Supplementary Material: Taxonomy Adaptive Cross-Domain Adaptation in Medical Imaging via Optimization Trajectory Distillation

Jianan Fan[1], Dongnan Liu[1], Hang Chang[2], Heng Huang[3], Mei Chen[4], and Weidong Cai[1]

[1]University of Sydney [2]Lawrence Berkeley National Laboratory [3]University of Maryland at College Park [4]Microsoft

In this supplementary, we provide additional illustration of our proposed method and proofs to support theoretical analysis, as well as dataset and implementation details. Extended experiments and analysis are performed to further verify the effectiveness and robustness of our method.

1. Overall Traininig Procedure

We summarize the workflow of our proposed method in Algorithm 1. The referred equation can be found in the main text.

Algorithm 1: Training Procedure for Optimization Trajectory Distillation

Input: Source dataset $\mathcal{D}_s$, target dataset $\mathcal{D}_t$, number of iterations I
Output: Optimized model parameters θ
Init: Gradient memory buffer $\boldsymbol{G}_s, \boldsymbol{G}_A, \boldsymbol{G}_{It}$
for $i \leftarrow 1$ **to** I **do**
  $\{(\mathbf{x}_s, \mathbf{y}_s)\} \leftarrow$ Image-label pairs sampled from$(\mathcal{D}_s)$
  $\{(\mathbf{x}_t^u)\}, \{(\mathbf{x}_t^l, \mathbf{y}_t^l)\} \leftarrow$ Sample from$(\mathcal{D}_t)$
  Generate online pseudo-label $\tilde{\mathbf{y}}_t^u$
  $\{(\mathbf{x}_t, \mathbf{y}_t)\} \leftarrow \{(\mathbf{x}_t^u, \tilde{\mathbf{y}}_t^u), (\mathbf{x}_t^l, \mathbf{y}_t^l)\}$
  Backpropagate gradients $\boldsymbol{g}_s, \boldsymbol{g}_t, \boldsymbol{g}_A, \boldsymbol{g}_N$ by Eq.(2)
  Cross-domain/class distillation:
  Update $\boldsymbol{G}_s$ and $\boldsymbol{G}_A$ with $\boldsymbol{g}_s$ and $\boldsymbol{g}_A$
  if $\boldsymbol{G}_s$ and $\boldsymbol{G}_A$ are full **then**
    Apply SVD to identify the principal subspace and form the corresponding projection matrix $\boldsymbol{M}_s, \boldsymbol{M}_A$ by Eq.(5)(6)
    Clear $\boldsymbol{G}_s$ and $\boldsymbol{G}_A$
  end
  Perform gradient projection by Eq.(7)
  Compute the overall training objective $\mathcal{L}$ by Eq.(8)
  Temporal self-distillation:
  Update $\boldsymbol{G}_{It}$ with $\boldsymbol{g}_{It}$
  if $\boldsymbol{G}_{It}$ is full **then**
    Form the projection matrix $\boldsymbol{M}_{It}$ by Eq.(5)(6)
    Clear $\boldsymbol{G}_{It}$
  end
  Compute the mini-batch gradients $\tilde{\boldsymbol{g}} \leftarrow \nabla_\theta \mathcal{L}(\theta)$
  Update model parameters θ by Eq.(11)
end
return θ

2. Details for Theoretical Analysis

2.1. Joint Characterization of Feature and Output Space

Here, we prove that in our gradient-based method, information in feature space and model-output space can be jointly modeled implicitly, which demonstrates the superiority of our method as a unified framework that jointly characterizes the feature and output space as well as the learning dynamics. We illustrate this property under cross-entropy loss and Dice loss.

Specifically, for the task-specific prediction layer, which is implemented as a convolutional layer with 1×1 kernel, its function can be mathematically expressed as:

$$\boldsymbol{u} = \boldsymbol{W}^\top \boldsymbol{z} + \boldsymbol{b}, \tag{1}$$

where $\boldsymbol{z} \in \mathbb{R}^{B \times m \times h \times w}$ denotes the input feature maps with batch size B, channel number m, and spatial size $h \times w$. $\boldsymbol{W} \in \mathbb{R}^{m \times c}$ is the convolutional kernel matrix with input channel m and output channel c. $\boldsymbol{b} \in \mathbb{R}^c$ indicates the bias tensor. $\boldsymbol{u} \in \mathbb{R}^{B \times c \times h \times w}$ is the logit predictions.

