TAKT: Target-Aware Knowledge Transfer for Whole Slide Image Classification

Conghao Xiong[1(✉)][0000−0003−0134−9397], Yi Lin[2][0000−0002−7635−2518], Hao Chen[2,3][0000−0002−8400−3780], Hao Zheng[4(✉)], Dong Wei[4], Yefeng Zheng[4][0000−0003−2195−2847], Joseph J. Y. Sung[5][0000−0003−3125−5199], and Irwin King[1][0000−0001−8106−6447]

[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
[2] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[3] Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[4] Jarvis Research Center, Tencent YouTu Lab, Shenzhen, China
[5] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore
chxiong21@cse.cuhk.edu.hk, howzheng@tencent.com

Abstract. Knowledge transfer from a source to a target domain is vital for whole slide image classification, given the limited dataset size due to high annotation costs. However, domain shift and task discrepancy between datasets can impede this process. To address these issues, we propose a Target-Aware Knowledge Transfer framework using a teacher-student paradigm, enabling a teacher model to learn common knowledge from both domains by actively incorporating unlabelled target images into the teacher model training. The teacher bag features are subsequently adapted to supervise the student model training on the target domain. Despite incorporating the target features during training, the teacher model tends to neglect them under inherent domain shift and task discrepancy. To alleviate this, we introduce a target-aware feature alignment module to establish a transferable latent relationship between the source and target features by solving an optimal transport problem. Experimental results show that models employing knowledge transfer outperform those trained from scratch, and our method achieves state-of-the-art performance among other knowledge transfer methods on various datasets, including TCGA-RCC, TCGA-NSCLC, and Camelyon16. Codes are released at https://github.com/BearCleverProud/TAKT.

Keywords: Knowledge Transfer · Computational Pathology · Whole Slide Image Classification.

1 Introduction

Whole **S**lide **I**mage (WSI) refers to the digitised glass slides containing histology tissues, which is crucial for cancer diagnosis. Consequently, WSI classification

C. Xiong and Y. Lin—Equal contribution.

has become a research focus [14,17,20,26], including skin [13], lung [5], breast [19], prostate [4], and pancreas [15] cancers, while its success heavily depends on the availability of a large set of labeled samples. However, annotating WSIs is labour-intensive, and in certain cases, the patient cohort is limited, resulting in fewer WSIs in one dataset [2], limiting the data-hungry deep learning techniques.

An effective solution is to acquire knowledge from other datasets. Malignant cells share similar morphological characteristics, such as enlarged nuclei, irregular size and shape, prominent nucleoli, and intense or pale cytoplasm [3]. The shared characteristics can be leveraged to alleviate the over-fitting problem in relatively small datasets [21,29,30]. However, in practice, knowledge transfer could suffer from two challenges, namely, domain shift and task discrepancy.

Domain shift is a change in distribution, caused by differences in tumour sizes and colour tones across datasets. Models trained on one dataset are often biased towards it [6], limiting knowledge transfer. Task discrepancy is inconsistency in tasks between domains, and task-specific features may not directly transfer to another task. Previous works address domain shift from the domain adaptation perspective [1,7]. However, they focus on style differences and assume that the label space remains the same, while the label spaces are different in our case.

To tackle these issues, we propose a **T**arget-**A**ware **K**nowledge **T**ransfer (TAKT) framework to enable the teacher model to learn common knowledge across two domains in a teacher-student paradigm. Within this framework, we actively incorporate unlabelled data from the target domain during the training of the teacher model, using our proposed **T**arget-**A**ware **D**ata **A**ugmentation (TADA) method. When augmenting a source feature, we retrieve the closest centroids generated by an unsupervised clustering algorithm from the target dataset and integrate them into the source features. These centroids are the most representative and relevant samples from the target domain, which enhance the diversity of the augmented dataset without compromising its original distribution. This approach has two rationales. Firstly, integrating target domain representations enhances the transferability of learned knowledge in a specified direction, *i.e.*, from the source domain to the target domain. Secondly, rich data augmentation inherently helps the model learn more generalised features.

