On Transferability of Histological Tissue Labels in Computational Pathology

Mahdi S. Hosseini[1 [0000−0002−9147−0731]], Lyndon Chan[1 [0000−0002−1185−7961]], Weimin Huang[1 [0000−0001−6041−9848]], Yichen Wang[1 [0000−0001−5969−3233]], Danial Hasan[1 [0000−0003−0620−4402]], Corwyn Rowsell[2,3 [0000−0002−0604−638X]], Savvas Damaskinos[4 [0000−0001−6030−1823]], and Konstantinos N. Plataniotis[1 [0000−0003−3647−5473]]

[1] Department of Electrical & Computer Engineering, University of Toronto
[2] Division of Pathology, St. Michaels Hospital, Toronto, ON, M4N 1X3, Canada
[3] Department of Laboratory Medicine and Pathobiology, University of Toronto
[4] Huron Digital Pathology, St. Jacobs, ON, N0B 2N0, Canada
{mahdi.hosseini, lyndon.chan, cheryl.huang, yichenk.wang, danial.hasan, kostas.plataniotis}@utoronto.ca
https://github.com/mahdihosseini/HistoLabelTransfer/

Abstract. Deep learning tools in computational pathology, unlike natural vision tasks, face with limited histological tissue labels for classification. This is due to expensive procedure of annotation done by expert pathologist. As a result, the current models are limited to particular diagnostic task in mind where the training workflow is repeated for different organ sites and diseases. In this paper, we explore the possibility of transferring diagnostically-relevant histology labels from a source-domain into multiple target-domains to classify similar tissue structures and cancer grades. We achieve this by training a Convolutional Neural Network (CNN) model on a source-domain of diverse histological tissue labels for classification and then transfer them to different target domains for diagnosis without re-training/fine-tuning (zero-shot). We expedite this by an efficient color augmentation to account for color disparity across different tissue scans and conduct thorough experiments for evaluation.

Keywords: Cancer Detection, Cancer Grade Classification, Deep Learning, Domain Adaptation, Zero-Shot Transfer, Color Augmentation

1 Introduction

Recent deep learning techniques have been achieving competitive (at times even superior) performances compared to medical pathologists when diagnosing disease from Whole Slide Images (WSI). Histology slides are collected from particular organs and annotated with a particular disease to solve a particular diagnostic task in mind and deep learning models are trained with these annotated images to produce accurate and diagnostically meaningful predictions [1,9,11,12,19,27,29]. While the latter approach largely solves specific diagnostic

problems, it also prevents ready application to other organs and diseases. There has been little research to date on generalizing these state-of-the-art deep learning models to other datasets with different but related organs and diseases. This is problematic, because without the ability to transfer relevant knowledge from other datasets, it becomes prohibitively *expensive* and *time-consuming* to collect the histological annotations needed to train a new deep learning model for each new application [30, 46].

There are two main bottlenecks to such knowledge transfer: (1) the lack of annotated labels with diverse and generalizable tissue types from different organs and diseases; and (2) the variation in WSI scanners and staining protocols. Firstly, most openly available datasets were collected to solve a particular diagnostic problem and hence only contain slides focusing on specific organs and diseases: colorectal [22, 23, 37], breast [2, 4, 6], brain [13, 14], and various organs [26, 35]. This restricts the scope of histopathology for representational learning. In particular, most images are provided as single-label patches referring to specific tissue/disease related components. This restricts the ability to train classifiers that can discriminate tissue components smaller than the patch. Secondly, the color fidelity of WSI scans varies considerably: (a) digital pathology scanners follow different optics configurations and camera sensor calibration guidelines, and (b) pathology laboratories adopt different standards for staining the histology slides [31, 33, 36]. This means the same histology slide prepared by different institutions and digitized by different scanners can vary drastically by color and illumination contrast, which causes enormous challenges for training generalizable computational pathology algorithms [3, 8, 25, 39–43].

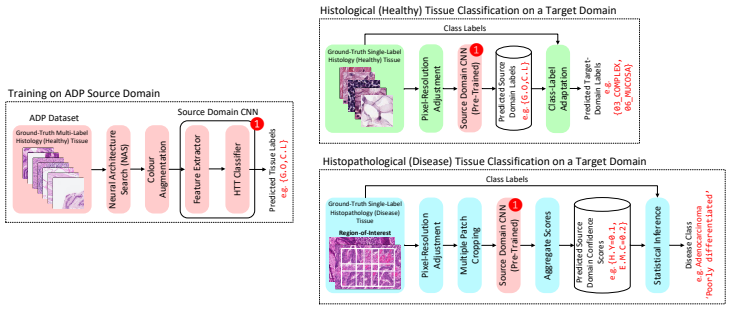


Fig. 1. Transferring HTT labels from source domain CNN (with labels available for training) into target domains for both tissue and disease classification (with labels unavailable for training) without re-training/fine-tuning. Labels are transferred based on prior histological/histopathological knowledge.

In this paper, we address these gaps by proposing a new approach to computational pathology - by training a *"universal"* model to recognize diverse histological tissue types (HTTs) from a source domain dataset of healthy slides (obtained from various organs), we can adapt the model to transfer diagnosti-

cally relevant labels for tissue and disease classification in target domain datasets without re-training or fine-tuning (i.e. zero-shot), see Figure 1 for demonstration. Unlike the existing domain adaptation methods, our approach requires training only once on the source domain dataset followed by a simple label adaptation to the target domain consulted by expert pathologist. For this purpose, we employ the Atlas of Digital Pathology (ADP) database [18] as our source domain dataset, where its diverse multi-label/multi-class set overlaps with many other existing datasets. To account for color disparities in WSI scans, we explore two different color augmentation methods for training using HSV [39, 42, 43] as well as a less complex color space transform i.e. YCbCr. Furthermore, we develop a simple and yet efficient Convolutional Neural Network (CNN) architecture called *"HistoNet"*, guided by the Reinforcement Learning (RL)-based Neural Architecture Search (NAS) [48] as a means to the end goal of domain adaptation. We further study optimum optical resolution that can produce acceptable performance in processing WSI scans. We define two tasks in target domain labels: (1) recognize tissues by matching labels in the source and target domains using prior histological knowledge; and (2) classify cancer grades by employing the confidence scores of diagnostically relevant labels as a surrogate for disease progression. Both tasks can be seen as variants of the "transductive" and "inductive" transfer learning problems without the target domain images or labels available during source domain training [32, 45]. Our results, for the first time, reveal that different but related histopathology datasets can be unified and efficiently represented by a deep learning model trained on a sufficiently diverse label set. Without retraining, our approach can form reasonable predictions of the tissue classes and diseases in unseen images.