We first consider the gradients for network parameter θ by backpropagating the cross entropy (CE) loss $\mathcal{L}_{CE}$:

$$\min_{\theta \sim [\boldsymbol{W}, \boldsymbol{b}]} -\sum_{n=1}^{p} \sum_{i=1}^{c} \boldsymbol{y}_i^n \log \frac{e^{\boldsymbol{u}_i^n}}{\sum_{j=1}^{c} e^{\boldsymbol{u}_j^n}}, \tag{2}$$

where $p = B \times h \times w$ is the total number of pixels, $\boldsymbol{y} \in \mathbb{R}^{B \times c \times h \times w}$ is the one-hot pixel-wise class label. Since our focus is on feature space $\boldsymbol{z}$ and output space $\boldsymbol{u}$, without loss of generality, we set $\boldsymbol{y}$ to follow the uniform class distribution that $\boldsymbol{y} = [1/c, 1/c, ..., 1/c]$. Then, for each pixel,

the derivative of the CE loss can be formulated as:

$$\begin{aligned}
&\frac{\partial \mathcal{L}_{CE}}{\partial \theta}\Big|_{\theta=[\theta_1,\theta_2,...,\theta_c]} = \frac{\partial \mathcal{L}_{CE}}{\partial \boldsymbol{u}}\Big|_{\boldsymbol{u}=[\boldsymbol{u}_1,\boldsymbol{u}_2,...,\boldsymbol{u}_c]} \cdot \frac{\partial \boldsymbol{u}}{\partial \theta} \\
&= -\frac{1}{c} \cdot \frac{\partial \left(\sum_{i=1}^{c} \log \frac{e^{\boldsymbol{u}_i}}{\sum_{j=1}^{c} e^{\boldsymbol{u}_j}}\right)}{\partial \boldsymbol{u}} \cdot \frac{\partial \boldsymbol{u}}{\partial \theta} \\
&= -\frac{1}{c} \cdot \frac{\partial \left(\sum_{i=1}^{c} \boldsymbol{u}_i - c \cdot \log \sum_{j=1}^{c} e^{\boldsymbol{u}_j}\right)}{\partial \boldsymbol{u}} \cdot \frac{\partial \boldsymbol{u}}{\partial \theta} \\
&= -\frac{1}{c} \cdot (1 - c \cdot \frac{[e^{\boldsymbol{u}_1},\, e^{\boldsymbol{u}_2},\, ...,\, e^{\boldsymbol{u}_c}]}{\sum_{j=1}^{c} e^{\boldsymbol{u}_j}}) \cdot \boldsymbol{z}.
\end{aligned} \tag{3}$$

Then we summarize the gradient magnitudes over all pixels and channels:

$$\begin{aligned}
&\sum_{i=1}^{c} \|\frac{\partial \mathcal{L}_{CE}}{\partial \theta_i}\| = \sum_{n=1}^{p}\sum_{i=1}^{c} \|\frac{\partial \mathcal{L}_{CE}}{\partial \boldsymbol{u}_i^n} \cdot \frac{\partial \boldsymbol{u}_i^n}{\partial \theta_i}\| \\
&= -\frac{1}{c} \cdot \underbrace{(\sum_{n=1}^{p}\sum_{i=1}^{c} \|1 - c \cdot \frac{e^{\boldsymbol{u}_i^n}}{\sum_{j=1}^{c} e^{\boldsymbol{u}_j^n}}\|)}_{\text{output space}} \cdot \underbrace{(\sum_{n=1}^{p} \|\boldsymbol{z}^n\|)}_{\text{feature space}}.
\end{aligned} \tag{4}$$

The result indicates that the gradients of CE loss characterize the information in both feature space and output space.

Similarly, for Dice loss $\mathcal{L}_{Dice}$:

$$\min_{\theta \sim [\boldsymbol{W}, \boldsymbol{b}]} \sum_{n=1}^{p}\sum_{i=1}^{c} (1 - \frac{2\boldsymbol{y}_i^n \texttt{act}(\boldsymbol{u}_i^n)}{\boldsymbol{y}_i^n + \texttt{act}(\boldsymbol{u}_i^n)}). \tag{5}$$