However, it is observed that due to the domain shift and task discrepancy, the teacher model still tend to overlook the target features after TADA, and hence the teacher model remains biased towards the source domain. To address this issue, we introduce a **T**arget-**A**ware **F**eature **A**lignment (TAFA) module, applied between features from source and target domains. This module establishes a transferable relationship between the source and target features by solving an **O**ptimal **T**ransport (OT) problem, providing a holistic view of the two distributions. This relationship enforces the teacher model to pay similar attention to the target features, enabling the model to learn more common knowledge across both domains. The contributions of this paper are summarised below:

1. We propose a TAKT framework to alleviate the impact of domain shift by actively incorporating unlabelled target data with the source domain.

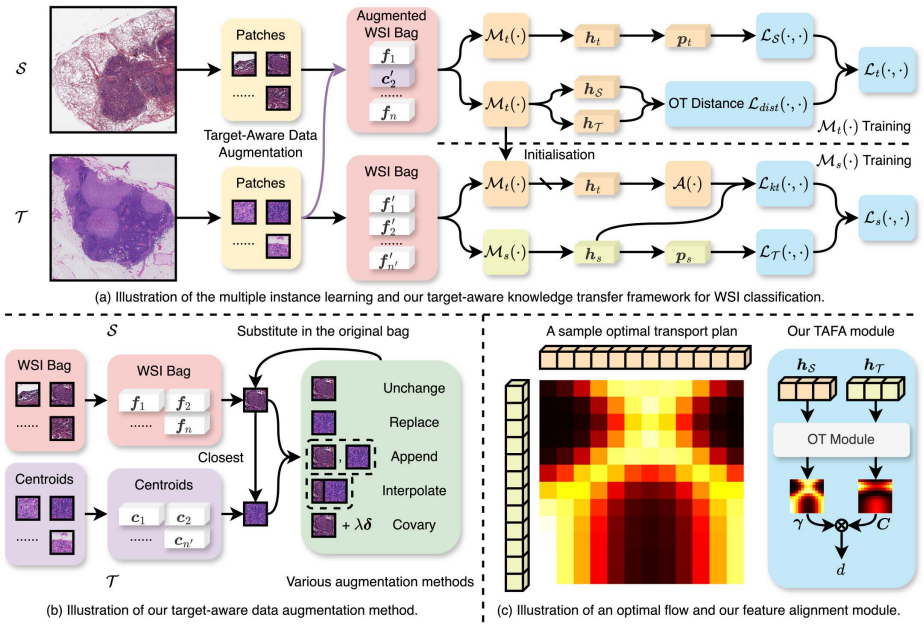

Fig. 1. Illustrations of (a) our target-aware knowledge transfer framework, (b) our target-aware data augmentation method and (c) our target-aware feature alignment module and a sample optimal transport flow. We first train the teacher model and then the student model. The light areas in (c) indicate regions with higher values.

2. We propose a TAFA module to mitigate bias towards the source domain by establishing a latent relationship between source and target features.
3. We conduct extensive experiments on three datasets, including RCC, NSCLC and Camelyon16. Results demonstrate the efficacy of our proposed method.

2 Methodology

2.1 Overview of the Target-Aware Knowledge Transfer

The TAKT framework, shown in Fig. 1(a), has two parts: a TADA method and a TAFA module. Following the teacher-student paradigm, we first train a teacher model $\mathcal{M}_t(\cdot)$ on a dataset generated with TADA and then train a student model $\mathcal{M}_s(\cdot)$ on the target dataset, supervised by the adapted teacher bag features. Mathematically, the loss function for the student model $\mathcal{L}_s$ is given as,

$$\mathcal{L}_s = \mathcal{L}_\mathcal{T} + \alpha\mathcal{L}_{kt} = \mathcal{L}_\mathcal{T} + \alpha \sum \left(\text{MHA}\left(\text{sign}(\boldsymbol{h}_t) | \frac{\boldsymbol{h}_t}{T} |^{\frac{1}{n}} \right) - \boldsymbol{h}_s \right)^2, \quad (1)$$

where $\mathcal{L}_\mathcal{T}$ is the loss function on the target domain $\mathcal{T}$ (in our case, cross entropy loss applied to the student logits $\boldsymbol{p}_s$), α is a coefficient, MHA$(\cdot)$ is the **M**ulti-**H**ead **A**ttention (MHA) [25], used to adapt source-specific features from the

teacher model to target-specific ones that are more suitable for the student model to learn, $\boldsymbol{h}_t, \boldsymbol{h}_s$ are the teacher and student bag features, respectively, and $T = 0.1, n = 3$ are coefficients in **P**ower **T**emperature **S**caling (PTS) [10]. The PTS norm together with an MHA module is the $\mathcal{A}(\cdot)$ module in Fig. 1(a).

