This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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
jfan6480@uni.sydney.edu.au dongnan.liu@sydney.edu.au hchang@lbl.gov
henghuanghh@gmail.com Mei.Chen@microsoft.com tom.cai@sydney.edu.au

Abstract

The success of automated medical image analysis depends on large-scale and expert-annotated training sets. Unsupervised domain adaptation (UDA) has been raised as a promising approach to alleviate the burden of labeled data collection. However, they generally operate under the closed-set adaptation setting assuming an identical label set between the source and target domains, which is over-restrictive in clinical practice where new classes commonly exist across datasets due to taxonomic inconsistency. While several methods have been presented to tackle both domain shifts and incoherent label sets, none of them take into account the common characteristics of the two issues and consider the learning dynamics along network training. In this work, we propose optimization trajectory distillation, a unified approach to address the two technical challenges from a new perspective. It exploits the low-rank nature of gradient space and devises a dual-stream distillation algorithm to regularize the learning dynamics of insufficiently annotated domain and classes with the external guidance obtained from reliable sources. Our approach resolves the issue of inadequate navigation along network optimization, which is the major obstacle in the taxonomy adaptive cross-domain adaptation scenario. We evaluate the proposed method extensively on several tasks towards various endpoints with clinical and open-world significance. The results demonstrate its effectiveness and improvements over previous methods. Code is available at https://github.com/camwew/TADA-MI.

1. Introduction

Automated and objective analysis of medical images is an important research topic and has been explored in various clinical applications [57, 13]. To relieve the burden of acquiring massive annotated data for model development on new domains, several unsupervised domain adaptation (UDA) methods [7, 38, 37] are proposed to mitigate the data distribution bias [51] between a richly labeled source domain and a target domain with no explicit supervision.

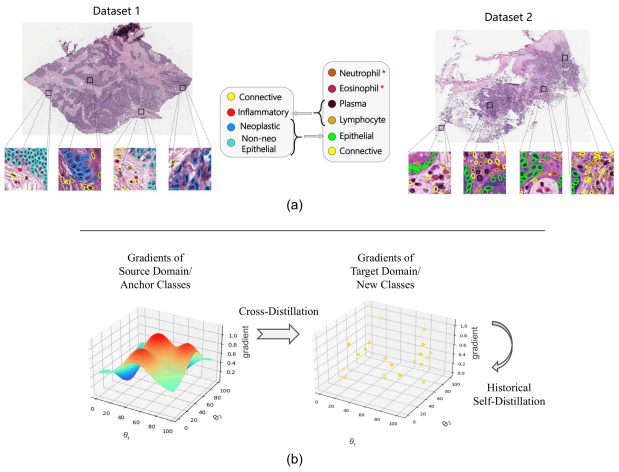

Figure 1. (a) Illustration of the issue of taxonomic inconsistency with nuclei recognition. The category label sets are incoherent across datasets. Different colours indicate the class each nucleus belongs to. The red asterisks denote novel classes only existing in Dataset 2. (b) Concept of the proposed method. We perform optimization trajectory distillation to deliver external navigation for learning target domain and new classes where the optimization steps tend to be restricted and unreliable.

However, these works assume a closed-set adaptation setting that the source and target domains should necessarily share the identical category label space and definition [44]. The restriction limits their practical applicability in *the clinical wild* [15]. Unlike natural images where the definitions of different entity categories (*e.g.*, cat v.s. dog) are structured and globally unified, in the medical domain, inconsistency in taxonomy across different countries or institutes is a common issue compromising the feasibility of cross-domain adaptation [12]. Take nuclei recognition in histology images as an example, the ambiguous biologi-

cal characteristics of cells and distinct clinical usages of the analysis outcomes result in a lack of a unified categorization schema [15, 20]. With specific clinical purposes and downstream applications, different datasets could categorize nuclei with disparate granularities of biological concepts. Besides, as medical research evolves rapidly, novel cell and disease classes could be discovered and form datasets with continually expanding category label sets [41]. This motivates us to develop a generalized adaptation method with the capability to tackle both data distribution bias and category gap for more flexible UDA in clinical practice.

To this end, we study the problem of taxonomy adaptive domain adaptation [19]. It assumes the target domain could adopt a label space different from the source domain. Following UDA, a labeled source dataset and unlabeled samples from the target domain are available during training. Additionally, to recognize the novel/fine-grained categories which are non-existent or unspecified in the source domain, a few samples of those target-private categories are annotated and utilized as exemplars. The technical challenge for this setting lies in how to alleviate domain shifts and concurrently learn new classes with very limited supervision. Recently, several methods are proposed towards a similar goal [32, 22]. However, they typically address the two issues individually in separate contexts and fail to design a unified paradigm according to their common characteristics [19], which could incur a subtle collision across issues since their objectives are non-relevant [60]. Besides, existing works either focus on cross-domain/class alignment in the feature space [40, 36] or resort to model output-based regularization such as self-supervision [45], whereas those approaches suffer from the equilibrium challenge of adversarial learning [1] and the low-quality pseudo-labels [54].

In this work, we present a unified framework for taxonomy adaptive cross-domain adaptation from a new perspective, *i.e.*, via optimization trajectory distillation. Our strategy is motivated by a common challenge existing in both cross-domain and small-data learning regimes, which is the inadequate navigation in the course of network optimization. For cross-domain learning, the unstable feature alignment procedure and noisy pseudo-labels tend to induce error accumulation along network training [62]. Similarly, with limited support samples in new class learning, the optimization of network is inclined to step towards restricted local minima and cannot cover a globally-generalizable space [55]. To this end, we propose to exploit the optimization trajectory from a reliable "teacher" to provide external guidance for regularizing the learning dynamics of insufficiently annotated domain and classes, as illustrated in Fig. 1(b).

Our method consists of two key components, *i.e.,* cross-domain/class distillation and historical self-distillation. **(i)** Cross-domain and cross-class distillation aim to leverage the optimization trajectory from the richly labeled domain and classes to serve as the "teacher". Motivated by the Neural Tangent Kernel (NTK) theory [27], we characterize the network optimization trajectory through gradient statistics. Then, given the observation that the subspace spanned by the gradients from most iterations is generally low-rank [35, 2], we design a gradient projection approach to suppress the noisy signals in stochastic gradients and rectify the distillation process. Thereafter constraints are imposed on the gradient statistics of target domain and new classes to calibrate their training dynamics towards domain-invariant and unbiased learning procedure. **(ii)** Historical self-distillation further drives the optimization paths of model to converge towards flat minima. It is found that the flatness of loss landscapes is strongly related to the model's robustness and generalizability [26], while how to take advantage of the insight to tackle domain shifts and limited supervision is under-explored. We propose to exploit the historical gradients to construct the informative low-rank subspaces and then perform gradient projection to alleviate loss sharpness. It compensates for the intense and out-of-order optimization updates incurred by inadequate regularization and leads to better generalization.