Here we omit the details of softmax layer and use $\texttt{act}$ to denote the activation function for simplicity. When c is large, the pixel-wise derivative can be approximated by:

$$\begin{aligned}
&\frac{\partial \mathcal{L}_{Dice}}{\partial \theta}\Big|_{\theta=[\theta_1,\theta_2,...,\theta_c]} \\
&= \frac{\partial \mathcal{L}_{Dice}}{\partial \texttt{act}(\boldsymbol{u})} \cdot \frac{\partial \texttt{act}(\boldsymbol{u})}{\partial \boldsymbol{u}} \cdot \frac{\partial \boldsymbol{u}}{\partial \theta} \\
&\approx \boldsymbol{\xi} \cdot \frac{[e^{\boldsymbol{u}_1},\, e^{\boldsymbol{u}_2},\, ...,\, e^{\boldsymbol{u}_c}]}{[\texttt{act}(\boldsymbol{u}_1)^2,\, \texttt{act}(\boldsymbol{u}_2)^2,\, ...,\, \texttt{act}(\boldsymbol{u}_c)^2]} \cdot \boldsymbol{z},
\end{aligned} \tag{6}$$

where $\boldsymbol{\xi}$ is a constant term. Its gradient magnitudes can be written as:

$$\begin{aligned}
\sum_{i=1}^{c} \|\frac{\partial \mathcal{L}_{Dice}}{\partial \theta_i}\| &= \sum_{n=1}^{p}\sum_{i=1}^{c} \|\frac{\partial \mathcal{L}_{Dice}}{\partial \texttt{act}(\boldsymbol{u}_i^n)} \cdot \frac{\partial \texttt{act}(\boldsymbol{u}_i^n)}{\partial \boldsymbol{u}_i^n} \cdot \frac{\partial \boldsymbol{u}_i^n}{\partial \theta_i}\| \\
&\approx \boldsymbol{\xi} \cdot \underbrace{(\sum_{n=1}^{p}\sum_{i=1}^{c} \|\frac{e^{\boldsymbol{u}_i^n}}{\texttt{act}(\boldsymbol{u}_i^n)^2}\|)}_{\text{output space}} \cdot \underbrace{(\sum_{n=1}^{p} \|\boldsymbol{z}^n\|)}_{\text{feature space}},
\end{aligned} \tag{7}$$

which proves that the proposition also holds true for Dice loss.

2.2. Impacts on Generalization Error

In this section, we prove the effectiveness of our method towards a tighter generalization error bound on the target domain and novel classes.

Firstly, we analyze the underlying mechanism for the cross-domain distillation module. Let $\mathcal{H}$ be a hypothesis space of VC-dimension d, for $h \in \mathcal{H}$, the correlations between the error on the target domain and the distance in gradient space across domains are established as [1, 8]:

$$\begin{aligned}
\epsilon_T(h) \leq \hat{\epsilon}_S(h) + \frac{4}{n_l}\sqrt{(d \log \frac{2en_l}{d} + \log \frac{4}{\delta})} + \Lambda \\
+ \mathbf{Div}_{\nabla}(\tilde{\boldsymbol{\mathcal{U}}}_{\boldsymbol{S}}, \tilde{\boldsymbol{\mathcal{U}}}_{\boldsymbol{T}}) + 4\sqrt{\frac{4\log(2n_u) + \log(\frac{4}{\delta})}{n_u}},
\end{aligned} \tag{8}$$

with probability at least $1 - \delta$. Here ϵ_T and $\hat{\epsilon}_S$ represent the true and empirically estimated error of the target and source domain, respectively. $Div_{\nabla}(\tilde{\mathcal{U}}_S, \tilde{\mathcal{U}}_T)$ is the distance between data distributions $\tilde{\mathcal{U}}_S$ and $\tilde{\mathcal{U}}_T$ in gradient space. n_l and n_u denote the number of labeled and unlabeled samples. Λ, δ, and e are constants. It implies that constraining the gradient descent trajectory of the target domain to approximate the source domain's, which reduces $Div_{\nabla}(\tilde{\mathcal{U}}_S, \tilde{\mathcal{U}}_T)$, could lead to lower cross-domain generalization error.

Furthermore, we demonstrate that the cross-class distillation module contributes to lower empirical error on novel classes from the multi-task learning perspective. Suppose that $\mathcal{L}$ is the empirical training loss and $\nabla_\theta \mathcal{L}_q(\theta)$ denotes its derivative *w.r.t.* class q. Given a set of anchor classes $\{a_i\}_{i=1}^{A}$ and a novel class q, with the first-order Taylor expansion, we have:

$$\mathcal{L}_q(\theta - \mu \cdot \Delta\theta^*) = \mathcal{L}_q(\theta) - \mu \cdot \nabla_\theta \mathcal{L}_q(\theta) \cdot \Delta\theta^* + \mathcal{O}(\mu), \tag{9}$$

