Adversarial Domain Adaptation for Classification of Prostate Histopathology Whole-Slide Images

Jian Ren[1], Ilker Hacihaliloglu[2], Eric A. Singer[3], David J. Foran[3], Xin Qi[3]

[1]Department of Electrical and Computer Engineering, Rutgers University, USA
[2]Department of Biomedical Engineering, Rutgers University, USA
[3]Rutgers Cancer Institute of New Jersey, USA

Abstract. Automatic and accurate Gleason grading of histopathology tissue slides is crucial for prostate cancer diagnosis, treatment, and prognosis. Usually, histopathology tissue slides from different institutions show heterogeneous appearances because of different tissue preparation and staining procedures, thus the predictable model learned from one domain may not be applicable to a new domain directly. Here we propose to adopt unsupervised domain adaptation to transfer the discriminative knowledge obtained from the source domain to the target domain without requiring labeling of images at the target domain. The adaptation is achieved through adversarial training to find an invariant feature space along with the proposed Siamese architecture on the target domain to add a regularization that is appropriate for the whole-slide images. We validate the method on two prostate cancer datasets and obtain significant classification improvement of Gleason scores as compared with the baseline models.

1 Introduction

Prostate cancer is the most common non-cutaneous malignancy and affects 1 in 7 men in the United States [1]. Gleason scores, graded from whole-slide images (WSIs), have been shown to serve as one of the best predictors for prostate cancer diagnosis [2]. Gleason grading is crucial for studying disease onset, progression and decision making for targeted therapy. However, Gleason grading is a time-consuming process due to the giga-pixel size of the WSIs. Furthermore, inter- and intra-observer variability errors often arise when pathologists make diagnosis based on WSIs. In order to provide an objective and quantitative Gleason grading score, computational methods have been applied for detection, extraction, and recognition of histopathological patterns. Methods based on convolutional neural networks (CNN) are considered state-of-the-art due to their high classication rates [3][4][5]. Most of these studies focus on the supervised classification. Histopathology WSIs obtained from different institutions usually present distinct glandular region distributions due to differences in appearance that may be caused by using different microscope scanners and staining procedures. These differences may render the supervised classification model used

for predicting the Gleason score for one annotated dataset (source domain) ineffective on another prostate dataset (target domain). A widely used approach to address the challenge is to label new images on the target domain and fine-tune the model trained on source domain [6]. Instead, methods that can learn from existing datasets and adapt to new target domains, without the need for additional labeling, are highly desirable.

Thus in this work, we aim to classify the newly given prostate datasets into low and high Gleason grade through unsupervised learning. To achieve this goal, we adopt the unsupervised domain adaptation paradigm to align the image distributions along the annotated source domain and the unlabeled target domain, where the two domains have the same number of high-level classes [7][8]. We apply adversarial training to minimize the distribution discrepancy at the feature space between the domains, with the loss function adopted from the Generative Adversarial Network (GAN) [9]. Furthermore, we developed a Siamese architecture for the target network to serve as a regularization of patches within the WSIs. The proposed method is validated on public prostate datasets and a newly collected local dataset. The experimental results show the approach significantly improves the classification accuracy of Gleason score as compared with the baseline model. To the best of our knowledge, this is the first study of domain adaptation for unsupervised prostate histopathology WSIs classification.

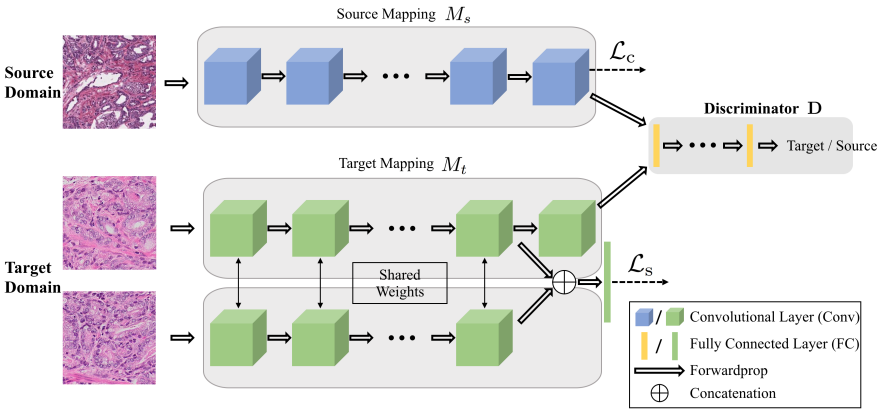


Fig. 1: The architecture of the networks for the unsupervised domain adaptation. The source network and the target network map the input samples into the feature space. The adaptation is accomplished by jointly training the discriminator and target network using the GAN loss to find the domain invariant feature. A Siamese network at target domain adds constrains for the WSIs.