2.2 Target-Aware Data Augmentation

Overview of the TADA Method. The illustration of TADA is shown in Fig. 1(b). The augmentation is performed at the instance level between WSI bags, while for target WSI bags, clustered centroids instead of raw patch features are leveraged. During the augmentation, we find the closest centroid for each source feature. The augmentation operations are designed based on ReMix [28].

Step 1: Clustering of the Target Dataset. Each WSI bag from the target domain $\mathcal{T}$ is clustered using K-means, which discovers highly representative features within that bag. Consequently, the feature vectors are transformed into more representative centroids that can effectively represent a set of features. The source dataset remains unchanged due to the potential adverse effects of clustering on the dataset, such as a significant reduction in the number of instances, in which case, augmentation techniques would greatly change the original distribution of the source dataset, thereby negatively affecting the performance of the teacher model and its transferability to downstream tasks.

Step 2: Cross-Dataset Mix Operation. The motivation behind our TADA is to enrich the source features by imparting them with a specific "direction" to transfer, *i.e.*, towards the target domain. Trained with the target data, the teacher model grasps the general target feature distribution, enabling it to learn source features that exhibit better generalisability compared to the teachers that have not been exposed to the target data. Specifically, for each WSI bag in the source dataset, we enumerate each feature and augment the feature with a fixed probability p. For each feature $\boldsymbol{f}$ selected for augmentation, we first identify the closest centroid vector $\boldsymbol{c}$ from the stacked centroid matrix of the target domain. With the closest centroid vector $\boldsymbol{c}$, we employ the following operations:

1. **Append.** $\boldsymbol{c}$ is appended to the bag.
2. **Replace.** $\boldsymbol{f}$ is replaced with $\boldsymbol{c}$ in the bag.
3. **Interpolate.** The interpolation vector $\boldsymbol{f}_I = (1-\lambda)\boldsymbol{f} + \lambda\boldsymbol{c}$ is appended to the bag, where $\lambda \in (0, 1)$ is the strength of the augmentation.
4. **Covary.** The vector $\boldsymbol{f}_C = \boldsymbol{f} + \lambda\boldsymbol{\delta}$ is appended to the bag. $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{c}})$, where $\Sigma_{\boldsymbol{c}}$ is the covariance matrix of $\boldsymbol{c}$ and $\mathcal{N}(\cdot, \cdot)$ is the normal distribution.
5. **Joint.** All the aforementioned methods are applied to the bag.

2.3 Target-Aware Feature Alignment Module

Most of the recent methods for WSI fall under the category of attention-based multiple instance learning [14,20,26]. These methods calculate attention scores

based on patch features using a gated attention network [14] to determine the importance of each patch in the final prediction. In our framework, we observed that attention scores for target features are often lower than those for the source domain, indicating the teacher model is still biased towards the source domain after TADA. This bias hampers the effectiveness of data augmentation and diminishes the transferability of teacher features. However, directly penalising the low attention scores is not reasonable since target features may not necessarily contain source-specific features. Therefore, we propose a target-aware feature alignment module to establish a transferable latent relationship between source and target features instead of pursuing complete consistency, and this also prevent the target features to be neglected during training.

As shown in Fig. 1, the inputs to TAFA module are bag features exclusively obtained from the source features ($\boldsymbol{h}_{\mathcal{S}}$) and target centroids ($\boldsymbol{h}_{\mathcal{T}}$). Within the TAFA module, how to establish the transferable latent relationship is treated as an OT problem. The OT distance is derived from the transportation plan with the lowest cost, providing a comprehensive view of the difference between two distributions in terms of shape, density, and spread, which are lacking in the traditional measures, such as Euclidean distance and cosine similarity. Mathematically, the input to the teacher model $\mathcal{M}_t(\cdot)$ is denoted as $\boldsymbol{F} = [\boldsymbol{F}_{\mathcal{S}}, \boldsymbol{C}_{\mathcal{T}}]$, where $\boldsymbol{F}_{\mathcal{S}}, \boldsymbol{C}_{\mathcal{T}}$ are the feature vectors and centroids from the source and target domains, respectively. The bag features from the teacher model only using $\boldsymbol{F}_{\mathcal{S}}$ and $\boldsymbol{C}_{\mathcal{T}}$ are denoted as $\boldsymbol{h}_{\mathcal{S}} = \mathcal{M}_t(\boldsymbol{F}_{\mathcal{S}})$ and $\boldsymbol{h}_{\mathcal{T}} = \mathcal{M}_t(\boldsymbol{C}_{\mathcal{T}})$, respectively. For each input pair $(\boldsymbol{h}_{\mathcal{S}}, \boldsymbol{h}_{\mathcal{T}})$, we solve for the unbalanced OT problem,

