Prototypical Cross-domain Knowledge Transfer for Cervical Dysplasia Visual Inspection

Yichen Zhang
National University of Singapore
zhang.yichen@u.nus.edu

Yifang Yin
Institute for Infocomm Research, A*STAR
yin_yifang@i2r.a-star.edu.sg

Ying Zhang
Northwestern Polytechnical University
izhangying@nwpu.edu.cn

Zhenguang Liu*
Zhejiang Gongshang University
liuzhenguang2008@gmail.com

Zheng Wang
Wuhan University
wangzwhu@whu.edu.cn

Roger Zimmermann
National University of Singapore
rogerz@comp.nus.edu.sg

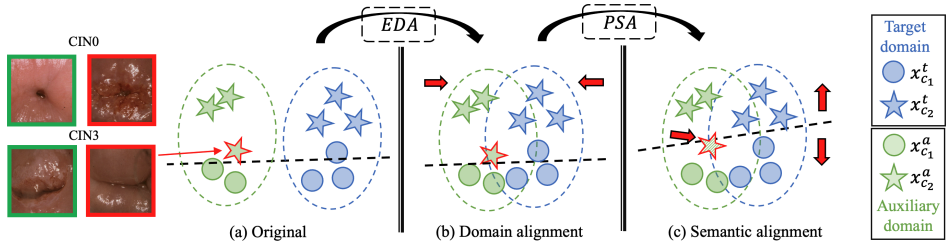


Figure 1: Illustration of our proposed method. (a) In the original feature space, direct supervised learning with auxiliary samples may degrade the model's performance in the target domain. We thus propose (b) an Early Domain Alignment (EDA) module to reduce the domain gap, and (c) a Prototypical Semantic Alignment (PSA) module to identify auxiliary samples with high-uncertainty labels (*i.e.*, red border) and reduce their impact when aligning the representations at the semantic level.

ABSTRACT

Early detection of dysplasia of the cervix is critical for cervical cancer treatment. However, automatic cervical dysplasia diagnosis via visual inspection, which is more appropriate in low-resource settings, remains a challenging problem. Though promising results have been obtained by recent deep learning models, their performance is significantly hindered by the limited scale of the available cervix datasets. Distinct from previous methods that learn from a single dataset, we propose to leverage cross-domain cervical images that were collected in different but related clinical studies to improve the model's performance on the targeted cervix dataset. To robustly learn the transferable information across datasets, we propose a novel prototype-based knowledge filtering method to estimate the transferability of cross-domain samples. We further optimize the shared feature space by aligning the cross-domain image representations simultaneously on *domain level* with early alignment and *class level* with supervised contrastive learning, which endows model training and knowledge transfer with stronger robustness. The empirical results on three real-world benchmark cervical image datasets show that our proposed method outperforms the state-of-the-art cervical dysplasia visual inspection by an absolute improvement of 4.7% in top-1 accuracy, 7.0% in precision, 1.4% in recall, 4.6% in F1 score, and 0.05 in ROC-AUC.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Cervical dysplasia visual inspection, medical image processing, colposcopic image, cross-domain learning, contrastive learning

ACM Reference Format:
Yichen Zhang, Yifang Yin, Ying Zhang, Zhenguang Liu, Zheng Wang, and Roger Zimmermann. 2023. Prototypical Cross-domain Knowledge Transfer for Cervical Dysplasia Visual Inspection. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3581783.3612000

1 INTRODUCTION

Cervical cancer is one of the most common cancers for women [59], posing serious risks to their health and spreading through direct or distant metastasis [28]. Especially in developing countries, it is the second most prevalent malignancy after breast cancer and the third dominant cause of cancer-related deaths [32], despite the fact that it is one of the most successfully treatable forms of cancer if diagnosed in an early stage [21]. Cervical dysplasia, also known as cervical intraepithelial neoplasia (CIN), is a precancerous change

*The corresponding author

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612000

indicating potential cervical cancer in an early stage. Although it can be detected via a few screening methods, most of them are conducted in a laboratory setting where special infrastructure and extensively trained personnel are needed. Such constraints significantly limit their wide deployment in low-resource regions. To accommodate the medical needs, visual inspection of the cervix after applying 5% acetic acid to the cervix epithelium (a method known in the medical community as VIA) has been advocated by the WHO because of its simplicity and low cost. In this paper, we focus on improving the performance of computational visual inspection to assist in faster and more accurate inspection. Note that colposcopic photographs from the VIA approach are referred to as cervical images in the rest of the paper.