2 Method

In this section, we present our approach on the unsupervised domain adaptation for the classification of prostate histopathology WSIs, as illustrated in Figure 1 above.

Problem formulation: Formally, we have a source domain distribution $\mathcal{S}$ that includes N_s labeled prostate histopathology images $\{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{N_s}$ where $\mathbf{y}_i^s$ is one-hot vector denoting the Gleason score, and a target domain distribution $\mathcal{T}$ contains N_t unlabeled prostate histopathology images $\{(\mathbf{x}_i^t)\}_{i=1}^{N_t}$. We use the source domain to generate a feature space through the mapping function M_s, and seek to find the mapping M_t at the target domain to obtain a similar feature space with the one from source domain. Thus the Gleason score prediction for the target domain is easily achieved by using the M_t.

Learning at source domain: Since the Gleason scores for the prostate images from the source domain are available, we train the network on the source domain to get the discriminative feature space using the supervised learning. In order to feed the WSIs into the network, we crop them into patches and adopt the cross-entropy loss $\mathcal{L}_{\mathrm{c}}$ to optimize the classifier $\mathbf{C}$, with weights as θ^S, to classify the images into low-grade (score as 6 and 7) and high-grade (score higher than 7) Gleason scores, which are highly related to clinical outcomes.

$$\mathcal{L}_{\mathrm{c}} = \mathbb{E}_{\mathbf{x}_s \sim \mathcal{S}} - \sum_{i=1}^{N_s} \mathbf{y}_i^s \cdot \log \mathbf{C}(M_s(\mathbf{x}_s; \theta^S)) \tag{1}$$

The majority vote is applied on the cropped patches within each WSI to obtain the final Gleason score for the WSIs.

Adversarial adaptation for target domain: Due to lack of annotations for the training set on the target domain, the $\mathcal{L}_{\mathrm{c}}$ is only applied on the source domain. To optimize the target network, we leverage the adversarial training to minimize the discrepancy between the feature space of the target domain and the one of the source domain. We perform an asymmetric adaptation where the network at the target domain is fine-tuned from the network of the source domain. Through optimization, the feature space of the target domain learns to mimic the distribution of the source feature space. Thus the target network is trained to extract the domain invariant features from input samples, which has the same distribution as the source domain.

Adversarial training is achieved by utilizing a GAN loss [9]. Two feature spaces generated from the source network and target network are fed into the discriminator $\mathbf{D}$. $\mathbf{D}$ is trained to map the input feature spaces into a binary domain label, where the true denotes the source domain and false denotes the target domain. Additionally, the target mapping M_t, is learned in an adversarial manner to purposely mislead the discriminator by reversing the domain label so that it cannot distinguish between the two feature spaces. Since the mapping parameterization of source model is determined before the adversarial training, we only optimize the target mapping. By using adversarial learning, we minimize the discrepancy between the two spaces. Therefore, estimating the Gleason scores for the images from target domain can be implemented by M_t. More specifically, the adversarial loss $\mathcal{L}_{\mathrm{adv}_{\mathbf{D}}}$ for optimizing the discriminator and the mapping loss $\mathcal{L}_{\mathrm{adv}_M}$ for optimizing the target mapping are represented as:

$$\min_{\mathbf{D}} \mathcal{L}_{\mathrm{adv}_{\mathbf{D}}} = -\mathbb{E}_{\mathbf{x}_s \sim \mathcal{S}} \log \mathbf{D}(M_s(\mathbf{x}_s; \theta^S); \theta^D) - \mathbb{E}_{\mathbf{x}_t \sim \mathcal{T}} \log(1 - \mathbf{D}(M_t(\mathbf{x}_t; \theta^T); \theta^D) \tag{2}$$

$$\min_{M_t} \mathcal{L}_{\text{adv}_M} = -\mathbb{E}_{\mathbf{x}_t \sim \mathcal{T}} \log(\mathbf{D}(M_t(\mathbf{x}_t; \theta^T); \theta^D)) \tag{3}$$

For the adversarial training, we optimize the $\mathcal{L}_{\text{a}}$, where $\mathcal{L}_{\text{a}} = \mathcal{L}_{\text{adv}_\mathbf{D}} + \mathcal{L}_{\text{adv}_M}$.