$$\min_{\boldsymbol{\gamma}} < \boldsymbol{\gamma}, \boldsymbol{C} >_F + r \cdot \Omega(\boldsymbol{\gamma}) + r_1 \cdot \mathrm{KL}(\boldsymbol{\gamma}\mathbf{1}, \boldsymbol{a}) + r_2 \cdot \mathrm{KL}(\boldsymbol{\gamma}^T\mathbf{1}, \boldsymbol{b}), \tag{2}$$

where the $(i,\ j)$-th element in the cost matrix $\boldsymbol{C}$ is $|\boldsymbol{h}_{\mathcal{S}_i} - \boldsymbol{h}_{\mathcal{T}_j}|$, $< \cdot, \cdot >_F$ is the Frobenius inner product, $\Omega(\boldsymbol{\gamma}) = \mathrm{KL}(\boldsymbol{\gamma}, \boldsymbol{a}\boldsymbol{b}^T)$ is the entropic regularisation term, r, r_1, r_2 are the regularisation coefficients, $\boldsymbol{a}, \boldsymbol{b} = [1/c, \cdots, 1/c]$ are two uniform distributions of size c, c is the dimension of $\boldsymbol{h}_{\mathcal{S}}$ and $\boldsymbol{h}_{\mathcal{T}}$, $\mathrm{KL}(\cdot, \cdot)$ is the Kullback–Leibler divergence, and $\boldsymbol{\gamma} \geq 0$ is the optimal transport flow. The optimal flow indicates how teacher features can be transformed to have the same distribution as the student feature at minimal cost, considering the amount of mass transferred. The calculated distance between these two features decreases as the mass moved decreases. The problem can be solved using the Sinkhorn-Knopp algorithm [8]. Having obtained the optimal flow $\boldsymbol{\gamma}$, we calculate the distance $d = \sum \boldsymbol{C}\boldsymbol{\gamma}$. The overall loss function of the teacher model $\mathcal{L}_t$ is given as,

$$\mathcal{L}_t = \mathcal{L}_{\mathcal{S}} + \beta\mathcal{L}_{dist} = \mathcal{L}_{\mathcal{S}} + \beta \sum \boldsymbol{C}(\boldsymbol{h}_{\mathcal{S}}, \boldsymbol{h}_{\mathcal{T}}) \cdot \boldsymbol{\gamma}(\boldsymbol{h}_{\mathcal{S}}, \boldsymbol{h}_{\mathcal{T}}), \tag{3}$$

where $\mathcal{L}_{\mathcal{S}}$ is the loss function on the source domain (in our case, cross entropy loss applied to the teacher logits $\boldsymbol{p}_t$), and $\beta > 0$ is the coefficient. Within this formulation, the latent relationship is represented as the optimal flow, and the optimisation of the regularisation term aims to enhance the transferability of features. The establishment of such a transferable relationship ensures the target centroids receive enough attention during the training of teacher model.

Table 1. Results on Camelyon16 with source domains being NSCLC or RCC. The best results are in red underlined, and the second best ones are in blue italic. The subscript in each cell is the standard deviation.