Our prime contributions are as follows: (1) We introduce a more generalized cross-domain adaptation paradigm for medical image analysis in which both data distribution bias and category gap exist across the source and target domains. (2) We leverage insights from recent learning theory research and propose a novel dual-stream optimization trajectory distillation method to provide external navigation in network training. We perform theoretical justifications from two perspectives to illustrate the merits of our method. (3) Experiments on various benchmarks validate the effectiveness and robustness of our proposed method and its improvements over existing approaches.

2. Related Works

Domain Adaptation Domain adaptation (DA) aims to mitigate the data distribution bias [16]. The classic semi-supervised/unsupervised DA focuses on the scenario where the target domain has an identical label space to the source domain [47, 14, 33]. The over-restricted precondition cannot suffice the demand of clinical usages [58]. There exist several efforts to take a step further than the closed-set adaptation, such as open-set DA [44] and universal DA [59]. However, the aim of those works is solely to detect the previously unseen classes, instead of learning to separately identify each new class with few supports. Some recent studies suggest to explicitly recognize the target-private categories. [32] performs feature projection and alignment based on the prototypical networks [50]. [19] combines pseudo-labeling and contrastive learning to combat domain bias and label gap. Nevertheless, those works are limited to tackling the two issues individually, which fails to exploit

their common characteristics and propose a unified solution.

Cross-Domain Few-Shot Learning The technical challenge we need to resolve is similar to an emerging research topic, *i.e.,* cross-domain few-shot learning [22]. The mainstream methods for this challenge include self-training [45, 25, 61], contrastive learning [11, 19], feature alignment/augmentation [53, 8, 24], and channel transformation [34, 39]. Those methods either focus on feature-level modulation or turn to output-level self-supervision and fine-tuning, which could bring instability along optimization [1, 54]. In this work, we propose a novel gradient-based framework to transfer knowledge in both cross-domain and few-shot settings. It implicitly characterizes the information in both feature space and output space.

Gradient Manipulation In current deep learning systems where gradient descent serves as the most popular training algorithm, the directions and amplitudes of network optimization steps are embedded in the gradient space [9]. Gradient manipulation refers to directly modulating the optimization gradients for improved learning process [42]. Recently, it is introduced to mitigate the data distribution shift and shows promises in UDA and domain generalization [14, 17, 49]. However, those works are still restricted by the closed-set condition and do not consider the low-rank nature of gradients [35], which make their approaches vulnerable to noises in the small-data learning regime.

Flatness in Optimization In previous machine learning research, the connection between convergence to flat minima in optimization and model generalizability has long been established [23, 31]. Theoretical and empirical studies prove that the flatness of loss landscapes highly correlates with generalization capability [4]. [5] averages the network weights of multiple checkpoints to prompt generalization across domains. Different from their post-hoc averaging approach operating on fixed models, we propose to regularize the intermediate dynamics of network optimization to achieve flat minima.

3. Method

3.1. Problem Statement

The taxonomy adaptive cross-domain adaptation problem can be formalized as follows. There exists a labeled source dataset $\mathcal{D}_s = \{(\mathbf{x}_s, \mathbf{y}_s)\}$, an unlabeled target dataset $\mathcal{D}_t^u = \{(\mathbf{x}_t^u)\}$, and a few-shot labeled support set from the target domain $\mathcal{D}_t^l = \{(\mathbf{x}_t^l, \mathbf{y}_t^l)\}$. Samples from source and target datasets are drawn from different data distributions, *i.e.*, $\mathbf{x}_s \sim \mathcal{P}_s$, $\mathbf{x}_t \sim \mathcal{P}_t$, $\mathcal{P}_s \neq \mathcal{P}_t$. The label sets of source and target datasets are denoted as $\mathcal{C}_s$ and $\mathcal{C}_t$. Due to the novel/fine-grained categories only existing in the target dataset, we have $\mathcal{C}_s \neq \mathcal{C}_t$. The shared and target-private new classes are indicated as $\overline{\mathcal{C}} = \mathcal{C}_s \cap \mathcal{C}_t$ and $\tilde{\mathcal{C}} = \mathcal{C}_t \backslash \overline{\mathcal{C}}$. We call $\overline{\mathcal{C}}$ existing in both datasets as anchor classes since they can be used as anchors to facilitate knowledge transfer. The goal is to train a model jointly on $\mathcal{D}_s, \mathcal{D}_t^u, \mathcal{D}_t^l$ which performs inferences on data from $\mathcal{P}_t$ with labels in $\mathcal{C}_t$.

3.2. Characterization of Optimization Trajectory

The technical challenge for the identified problem lies in overcoming both domain shifts and overfitting incurred by scarce training samples. As discussed in Section 1, those issues could lead to unreliable navigation during neural network optimization [18, 10]. Therefore, we propose to distill the optimization trajectory knowledge from well-annotated sources to regularize the training dynamics of target domain and new classes.

Recently, NTK theory [27] has been regarded as a tool to dictate the evolution of neural networks. For a network f parameterized by θ, its training dynamics can be described by a kernel K. Given a training set with N samples, the kernel can be defined as an $N \times N$ matrix with $K(i, j; \theta) = \nabla_\theta f(x^i; \theta)^\top \cdot \nabla_\theta f(x^j; \theta)$, where x^i and x^j are two input samples. It indicates the strong correlations between the network's optimization trajectory and the gradients *w.r.t.* its parameters. We thus propose to collect a set of intermediate gradients during training and capture their statistics to characterize the optimization trajectory. Specifically, the second-order Taylor expansion shows that:

$$\begin{aligned} f(x;\theta) =& f(x;\theta^*) + (\theta-\theta^*)^\top \cdot \nabla_\theta f(x;\theta^*) \\ &+\frac{1}{2}(\theta-\theta^*)^\top \cdot \nabla_\theta^2 f(x;\theta^*) \cdot (\theta-\theta^*) \\ &+\mathcal{O}(\|\theta-\theta^*\|^2), \end{aligned} \tag{1}$$

which motivates us to model the first- and second-order derivatives (Hessian) of f to represent its learning dynamics. We use the mean of gradients to indicate the first-order derivative and follow [28] to approximate the second-order derivative with gradient variances.