where the optimization step $\Delta\theta^*$ is characterized by $\sum_{i=1}^{A} \nabla_\theta \mathcal{L}_{a_i}(\theta) + \nabla_\theta \mathcal{L}_q(\theta)$, μ is a small value. Then:

$$\begin{aligned}
\mathcal{L}_q(\theta - \mu \cdot \Delta\theta^*) - \mathcal{L}_q(\theta) = -\mu \cdot \Big\{ \|\nabla_\theta \mathcal{L}_q(\theta)\|^2 \\
+ \sum_{i=1}^{A} \left[\boldsymbol{\nabla_\theta \mathcal{L}_q(\theta) \cdot \nabla_\theta \mathcal{L}_{a_i}(\theta)}\right] \Big\} + \mathcal{O}(\mu).
\end{aligned} \tag{10}$$

It indicates that by enforcing the similarity between the gradients *w.r.t.* novel and anchor classes, we could drive the model to reduce the empirical loss on novel classes along optimization and thereby attain a well-generalizable solution.

3. Implementation Details

3.1. Nuclei Segmentation and Recognition

3.1.1 Datasets and Preprocessing

Accurate detection, segmentation, and classification of nuclei serve as essential prerequisites for various clinical and research studies within the digital pathology field [11]. Inconsistent taxonomy for nuclei categorization is common across different institutes, which results in the unmatched label sets among datasets. In this regard, we use PanNuke [6] and Lizard [10] as the source and target dataset, respectively. **PanNuke** contains 481 visual fields cropped from whole-slide images along with 189,744 annotated nuclei. It follows a categorization schema where nuclei are divided into five classes, including neoplastic, non-neoplastic epithelial, inflammatory, connective, and dead cells. We discard the "dead" class as it does not exist in most image patches. To ensure the dataset has a uniform data distribution, we use all images from the breast tissue to formulate the source dataset. **Lizard** consists of 291 image regions with an average size of 1016×917 pixels from the colon tissue and annotates 495,179 nuclei. It adopts a categorization schema different to PanNuke that there are six classes in total, *i.e.*, neutrophil, eosinophil, plasma, lymphocyte, epithelial, and connective cells. We use the Dpath subset as the target dataset. For preprocessing, all the visual fields with divergent size are randomly cropped into image patches of 128×128 pixels. CutMix [27] is used to augment the target dataset.

3.1.2 Network Architectures and Parameter Settings

We employ the widely used Hover-Net [11] architecture with a standard ResNet-50 backbone as the base model. The optimizer is Adam with a learning rate of $1e-4$ and $(\beta_1, \beta_2) = (0.9, 0.999)$, and the batch size is set as 4. To supervise the classification and segmentation branches, we adopt a combined loss of $\mathrm{CrossEntropyLoss} + \mathrm{DiceLoss}$. λ in Eq.(8) and κ in Eq.(11) are empirically set to 1000 and 10, respectively.

3.1.3 Evaluation Metrics

F1 score is a popular metric to evaluate classification performance. It measures both precision and recall harmonically. We report the class-averaged score to indicate the overall accuracy. Panoptic quality (PQ) [11] is a unified metric for the instance segmentation task which models the quality of both detection and segmentation results concurrently:

$$\mathcal{PQ} = \underbrace{\frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}\mathrm{FP} + \frac{1}{2}\mathrm{FN}}}_{\text{Detection Quality}} \cdot \underbrace{\frac{\sum_{(y,\hat{y}) \in \mathrm{TP}} IoU(y, \hat{y})}{\mathrm{TP}}}_{\text{Segmentation Quality}}. \quad (11)$$

where $\mathrm{TP}, \mathrm{FP}, \mathrm{FN}$ are the true positive, false positive, and false negative detection predictions, respectively. $(y, \hat{y})$ represents the pair of ground truth and predicted segmentation mask. IoU is the intersection over union score.

3.2. Cancer Tissue Phenotyping

3.2.1 Datasets and Preprocessing

Identifying distinct tissue phenotypes is an essential step towards systematic profiling of the tumor microenvironment in pathological examination [13]. Previous works are mostly limited to the discrimination of two classes of tissue: tumor and stroma [19], while recent studies argue that recognizing more heterogeneous tissues brings clinical value [15]. We therefore propose to perform adaptation from a dataset with only two categories of tissue to another dataset with several novel classes. In particular, we select images of tumor and stroma tissue from the CRC-TP [13] to form the source dataset. **CRC-TP** contains 20 H&E-stained colorectal cancer (CRC) slides obtained from 20 different patients. Region-level tissue phenotype annotations are provided by expert pathologists. To ensure a unique category label can be assigned to each image, patches are extracted at $20\times$ magnification with the size of 150×150 pixels. The patch-wise tissue phenotypes are decided based on the majority of their content. **Kather** [15] is then regarded as the target dataset. It consists of 5000 150×150 pathology image patches sampled from 10 anonymized CRC tissue slides. Other than tumor and stroma tissue, Kather includes six novel tissue types, *i.e.*, complex stroma, immune cells, debris, normal mucosal glands, adipose tissue, background, and thus poses an 8-class classification problem.