Despite deep neural networks having been widely adopted in computer vision, attaining state-of-the-art performance usually requires vast quantities of labeled data. Unlike natural images, medical image acquisition, annotation, and analysis require significant efforts of human expertise [33] and are traditionally part of localized medical studies. Existing methods mostly perform transfer learning based on models pre-trained on natural images, particularly ImageNet [15], to alleviate this situation. While this may work well in some general instances, recent research shows that such task-agnostic transfer learning alone does not necessarily result in performance improvements for medical applications, due to the considerable visual differences between natural and medical images [48]. A dearth of large task-specific datasets still stands in the way of achieving outstanding model performance.

The above findings motivate us to look for new auxiliary data sources to facilitate medical image analysis. In the field of cervical dysplasia visual inspection, we observe that multiple image datasets exist (*e.g.*, NHS [26] and ALTS [22]), which are relevant but differ significantly in their collection environment. Intuitively, the knowledge learned from one dataset (*e.g.*, ALTS) will be helpful to improve the robustness of a model trained on another dataset (*e.g.*, NHS), which is, however, ignored by previous methods. We also observe that a direct utilization of existing domain adaptation/generalization methods performs unsatisfactorily due to not only (1) *domain shift* — datasets are collected using different devices in different environments; but also (2) *criterion mismatch* — the standards for ground-truth annotation can be different due to the subjective variance — the diagnosis was made by a single medical staff (*e.g.*, nurse, doctor) purely based on visual inspection without confirming laboratory tests.

To tackle the above challenges, we present the first prototypical cross-domain knowledge transfer framework for cervical dysplasia visual inspection, which learns transferable information from an auxiliary dataset to improve the performance on the target dataset. As illustrated in Figure 1, the framework has an edge in conducting simultaneous feature alignment under two distinct levels: domain level and class level. The *Early Domain Alignment* (*EDA*) module is presented to generate domain-aligned intermediate features, followed by the *Prototypical Semantic Alignment* (*PSA*) module producing semantically-consistent high-level representations across domains. Moreover, *PSA* tackles the *criterion mismatch* challenge by identifying and reducing the impact of the auxiliary samples with high-uncertainty labels. Specifically, *PSA* first computes the class prototypes (*i.e.*, the feature centroid of each class) in the target domain as the reference to generate soft assignments for auxiliary samples. Next, it measures the cross-domain label consistency by comparing the soft assignments with the ground-truth labels of the auxiliary samples. By thresholding the consistency score, we select reliable auxiliary samples and apply the supervised contrastive loss to pull together samples of the same class and push apart samples of different classes in the shared semantic space. Thereafter, semantically-consistent representations are learned across domains, which brings significant benefits for model optimization and knowledge transfer from the auxiliary to the target domain. Here we summarize the key contributions of this paper as follows:

- To the best of our knowledge, we present the first cross-domain cervical dysplasia visual inspection method, which effectively transfers knowledge from the auxiliary to the target domain. We propose to simultaneously align the intermediate features on both *domain level* and *class level* to learn transferable representations across domains.
- We propose a novel prototype-based method to estimate the transferability of samples in the auxiliary domain. The impact of inconsistent labels can thus be reduced by weighting the auxiliary samples according to their transferability estimated based on the distance to the class prototypes of the target domain.
- We have performed extensive experiments on three benchmark cervical image datasets. The experimental results show that our proposed method outperforms the state-of-the-art cervical cancer visual inspection methods by a significant margin.
- We have presented additional experiments on the Visda-2017 dataset. Results demonstrate the effectiveness of our method in general image analysis in addition to the cervical domain.