Method	NSCLC → Camelyon16			RCC → Camelyon16		
	AUC↑	F1↑	Accuracy↑	AUC↑	F1↑	Accuracy↑
CLAM [20]	0.814 $_{0.010}$	0.764 $_{0.032}$	0.801 $_{0.027}$	0.814 $_{0.0010}$	0.764 $_{0.032}$	0.801 $_{0.027}$
Fine-tuning	*0.885* $_{0.012}$	*0.831* $_{0.027}$	*0.853* $_{0.021}$	*0.872* $_{0.044}$	*0.844* $_{0.035}$	*0.855* $_{0.024}$
ST [11]	0.803 $_{0.016}$	0.747 $_{0.017}$	0.793 $_{0.012}$	0.820 $_{0.012}$	0.742 $_{0.012}$	0.788 $_{0.009}$
NST [12]	0.824 $_{0.080}$	0.758 $_{0.063}$	0.796 $_{0.043}$	0.814 $_{0.023}$	0.742 $_{0.007}$	0.788 $_{0.005}$
AT [16]	0.821 $_{0.011}$	0.760 $_{0.030}$	0.804 $_{0.022}$	0.847 $_{0.026}$	0.763 $_{0.045}$	0.804 $_{0.024}$
PKT [22]	0.794 $_{0.022}$	0.739 $_{0.016}$	0.788 $_{0.009}$	0.813 $_{0.025}$	0.743 $_{0.012}$	0.791 $_{0.008}$
CC [23]	0.858 $_{0.054}$	0.801 $_{0.055}$	0.827 $_{0.040}$	0.839 $_{0.063}$	0.777 $_{0.064}$	0.814 $_{0.040}$
SP [24]	0.819 $_{0.012}$	0.797 $_{0.023}$	0.827 $_{0.016}$	0.813 $_{0.025}$	0.743 $_{0.012}$	0.791 $_{0.008}$
PTS [10]	0.835 $_{0.037}$	0.780 $_{0.031}$	0.811 $_{0.020}$	0.795 $_{0.020}$	0.754 $_{0.006}$	0.793 $_{0.005}$
TAKT	0.952 $_{0.011}$	0.896 $_{0.012}$	0.904 $_{0.012}$	0.926 $_{0.010}$	0.854 $_{0.013}$	0.866 $_{0.012}$
p-value	0.0006	0.0049	0.0049	0.0027	0.1598	0.1340

3 Experiments and Results

3.1 Dataset Descriptions

Camelyon16. The Camelyon16 dataset [18] contains 399 WSIs of lymph nodes from women with breast cancer. The purpose of this dataset is metastasis detection. The training and test sets contain 270 and 129 WSIs, respectively. We further split the provided training set into training and validation datasets by 8:2 and compare the performances with other methods on the official test dataset.

TCGA-RCC (RCC). The RCC[1] dataset contains 940 WSIs, with 121 WSIs from 109 cases of Kidney Chromophobe Renal Cell Carcinoma (KICH), 519 WSIs from 513 cases of Kidney Renal Clear Cell Carcinoma (KIRC), and 300 WSIs from 276 cases of Kidney Renal Papillary Cell Carcinoma (KIRP). The dataset is split into training, validation, and test sets by the ratio of 6:1.5:2.5.

TCGA-NSCLC (NSCLC). The NSCLC[1] dataset includes 1,053 WSIs: 512 from 478 Lung Squamous Cell Carcinoma (LUSC) cases and 541 from 478 Lung Adenocarcinoma (LUAD) cases. The dataset split is identical to RCC.

3.2 Implementation Details

Evaluation Metrics. **A**rea **U**nder the **C**urve (AUC), F1 and accuracy are the evaluation metrics. These metrics can holistically reflect the overall performances of the models. The thresholds of F1 and accuracy scores are set to 0.5. We also conduct a significance test to further assess the significance of the difference between the means of the highest metrics and those of the second-highest metrics.

[1] https://www.cancer.gov/tcga

Table 2. Results on NSCLC and RCC transferring to each other.