3.2.2 Network Architectures and Parameter Settings

For experiments, we employ ResNet-101 as the image encoder and thereupon add two classification heads on top to perform 2-class and 8-class discrimination, respectively. During training, cross-entropy loss and Adam optimizer with learning rate $1e-4$ are used to optimize the model with a batch size of 4. λ and κ are set to 10 and 100.

3.3. Skin Lesion Diagnosis

3.3.1 Datasets and Preprocessing

Automatic fine-grained skin lesion recognition remains a global challenge in dermatology. By taking a step further than the basic benign/malignant differentiation, identifying the specific subtypes of lesions demonstrates significant diagnostic value [24]. We hereby assign a benign/malignant discrimination dataset as the source domain, and a fine-grained multi-disease dataset as the target domain. **HAM10000** is a dermatoscopic image dataset collected from different populations and modalities [25].

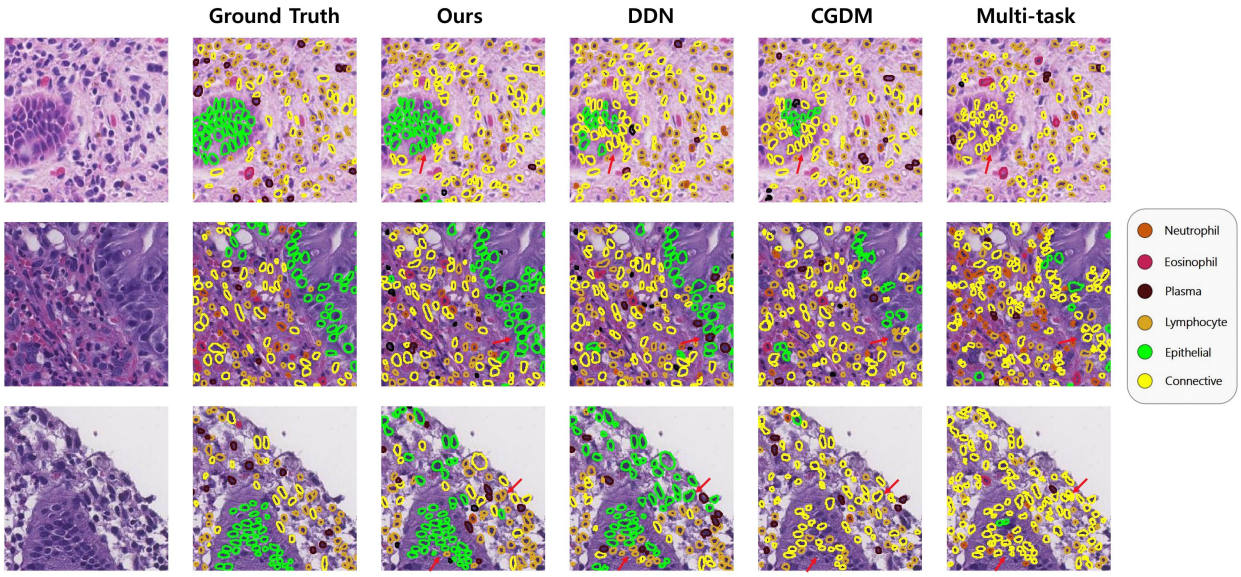

Figure 1. Visual comparisons with other methods on the nuclei segmentation and recognition benchmark.

After preprocessing procedures including histogram correction, sample filtering, and center crop, 10015 dermatoscopic images with lesions of seven diagnostic categories in total are provided. It contains four subtypes of benign lesions (melanocytic nevi (NV), benign keratinocytic lesions (BKL), dermatofibromas (DF), and vascular lesions (VASC)) and three subtypes of malignant ones (melanomas (MEL), basal cell carcinomas (BCC), and actinic keratoses intraepithelial carcinomas (AKIEC)). We use the face subset with only coarse two-class annotations as the source domain and the lower extremity subset with fine-grained seven-class annotations as the target domain. All images are randomly cropped to the size of 160×160 pixels before being forwarded to the network.