The summary of our main contributions is as follows.

1. We introduce a new label transferring solution in computational pathology from the ADP source domain to different target domain based on diagnostically relevant labels for tissue type and cancer grade classification
2. We explore the possibility of efficient CNN training that can be robust and optimized toward color disparities and pixel-resolution for tissue recognition in WSI scans.
3. We provide thorough adaptation experiments on variety of target domain datasets and show the strengths of source domain for generalization

1.1 Related Works

The problem of domain adaptation is widely used in machine learning such as vision and language [15,16,32,45,47] to study the problem of knowledge transfer from a source domain to perform a predefined target domain task. Most knowledge transfer methods are done in an unsupervised fashion where the target-domain data is unlabeled [15] but its representation is employed during the training such as adversarial training in [16, 47]. This concept has been recently investigated in computational pathology to train a classifier using the source domain labels and adversaries from target-domain [7, 28, 31, 34]. A common disadvantage to such methods is the target domain images must be available for

training in the source domain. Therefore, retraining must be done independently for each target domain task.

Color augmentation is noticeably becoming a key factor in computational pathology domain adaptation, since the color disparities have strong effects on CNN predictions [3,39,41–43,46]. Methods such as HSV transformation [3,39,43] randomly perturbs the original images during training to expand the color space and hence generalize the model to images with colors not seen in the original training set. Stain normalization is also used to map the color distribution in the target domain to the source domain [5, 7, 8, 21, 28, 31, 34]. The downside of the latter approaches is that HSV performs a non-linear mapping and produces color residues during transformation which tend to produce misleading results [3]. Also, the stain normalization performs a one-to-one mapping between two histological stains (usually H&E). Hence it is unsuitable for training domain adaptation datasets such as ADP which contain multiple stains and using it would limit model generalizability to other stains.

2 Transferring Diagnostically-Relevant Labels

In this section, we describe a useful tissue label mapping (a rule based approach) that can be adapted between the HTT labels from ADP source domain and other target domain labels. This adaptation is shown in Figure 2 for both healthy and disease tissue labels, where the connections are consulted by expert pathologist for label mapping. If the target task is tissue classification, we only map the source labels to the target label set using its prior histological knowledge. If the target task is disease classification problem, the corresponding source labels correlated to disease level are identified and use their inverse confidence scores for statistical inference of disease class. Note that the primary site in computational pathology (i.e. the type of organ) is usually given as a prior knowledge. Our hypothesis here is if ADP contains such organ tissues, then with good probability, the relevant tissue type and disease class(es) can be well predicted through the inference of HTTs. For instance, in GlaS (Colon tissues) dataset, we employ two HTTs from ADP, i.e. E.M.C and H.Y, which are highly relevant for cancer detection in Colon tissues.

While our approach is indeed rule-based, histological tissues exhibit many superficial visual similarities, so taking a data-driven approach would be counter-productive. Our label mapping was derived through consultation with a medical pathologist and is a quick, one-step procedure. Our approach is more explainable (and hence trustworthy) to pathologists because it mimics their own diagnostic workflow better than a black-box solution. Pathologists cannot exhaustively learn to diagnose all possible cases, so they learn from labeled educational examples (i.e. train on source domain healthy samples), transfer their knowledge to diagnostic task at hand by searching for abnormalities (i.e. label mapping), and diagnose each new case by classifying visual appearance of cancer diseases (i.e. predict on target domain for grading).

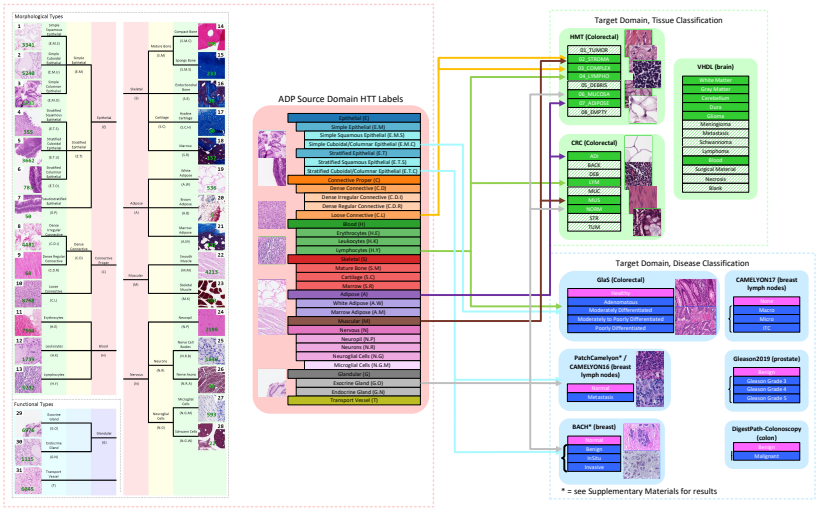


Fig. 2. Transferring Histological Tissue Type (HTT) labels from ADP source domain to target domains consulted by expert pathologist. For the tissue classification case (top), transferable target classes are in solid color and nontransferable classes are in striped color. For the disease classification case (bottom), the normal classes are indicated in magenta and the diseased classes in blue.

