Peru **2020**, pp. 447–456
- [23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1.
- [24] X. Liu, X. Guo, Y. Liu, Y. Yuan, *Med. Image Anal.* **2021**, *71*, 102052.
- [25] B. Sun, K. Saenko, in *Computer Vision – ECCV 2016 Workshops*, Amsterdam, The Netherlands **2016**, pp. 443–450.
- [26] a) G. Wei, C. Lan, W. Zeng, Z. Chen, in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN **2021**, pp. 16643–16653; b) J. Liu, H. Liu, S. Gong, Z. Tang, Y. Xie, H. Yin, J. P. Niyoyita, *Med. Image Anal.* **2021**, *72*, 102135.
- [27] A. P. Kiraly, C. Abi Nader, A. Tuysuzoglu, R. Grimm, B. Kiefer, N. El-Zehiry, A. Kamen, in *Med. Image Comput. Assist. Interv.*, Quebec, Canada **2017**, pp. 489–497.
- [28] H. Guan, Y. Liu, E. Yang, P.-T. Yap, D. Shen, M. Liu, *Med. Image Anal.* **2021**, *71*, 102076.
- [29] Z. Wang, C. Liu, D. Cheng, L. Wang, X. Yang, K.-T. Cheng, *IEEE Trans. Med. Imaging* **2018**, *37*, 1127.
- [30] P. Mehta, M. Antonelli, H. U. Ahmed, M. Emberton, S. Punwani, S. Ourselin, *Med. Image Anal.* **2021**, *73*, 102153.
- [31] A. Choudhary, L. Tong, Y. Zhu, M. D. Wang, *Yearb. Med. Inf.* **2020**, *29*, 129.
- [32] Y. Shao, J. Wang, B. Wodlinger, S. E. Salcudean, *IEEE Trans. Med. Imaging* **2020**, *39*, 3148.
- [33] J. Ren, I. Hachililoglu, E. A. Singer, D. J. Foran, X. Qi, *Front. Bioeng. Biotechnol.* **2019**, *7*, 102.
- [34] G. Lemaître, R. Marti, J. Freixenet, J. C. Vilanova, P. M. Walker, F. Meriaudeau, *Comput. Biol. Med.* **2015**, *60*, 8.
- [35] H.-S. Tong, Y.-L. Ng, Z. Liu, J. D. Ho, P.-L. Chan, J. Y. Chan, K.-W. Kwok, *Int. J. Comput. Assist. Radiol. Surg.* **2021**, *16*, 731.
- [36] G. Fang, X. Wang, J. D. Ho, K. Wang, C.-K. Chow, K.-H. Lee, X. Xie, W. L. Tang, L. Liang, H.-C. Chang, *Adv. Intell. Syst.* **2022**, *4*, 2200197.
- [37] J. Xiao, Q. Dai, X. Xie, Q. Dou, K.-W. Kwok, J. Lam, *IEEE Trans. Netw. Sci. Eng.* **2022**, *10*, 35.
- [38] Q. Liu, Q. Dou, L. Yu, P. A. Heng, *IEEE Trans. Med. Imaging* **2020**, *39*, 2713.
- [39] Z. Wang, D. Acuna, H. Ling, A. Kar, S. Fidler, in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, New York City, United States **2019**, pp. 7500–7508.
- [40] K. He, X. Zhang, S. Ren, J. Sun, in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV **2016**, pp. 770–778.
- [41] Y. Ganin, V. Lempitsky, in *Int. Conf. Mach. Learn.*, Lille, France **2015**, pp. 1180–1189.
- [42] L. Van der Maaten, G. Hinton, *J. Mach. Learn. Res.* **2008**, *9*, 2579.
- [43] K. He, G. Gkioxari, P. Dollár, R. Girshick, in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, New York City, United States **2017**, pp. 2961–2969.

- [44] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, New York City, United States **2017**, pp. 2117–2125.
- [45] Z. Liu, W. Jiang, K.-H. Lee, Y.-L. Lo, Y.-L. Ng, Q. Dou, V. Vardhanabhuti, K.-W. Kwok, *Artificial Intelligence in Radiation Therapy* (Eds: L. X. Dan Nguyen, S. Jiang), Springer International Publishing, Cham, Shenzhen, China **2019**, Ch. 6.
- [46] A. Khoreva, R. Benenson, J. Hosang, M. Hein, B. Schiele, in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, New York City, United States **2017**, pp. 876–885.
- [47] C. Li, C. Xu, C. Gui, M. D. Fox, *IEEE. Trans. Image Process.* **2010**, *19*, 3243.
- [48] F. Chollet, *Deep Learning with Python*, Simon and Schuster, New York, NY **2021**.