2 RELATED WORKS

Cervical Dysplasia Visual Inspection. A significant number of machine-learning-based methods for cervical dysplasia visual inspection have been proposed in recent years [10, 16, 41, 55, 60]. CYENet [9] and ColpoNet [50] were network architectures tailored for cervical cancer detection with cross-norm operations. Zhang *et al.* [8, 65] introduced a split-and-aggregation framework to process the high-resolution cervical images and provided classification results by summarizing patch features. One alternative solution to leverage the high-resolution input is to train a cervix detector to generate the region of interest from the original image. Faster-RCNN [49] was adopted by Hu *et al.* [27] as the detector, which was trained based on their self-annotated bounding box labels. Alyafeai *et al.* [3] proposed a more general pipeline for detector training, following which cervical detectors can be trained using a public dataset. Park *et al.* [43] applied multiple augmentation schemes to the cropped images, together with a ResNet-50 structure initialized with an ImageNet pre-trained model. Some studies focused on the information integration from metadata such as Pap results [16] and HPV signal strength [60]. However, only a small number of cervical images are associated with such metadata, which significantly limits the feasibility of such approaches.

Domain Adaptation. Domain adaptation (DA) focuses on transferring label information from the source domain to the target domain. Existing DA methods achieved it mainly based on statistical metrics [19, 37, 38, 46, 57, 63], semantic clustering [4, 23, 40, 42,

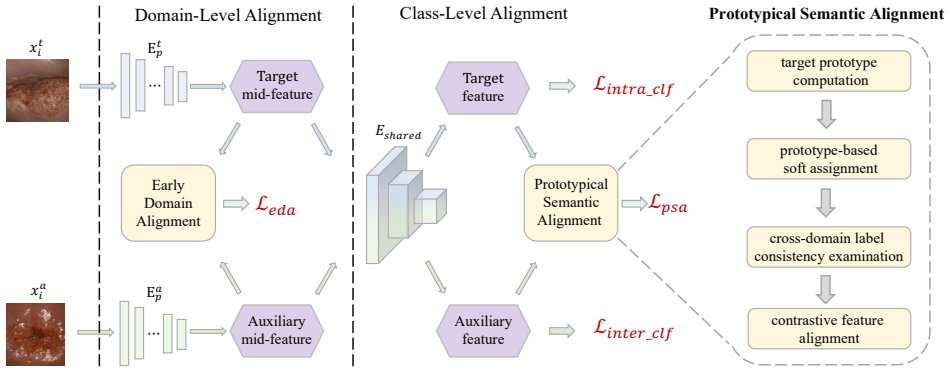


Figure 2: The overall architecture of our proposed Prototypical Cross-domain Knowledge Alignment and Transfer. E_p and E_{shared} denote the domain-private encoder and the shared encoder, respectively.

52, 62, 64], adversarial learning [11, 17, 18, 29, 36, 51, 53], or reconstruction [5, 6, 34, 35, 45, 61]. For example, CCSA [40] matched the cross-domain semantic space by aligning features based on their labels. DSN [5] proposed a disentanglement-based complex framework to separate style and content information. BrAD [23] designed an auxiliary bridge to narrow the gap between different domains. JCL [42] adopted a MoCo-like [24] structure to align unlabeled data and PAC [39] introduced a pre-training stage for model training. Since our goal differs from the DA task but shares similar properties, we select some of the existing works for comparison in our experiment. However, the asymmetrical designs of DA methods make them inappropriate to deploy in our setting, leading to worse performance compared to our framework.

Contrastive Learning. Contrastive learning was initially proposed to learn high-quality representations in a self-supervised manner where the positive pairs are constructed as the multiple augmentation views of the same sample [12–14, 24]. For example, MoCo [24] proposed to use a momentum encoder and a large dictionary to improve model stability. SimCLR [12] adopted a non-linear projection head to calculate the NT-Xent loss within the latent space. Recently, contrastive learning has also been investigated under the supervised configuration, where positive pairs are defined as the same-class samples in a mini-batch [30]. Multiple positive pairs are considered jointly in the calculation. In this paper, we further investigate supervised contrastive learning in a cross-domain setting for feature alignment.