Method	RCC → NSCLC			NSCLC → RCC		
	AUC↑	F1↑	Accuracy↑	AUC↑	F1↑	Accuracy↑
CLAM [20]	0.936 $_{0.015}$	0.859 $_{0.007}$	0.859 $_{0.007}$	0.970 $_{0.003}$	0.859 $_{0.010}$	0.886 $_{0.007}$
Fine-tuning	0.944 $_{0.002}$	0.876 $_{0.008}$	0.876 $_{0.008}$	0.979 $_{0.002}$	*0.874* $_{0.010}$	0.899 $_{0.009}$
ST [11]	0.942 $_{0.004}$	0.865 $_{0.004}$	0.866 $_{0.004}$	0.977 $_{0.006}$	0.860 $_{0.006}$	0.889 $_{0.007}$
NST [12]	0.943 $_{0.007}$	0.874 $_{0.007}$	0.875 $_{0.007}$	*0.980* $_{0.003}$	0.871 $_{0.006}$	*0.900* $_{0.003}$
AT [16]	*0.949* $_{0.001}$	*0.877* $_{0.004}$	*0.877* $_{0.004}$	0.977 $_{0.002}$	0.863 $_{0.003}$	0.893 $_{0.000}$
PKT [22]	0.944 $_{0.001}$	0.872 $_{0.012}$	0.872 $_{0.012}$	0.978 $_{0.005}$	0.865 $_{0.014}$	0.895 $_{0.011}$
CC [23]	0.945 $_{0.004}$	0.872 $_{0.012}$	0.872 $_{0.012}$	0.977 $_{0.005}$	0.864 $_{0.008}$	0.893 $_{0.004}$
SP [24]	0.937 $_{0.004}$	*0.877* $_{0.006}$	*0.877* $_{0.006}$	0.979 $_{0.001}$	*0.874* $_{0.003}$	0.896 $_{0.003}$
PTS [10]	0.942 $_{0.003}$	0.864 $_{0.009}$	0.864 $_{0.009}$	0.970 $_{0.002}$	0.865 $_{0.004}$	0.885 $_{0.004}$
TAKT	0.955 $_{0.003}$	0.890 $_{0.010}$	0.890 $_{0.010}$	0.984 $_{0.002}$	0.890 $_{0.008}$	0.909 $_{0.007}$
p-value	0.0048	0.0288	0.0288	0.0429	0.0072	0.0253

Training Settings. Methods comparing with ours include no knowledge transfer, fine-tuning, **S**oft **T**arget (ST) [11], **N**euron **S**electivity **T**ransfer (NST) [12], **A**ttention **T**ransfer (AT) [16], **P**robabilistic **K**nowledge **T**ransfer (PKT) [22], **C**orrelation **C**ongruence (CC) [23], **S**imilarity **P**reserving (SP) [24] and PTS norm [10]. The base model is **CL**ustering-constrained-**A**ttention **M**ultiple-instance learning (CLAM) [20]. The student model in our method is initialised with the teacher model. Both the student and teacher models are trained up to 200 and no less than 50 epochs. The training is ceased when the validation loss stops decreasing for 20 epochs. All experiments are repeated three times with different seeds. The means and standard deviations of the performances are reported.

Hyper-parameters. The learning rate, weight decay and dropout are set to 2×10^{-4}, 1×10^{-5}, and 0.25, respectively [20]. Adam optimiser is used [20]. The probability of augmenting each feature p and the strength of augmentation λ is set to 0.3 and 0.5, respectively [28]. The number of MHA heads is 8. WSIs are split into non-overlapping patches of 256 × 256 pixels at 20× magnification [20]. ResNet-50 [9] pre-trained on ImageNet is used to extract features from them. r, r_1, r_2 are set to 0.1, 0.5, 0.5, respectively [27]. We perform sensitivity analysis on α, β in supplementary materials, and they are set to 0.1, 0.2, respectively.

3.3 Comparison Results

We compare our method with other related knowledge transfer methods in four settings: NSCLC to Camelyon16, RCC to Camelyon16, RCC to NSCLC and NSCLC to RCC. The experimental results are reported in Table 1 and Table 2. Our method achieves the best performance across every metric, especially on Camelyon16, where our method outperforms other methods by a large margin. In addition, the p-values comparing the best (ours) and second-best metrics indicate that our method significantly outperforms second-best performing

Table 3. Ablation studies on NSCLC to Camelyon16 transfer. Left: Ablation study results on our "replace" augmentation method (R), MSE, cosine similarity (cos), and OT distance with dimension reduction. Right: The augmentation methods are denoted as: Append (A), Replace (R), Interpolate (I), Covary (C), and Joint (J).