3.4. Overall Experiment Settings

For all experiments, we implement our method with Pytorch and conduct training on a NVIDIA GeForce RTX 3090 GPU with 24GB of memory. Gradient backpropagation is performed for each mini-batch using BackPACK [4]. Following previous works in UDA [2], each dataset is randomly split into $80\%/20\%$ as the training/test sets. For novel classes in the target dataset, we sample few (5/10) samples with corresponding labels to formulate the support set. The remaining target data is left unlabeled. Data augmentation techniques such as rotation and horizontal/vertical flip are employed during training. Please refer to the source code for more details.

It is noted that although from the technical perspective, skin lesion diagnosis is a multi-class classification problem similar to cancer tissue phenotyping, they differ largely in the task context. Specifically, skin lesion diagnosis is more like an object recognition task where its decision is dominated by the local attributes of lesions, while cancer tissue phenotyping relies on the global structure of the whole pathology images, instead of focusing on a salient object.

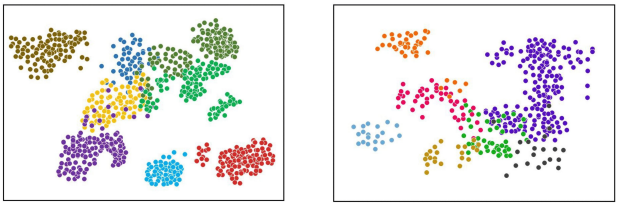

Figure 2. Qualitative visualizations of our proposed method on the cancer tissue phenotyping and skin lesion diagnosis benchmark with t-SNE plot. Varied colours of points indicate the samples of different classes.

4. Additional Experiment Results

4.1. Visualization

We provide additional qualitative results on the three benchmarks. The comparison results shown in Fig. 1 demonstrate the superiority of our method to detect each nucleus and delineate its boundary, as well as differentiating nuclei of various types with their detailed biological features. The t-SNE visualization in Fig. 2 shows that our method could discover the underlying embedding structures of various classes even with very limited labeled data.

4.2. Results with More Annotations

In this section, we compare our method with previous state-of-the-art approaches for cross-domain adaptation

Table 1. Comparison results of our proposed method against other state-of-the-art methods for nuclei segmentation and recognition with 30-shot labeled target samples. The best and second-best results are highlighted in bold and brown, respectively.

Methods	30-shot			
	mF1	mF1*	mPQ	mPQ*
Sup-only	$43.78_{0.56}$	$36.73_{0.34}$	$21.83_{0.54}$	$21.27_{0.38}$
Multi-task [23]	$43.23_{1.02}$	$30.39_{1.37}$	$23.55_{0.49}$	$17.78_{0.72}$
DANN [7]	$41.70_{1.41}$	$28.25_{1.63}$	$22.38_{0.57}$	$15.99_{0.80}$
CGDM [5]	$43.82_{0.98}$	$34.36_{0.91}$	$24.92_{0.34}$	$20.48_{0.26}$
LETR [20]	$40.04_{1.13}$	$25.80_{0.82}$	$21.96_{0.70}$	$14.89_{0.44}$
FT-CIDA [16]	$40.59_{0.87}$	$32.95_{0.54}$	$22.37_{0.58}$	$19.34_{0.30}$
STARTUP [21]	$49.00_{0.74}$	$36.32_{1.10}$	$28.78_{0.46}$	$22.57_{0.43}$
DDN [12]	$48.24_{0.79}$	$36.77_{0.85}$	$27.90_{0.61}$	$22.64_{0.28}$
TSA [18]	$44.96_{0.60}$	$33.13_{0.97}$	$25.69_{0.91}$	$21.40_{0.63}$
TACS [9]	$47.55_{0.81}$	$37.38_{1.55}$	$28.47_{0.83}$	$23.18_{1.04}$
Ours	$\mathbf{51.69}_{0.48}$	$\mathbf{41.63}_{0.65}$	$\mathbf{29.42}_{0.55}$	$\mathbf{25.34}_{0.67}$

Table 2. Comparison results of our proposed method against other state-of-the-art methods on three diverse tasks.