3.3. Cross-Domain and Cross-Class Distillation

Gradient-based optimization algorithms depend on a large number of update steps supervised by sufficient labeled data [46]. In domain-shifted and small-data regimes, those algorithms suffer from noisy update signals and overfitting, without guarantee to converge towards generalizable minima [22]. To this end, we propose to distill the training dynamics from the label-rich source domain and anchor classes to the target domain and new classes for external regularization by leveraging their intermediate gradients. We firstly compute the gradients $\boldsymbol{g}$ for each domain and class by backpropagating the obtained losses. With a general backbone $\mathbf{F}$ for feature extraction and a task-specific predictor $\mathbf{C}$ for sample-/pixel-wise classification, the individual gradients can be expressed as:

$$\boldsymbol{g} := [\frac{\partial \mathcal{L}(\mathbf{C}(\mathbf{F}(x)), y)}{\partial \theta_{\mathbf{C}_1}}, \cdots, \frac{\partial \mathcal{L}(\mathbf{C}(\mathbf{F}(x)), y)}{\partial \theta_{\mathbf{C}_w}}], \tag{2}$$

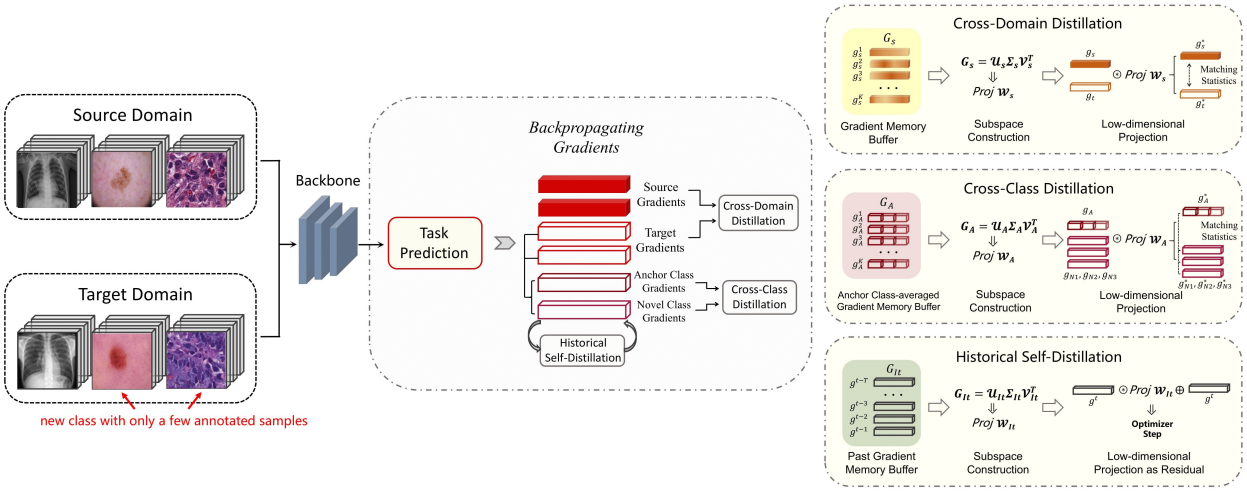

Figure 2. Overview of our proposed method. We devise the cross-domain and cross-class distillation module as well as the historical self-distillation module to transfer the model optimization knowledge from reliable sources by imposing constraints on the gradient descent steps *w.r.t.* the insufficiently supervised target domain and new classes.

where $\mathcal{L}$ is the loss function, w indicates the total number of parameters in $\mathbf{C}$, x and y are training data and the corresponding label. For unlabeled samples $\mathbf{x}_t^u$ in the target dataset, online pseudo-labels $\tilde{\mathbf{y}}_t^u$ are used for loss calculation. Then, as shown in Fig. 2, we aggregate the individual-level gradients in a memory buffer and retrieve the domain- and class-level gradient statistics to model the global first- and second-order derivatives. According to Section 3.2, for gradients *w.r.t.* domain/class π collected from a training set with n_π samples, we derive their mean and variance as:

$$\boldsymbol{g}_\pi^m = \frac{1}{n_\pi}\sum\nolimits_{i=1}^{n_\pi} \boldsymbol{g}_\pi^i, \tag{3}$$

$$\boldsymbol{g}_\pi^v = \frac{1}{n_\pi - 1}\sum\nolimits_{i=1}^{n_\pi} (\boldsymbol{g}_\pi^i - \boldsymbol{g}_\pi^m)^2. \tag{4}$$

Those metrics are measured for each domain and class separately.

However, the gradients obtained from insufficiently annotated domain and classes are dubious and error-prone [55]. To alleviate the issue, we take inspiration from recent learning theory findings that stochastic gradients along the optimization steps are generally low-rank [35], which indicates that the optimization process is governed by the top gradient eigenspace. Specifically, we devise a gradient projection approach to suppress the noisy model update signals in particular to the target domain and new classes by projecting their gradients on the reliable subspaces identified by the source domain and anchor classes with sufficient supervision. Our approach involves three iterative steps, *i.e.*, subspace construction, gradient projection, and gradient statistics matching. Firstly, to identify the principal eigenspace, we perform Singular Value Decomposition (SVD) on the aggregated gradients from the source domain and anchor classes. We denote gradients collected from source domain, target domain, anchor class, and new class as $\boldsymbol{g}_s$, $\boldsymbol{g}_t$, $\boldsymbol{g}_A$, and $\boldsymbol{g}_N$. Given the sets of gradients $\boldsymbol{G}_s = [\boldsymbol{g}_s^1, \boldsymbol{g}_s^2, \cdots, \boldsymbol{g}_s^K]$ and $\boldsymbol{G}_A = [\boldsymbol{g}_A^1, \boldsymbol{g}_A^2, \cdots, \boldsymbol{g}_A^K]$ stored in the memory buffer of volume K, we apply SVD:

$$\boldsymbol{G}_s = \boldsymbol{U}_s \boldsymbol{\Sigma}_s \boldsymbol{V}_s^\top, \qquad \boldsymbol{G}_A = \boldsymbol{U}_A \boldsymbol{\Sigma}_A \boldsymbol{V}_A^\top, \tag{5}$$

where $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{\Sigma}$ contain left singular vectors $\boldsymbol{u}^i$, right singular vector $\boldsymbol{v}^i$, and non-negative singular values $\boldsymbol{\sigma}^i$, respectively. To capture the overall characteristics, we use the gradients averaged along all anchor classes to represent $\boldsymbol{G}_A$. Next, we adopt the low-rank matrix approximation to select the top-r significant left singular vectors in $\boldsymbol{U}_s$ and $\boldsymbol{U}_A$ based on the following criteria:

$$\|(\boldsymbol{G}_s)^{r_s}\|_F^2 \geq \tau \|\boldsymbol{G}_s\|_F^2, \quad \|(\boldsymbol{G}_A)^{r_A}\|_F^2 \geq \tau \|\boldsymbol{G}_A\|_F^2, \tag{6}$$

where $(\boldsymbol{G})^r = \sum_{i=1}^r \boldsymbol{\sigma}^i \boldsymbol{u}^i \boldsymbol{v}^{i\top}$, $\|\cdot\|_F$ denotes the Frobenius norm, τ is a threshold to ensure most information is preserved and is set to 0.98. The principal subspace and the corresponding projection matrix are thereby constructed as $\boldsymbol{M}_s = [\boldsymbol{u}_s^1, \boldsymbol{u}_s^2, \cdots, \boldsymbol{u}_s^{r_s}]$ and $\boldsymbol{M}_A = [\boldsymbol{u}_A^1, \boldsymbol{u}_A^2, \cdots, \boldsymbol{u}_A^{r_A}]$. Afterward, all gradients are projected on the identified subspace as shown in Fig. 2:

$$\begin{aligned} \boldsymbol{g}_s^* &= \boldsymbol{M}_s \boldsymbol{M}_s^\top \boldsymbol{g}_s, \qquad \boldsymbol{g}_t^* = \boldsymbol{M}_s \boldsymbol{M}_s^\top \boldsymbol{g}_t, \\ \boldsymbol{g}_A^* &= \boldsymbol{M}_A \boldsymbol{M}_A^\top \boldsymbol{g}_A, \qquad \boldsymbol{g}_N^* = \boldsymbol{M}_A \boldsymbol{M}_A^\top \boldsymbol{g}_N. \end{aligned} \tag{7}$$

Then following Eq.(3)(4), we minimize the discrepancy between the statistics (*i.e.*, mean and variance) of projected gradients to distill the learning dynamics from the source domain and anchor classes to the target domain and new classes. The overall training objective can be formulated as:

$$\begin{aligned} \min_\theta \ \mathcal{L}_{ERM} + \lambda \Big\{ &\|(\boldsymbol{g}_s^{m*} - \boldsymbol{g}_t^{m*})\|_2^2 + \|(\boldsymbol{g}_s^{v*} - \boldsymbol{g}_t^{v*})\|_2^2 \\ &+ \frac{1}{n_{new}} \sum\nolimits_{i=1}^{n_{new}} \left[\|(\boldsymbol{g}_A^{m*} - \boldsymbol{g}_{Ni}^{m*})\|_2^2 + \|(\boldsymbol{g}_A^{v*} - \boldsymbol{g}_{Ni}^{v*})\|_2^2 \right] \Big\}. \end{aligned} \tag{8}$$

Here $\mathcal{L}_{ERM}$ denotes the empirical risk loss, which is implemented with cross-entropy loss for classification and Dice loss [43] for segmentation. λ is a balancing coefficient, n_{new} is the number of new classes in the target domain. It is noted that there exists discrepancy between the anchor and new classes in semantic concepts, which makes feature alignment across classes harmful due to negative transfer [56]. However, aligning class-wise learning dynamics by enforcing the similarity between their gradients could contribute to lower empirical error on new classes. We prove this property theoretically in the supplementary material.

3.4. Historical Self-distillation

Inferior generalization behaviour is an outstanding challenge for taxonomy adaptive cross-domain adaptation [19]. When domain shifts and unseen classes exist, a trained model could only maintain its performance on test samples sharing multiple similar attributes with the training data but cannot generalize well to the disparate ones [63]. Motivated by the connection between flat minima and generalizability [31], we devise the historical self-distillation module to smooth the optimization path towards a well-generalizable solution by calibrating gradient descent steps.

To reach flat loss landscapes [6], assuming a local loss function ψ and a network f parameterized by θ, instead of minimizing the original loss $\psi(f(x;\theta))$ only, the optimization objective should be:

$$\min_{\theta} -\log \int_{\|\theta-\theta^*\|\leq\Gamma} \exp(-\psi(f(x;\theta^*))-\gamma\,\|\theta-\theta^*\|)\,d\theta^*, \tag{9}$$

where x indicates random input samples, Γ is defined as the width of a local parameter valley, γ is a balancing weight. In other words, when θ and θ^* are neighbouring, the following criteria is required to be satisfied:

$$\|\nabla_\theta\psi(f(x;\theta)) - \nabla_\theta\psi(f(x;\theta^*))\| \leq \beta\,\|\theta-\theta^*\|\,, \tag{10}$$

where β is a smoothness factor. To this end, we propose to impose regulations on the optimizer steps by enforcing minima with uniformly-distributed gradients. Given that the gradients from most iterations are generally low-rank [35, 2], we exploit past gradients to identify the low-dimensional subspace which indicates the principal gradient distribution of local minima and then project current gradients on the constructed subspace to exclude sharp and noisy signals. Specifically, we collect and store individual gradients for each iteration along the training procedure, denoted as $\boldsymbol{G}_{It} = [\boldsymbol{g}_{It}^{t-T}, \boldsymbol{g}_{It}^{t-T+1}, \cdots, \boldsymbol{g}_{It}^{t-2}, \boldsymbol{g}_{It}^{t-1}]$, where t is the index of current iteration, T is the size of memory buffer. Thereafter we perform SVD on the set of gradients $\boldsymbol{G}_{It}$ to construct the historical subspace. Similar to Eq.(5)(6)(7), the top-r left singular vectors of the decomposed gradient set are concatenated to formulate the projection matrix $\boldsymbol{M}_{It}$. Then for upcoming gradients obtained from back-propagation in each iteration, we project the gradients on the identified low-rank subspace and consider them as residuals. The subsequent optimizer step is conducted with the sum of the original and rectified gradients, as shown in Fig. 2. Take vanilla stochastic gradient descent as an example, with mini-batch gradients $\tilde{\boldsymbol{g}}$ in each iteration, the parameters θ of the optimized network are updated by:

$$\theta := \theta - \frac{1}{\kappa}\cdot\eta\cdot(\boldsymbol{M}_{It}\,\boldsymbol{M}_{It}^{\top}+\kappa)\cdot\tilde{\boldsymbol{g}}, \tag{11}$$

where η is the learning rate, κ is a trade-off parameter. As training proceeds, the principal subspace and the corresponding projection matrix are periodically renewed by applying SVD on the recently collected gradients. The overall training procedure is summarized in the supplementary material.