Algorithm 1: Learning Algorithm for the Network at Target Domain

Input: Initialized target network from source network with weights $\theta^T = \theta^S$

for *number of training iterations* **do**
- sample two same number of mini-batches $\mathbf{x_s} \sim \mathcal{S}$, $\mathbf{x_t} \sim \mathcal{T}$;
- obtain the estimation $\mathbf{y} = M_s(\mathbf{x}_s; \theta^S)$, $\mathbf{y}' = M_t(\mathbf{x}_t; \theta^T)$;
- $\theta^D \leftarrow$ back propagate with stochastic gradient $\nabla \mathcal{L}_{\text{adv}_\mathbf{D}}(\mathbf{y}, \mathbf{y}')$;
- $\theta^T \leftarrow$ back propagate with stochastic gradient $\nabla \mathcal{L}_{\text{adv}_\mathbf{M}}(\mathbf{y}')$;
- sample mini-batches with paired of images $\mathbf{x_t^1}, \mathbf{x_t^2} \sim \mathcal{T}$;
- obtain the estimation $\mathbf{y_f} = f(\mathbf{x_t^1}, \mathbf{x_t^2}; \theta^F)$;
- $\theta^F \leftarrow$ back propagate with stochastic gradient $\nabla \mathcal{L}_{\text{s}}(\mathbf{y_f})$;

Siamese architecture at target domain: Although there are no annotations for the prostate WSIs at the target domain, the cropped patches from the same WSI should still be predicted with the same Gleason score by the target network. While the adversarial loss forces the distribution across two domains to be similar, it can not constrain the target network to determine the similarity of the input patches. Therefore, we introduce a Siamese architecture at target domain to explicitly regularize patches from the same WSI to have the same Gleason score. As shown in Figure 1, the two identical networks share the same weights with the input as a pair of images $(\mathbf{x}_t^1, \mathbf{x}_t^2) \subseteq \mathcal{T} \times \mathcal{T}$. The feature maps obtained from the second to the last layer of the two networks are concatenated to serve as the input for a one-layer perceptron to classify the features. Therefore, the input samples are classified by the function $f(\mathbf{x}_t^1, \mathbf{x}_t^2; \theta^F)$, that $f : \mathcal{T} \times \mathcal{T} \mapsto 0, 1$, where 1 indicates input patches belong to the same WSI while 0 denotes not. We learn the binary classifier f using cross-entropy loss $\mathcal{L}_{\text{s}}$.

To learn the network at target domain, we adopt a two-stage training process. For the first stage, we train the network at source domain. For the second stage, we optimize the Siamese network at target domain by applying $\mathcal{L}_{\text{t}}$ where $\mathcal{L}_{\text{t}} = \mathcal{L}_{\text{a}} + \mathcal{L}_{\text{s}}$. The learning algorithm for the target network is shown in Algorithm 1.

3 Experimental Validation and Results

Validation of the proposed method is performed in two datasets: (1) publicly available The Cancer Genome Atlas (TCGA) dataset [10], and (2) a local data set collected from Cancer Institute of New Jersey (CINJ) after obtaining the institutional review board (IRB) approval.

Dataset In the first unsupervised domain adaptation experiment, we only use the TCGA dataset. The TCGA prostate cancer data includes histopathology WSIs uploaded from 32 institutions that have been acquired at 40× and 20× magnifications. We crop the WSIs into patches by the size of 2048×2048. We

	Gleason 6	Gleason 7	Gleason 8	Gleason 9	Gleason 10
# WSIs	115 (32)	395 (95)	94 (20)	128 (24)	4 (0)
# Patches	16293 (6517)	67162 (26583)	16204 (4968)	23978 (9606)	342 (0)

Table 1: The number of WSIs and patches for the prostate histopathology images from TCGA under different Gleason scores. The images from University of Pittsburgh (UP) are shown in parentheses.

	Gleason 6	Gleason 8
# WSIs	57	26
# Patches	3933	666

Table 2: The number of WSIs and patches for the dataset from CINJ under different Gleason grades.