Method	NSCLC → Camelyon16 AUC↑	F1↑	Acc.↑	Method	NSCLC → Camelyon16 AUC↑	F1↑	Acc.↑
No Aug.	$0.921_{0.019}$	$0.855_{0.046}$	$0.871_{0.038}$	+A,OT	$0.939_{0.018}$	*$0.885_{0.031}$*	*$0.897_{0.025}$*
+R	$0.932_{0.013}$	$0.862_{0.016}$	$0.873_{0.016}$	+R,OT	$\underline{0.952}_{0.011}$	$\underline{0.896}_{0.012}$	$\underline{0.904}_{0.012}$
+R,MSE	$0.942_{0.007}$	$0.861_{0.012}$	$0.873_{0.012}$	+I,OT	$0.933_{0.021}$	$0.854_{0.038}$	$0.871_{0.031}$
+R,cos	*$0.943_{0.004}$*	*$0.877_{0.006}$*	*$0.889_{0.005}$*	+C,OT	$0.931_{0.012}$	$0.854_{0.018}$	$0.868_{0.013}$
+R,OT	$\underline{0.952}_{0.011}$	$\underline{0.896}_{0.012}$	$\underline{0.904}_{0.012}$	+J,OT	*$0.941_{0.004}$*	$0.865_{0.013}$	$0.876_{0.013}$

methods. Furthermore, due to the small tumour size and the limited number of samples, Camelyon16 is more difficult than TCGA datasets, which is reflected in the absolute value of the metrics. The results prove that our method can effectively transfer knowledge from a simpler dataset to a harder one. Another observation is that methods with additional supervision signals perform better on TCGA datasets, and fine-tuning performs better when transferring knowledge from TCGA to the Camelyon16 dataset. The potential reason for this is that TCGA datasets are more similar, as evidenced by the maximum mean discrepancy scores between these datasets. The tumour size is drastically different between Camelyon16 and TCGA datasets, leading to significant differences in attention distributions. Methods with additional supervision signals (features, attention scores and logits) all rely on the attention scores as bag features are eventually derived from them. Therefore, these methods may consistently introduce bias during student training, resulting in lower performance.

3.4 Ablation Studies

TAFA Module. We investigate the impact of "replace" augmentation with different alignment losses on the final performance, including **M**ean **S**quared **E**rror (MSE), cosine similarity (cos) and OT distance. The experimental results are shown in the left part of Table 3. When we incorporate the "replace" augmentation during teacher model training, the AUC score increases by 1.10%. With MSE, cos and OT, the AUC score further increases by 1.00% and 1.10% and 2.00%, respectively, demonstrating the effectiveness of OT distance.

Augmentation Methods. We conduct experiments using the aforementioned augmentation methods. The experimental results are presented in the right part of Table 3. Incorporating augmentation methods results in improved AUC scores compared to the absence of augmentation, with a minimum increase of 1.00%, highlighting their effectiveness. Particularly, the "replace" augmentation method achieved the highest performance, surpassing other methods by a significant margin (with a best average AUC score of 95.2%).

4 Conclusion

In this work, we presented a TAKT framework specifically designed for WSI classification. The framework contains two main components: the TADA method and the TAFA module. The TADA method augments the source dataset by actively retrieving the closest centroids from the target domain, facilitating the teacher model to acquire common knowledge from both domains. To mitigate the bias of teacher model towards the source domain, we proposed a TAFA module to establish a latent relationship between source and target features by solving a OT problem, enforcing the teacher model to pay similar attention to features from both the source and target domains. Our experimental results demonstrated that models trained with knowledge transfer techniques outperformed those trained from scratch and our method achieved state-of-the-art performance among other adapted knowledge transfer methods.

Acknowledgments. This research was done with Jarvis Research Center, Tencent YouTu Lab and partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK 14222922, RGC GRF 2151185).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahn, E., Kumar, A., Fulham, M., Feng, D., Kim, J.: Unsupervised domain adaptation to classify medical images using zero-bias convolutional auto-encoders and context-based feature augmentation. IEEE transactions on medical imaging **39**(7), 2385–2394 (2020) 2
2. Aumpan, N., Vilaichone, R.k., Pornthisarn, B., Chonprasertsuk, S., Siramolpiwat, S., Bhanthumkomol, P., Nunanan, P., Issariyakulkarn, N., Ratana-Amornpin, S., Miftahussurur, M., et al.: Predictors for regression and progression of intestinal metaplasia (im): a large population-based study from low prevalence area of gastric cancer (im-predictor trial). PloS one **16**(8), e0255601 (2021) 2
3. Baba, A.I., Câtoi, C.: Comparative oncology. Publishing House of the Romanian Academy (2007) 2
4. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., 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) 2
5. Chen, C.L., Chen, C.C., Yu, W.H., Chen, S.H., Chang, Y.C., Hsu, T.I., Hsiao, M., Yeh, C.Y., Chen, C.Y.: An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. Nature communications **12**(1), 1193 (2021) 2
6. Cui, Y., Liu, Z., Chen, Y., Lu, Y., Yu, X., Liu, X.S., Kuo, T.W., Rodrigues, M., Xue, C.J., Chan, A.: Retrieval-augmented multiple instance learning. Advances in Neural Information Processing Systems **36** (2024) 2