2.1 Source Domain: ADP

The Atlas of Digital Pathology (ADP) [18] is a database of patch images extracted from 100 healthy slides from the same medical institution scanned with a TissueScope LE1.2 at 0.25μm/px resolution. Each patch in the database is annotated with up to 33 hierarchical tissue types (modified from the 42 types in the original release - see the Supplementary Materials for details) in multi-labeled class format, covering a diverse set of morphological and functional types across different organs, such as stomach, colon, and thyroid. For visual presentation of this hierarchy, please refer to Figure 2. The label set in ADP contains tissue types observed in different organs and hence overlaps with the healthy tissue types in other histopathology databases.

2.2 Task 1: Tissue Classification

We simply map those target labels representing the same healthy tissue types as in the source domain set. Our approach is both simpler (no retraining is required) and requires less information (only the label set) from the target domain. We evaluate our approach for tissue classification in the ColoRectal Cancer dataset CRC [22] and Histology Multiclass Texture (HMT) [23] datasets. The CRC consists of patch images extracted from cancerous colorectal slides, each labeled with one of nine tissue type labels. Five labels are largely healthy tissue

types, two are disease types, and two are non-tissue types. We evaluate on the un-normalized validation set of this dataset. The ADP labels are mapped to the healthy types in CRC demonstrated in Table 1. The HMT consists of patch images also extracted from cancerous colorectal slides, each labeled with one of eight tissue type labels. Five of these are healthy tissue types, one is diseased, and two are non-tissue types. The ADP labels are mapped to the healthy types in HMT shown in the same Table 1.

Table 1. Mapping ADP Source-domain labels to CRC and HMT target datasets.

CRC		HMT	
Source Label	Target Label	Source Label	Target Label
A	→ ADI (adipose)	max(C.L, M)	→ 02_STROMA (simple stroma)
H.Y	→ LYM (lymphocyte)	C.L	→ 03_COMPLEX (complex stroma)
M	→ MUS (muscle)	H.Y	→ 04_LYMPHO (lymhocyte)
C.L	→ NORM (normal stroma)	G.O	→ 06_MUCOSA (mucosa)
		A	→ 07_ADIPOSE (adipose)

2.3 Task 2: Disease Classification

In preliminary experiments, we noticed that if the disease classes are quantified between 0 (normal) and 1 (most diseased), the confidence scores of diagnostically relevant classes would deteriorate with worsening disease. Here, we study three diseased datasets i.e. Gland Segmentation (GlaS) challenge [37], PatchCamelyon [44] extracted from the WSI scans of the original Camelyon16 challenge dataset [6], and Grand Challenge on Breast Cancer Histology (BACH) [2] all listed in Table 2 including the diagnostic features of each dataset. For each dataset, the diagnostically relevant HTT labels are identified from ADP source domain. Note that GlaS is comprised of Colon tissues which already exist in the ADP source domain. However, both PatchCamelyon and BACH are from breast tissues missing from ADP. We will show in experiments that, in fact, classification success is directly related to the type of organ tissues contained in the source domain. Once an organ type is included in ADP, it will be accurately classified through domain adaptation, e.g. GlaS but not for PatchCamelyon/BACH.

Table 2. HTT labels from source domain ADP identified as being diagnostically relevant in three different disease datasets i.e. GlaS, PatchCamelyon and BACH. For description of the labels please refer to label adaptation in Figure 2.

	GlaS	PatchCamelyon	BACH
Type of Data	patches extracted from colorectal slides	patches extracted from lymph node sections of breast slides	patches extracted from breast slides
Disease Class(es)	5-classes: 1 normal + 4 progressive cancer grades	2-classes: normal + tumorous	4-classes: 1 normal + 3 progressive cancer grades
Diagnostically Relevant Labels	"E.M.C" and "H.Y"	"E.T.C" and "G.O"	"E.T.C" and "G.O"
Availability of Primary Organ in ADP	Included	Not Included	Not Included

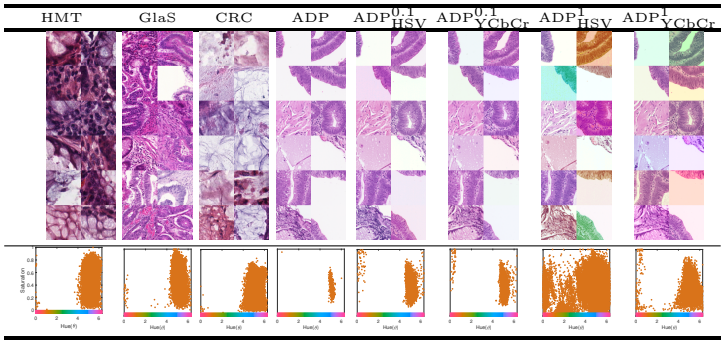
3 YCbCr Color Augmentation

ADP contains a wide range of stained tissue colors which extends beyond H&E spectrum. The goal is to develop simple and efficient color augmentation for CNN training that accounts for such wide variations. HSV color augmentation from [39, 42, 43] could be an alternative solution since it perturbs wide color range but it is (a) slow for CNN training; and (b) produces color residuals during forward/backward color conversion. Our solution to solve these issues is to switch to YCbCr using a linear transformation.

$$\begin{bmatrix} Y \\ C_\mathrm{b} \\ C_\mathrm{r} \end{bmatrix} = \begin{bmatrix} 0.2568 & 0.5041 & 0.0979 \\ -0.1482 & -0.2910 & 0.4392 \\ 0.4392 & -0.3678 & -0.0714 \end{bmatrix} \begin{bmatrix} I_\mathrm{R} \\ I_\mathrm{G} \\ I_\mathrm{B} \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \tag{1}$$