	Accuracy (%)
Previous Study [11]	73.5
TCGA (w/o UP)	**76.9**
TCGA	**83.0**

Table 3: The network performance at the source domain. The two source networks both have better performance than [11].

calculate the tissue area on the grayscale images and remove the images with tissue area less than the half of the patch size. The dataset includes the Gleason scores annotated by pathologists ranging from 6 to 10. As the University of Pittsburgh (UP) has contributed more images than other institutions, we treat the UP as the target domain where the annotations are withheld and the images from other institutions as the source domain, which we denote it as TCGA (w/o UP). We show the total number of WSIs and the cropped patches from TCGA in Table 1 and UP in the parentheses. We denote the adaptation as TCGA (w/o UP) $\rightarrow$ UP. For the second unsupervised domain adaptation experiment, we use all the images from TCGA as the source domain, and the images from CINJ as the target domain. The images from CINJ are acquired at $20\times$ magnification. More details of the CINJ dataset is shown in Table 2. The dataset is labeled by one pathologist with the Gleason scores as 6 or 8. We denote the adaptation as TCGA $\rightarrow$ CINJ.

Implementation Details For the two sets of experiments, we aim to optimize the network at target domain that could classify the WSIs into low and high Gleason scores. Thus we divide the TCGA dataset into low Gleason grade for the WSIs with score as 6 and 7, and high Gleason grade for the WSIs with score as 8, 9 and 10. For the CINJ dataset, the WSIs with Gleason score of 6 belong to the low Gleason grade and Gleason score of 8 belong to high Gleason grade. The training process is composed of two steps. We first train the binary classification network using the data from the source domain. We use a modified fully convolutional AlexNet [12], which only contains convolutional layers, as the network for the classification task. All the convolutional layers are followed by the Batch Normalization layer except the last one that gives the prediction. The data from source domain is randomly divided into the training and the testing sets at a ratio of 80% (validation set is selected from the training set) / 20%. The patients with more than one WSIs can only contribute the images to the training set or

	Accuracy (%)
Baseline	54.3
$\mathcal{L}_{\mathrm{a}}$ only	71.4
$\mathcal{L}_{\mathrm{t}}$	**77.1**

Table 4: The unsupervised adaptation of TCGA (w/o UP) → UP.

	Accuracy (%)
Baseline	56.3
$\mathcal{L}_{\mathrm{a}}$ only	62.5
$\mathcal{L}_{\mathrm{t}}$	**75.0**

Table 5: The unsupervised adaptation of TCGA → CINJ.

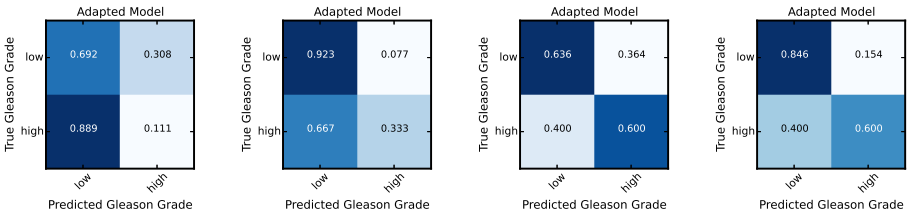
the testing set. During the training process, the images are resized as 256×256 and randomly cropped to 224×224 to feed into the network. And we train the network from scratch. The second step is to optimize the Siamese network at target domain. During the second step, we fix the parameters of the source network, and train the target network and the discriminator network at the same time. The feature vectors from the two domains are sent into the discriminator network that contains three fully connected layers. And the last layer gives the domain label estimation for the input feature samples. The prostate images at the target domain are randomly divided into the training and the testing sets at a ratio of 80% / 20%.

Source network performance As the training process contains two steps, we first show the performance of the network at the source domain. The comparison between the source network and the previous study [11] is shown in Table 3. From the results, we can see both of our models have better performance than [11]. However, the study at [11] uses less WSIs than ours and the network with the best performance reported in [11] is wider and deeper than our study. Although such differences lead to biased comparison, it still demonstrates the source domain network is well trained to classify the TCGA prostate images into low Gleason score and high Gleason score.

Adaptation of TCGA (w/o UP) → UP In order to prove the effectiveness of the knowledge transfer from source domain to the target domain, we show the quantitative results for TCGA (w/o UP) → UP in Table 4. We can see that due to the different image distribution for the TCGA (w/o UP) and UP, the network learned from TCGA (w/o UP) is not working appropriately on UP. But through the unsupervised adaptation, we could effectively adapt the discriminative knowledge from TCGA (w/o UP) to the UP without requiring additional annotations. We further calculate the statistically significance of the accuracy improvement between the adapted network and the baseline network using McNemar Test [13] and demonstrates the improvement of classification accuracy is statistically significant with a p-value as 0.039. In addition, we show the result of the ablation study in Table 4 that using $\mathcal{L}_{\mathrm{t}}$ achieves better classification accuracy than $\mathcal{L}_{\mathrm{a}}$ only. The confusion matrices for the adaptation are shown in Figure 2a-2b. After the adaptation, the classification accuracy for both WSIs of low and high Gleason scores are significantly improved.