3.5. Theoretical Analysis

In this section, we motivate our proposed method theoretically and provide mathematical insights on its merits.

Joint Characterization of Feature and Output Space While our optimization trajectory distillation approach proposes a novel perspective that differs from existing works imposing regularization terms in feature and model-output space [32, 19], those information is still captured and implicitly modeled in the gradient-based framework. We prove this property for the classification and segmentation tasks under two commonly adopted loss functions. The detailed proof is presented in the supplemental material. It illustrates the superiority of our method as a unified framework that jointly characterizes the feature and output space as well as the learning dynamics.

Impacts on Generalization Error In our method, we propose to minimize the discrepancy between the gradient statistics from different domains and classes in the identified principal subspace. We hereby prove its effectiveness towards a tighter generalization error bound on the target domain and new classes [3]. Details can be found in the supplemental material.

4. Experiments

4.1. Datasets and Experiment Settings

To evaluate the effectiveness of our method and its potential clinical implications, we conduct experiments on five medical image datasets in regard to three important downstream tasks from different clinical scenarios. Extended experiments on more diverse tasks, including radiology and fundus analysis, as well as general visual task, are presented in the supplemental material to substantiate the broader applicability of our method.

Table 1. Comparison results of our proposed method against other state-of-the-art methods for nuclei segmentation and recognition. The metrics with asterisk (*) are calculated on target-private new classes only. The subscripts denote the standard deviations. The best and second-best results are highlighted in bold and brown, respectively.

Methods	5-shot				10-shot			
	mF1	mF1*	mPQ	mPQ*	mF1	mF1*	mPQ	mPQ*
Sup-only	$28.08_{0.89}$	$21.46_{0.36}$	$12.35_{0.53}$	$12.23_{0.45}$	$30.90_{0.66}$	$25.23_{0.83}$	$14.18_{0.24}$	$14.94_{0.32}$
Multi-task [48]	$33.85_{1.26}$	$18.15_{1.08}$	$18.58_{0.42}$	$10.77_{0.35}$	$35.15_{0.94}$	$21.29_{0.62}$	$19.14_{0.27}$	$12.89_{0.34}$
DANN [16]	$32.76_{1.41}$	$16.52_{1.10}$	$18.31_{0.66}$	$10.40_{0.59}$	$35.76_{1.35}$	$23.12_{0.88}$	$19.70_{0.60}$	$14.14_{0.42}$
CGDM [14]	$35.03_{1.28}$	$21.04_{0.99}$	$19.59_{0.50}$	$13.22_{0.32}$	$37.13_{1.57}$	$24.22_{1.15}$	$19.78_{0.71}$	$13.90_{0.40}$
LETR [40]	$27.78_{1.44}$	$15.36_{1.02}$	$16.68_{0.54}$	$10.03_{0.51}$	$34.85_{1.54}$	$20.03_{0.73}$	$18.18_{0.59}$	$11.79_{0.27}$
FT-CIDA [32]	$30.51_{0.92}$	$16.55_{0.80}$	$15.97_{0.39}$	$9.42_{0.40}$	$32.63_{1.21}$	$21.08_{0.95}$	$17.09_{0.47}$	$11.96_{0.49}$
STARTUP [45]	$37.03_{0.58}$	$22.44_{0.76}$	$20.40_{0.33}$	$13.82_{0.30}$	$40.85_{0.50}$	$25.92_{0.91}$	$23.37_{0.19}$	$16.64_{0.24}$
DDN [25]	$37.92_{0.86}$	$23.71_{1.05}$	$20.62_{0.57}$	$13.87_{0.43}$	$40.16_{0.72}$	$26.94_{1.08}$	$22.79_{0.32}$	$16.82_{0.50}$
TSA [34]	$34.03_{1.17}$	$19.55_{1.11}$	$18.65_{0.43}$	$12.06_{0.37}$	$34.27_{1.05}$	$22.58_{1.20}$	$19.78_{0.61}$	$14.24_{0.73}$
TACS [19]	$36.73_{1.65}$	$22.26_{1.29}$	$18.13_{0.80}$	$12.57_{0.62}$	$39.29_{1.48}$	$25.84_{1.13}$	$22.27_{0.68}$	$15.98_{0.56}$
Ours	$\mathbf{40.26}_{0.90}$	$\mathbf{27.14}_{0.97}$	$\mathbf{21.78}_{0.49}$	$\mathbf{14.96}_{0.22}$	$\mathbf{43.88}_{0.52}$	$\mathbf{31.43}_{0.78}$	$\mathbf{24.81}_{0.26}$	$\mathbf{19.35}_{0.33}$

Nuclei Segmentation and Recognition Technically, this task can be formalized as a simultaneous instance segmentation and classification problem [21]. We evaluate our method by considering PanNuke [15] and Lizard [20] as the source and target datasets. They contain 481 and 291 visual fields cropped from whole-slide images with 189,744 and 495,179 annotated nuclei, respectively. There are six types of nuclei in Lizard and only two of them overlap with the five classes in PanNuke, which suggests four new classes in the target domain. We employ the widely used HoverNet [21] architecture with a standard ResNet-50 backbone as the base model. For quantitative evaluation, we follow [20] and report the average F1 and PQ (panoptic quality) scores for all classes (denoted as mF1 and mPQ) to provide a comprehensive assessment measuring the accuracy of nuclei detection, segmentation, and classification. Additionally, we also report the mF1* and mPQ* scores which are solely computed on new classes in the target domain.

Cancer Tissue Phenotyping In this setting, we perform adaptation from a two-class discrimination task to an eight-class one on CRC-TP [29] and Kather [30] datasets. Kather consists of 5000 150×150 pathology image patches and includes six novel classes other than the two common classes also existing in the source domain. The experiments are conducted with ResNet-101 as backbone. For evaluation, we exploit the accuracy and F1 scores as metrics to indicate the classification performance. The average scores for all classes (mAcc and mF1) and particularly the ones on new classes only (mAcc* and mF1*) are reported.

Skin Lesion Diagnosis We construct a fine-grained taxonomy adaptation setting from two coarse classes to seven subclasses with HAM10000 [52], which contains 10015 dermatoscopic images sampled from different body structures. The adopted experiment settings and evaluation metrics are the same as the cancer tissue phenotyping task. Since the label space of the source and target datasets do not overlap, we only report the mAcc* and mF1* scores which are computed on target classes.