Consider the raw input image to the network (without channel normalization) I in RGB color space to be augmented in YCbCr space through random perturbations of both red and blue chroma channels $I_{[R][G][B]} \mapsto I_{[Y][C_b][C_r]} \mapsto I_{[Y][C_b+\mathcal{N}(0,s\sigma_\mathrm{b})][C_r+\mathcal{N}(0,s\sigma_\mathrm{r})]} \mapsto \tilde{I}_{[R][G][B]}$ where σ_b and σ_r are the standard deviation of the chroma channels, and s is the significance of perturbation. This is similar to the ratio defined in [43] where $s = 0.1$ is referred to as "Light" and $s = 1$ as "Strong" augmentation. The main advantages of this transformation are: (a) it separates the color chroma spaces (i.e. tuples of C_b and C_r) from the Luma channel Y, enabling direct manipulation of the stain colors without affecting the tissue illumination stimulated by WSI scanner condenser; and (b) the computational cost of such transform is much lower due to simple multiplication and addition operations.

Table 3. Distribution of color space shown in Hue-Saturation domain for five different datasets: (a) HMT, (b) GlaS, (c) CRC, and (d) ADP. Two different augmentation methods i.e. HSV [43] and YCbCr (Proposed) are considered to perturb ADP pixels shown from (e) to (h) using scaling factors $s = 0.1$ (light) and $s = 1$ (strong).



We demonstrate the effectiveness of YCbCr compared to the HSV augmentations in Table 3 by randomly perturbing several image patches from ADP [18]

and visually comparing them to patch examples from other datasets, i.e. HMT [23], GlaS [37], PCam [6, 44], and CRC [22]. Notice how applying different color augmentation methods (using "Light"/"Strong") can match color distribution of other datasets. HSV-Strong method creates random outliers while YCbCr-Strong remains more stable.

4 HistoNet for ADP Source Domain

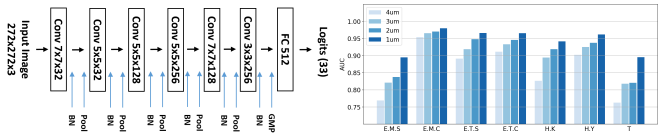
Here, we design an efficient CNN model optimized on ADP database for HTT classification. This is highly preferable in computational pathology for processing thousands of image-patches cropped from GigaPixel WSI scans. Therefore, both computational complexity and precision are equally important for practical considerations.

4.1 Neural Architecture Search (NAS) for HistoNet

We design a CNN model with six sequential convolutional layers and one fully connected layer, followed by a sigmoid layer for multi-label class activation. After each convolutional layer, ReLU activation, Batch Normalization (BN), and max pooling (2×2) are applied. Global max pooling is used at the end of the sixth layer. The parameters we explore include the kernel size $\{w_\ell \times w_\ell\}_{\ell=1}^{6}$ and the number of filters $\{D_\ell\}_{\ell=1}^{6}$ of the convolutional weights (tensor) for each layer i.e. $\Phi_\ell \in \mathbb{R}^{w_\ell \times w_\ell \times D_{\ell-1} \times D_\ell}$. We recast the HistoNet design in two phases. First, we seek the optimal configuration for $\{w_\ell\}_{\ell=1}^{6}$ and $\{D_\ell\}_{i=1}^{6}$ using the neural architecture search with reinforcement learning algorithm introduced in [48]. A controller RNN is used to generate child CNN architectures with different kernel sizes and kernel number configurations. The child CNN model is trained, and the classification accuracy of child model is used as a reward signal to update the controller to sample configurations for the next step. We explore the search space of the RNN controller using different kernel size $w_\ell \in \{1, 3, 5, 7\}$ and channel depth $D_\ell \in \{32, 64, 128, 192, 256\}$. We run the controller RNN to generate 250 child CNN architectures that have different configurations, and train each child CNN for 20 epochs on ADP patches of size 224×224. The 250 trained CNN models are applied to the ADP test dataset and choose the seven best configurations (by F1 score) for kernel size and number of filters. In the second phase, we trained all the seven HistoNet models for 100 epochs and selected the best configuration to finalize the network. The overall layout of the optimized architecture is shown in Figure 3(a).

4.2 Choice of Pixel Resolution

Although the input image size studied to optimize the HistoNet is 224×224 (@$1.21\mu m$/pixel resolution), it is important to understand how the network training corresponds to different scan resolutions. To study this, we downsize the original ADP image from 1088×1088 (@$0.25\mu m$/pixel resolution) into four



(a) HistoNet Architecture (b) Scan Resolution Performance

Fig. 3. (a) HistoNet serial architecture. The network consists of six convolutional layers, followed by a fully-connected and sigmoid-activation layers for ADP multil-label HTT classification. The kernel size and channel depth are optimized using the NAS-Reinforcement Learning method in [48]; and (b) Classification performance (AUC) of HistoNet on selected HTTs at different scan resolutions.

different pixel resolutions of $\{1\mu m, 2\mu m, 3\mu m, 4\mu m\}$. As shown in Figure 3(b), decreasing the pixel resolution improves the AUC for all classes, especially for smaller tissues such as simple squamous epithelium (E.M.S), leukocytes (H.K), and transport vessels (T) (see Figure 2 for full HTT names).

The overall HistoNet performance for different scan resolutions and augmentation methods is demonstrated in Table 4. While the network yields better classification on higher resolutions, the choice of color augmentation impacts the overall results. For instance, employing YCbCr-Strong augmentation improves about 0.5% compared to no augmentation at the 1 μm scan resolution. While both YCbCr and HSV provide similar performances, HSV adds about 40% computational overhead per epoch during training.

Table 4. HistoNet test set performance applied to HTT classification on ADP database using different pixel-resolution scan and color augmentation methods. The performances are reported by Area Under the Curve (AUC) of the ROC and F_1 measure.