3 PROBLEM FORMULATION

The cervical dysplasia visual inspection is usually formulated as an image classification problem based on the CIN grades (CIN0 ~ CIN4). Such an AI medical system can behave as a useful and efficient tool in alerting potential patients to take further medical examinations in real life, especially in low-resource regions where medical resources are deficient. However, the performance of existing deep learning models for cervical dysplasia visual inspection is generally limited by small-scale cervical datasets. Moreover, the integration of multiple datasets will possibly lead to even worse performance if the aforementioned challenges of *domain shift* and *criterion mismatch* are not properly addressed. Following this path, we focus on leveraging data from two different but relevant datasets (domains[1]) to perform a more robust cervical dysplasia visual inspection. Given a target domain $X_t = \{x_1^t, x_2^t, \ldots, x_{N^t}^t\}$ with labels $Y_t = \{y_1^t, y_2^t, \ldots, y_{N^t}^t\}$ and an auxiliary domain $X_a = \{x_1^a, x_2^a, \ldots, x_{N^a}^a\}$ with labels $Y_a = \{y_1^a, y_2^a, \ldots, y_{N^a}^a\}$, our goal is to improve the performance of the model on the target domain X_t with the facilitation of the auxiliary domain X_a.

Recall that the annotation quality of our auxiliary domain may not meet the standard of the target domain due to the *criterion mismatch* challenge. The auxiliary labels Y_a cannot be directly used for training. We thus propose a novel prototypical cross-domain knowledge alignment and transfer framework. Without loss of generality, we sample $|S^t|$ and $|S^a|$ images from X_t and X_a in each iteration, where S^t and S^a represent the target domain mini-batch and the auxiliary domain mini-batch, respectively. Next, we introduce our proposed model architecture and optimization objective based on $S = S^t \cup S^a$ in each iteration.

4 PROTOTYPICAL CROSS-DOMAIN KNOWLEDGE ALIGNMENT AND TRANSFER

The architecture overview of our proposed prototypical cross-domain knowledge alignment and transfer framework is illustrated in Figure 2. As aforementioned, our framework consists of an Early Domain Alignment (*EDA*) module for domain-level feature alignment and a Prototypical Semantic Alignment (*PSA*) module for class-level feature alignment. The *PSA* module further estimates the transferability of the auxiliary samples to reduce the impact of label inconsistency. By jointly optimizing the feature alignment and the classification objectives, cross-domain transferable knowledge can be effectively learned and transferred to the target domain.

4.1 Early Domain Alignment

Intuitively, a shared encoder is preferred for performance improvement if we try to introduce auxiliary data for training. However, different domains are generally occupied with differences in local feature distributions. Therefore, we adopt a Y-shape domain-adapted architecture as illustrated in Figure 2 to deal with the *domain shift*. It consists of two domain-private encoders (E_p^t and E_p^a) for local

[1]We use these two terms interchangeably in this paper.

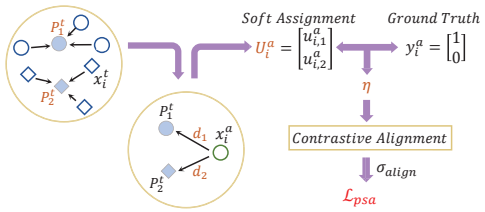


Figure 3: The pipeline of proposed PSA module.

information extraction, a shared encoder for high-level semantic deduction (E_{shared}). On top of that, an early domain alignment module is introduced to reduce the gap between the intermediate representations extracted by the two domain-private encoders.

To obtain domain-invariant features, we have investigated two major approaches to narrow the gap between domains:

Adversarial-based. Following [18], the adversarial-based alignment is achieved by minimizing the domain classification loss for the domain classifier g, while maximizing this loss for the encoders, with the help of a gradient reversal layer. We formulate the adversarial-based objective for early domain alignment as

$$\max_{E_p} \min_{g} \mathcal{L}_{eda} = -\frac{1}{|S|} \sum_{i=1}^{|S|} (y_i^d \log(\hat{y}_i^d) + (1 - y_i^d) \log(1 - \hat{y}_i^d)), \quad (1)$$

where $\hat{y}_i^d$ is the output of the domain classifier, and y_i^d is the domain index of the input image (*i.e.*, 0 for target and 1 for auxiliary).