Following previous works in UDA [7], for all experiments, each dataset is randomly split into $80\%/20\%$ for training/testing. To explicitly recognize the target-private new classes, we sample few (5/10) samples with corresponding labels to formulate the support set. The remaining target data is left unlabeled. More details can be found in the supplemental material.

4.2. Results

We compare our method with state-of-the-art approaches for cross-domain adaptation, including UDA methods DANN [16] and CGDM [14] which do not consider the taxonomic inconsistency, as well as LETR [40], FT-CIDA[32], STARTUP [45], DDN [25], TSA [34], and TACS [19] which are designed for both domain shifts and new class learning. For UDA methods, we jointly perform domain alignment and train the model with the labeled target data. In addition, we conduct comparisons with multi-task learning approach [48], which trains the model on all domains and label sets simultaneously. "Sup-only" indicates the baseline where only the supervised loss on the labeled target dataset is used for training.

Table 1 shows the comparison results on the cross-domain nuclei recognition and segmentation task. All methods adopt the same ResNet-50-HoverNet as base model following [21]. It can be observed that our proposed method outperforms the state-of-the-art cross-domain adaptation approaches by a large margin. We also notice an important point that compared with the "Sup-only" baseline, most competing methods fail to attain better performance on the metrics specific to new classes (*i.e.,* mF1* and mPQ*). For example, under the 5-shot setting, despite a 4.68% improvement is achieved by DANN on mF1, it achieves inferior performance on mF1* which is 4.94% worse than "Sup-only".

Table 2. Comparison results of our proposed method against other state-of-the-art methods for cancer tissue phenotyping.

Methods	5-shot				10-shot			
	mAcc	mAcc*	mF1	mF1*	mAcc	mAcc*	mF1	mF1*
Sup-only	$46.82_{1.30}$	$42.26_{1.26}$	$44.64_{1.49}$	$42.83_{1.08}$	$48.52_{1.02}$	$45.93_{1.67}$	$45.34_{1.45}$	$46.15_{1.83}$
Multi-task [48]	$48.35_{0.78}$	$42.02_{1.34}$	$45.55_{1.12}$	$41.68_{1.63}$	$48.56_{0.69}$	$45.00_{1.72}$	$46.03_{0.64}$	$44.46_{1.21}$
DANN [16]	$47.46_{0.67}$	$42.53_{1.22}$	$43.08_{1.14}$	$43.13_{1.30}$	$48.42_{0.48}$	$45.31_{1.27}$	$46.50_{0.81}$	$44.02_{0.64}$
CGDM [14]	$49.02_{0.82}$	$43.53_{1.56}$	$46.29_{0.87}$	$43.58_{0.69}$	$50.54_{0.84}$	$44.57_{1.35}$	$48.17_{0.66}$	$45.23_{0.40}$
LETR [40]	$45.47_{1.95}$	$41.33_{1.47}$	$43.17_{1.88}$	$41.20_{2.02}$	$45.94_{0.56}$	$40.13_{1.30}$	$41.16_{1.13}$	$39.61_{0.90}$
FT-CIDA [32]	$43.72_{0.88}$	$39.06_{0.79}$	$40.91_{1.02}$	$38.73_{0.84}$	$44.65_{0.53}$	$38.80_{0.94}$	$40.94_{0.96}$	$38.96_{1.07}$
STARTUP [45]	$49.52_{0.63}$	$42.93_{1.83}$	$43.59_{1.37}$	$39.97_{0.59}$	$49.40_{0.51}$	$46.91_{1.87}$	$46.38_{0.70}$	$47.85_{1.36}$
DDN [25]	$47.14_{1.75}$	$43.25_{0.99}$	$42.92_{2.29}$	$40.76_{0.50}$	$48.27_{0.67}$	$44.08_{1.22}$	$46.03_{1.14}$	$45.44_{1.18}$
TSA [34]	$48.22_{1.49}$	$42.13_{0.78}$	$44.39_{1.84}$	$41.92_{1.13}$	$47.18_{1.02}$	$41.26_{1.10}$	$45.34_{1.72}$	$44.69_{1.89}$
TACS [19]	$45.30_{1.07}$	$40.28_{1.11}$	$44.74_{1.26}$	$42.04_{0.95}$	$46.95_{1.16}$	$43.69_{1.74}$	$47.46_{1.51}$	$47.51_{1.25}$
Ours	$\mathbf{50.14}_{0.76}$	$\mathbf{47.66}_{1.54}$	$\mathbf{47.41}_{1.10}$	$\mathbf{46.45}_{1.15}$	$\mathbf{55.86}_{0.98}$	$\mathbf{51.27}_{1.02}$	$\mathbf{52.98}_{0.49}$	$\mathbf{52.12}_{0.94}$

Table 3. Comparison results of our proposed method against other state-of-the-art methods for skin lesion diagnosis.

Methods	5-shot		10-shot	
	mAcc*	mF1*	mAcc*	mF1*
Sup-only	$35.71_{0.79}$	$19.89_{0.22}$	$38.50_{0.38}$	$22.36_{0.67}$
Multi-task [48]	$33.80_{1.20}$	$19.06_{0.31}$	$37.85_{0.40}$	$23.32_{0.32}$
DANN [16]	$37.22_{0.57}$	$23.99_{0.92}$	$40.06_{0.94}$	$24.32_{0.72}$
CGDM [14]	$41.50_{1.03}$	$24.26_{0.80}$	$47.22_{0.70}$	$24.00_{0.23}$
LETR [40]	$36.04_{0.42}$	$21.39_{0.66}$	$40.87_{1.22}$	$23.50_{0.49}$
FT-CIDA [32]	$31.95_{0.77}$	$18.13_{0.48}$	$34.60_{0.90}$	$20.96_{0.88}$
STARTUP [45]	$43.21_{1.15}$	$27.17_{0.92}$	$43.88_{1.35}$	$25.59_{0.60}$
DDN [25]	$36.19_{1.05}$	$21.50_{0.34}$	$42.46_{1.29}$	$23.92_{0.28}$
TSA [34]	$40.41_{1.36}$	$24.05_{1.01}$	$42.07_{1.51}$	$25.44_{1.06}$
TACS [19]	$45.05_{0.46}$	$25.60_{0.65}$	$46.73_{0.64}$	$25.96_{0.79}$
Ours	$\mathbf{50.79}_{0.69}$	$\mathbf{30.05}_{0.81}$	$\mathbf{52.14}_{0.98}$	$\mathbf{30.52}_{0.83}$

This phenomenon is incurred by the failure of cross-class generalization. In contrast, our method yields substantial improvements on those new class-specific metrics, which indicates the superior generalizability of our method. The observation is further verified through visual comparisons in the supplementary material.