	Sec/Epoch			AUC					F_1				
					HSV		YCbCr			HSV		YCbCr	
	None	HSV	YCbCr	None	Light	Strong	Light	Strong	None	Light	Strong	Light	Strong
$4\mu m$	72	92	72	0.9136	0.9310	0.9205	0.9277	0.9208	0.7485	0.7780	0.7536	0.7699	0.7658
$3\mu m$	94	129	95	0.9276	0.9367	0.9325	0.9374	0.9369	0.7670	0.7836	0.7790	0.7895	0.7943
$2\mu m$	156	236	155	0.9454	0.9539	0.9526	0.9511	0.9525	0.8085	0.8247	0.8215	0.8203	0.8204
$1\mu m$	567	937	569	0.9594	0.9650	0.9638	0.9646	0.9645	0.8378	0.8534	0.8476	0.8499	0.8537

4.3 Network Performance Comparison

We further compare the performance of HistoNet to three CNNs, i.e. ResNet18 [17], MobileNet [20] and Xception [10]. Note that we trimmed number of middle flow blocks in Xception from eight (baseline) to one to reduce the network parameters - we call this Xception-1. The criteria of our comparison networks selection here is mainly based on the simplicity of architectures for practical

implementations. All models are trained at 1 μm scan resolution following the multi-label class weighting suggested in [18] as well as using Cyclical Learning Rate [38] with an initial learning rate of 0.1, batch size of 32, and termination after 100 epochs. Table5 demonstrates the predictive performance over different networks and color augmentation methods. The rank performance of HistoNet is preserved compared to the other CNNs, while consuming the lowest complexity with 3M parameters.

Table 5. Test set performance of four different networks for HTT classification on ADP database at $1\mu m$/pixel with $272\mu m$ field-of-view scan (272×272 input image).

	# Conv Layers	Params	AUC			F_1		
			NA	HSV	YCbCr	NA	HSV	YCbCr
ResNet18 [17]	18	11.20M	0.9533	0.9511	0.9521	0.8244	0.8200	0.8209
Xception-1 [10]	15	9.60M	0.9576	0.9576	0.9567	0.8363	0.8362	0.8339
MobileNet [20]	14	3.26M	0.9521	0.9503	0.9518	0.8233	0.8200	0.8181
HistoNet/NAS [48]	6	3.00M	0.9594	0.9638	0.9645	0.8378	0.8476	0.8537

5 Experiments on HTT Transferability

To evaluate the transferability of HTT labels from ADP source domain into different target domains, we adopt the CNN models trained in previous Section 4 w- and w/o- color augmentation. Then, we transfer the models to solve two datasets in tissue classification and three datasets in cancer detection and cancer grade classification tasks. For tissue classification we compare the ROC performance of the domain-adapted CNNs; for cancer detection we compare the ROC performance of the relevant inverted source class scores against the target cancer labels; and finally for cancer classification we compare the statistical correlation of inverted source class scores against the target cancer grades. For details on datasets, visual examples, and additional results, see the Supplementary Materials. Codes and trained models are available on GitHub[5].

5.1 Image Modifications and Pixel-Resolution Adjustment

The source-domain CNN is trained on ADP square patch images $\mathbf{X_s}$ of size $W_s \times W_s$ pixels with a ρ_s resolution (in μm/px) and a fixed Field-Of-View (FOV) of 272μm, so to ensure the target-domain images $\mathbf{X_t}$ of size $H_t \times W_t$ and ρ_t resolution have the same pixel resolution and FOV, a few modifications must be performed. To ensure the same pixel resolution, they are resized by a constant factor $\alpha = \rho_t/\rho_s$, such that $(W_t, H_t) \leftarrow \alpha \cdot (W_t, H_t)$. The target image must be also padded (if the FOV is too small) and/or cropped (if the FOV is too large). It is symmetrically padded by $(k_x, k_y) = (\texttt{max}(W_s - W_t, 0), \texttt{max}(W_s - H_t, 0))$,

[5] https://github.com/mahdihosseini/HistoLabelTransfer/

such that $(W_t, H_t) \leftarrow (W_t + k_x, H_t + k_y)$. Then, if $W_s < W_t$ or $W_s < H_t$, crops of size $W_s \times W_s$ are extracted and the confidence scores of the cropped patches are later aggregated back to the image level by taking their average.

5.2 Transferability in CRC Dataset for Tissue Classification

The images in CRC are colorectal tissues which partially correlates with primary sites compiled in ADP. The stain color distributions of CRC, however, differs from ADP (see Figure 3). Figure 4 (top) shows that HistoNet performs well in all adapted classes except for the LYM (Lymphocyte) class ([22] reports a mean 4-class AUC of 0.995). Here, the "*lighter*" forms of color augmentation are better. Furthermore, class performance in CRC is heavily dependent on the network architecture used. Figure 4 (bottom) shows that HistoNet performs best in the large tissue classes (ADI, MUS, and NORM) and that the other networks progressively get better performance with increasing depth.

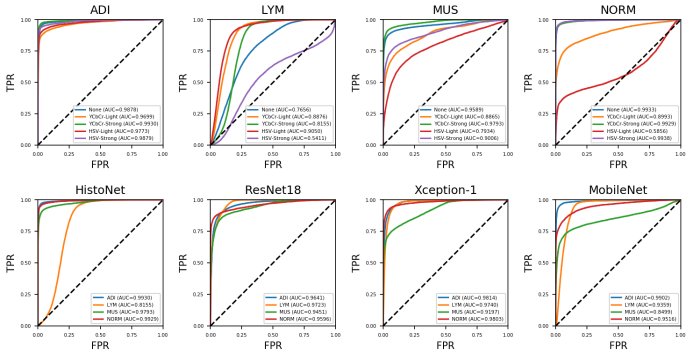


Fig. 4. Network ROC performance on CRC trained with different color augmentations and evaluated on (a) four classes of ADI, LYM, MUS, and NORM (top row); and (b) four different CNNs of HistoNet, ResNet18, Xception-1, and MobileNet (bottom row).