Divergence-based. An alternative for domain-level alignment is divergence-based approaches, the goal of which is to minimize the distance between representations of samples from different domains. Here we investigate a widely used distance metric termed MK-MMD [37]. Thereby, the divergence-based objective is to minimize the *MKMMD* distance between the intermediate features of x^t and x^a to narrow the gap across domains:

$$\min_{E_p} \mathcal{L}_{eda} = MKMMD(GAP(E_p^t(x_i^t)), GAP(E_p^a(x_i^a))), \quad (2)$$

where *GAP* is the global average pooling that maps the output of the private encoders E_p^t and E_p^a into a vector.

4.2 Prototypical Semantic Alignment

Recall that in our auxiliary dataset, labels are generally provided by a single medical staff from a local clinic. The large variance of technical skills and subjective deviations lead to the *criterion mismatch* challenge. To reduce the impact of such label inconsistency, we present a *Prototypical Semantic Alignment* (*PSA*) module to align the semantics of the high-level feature representations (*i.e.*, output of E_{shared}) between the target and auxiliary domains. As shown in Figure 3, it consists of target prototype computation, prototype-based soft assignment, cross-domain label consistency examination, and contrastive feature alignment.

Target Prototype Computation. We propose to compute the per-class prototypes in the target domain, and use them as references to deduce reliable and transferable information from the auxiliary domain. A prototype is defined as the center of a semantic cluster, consisting of features with the same semantic label. Compared with instance-to-instance matching [31, 40], where matching is performed between cross-domain instance pairs, instance-to-prototype matching is more robust to abnormal instances, thus providing a better foundation for the following steps.

Specifically, we append a classification head that consists of two fully-connected layers (*i.e.*, FC_1 and FC_2) on top of the shared encoder, and compute the prototype of each target class k after every epoch as the average of all the features from this class by

$$P_k^t = \frac{\sum_{i=1}^{N^t} f^t(x_i^t) \mathbb{1}_{(y_i^t == k)}}{\sum_{i=1}^{N^t} \mathbb{1}_{(y_i^t == k)}}, \quad (3)$$

where $f^t(x^t) = FC_1(E_{shared}(E_p^t(x^t)))$ maps target image x^t to the feature before the last classification layer. $P_k^t \in \mathbb{R}^{256}$ represents the prototype of class k, and $N^t = |X_t|$ is the number of target training samples. The prototypes are denoted as $P^t = [P_1^t, P_2^t, \ldots, P_K^t]$, where K is the total number of classes.

Prototype-based Soft Assignment. The target prototypes P^t are next utilized to calculate a distance-based soft assignment for each auxiliary sample as follows

$$u_{i,k}^a = \frac{exp(-||f^a(x_i^a) - P_k^t||_2)}{\sum_{c=1}^{K} exp(-||f^a(x_i^a) - P_c^t||_2)}, \quad (4)$$

where $U_i^a = [u_{i,1}^a, u_{i,2}^a, \ldots, u_{i,K}^a] \in \mathbb{R}^K$ is a K-dimensional vector representing the probability of x_i^a belonging to target class k. $f^a(x^a) = FC_1(E_{shared}(E_p^a(x^a)))$ maps auxiliary image x^a to the shared feature space, $||f^a(x_i^a) - P_k^t||_2$ computes the l_2 distance between $f^a(x_i^a)$ and P_k^t, and softmax is applied to normalize U_i^a by ensuring $\sum_k u_{i,k}^a = 1$. Thereby, an auxiliary sample x_i^a will be assigned with a large $u_{i,k}^a$ if it is close to the prototype of class k in the shared feature space.