The overall quantitative results on the cancer tissue phenotyping and skin lesion diagnosis tasks are presented in Table 2 and 3, respectively. All methods are implemented based on the ResNet-101 backbone. Under each few-shot setting, our method shows higher classification scores compared with previous approaches. In particular, we achieve 0.62%∼5.32% and 1.12%∼4.81% improvements against the second-best results in terms of mAcc and mF1 for cancer tissue phenotyping, as well as 4.92%∼7.58% and 2.88%∼4.93% improvements in terms of mAcc* and mF1* for skin lesion diagnosis. The higher accuracy consistently attained by our method under different cross-domain adaptation settings and clinical tasks is attributed to the strong generalizability brought by the proposed optimization trajectory distillation approach.

4.3. Analysis and Discussion

Ablation of Key Components Table 4 shows the ablation study results to illustrate the effectiveness of key components. "projection" denotes the gradient projection approach introduced in Section 3.3. For the "*w/o* projection" ablation, we perform cross-domain and cross-class distillation by directly minimizing the statistics discrepancy between unrectified gradients (g_s, g_t) and (g_A, g_N) in Eq. (7). It is observed that each component is indispensable and could improve the overall performance, which elaborates the importance of external navigation for learning target domain and new classes in network training. Interestingly, we notice that the impacts of cross-domain and cross-class distillation could be disparate across tasks. For example, cross-class distillation brings higher performance gain compared with cross-domain distillation on the cancer tissue phenotyping benchmark, while the situation is reversed for skin lesion diagnosis. It is caused by the different natures of each task. Class-imbalance is a serious issue for skin lesion diagnosis in that the class distribution is highly biased [52], which compromises the effectiveness of the cross-class distillation module. Differently, for cancer tissue phenotyping, classes are uniformly distributed [30] and could hence ensure the representativity of anchor classes and their exemplarity to be distilled to new classes.

Hyperparameter Sensitivity To investigate the effect of hyperparameters λ in Eq. (8) and κ in Eq. (11), we present the performance comparison by varying the choices of their values for cancer tissue phenotyping, as shown in Fig. 3. We only change one hyperparameter while keeping the other fixed at a time. The experimental results show that our framework is quite stable for hyperparameters in a wide interval, yet setting λ and κ to a large value is detrimental to the classification accuracy. Specifically, λ is a balancing coefficient between class discriminativeness and gradient alignment. When it is too large, the trained model is biased to matching learning dynamics and would lose es-

Table 4. Key component analysis of our method on three cross-domain adaptation benchmarks under 10-shot scheme. The best results are highlighted in bold.

Methods	Nuclei				Cancer				Skin	
	mF1	mF1*	mPQ	mPQ*	mAcc	mAcc*	mF1	mF1*	mAcc*	mF1*
Full Framework	**43.88**	**31.43**	**24.81**	**19.35**	**55.86**	**51.27**	**52.98**	**52.12**	**52.14**	**30.52**
w/o cross-domain	39.96	30.08	22.29	18.10	52.81	49.93	51.62	50.64	44.34	25.28
w/o cross-class	42.30	27.82	23.76	17.05	48.90	47.22	49.85	44.76	49.06	28.24
w/o historical	41.54	30.47	23.00	17.59	50.17	48.05	48.19	45.12	45.83	25.97
w/o projection	40.06	28.80	22.10	17.31	53.58	45.78	49.01	49.20	47.29	28.02

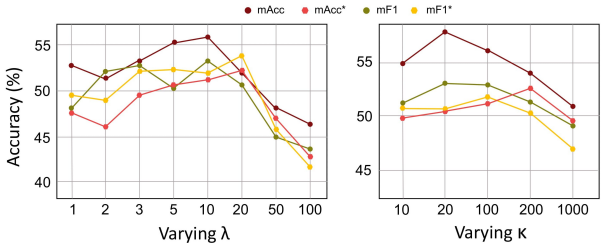

Figure 3. Performance comparison between different choices of hyperparameters λ and κ on the cancer tissue phenotyping benchmark under 10-shot scheme.

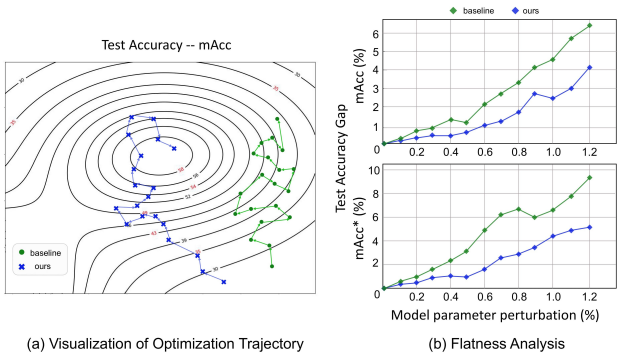

Figure 4. Further analysis of our proposed method. (a) We plot the optimization trajectory of our method and the multi-task baseline on the test accuracy surface. (b) We visualize the local flatness of our method and the baseline via test accuracy gap by varying the model parameter perturbation rate. Each point is computed by Monte-Carlo approximation over 50 random samples. All the experiments are conducted on the cancer tissue phenotyping benchmark under 10-shot scheme.

sential semantic knowledge. For κ, it controls the proportions of unrectified gradients. Its large value indicates that most original gradients are preserved, which would incur noisy optimization steps and loss of flatness. Therefore, we decide to set $\lambda = 10$ and $\kappa = 100$. It achieves the best performance under most metrics.

Impacts on Optimization Trajectory and Model Flatness To better understand the effectiveness of our method, we perform in-depth analysis about its impacts on model's training trajectory and flatness of the converged local minima. Firstly, as shown in Fig. 4(a), we visualize the optimization paths of our method and the multi-task baseline on the test accuracy surface with a contour plot. The details of plot procedure can refer to [26]. It can be observed that the optimization of the multi-task baseline fluctuates sharply and is restricted in a sub-optimal area with a relatively low test accuracy. This is induced by the biased navigation in the course of training due to domain shifts and category gap. With the devised dual-stream optimization trajectory distillation module, our method could by contrast suppress the noisy update signals and proceed consistently towards the optimum.

Besides, we perform analysis about the solutions found by our method and the baseline in terms of flatness [5]. Specifically, the local flatness of a model parameter θ is quantified by the expected changes of test accuracy between θ and a neighbouring θ^* with perturbation rate ϱ: $\mathcal{F}_\varrho(\theta) = \mathbb{E}_{\|\theta^*\|=(1+\varrho)\|\theta\|}\big[\mathrm{Acc}(\theta^*) - \mathrm{Acc}(\theta)\big]$. We approximate the expected value by Monte-Carlo sampling over 50 samples. As demonstrated in Fig. 4(b), comparing with the baseline approach, our method shows stronger robustness against model parameter perturbation. It proves that our method successfully converges to minima with better local flatness and thus possesses superior generalizability [31].