5.3 Transferability in HMT Dataset for Tissue Classification

The HMT images are very similar to CRC in the organ of origin but have a smaller FOV and a different color distribution (see Figure 3). Figure 5 (top) shows that HistoNet struggles to classify the smaller tissues (02_STROMA, 03_COMPLEX, 04_LYMPHO) and excels in the larger tissues (06_MUCOSA, 07_ADIPOSE). We hypothesize this phenomena due to stain normalization used in original dataset for training where color representation of tissues are deteriorated from their original spectrum. ([23] reports a mean 8-class AUC of 0.976). YCbCr-Strong color augmentation is the only method to consistently improve upon the unaugmented case. Again, class performance in HMT depends on the

architecture used. Figure 5 (bottom) shows that HistoNet is superior in three of five classes, with the other networks performing better with decreasing depth.

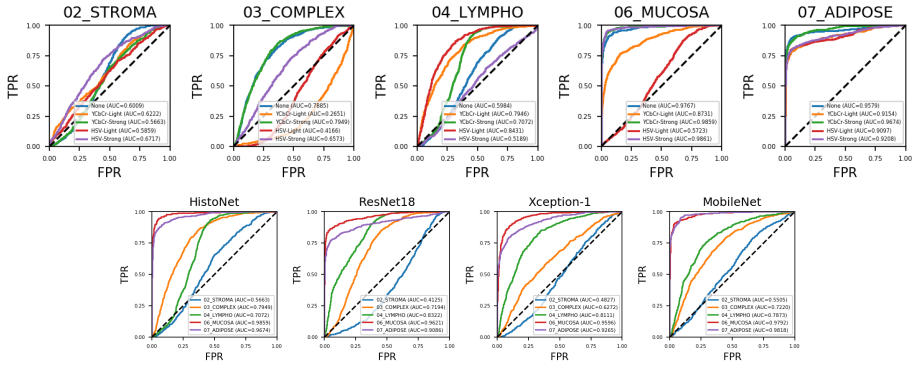


Fig. 5. Network ROC performance on HMT trained with different color augmentations and evaluated on (a) five classes: 02_STROMA, 03_COMPLEX, 04_LYMPHO, 06_MUCOSA, and 07_ADIPOSE (top); and (b) four different CNNs of HistoNet, ResNet18, Xception-1, and MobileNet (bottom).

5.4 Transferability in GlaS Dataset for Cancer Classification

Following the cancer mapping guideline for Colon tissues (i.e. GlaS) from Table 2, two ADP classes - simple cuboidal/columnar epithelium (E.M.C) and lymphocytes (H.Y) - are diagnostically relevant labels for classification. In Figure 6 (left) we demonstrate the statistical correlation measures using PLCC (Pearson Linear Correlation Coefficient), SROC (Spearman Rank Order Coefficient), and KROC (Kendall Rank Order Coefficient) calculated between predicted confidence scores (inverted) from network and five cancer grades for classification. Both YCbCr-Strong and HSV-Strong yield higher correlation results. This observation is interesting because it matches prior knowledge of colorectal cancer grades: high tumor differentiation distorts epithelial cell boundaries and absence of tumor infiltrating lymphocytes (TILs) has been linked to poor cancer prognosis. We further demonstrate the ROC analysis in Figure 6 (right) on cancer detection (binarizing healthy versus cancer) in Glas using E.M.C and H.Y. The HistoNet performs excellently here for all color augmentation methods achieving 0.95 AUC as the best result for lighter augmentation.

5.5 Alternative Source Domain Choice

While the ADP is mainly studied here to highlight the importance of transferring pathologists knowledge in the form of annotated labels, we further investigate

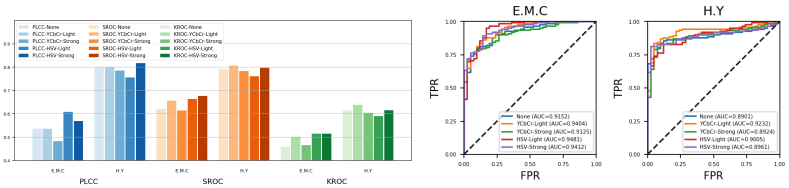


Fig. 6. Cancer classification on GlaS dataset using two HTTs: (a) Correlation measures (PLCC, SROC, and KROC) between select inverse HistoNet confidence scores (E.M.C and H.Y) and five quantified disease classes in GlaS (left plot); and ROC curves of HistoNet in GlaS on classifying normal/cancerous (right plot).

the reproducibility of similar results using other source domain choice. For this purpose, we directly train the HistoNet on CRC dataset using 1μm resolution, five different color augmentations, and 6:1:3 train-validation-test split. The obtained class AUCs for all healthy labels (see Table 1) achieved almost perfect classification i.e. AUC$>$ 0.99 similar to what original authors reported in [22]. Two labels are selected from CRC as being diagnostically relevant in Colon cancer i.e. LYM (lymphocyte) and NORM (normal stroma) and transfer into GlaS for cancer classification. Note that the cancer cells alter and embed in normal stroma [24]. The results are shown in Figure 7, where inferior performances are achieved for cancer grade classification comparing to ADP source domain, implying that the annotated labels in CRC is less comprehensive compared to ADP. This is in spite the fact that the train data size in CRC is statistically significant (70K patches) compared to ADP ($\sim$ 14.3K patches).