Adaptation of TCGA → CINJ The results showing in Table 5 also proves $\mathcal{L}_{\mathrm{t}}$ could achieve the best adaptation performance. The confusion matrices are shown in Figure 2c-2d. We further show the qualitative results in Figure 3. We use the probability predicted by the network on the patches to generate

a Gaussian heatmap and overlay the heatmap on the original image. The red color indicates the high Gleason grade and blue color indicates the low Gleason grade. Figure 3a shows an example prostate WSI from CINJ with the high Gleason grade (Gleason score 8) and the ground-truth heatmap overlaid on it. The heatmap generated from the baseline network is shown in Figure 3b. The heatmap indicates many low Gleason grade areas, which are misclassified. The heatmap obtained from the target network that optimized by $\mathcal{L}_{\mathrm{a}}$ is shown in Figure 3c, which presents less low Gleason grade areas. Using $\mathcal{L}_{\mathrm{t}}$, the target network could correctly classify all the patches into high Gleason grade, as demonstrated in Figure 3d.



(a) The confusion matrix for UP before adaptation.

(b) The confusion matrix for UP after adaptation.

(c) The confusion matrix for CINJ before adaptation.

(d) The confusion matrix for CINJ after adaptation.

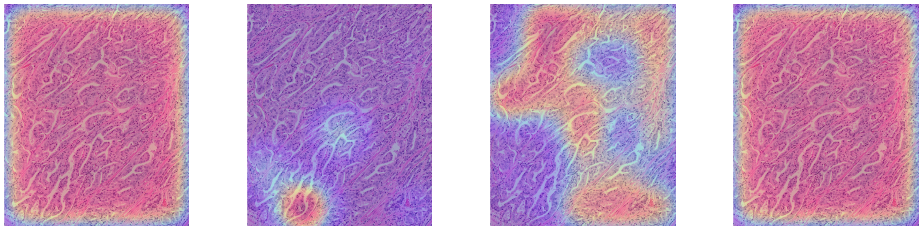
Fig. 2: The confusion matrix of the target network before and after the adaptation for TCGA (w/o UP) $\rightarrow$ UP and TCGA $\rightarrow$ CINJ .

4 Conclusion

In this work, we adopt an adversarial training and Siamese architecture to improve the classification performance of a target network in an unsupervised manner. We show that by using the proposed domain adaptation method statistically significant classification results can be achieved. Future work will include improvement of the method by using extensive datasets and extension to a wide range of histopathology image classification problems.

References

1. Ferlay, J., et al.: Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. International journal of cancer **136**(5) (2015)
2. Epstein, J.I., Zelefsky, M.J., Sjoberg, et al.: A contemporary prostate cancer grading system: a validated alternative to the gleason score. European urology **69**(3) (2016) 428–435
3. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: CVPR. (2016) 2424–2433



(a) Example image from CINJ with Gleason grade 8.

(b) Heatmap generated from the baseline model.

(c) Heatmap generated from the model optimized with $\mathcal{L}_{\mathrm{a}}$.

(d) Heatmap generated from the model optimized with $\mathcal{L}_{\mathrm{t}}$.

Fig. 3: We show an example image from CINJ with high Gleason grade and the heatmap generated from the prediction models.

4. Litjens, G., Sánchez, C.I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-Van De Kaa, C., Bult, P., Van Ginneken, B., Van Der Laak, J.: Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific reports **6** (2016) 26286
5. Otálora, S., Cruz-Roa, A., Arevalo, J., et al.: Combining unsupervised feature learning and riesz wavelets for histopathology image representation: application to identifying anaplastic medulloblastoma. In: MICCAI, Springer (2015) 581–588
6. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks **61** (2015) 85–117
7. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR. Volume 1. (2017) 4
8. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**(59) (2016) 1–35
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. (2014) 2672–2680
10. Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al.: Mutational landscape and significance across 12 major cancer types. Nature **502**(7471) (2013) 333
11. Jimenez-del Toroab, O., Atzoria, M., Otáloraab, S., Anderssonc, M., Eurénc, K., Hedlundc, M., Rönnquistc, P., Müllerabd, H.: Convolutional neural networks for an automatic classification of prostate tissue slides with high–grade gleason score. In: Proc. of SPIE Vol. Volume 10140. (2017) 101400O–1
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012) 1097–1105
13. Fagerland, M.W., Lydersen, S., Laake, P.: The mcnemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. BMC medical research methodology **13**(1) (2013) 91