Methods	Radiology		Fundus		OfficeHome	
	mF1	mF1*	mF1	mF1*	mF1	mF1*
Baseline	41.87	19.20	41.03	29.07	44.87	46.54
CIDA [33]	42.55	23.48	39.90	27.32	43.29	42.80
TACS [19]	46.36	22.13	44.84	32.68	42.25	47.71
Ours	**49.23**	**26.54**	**46.26**	**37.71**	**50.08**	**54.67**

when more labeled samples are available in the target domain. The results for nuclei segmentation and recognition with 30 labeled target samples are shown in Table 1. The overall improvements of our method under this setting are consistent with previous experiment results. It validates the effectiveness of our method under different levels of support.

4.3. Extended Experiments on Diverse Tasks

We further evaluate our method on two medical image tasks beyond pathology analysis and one general visual recognition task, where the medical image tasks include pneumonia screening in radiology and diabetic retinopathy grading in fundus photography. For radiology analysis, we adopt covid-kaggle [22] and Chest-Montreal [3] for TADA from normal/pneumonia coarse screening to fine-grained pneumonia diagnosis. For fundus, DDR [17] and APTOS19 [14] are used to construct a TADA setting with two novel classes (grade level $3, 4$). In these settings, distribution shifts exist across domains due to differences in image acquisition protocols among multiple cohorts. For general visual task, we adopt OfficeHome [26] and evaluate on "Artist" and "Real-world" domains. Experiments are conducted in 10-shot regime and the corresponding results are presented in Table 2. Through comparison with SOTA methods, the effectiveness and broader applicability of our method are proved.

4.4. Extended Key Component Analysis

In Table 3, we demonstrate the effectiveness of our method's key components. It complements the ablation study performed in the main text from a cumulative perspective. From the results, we observe that all the proposed modules are beneficial for improving the cross-domain adaptation performance. In particular, the employment of the cross-class distillation module contributes to a significant performance gain for target-private classes. For instance, it attains 3.48% and 2.18% improvements in terms of mF1* and mPQ* under 10-shot scheme. It verifies the effectiveness of our method to perform optimization trajectory distillation across domains and classes towards strong model generalization.

4.5. Extended Hyperparameter Sensitivity Analysis

We further analyze the choices of hyperparameters and their impacts on model performance in Fig. 3. We vary the values of K and T, which denote the volumes of memory bank employed in the cross-domain/class distillation and historical self-distillation modules. The choice of τ in Eq. (6) which coordinates the identification of principle subspace is also studied. The results indicate that the choices of those hyperparameters do not have a significant influence as long as they are set within reasonable intervals. Compared with them, the choices of λ in Eq. (8) and κ in Eq. (11) demonstrate more importance and are required to be carefully decided.

4.6. Robustness to Support Sample Selection

To evaluate the robustness of our method against the randomness during few-shot sample selection, we run the experiments for 10 times with different sets of labeled samples

Table 3. Ablation study by evaluating the performance gain of adding each key component. We start from the multi-task baseline [23] and add components cumulatively. Experiments are conducted on the nuclei segmentation and recognition benchmark.

Methods	5-shot				10-shot			
	mF1	mF1*	mPQ	mPQ*	mF1	mF1*	mPQ	mPQ*
Baseline	33.85	18.15	18.58	10.77	35.15	21.29	19.14	12.89
+cross-domain	36.56	21.41	20.09	12.41	38.94	23.78	21.93	13.95
+cross-class	36.40	22.84	20.15	13.30	40.52	27.26	22.78	16.13
+historical	37.38	24.90	19.71	13.17	40.06	28.80	22.10	17.31
+projection	**40.26**	**27.14**	**21.78**	**14.96**	**43.88**	**31.43**	**24.81**	**19.35**

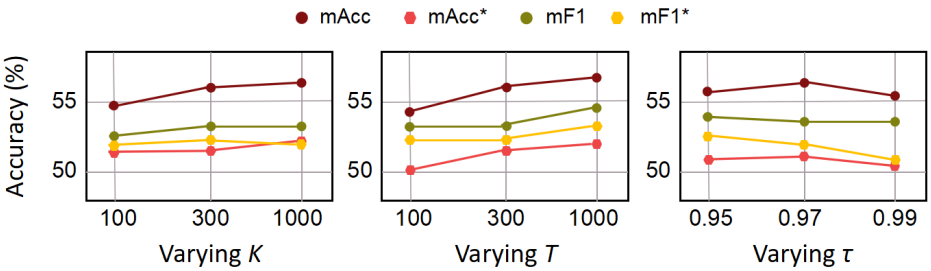

Figure 3. Performance comparison between different choices of hyperparameters K, T and τ on the cancer tissue phenotyping benchmark under 10-shot scheme.