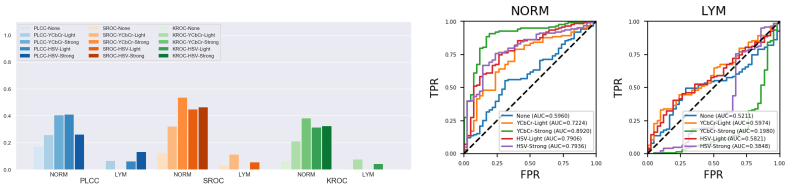
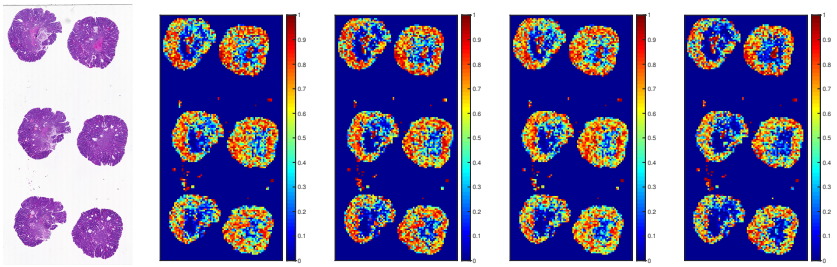


Fig. 7. (Cancer classification on GlaS dataset using two HTTs: (a) Correlation measures (PLCC, SROC, and KROC) between select inverse HistoNet confidence scores (E.M.C and H.Y) and five quantified disease classes in GlaS (left plot); and ROC curves of HistoNet in GlaS on classifying normal/cancerous (right plot).

6 Cancer Detection on WSI Level

In this section, we analyze a Colon tissue organ for cancer detection on the WSI level shown in Figure 8. The slide is mosaiced into multiple patches and classified by HistoNet with $1\mu m$/pixel resolution. We construct the heatmap for the

best color augmentation result shown in Figure 8 corresponding to inverse prediction score of Stratified Cuboidal/Columnar Epithelial (E.T.C). The pathologist's evaluation reads as follows: This WSI depicts an adenomatous polyp of the colon, shown in Figure 8(a). The majority of the epithelium in this slide is abnormal (neoplastic, precancerous), but there is an area of muscularis mucosa and normal epithelium at area where the polyp was removed. The heatmaps in Figure 8 show a high probability of abnormality in the areas of adenomatous epithelium (yellow, orange, and red), and indicate a low probability of abnormality in the regions with muscularis mucosa and normal epithelium (blue). The HSV Strong protocol appears to show the strongest correlation with histologic findings, followed by HSV Light, then YCbCr Light. The YCbCr Strong protocol shows the least correlation (while it still correctly indicates the normal areas, it appears to be less sensitive in identifying areas of abnormality compared to the other methodologies).



(a) WSI Scan (b) HSV-Light (c) HSV-Strong (d) YCbCr-Light (e) YCbCr-Strong

Fig. 8. A Colon tissue organ is selected for processing and diagnosed by pathologist. The heatmaps of abnormality is based on the inverse prediction of Stratified Cuboidal/Columnar Epithelial (E.T.C).

7 Concluding Remarks

In this paper, a new tissue label transferring method is proposed to classify different histological tissue structures and cancer grades across diverse target domains. The method is based on training a CNN model on the source domain dataset (using the Atlas of Digital Pathology's multi-label tissue types) and then transferring those labels to the relevant target labels using prior histological knowledge. The ability of proposed method is demonstrated to produce reasonable predictions in related tissue classification datasets. Furthermore, the confidence prediction scores of diagnostically relevant labels are inferred as a surrogate model for cancer grade progression. The results suggested that the diagnostically relevant labels can be better transferred by adopting an appropriate source domain with broad spectrum of tissue structures for classification.

References

1. Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., Campilho, A.: Classification of breast cancer histology images using convolutional neural networks. PloS one **12**(6), e0177544 (2017)
2. Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., et al.: Bach: Grand challenge on breast cancer histology images. Medical image analysis (2019)
3. Arvidsson, I., Overgaard, N.C., Åström, K., Heyden, A.: Comparison of different augmentation techniques for improved generalization performance for gleason grading. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 923–927. IEEE (2019)
4. Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., et al.: From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE transactions on medical imaging **38**(2), 550–560 (2018)
5. Bejnordi, B.E., Litjens, G., Timofeeva, N., Otte-Höller, I., Homeyer, A., Karssemeijer, N., van der Laak, J.A.: Stain specific standardization of whole-slide histopathological images. IEEE transactions on medical imaging **35**(2), 404–415 (2015)
6. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017)
7. Brieu, N., Meier, A., Kapil, A., Schoenmeyer, R., Gavriel, C.G., Caie, P.D., Schmidt, G.: Domain adaptation-based augmentation for weakly supervised nuclei detection. In: MICCAI 2019 Computational Pathology Workshop COMPAY (2019)
8. Bug, D., Schneider, S., Grote, A., Oswald, E., Feuerhake, F., Schüler, J., Merhof, D.: Context-based normalization of histological stains using deep convolutional features. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 135–142. Springer (2017)
9. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
10. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
11. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A.: Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine **24**(10), 1559 (2018)
12. Djuric, U., Zadeh, G., Aldape, K., Diamandis, P.: Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. NPJ precision oncology **1**(1), 22 (2017)
13. Faust, K., Bala, S., van Ommeren, R., Portante, A., Al Qawahmed, R., Djuric, U., Diamandis, P.: Intelligent feature engineering and ontological mapping of brain tumour histomorphologies by deep learning. Nature Machine Intelligence **1**(7), 316–321 (2019)