5. Conclusion

In this paper, we study the taxonomy adaptive cross-domain adaptation paradigm, which allows incoherent label sets between source and target datasets. To jointly address the data distribution bias and category gap, we propose a unified framework which performs cross-domain and cross-class optimization trajectory distillation to calibrate the learning dynamics of insufficiently annotated domain and classes. Historical self-distillation is further devised to attain a well-generalizable solution via regulating gradient distributions. Extensive experiments on multiple benchmarks with different task contexts demonstrate the effectiveness and generalization capability of our method. In the future, we will extend this work to further discover the underlying mechanism of optimization trajectory distillation for model generalization and evaluate its potential in more adaptation scenarios.

References

[1] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*, pages 224–232. PMLR, 2017.

[2] Sheikh Shams Azam, Seyyedali Hosseinalipour, Qiang Qiu, and Christopher Brinton. Recycling model updates in federated learning: Are gradient subspaces low-rank? In *International Conference on Learning Representations*, 2022.

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

[4] Debora Caldarola, Barbara Caputo, and Marco Ciccone. Improving generalization in federated learning by seeking flat minima. In *European Conference on Computer Vision*, pages 654–672. Springer, 2022.

[5] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

[6] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations*, 2017.

[7] 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.

[8] Wentao Chen, Zhang Zhang, Wei Wang, Liang Wang, Zilei Wang, and Tieniu Tan. Cross-domain cross-set few-shot learning via learning compact and aligned representations. In *European Conference on Computer Vision*, pages 383–399. Springer, 2022.

[9] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018.

[10] Ali Cheraghian, Shafin Rahman, Sameera Ramasinghe, Pengfei Fang, Christian Simon, Lars Petersson, and Mehrtash Harandi. Synthesized feature based few-shot class-incremental learning on a mixture of subspaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8661–8670, 2021.

[11] Debasmit Das, Sungrack Yun, and Fatih Porikli. ConfeSS: A framework for single source cross-domain few-shot learning. In *International Conference on Learning Representations*, 2022.

[12] Susanna Dodd, Mike Clarke, Lorne Becker, Chris Mavergames, Rebecca Fish, and Paula R Williamson. A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery. *Journal of clinical epidemiology*, 96:84–92, 2018.

[13] C Domínguez Conde, C Xu, LB Jarvis, DB Rainbow, SB Wells, T Gomes, SK Howlett, O Suchanek, K Polanski, HW King, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594), 2022.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *International Conference on Learning Representations*, 2020.

[19] 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.

[20] 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.

[21] 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.

[22] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European conference on computer vision*, pages 124–141. Springer, 2020.

[23] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[24] Yanxu Hu and Andy J Ma. Adversarial feature augmentation for cross-domain few-shot classification. In *European Conference on Computer Vision*, pages 20–37. Springer, 2022.

[25] Ashraful Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, and Richard J Radke. Dy-

namic distillation network for cross-domain few-shot recognition with unlabeled data. *Advances in Neural Information Processing Systems*, 34:3584–3595, 2021.

[26] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2018.

[27] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[28] Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

[29] 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.

[30] 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.

[31] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

[32] 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.

[33] Canran Li, Dongnan Liu, Haoran Li, Zheng Zhang, Guangming Lu, Xiaojun Chang, and Weidong Cai. Domain adaptive nuclei instance segmentation and classification via category-aware feature alignment and pseudo-labelling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 715–724. Springer, 2022.

[34] 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.

[35] Xinyan Li, Qilong Gu, Yingxue Zhou, Tiancong Chen, and Arindam Banerjee. Hessian based analysis of sgd for deep nets: Dynamics and generalization. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 190–198. SIAM, 2020.

[36] Hanwen Liang, Qiong Zhang, Peng Dai, and Juwei Lu. Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9424–9434, 2021.

[37] Dongnan Liu, Chaoyi Zhang, Yang Song, Heng Huang, Chenyu Wang, Michael Barnett, and Weidong Cai. Decompose to adapt: Cross-domain object detection via feature disentanglement. *IEEE Transactions on Multimedia*, 25:1333–1344, 2022.

[38] Dongnan Liu, Donghao Zhang, Yang Song, Fan Zhang, Lauren O'Donnell, Heng Huang, Mei Chen, and Weidong Cai. Unsupervised instance segmentation in microscopy images via panoptic domain adaptation and task re-weighting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4243–4252, 2020.

[39] Xu Luo, Jing Xu, and Zenglin Xu. Channel importance matters in few-shot image classification. In *International Conference on Machine Learning*, pages 14542–14559. PMLR, 2022.

[40] 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.

[41] Deborah Lupton. *Medicine as culture: illness, disease and the body*. Sage, 2012.

[42] Lucas Mansilla, Rodrigo Echeveste, Diego H Milone, and Enzo Ferrante. Domain generalization via gradient surgery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6630–6638, 2021.

[43] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[44] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 754–763, 2017.

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

[46] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2017.

[47] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019.

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

[49] Yuge Shi, Jeffrey Seely, Philip Torr, Siddharth N, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *International Conference on Learning Representations*, 2022.

[50] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[51] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[52] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source der-

matoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[53] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *International Conference on Learning Representations*, 2020.

[54] Qian Wang and Toby Breckon. Unsupervised domain adaptation via structured prediction based selective pseudo-labeling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6243–6250, 2020.

[55] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[56] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11293–11302, 2019.

[57] Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155(10):1135–1141, 2019.

[58] Mengya Xu, Mobarakol Islam, Chwee Ming Lim, and Hongliang Ren. Class-incremental domain adaptation with smoothing and calibration for surgical report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 269–278. Springer, 2021.

[59] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2720–2729, 2019.

[60] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020.

[61] Wang Yuan, Zhizhong Zhang, Cong Wang, Haichuan Song, Yuan Xie, and Lizhuang Ma. Task-level self-supervision for cross-domain few-shot learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):3215–3223, 2022.

[62] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *Advances in neural information processing systems*, 32, 2019.

[63] Lin Zuo, Mengmeng Jing, Jingjing Li, Lei Zhu, Ke Lu, and Yang Yang. Challenging tough samples in unsupervised domain adaptation. *Pattern Recognition*, 110:107540, 2021.