7. Feng, Y., Xu, X., Wang, Y., Lei, X., Teo, S.K., Sim, J.Z.T., Ting, Y., Zhen, L., Zhou, J.T., Liu, Y., et al.: Deep supervised domain adaptation for pneumonia diagnosis from chest x-ray images. IEEE Journal of Biomedical and Health Informatics **26**(3), 1080–1090 (2021) 2
8. Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M., Poggio, T.A.: Learning with a wasserstein loss. In: Advances in Neural Information Processing Systems. pp. 2053–2061 (2015) 5
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016) 7
10. He, R., Sun, S., Yang, J., Bai, S., Qi, X.: Knowledge distillation as efficient pre-training: Faster convergence, higher data-efficiency, and better transferability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9161–9171 (2022) 4, 6, 7
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv:1503.02531 (2015) 6, 7
12. Huang, Z., Wang, N.: Like what you like: Knowledge distill via neuron selectivity transfer. arXiv:1707.01219 (2017) 6, 7
13. Ianni, J.D., Soans, R.E., Sankarapandian, S., Chamarthi, R.V., Ayyagari, D., Olsen, T.G., Bonham, M.J., Stavish, C.C., Motaparthi, K., Cockerell, C.J., et al.: Tailored for real-world: a whole slide image classification system validated on uncurated multi-site data emulating the prospective pathology workload. Scientific Reports **10**(1), 3217 (2020) 2
14. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning. pp. 2127–2136 (2018) 2, 4, 5
15. Keikhosravi, A., Li, B., Liu, Y., Conklin, M.W., Loeffler, A.G., Eliceiri, K.W.: Non-disruptive collagen characterization in clinical histopathology using cross-modality image synthesis. Communications biology **3**(1), 414 (2020) 2
16. Komodakis, N., Zagoruyko, S.: Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In: 5th International Conference on Learning Representations, ICLR 2017 (2017) 6, 7
17. Lin, Y., Zhu, Z., Cheng, K.T., Chen, H.: Prompt-guided adaptive model transformation for whole slide image classification. arXiv:2403.12537 (2024) 2
18. Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., et al.: 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. GigaScience **7**(6), giy065 (2018) 6
19. 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**(1), 26286 (2016) 2
20. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021) 2, 4, 6, 7
21. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2009) 2
22. Passalis, N., Tefas, A.: Learning deep representations with probabilistic knowledge transfer. In: European Conference on Computer Vision. pp. 268–284 (2018) 6, 7
23. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: the IEEE/CVF International Conference on Computer Vision. pp. 5007–5016 (2019) 6, 7

24. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: the IEEE/CVF international conference on computer vision. pp. 1365–1374 (2019) 6, 7
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017) 3
26. Xiong, C., Chen, H., Sung, J.J.Y., King, I.: Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. In: International Joint Conference on Artificial Intelligence. pp. 1587–1595 (2023) 2, 4
27. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: IEEE/CVF International Conference on Computer Vision. pp. 21241–21251 (October 2023) 7
28. Yang, J., Chen, H., Zhao, Y., Yang, F., Zhang, Y., He, L., Yao, J.: Remix: A general and efficient framework for multiple instance learning based whole slide image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 35–45. Springer (2022) 4, 7
29. Yu, X., Wang, J., Hong, Q.Q., Teku, R., Wang, S.H., Zhang, Y.D.: Transfer learning for medical images analyses: A survey. Neurocomputing **489**, 230–254 (2022) 2
30. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proceedings of the IEEE **109**(1), 43–76 (2020) 2