14. Faust, K., Xie, Q., Han, D., Goyle, K., Volynskaya, Z., Djuric, U., Diamandis, P.: Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction. BMC bioinformatics **19**(1), 173 (2018)
15. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research **17**(1), 2096–2030 (2016)
16. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
18. Hosseini, M.S., Chan, L., Tse, G., Tang, M., Deng, J., Norouzi, S., Rowsell, C., Plataniotis, K.N., Damaskinos, S.: Atlas of digital pathology: A generalized hierarchical histological tissue type-annotated database for deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11747–11756 (2019)
19. Hou, L., Agarwal, A., Samaras, D., Kurc, T.M., Gupta, R.R., Saltz, J.H.: Robust histopathology image analysis: to label or to synthesize? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8533–8542 (2019)
20. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
21. Karimi, D., Nir, G., Fazli, L., Black, P.C., Goldenberg, L., Salcudean, S.E.: Deep learning-based gleason grading of prostate cancer from histopathology images-role of multiscale decision aggregation and data augmentation. IEEE Journal of Biomedical and Health Informatics (2019)
22. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS medicine **16**(1), e1002730 (2019)
23. Kather, J.N., Weis, C.A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Marx, A., Zöllner, F.G.: Multi-class texture analysis in colorectal cancer histology. Scientific reports **6**, 27988 (2016)
24. Kaukonen, R., Mai, A., Georgiadou, M., Saari, M., De Franceschi, N., Betz, T., Sihto, H., Ventelä, S., Elo, L., Jokitalo, E., et al.: Normal stroma suppresses cancer cell proliferation via mechanosensitive regulation of jmjd1a-mediated transcription. Nature communications **7**(1), 1–15 (2016)
25. Lafarge, M., Pluim, J., Eppenhof, K., Veta, M.: Learning domain-invariant representations of histological images. Frontiers in medicine **6**, 162 (2019)
26. Li, J., Yang, S., Huang, X., Da, Q., Yang, X., Hu, Z., Duan, Q., Wang, C., Li, H.: Signet ring cell detection with a semi-supervised learning framework. In: International Conference on Information Processing in Medical Imaging. pp. 842–854. Springer (2019)
27. 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**, 26286 (2016)
28. Mahmood, F., Borders, D., Chen, R., McKay, G.N., Salimian, K.J., Baras, A., Durr, N.J.: Deep adversarial training for multi-organ nuclei segmentation in histopathology images. IEEE transactions on medical imaging (2019)

29. Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D.A., Barnholtz-Sloan, J.S., Vega, J.E.V., Brat, D.J., Cooper, L.A.: Predicting cancer outcomes from histology and genomics using convolutional networks. Proceedings of the National Academy of Sciences **115**(13), E2970–E2979 (2018)
30. Niazi, M.K.K., Parwani, A.V., Gurcan, M.N.: Digital pathology and artificial intelligence. The Lancet Oncology **20**(5), e253–e261 (2019)
31. Otálora, S., Atzori, M., Andrearczyk, V., Khan, A., Müller, H.: Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology. Frontiers in Bioengineering and Biotechnology **7**, 198 (2019)
32. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2009)
33. Pantanowitz, L., Sinard, J.H., Henricks, W.H., Fatheree, L.A., Carter, A.B., Contis, L., Beckwith, B.A., Evans, A.J., Lal, A., Parwani, A.V.: Validating whole slide imaging for diagnostic purposes in pathology: guideline from the college of american pathologists pathology and laboratory quality center. Archives of Pathology and Laboratory Medicine **137**(12), 1710–1722 (2013)
34. Ren, J., Hacihaliloglu, I., Singer, E.A., Foran, D.J., Qi, X.: Unsupervised domain adaptation for classification of histopathology whole-slide images. Frontiers in bioengineering and biotechnology **7** (2019)
35. Riordan, D.P., Varma, S., West, R.B., Brown, P.O.: Automated analysis and classification of histological tissue features by multi-dimensional microscopic molecular profiling. PloS one **10**(7), e0128975 (2015)
36. Rolls, G., et al.: 101 steps to better histology. Melbourne: Leica Microsystems **7** (2008)
37. Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al.: Gland segmentation in colon histology images: The glas challenge contest. Medical image analysis **35**, 489–502 (2017)
38. Smith, L.N.: Cyclical learning rates for training neural networks. In: Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on. pp. 464–472. IEEE (2017)
39. Stacke, K., Eilertsen, G., Unger, J., Lundström, C.: A closer look at domain shift for deep learning in histopathology. MICCAI 2019 Computational Pathology Workshop COMPAY (2019)
40. Takahama, S., Kurose, Y., Mukuta, Y., Abe, H., Fukayama, M., Yoshizawa, A., Kitagawa, M., Harada, T.: Multi-stage pathological image classification using semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10702–10711 (2019)
41. Tellez, D., Balkenhol, M., Karssemeijer, N., Litjens, G., van der Laak, J., Ciompi, F.: H and e stain augmentation improves generalization of convolutional networks for histopathological mitosis detection. In: Medical Imaging 2018: Digital Pathology. vol. 10581, p. 105810Z. International Society for Optics and Photonics (2018)
42. Tellez, D., Balkenhol, M., Otte-Höller, I., van de Loo, R., Vogels, R., Bult, P., Wauters, C., Vreuls, W., Mol, S., Karssemeijer, N., et al.: Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. IEEE transactions on medical imaging **37**(9), 2126–2136 (2018)
43. Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., van der Laak, J.: Quantifying the effects of data augmentation and stain color normaliza-

tion in convolutional neural networks for computational pathology. Medical Image Analysis **58**, 101544 (2019)
44. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant cnns for digital pathology. In: International Conference on Medical image computing and computer-assisted intervention. pp. 210–218. Springer (2018)
45. Wilson, G., Cook, D.J.: A survey of unsupervised deep domain adaptation. arXiv preprint arXiv:1812.02849 (2019)
46. Wu, B., Zhao, S., Sun, G., Zhang, X., Su, Z., Zeng, C., Liu, Z.: P3sgd: Patient privacy preserving sgd for regularizing deep cnns in pathological image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2099–2108 (2019)
47. Zhang, Y., Barzilay, R., Jaakkola, T.: Aspect-augmented adversarial networks for domain adaptation. Transactions of the Association for Computational Linguistics **5**, 515–528 (2017)
48. Zoph, B., Le, Q.V.: Neural architecture search with reinforcement learning. In: International Conference on LearningRepresentations (2017)