Table 4. Comparisons of our proposed method against other state-of-the-art methods for nuclei segmentation and recognition by averaging the results from 10 random selections of support sets in the target domain. The best results are highlighted in bold.

Methods	5-shot				10-shot				30-shot			
	mF1	mF1*	mPQ	mPQ*	mF1	mF1*	mPQ	mPQ*	mF1	mF1*	mPQ	mPQ*
Multi-task [23]	30.04	16.17	16.51	9.20	33.58	20.73	18.45	12.40	39.52	28.16	22.21	17.25
CGDM [5]	33.30	19.88	19.08	12.46	37.63	26.69	20.61	15.21	42.53	35.11	23.96	20.82
STARTUP [21]	35.91	21.28	19.66	13.53	39.10	24.65	21.77	15.34	45.71	32.38	26.41	20.33
Ours	**37.73**	**25.36**	**20.18**	**14.29**	**44.04**	**32.52**	**24.70**	**19.49**	**49.12**	**40.17**	**28.44**	**25.31**

in the target domain. The averaged results are presented in Table 4. It demonstrates that our method consistently outperforms the competing approaches under diverse settings, which indicates its strong robustness.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

[2] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 865–872, 2019.

[3] Joseph Paul Cohen. Covid-19 image data collection: Prospective predictions are the future. *arXiv:2006.11988*, 2020.

[4] Felix Dangel, Frederik Kunstner, and Philipp Hennig. Backpack: Packing more into backprop. In *International Conference on Learning Representations*, 2020.

[5] Zhekai Du, Jingjing Li, Hongzu Su, Lei Zhu, and Ke Lu. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3937–3946, 2021.

[6] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *European congress on digital pathology*, pages 11–19. Springer, 2019.

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[8] Zhiqiang Gao, Shufei Zhang, Kaizhu Huang, Qiufeng Wang, and Chaoliang Zhong. Gradient distribution alignment certificates better adversarial domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8937–8946, 2021.

[9] Rui Gong, Martin Danelljan, Dengxin Dai, Danda Pani Paudel, Ajad Chhatkuli, Fisher Yu, and Luc Van Gool. Tacs:

Taxonomy adaptive cross-domain semantic segmentation. In *European Conference on Computer Vision*, pages 19–35. Springer, 2022.

[10] Simon Graham, Mostafa Jahanifar, Ayesha Azam, Mohammed Nimir, Yee-Wah Tsang, Katherine Dodd, Emily Hero, Harvir Sahota, Atisha Tank, Ksenija Benes, et al. Lizard: a large-scale dataset for colonic nuclear instance segmentation and classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 684–693, 2021.

[11] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019.

[12] Ashraful Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, and Richard J Radke. Dynamic distillation network for cross-domain few-shot recognition with unlabeled data. *Advances in Neural Information Processing Systems*, 34:3584–3595, 2021.

[13] Sajid Javed, Arif Mahmood, Muhammad Moazam Fraz, Navid Alemi Koohbanani, Ksenija Benes, Yee-Wah Tsang, Katherine Hewitt, David Epstein, David Snead, and Nasir Rajpoot. Cellular community detection for tissue phenotyping in colorectal cancer histology images. *Medical image analysis*, 63:101696, 2020.

[14] Maggie Karthik. Aptos 2019 blindness detection. *Kaggle*.

[15] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6(1):1–11, 2016.

[16] Jogendra Nath Kundu, Rahul Mysore Venkatesh, Naveen Venkat, Ambareesh Revanur, and R Venkatesh Babu. Class-incremental domain adaptation. In *European Conference on Computer Vision*, pages 53–69. Springer, 2020.

[17] Tao Li. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 2019.

[18] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Cross-domain few-shot learning with task-specific adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7161–7170, 2022.

[19] Nina Linder, Juho Konsti, Riku Turkki, Esa Rahtu, Mikael Lundin, Stig Nordling, Caj Haglund, Timo Ahonen, Matti Pietikäinen, and Johan Lundin. Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. *Diagnostic pathology*, 7:1–11, 2012.

[20] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li F Fei-Fei. Label efficient learning of transferable representations acrosss domains and tasks. *Advances in neural information processing systems*, 30, 2017.

[21] Cheng Perng Phoo and Bharath Hariharan. Self-training for few-shot transfer across extreme task differences. In *International Conference on Learning Representations*, 2021.

[22] Pranav Raikote. Covid-19 image dataset. *Kaggle*.

[23] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

[24] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020.

[25] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[26] Hemanth Venkateswara. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017.

[27] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.