Cross-domain Label Consistency Examination. For an auxiliary sample x_i^a with label $y_i^a = k$, it should be close to P_k^t in the feature space if it is well aligned with the target domain. Based on this observation, we propose to compute a cross-domain label consistency score for each auxiliary sample to measure the reliability of the knowledge learned from it. Given the calculated soft assignment probabilities U_i^a and the original auxiliary ground truth y_i^a, the cross-domain label consistency is formally defined as

$$\eta_i = [U_i^a]^\top \cdot y_i^a, \quad (5)$$

where $\eta_i \in [0, 1]$ is the consistency score, and operator $\cdot$ represents the dot product of two vectors. The closer η_i is to 1, the higher the confidence that this auxiliary sample is within the same decision boundary of the target domain.

Contrastive Feature Alignment. Given the consistency score η_i calculated using Eqn. 5 and a predefined threshold σ_{align}, we filter auxiliary samples by only keeping those with consistency scores $\eta_i \geq \sigma_{align}$ to align the high-level features. Following SimCLR [12], we apply a projection head on top of the shared encoder and perform supervised contrastive learning in the projection space [30]. By pulling together samples of the same class and pushing apart samples of different classes in the projection space, it introduces consistent performance gain for classification models. In our implementation, positive pairs (x_i, x_j) are defined as images that belong to the same semantic class (*i.e.*, $y_i = y_j$), while negative pairs are images that belong to different semantic classes (*i.e.*, $y_i \neq y_j$). Both cross-domain (*i.e.*, (x_i^t, x_j^t) or (x_i^a, x_j^a)) and intra-domain (x_i^t, x_j^a)

Table 1: Performance comparison between our method and state-of-the-art methods on the NHS dataset.

Methods	Aux data	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC-AUC
2conv-Alyafeai *et al.* [3]	✗	71.02±2.15	64.10±1.20	68.49±6.76	66.22±3.49	0.7327±0.0054
3conv-Alyafeai *et al.* [3]	✗	72.16±1.50	62.24±2.30	83.56±2.37	71.34±1.16	0.7442±0.0143
CYENet [9]	✗	75.57±1.74	72.06±1.99	67.12±3.45	69.50±1.86	0.7961±0.0200
Vasudha *et al.* [58]	✗	77.27±0.33	67.74±4.00	81.30±9.65	75.90±4.64	0.7885±0.0119
Zhang *et al.* [65]	✗	81.82±1.14	73.38±1.19	88.12±1.58	80.07±1.26	0.8351±0.0071
PAC [39]	✓	79.55±1.01	71.33±0.92	84.96±3.23	77.51±1.16	0.8411±0.0179
BrAD [23]	✓	81.25±1.20	76.32±5.91	79.45±7.75	77.85±0.29	0.8256±0.0031
JCL [42]	✓	81.39±0.98	75.64±2.58	80.82±2.74	78.14±0.52	0.8493±0.0124
CCSA [40]	✓	81.65±1.14	73.56±1.36	87.67±1.58	79.99±1.33	0.8564±0.0052
DSN [5]	✓	82.38±1.43	79.17±4.32	78.08±3.62	78.62±0.60	0.8577±0.0249
Ours-mkmmd	✓	85.80±0.57	79.27±0.87	84.04±3.72	83.87±1.18	**0.8832±0.0093**
Ours-adv	✓	**86.55±0.88**	**80.42±2.09**	**89.49±2.09**	**84.67±0.78**	0.8822±0.0111

pairs are jointly considered to compute the supervised contrastive loss $\mathcal{L}_{sc}$. For simplicity, we omit the superscript in the formulation of $\mathcal{L}_{sc}$, which is given as

$$\mathcal{L}_{sc}(x_i) = -\log \frac{1}{|Q^+(i)|} \sum_{x_p} \frac{exp(sim(\mathbf{z}_i, \mathbf{z}_p))}{\sum_{x_q} exp(sim(\mathbf{z}_i, \mathbf{z}_q))}, \tag{6}$$

where $\mathbf{z}_i$ is the output of the projection head corresponding to x_i, and $sim()$ computes the cosine similarity. $x_q \in Q(i) \equiv \{I - x_i\}$ where I is the union of the target batch and the filtered auxiliary batch based on σ_{align}. $x_p \in Q^+(i) \equiv \{x_q : y_q = y_i\}$ contains all the positive pairs (x_i, x_p) for x_i in $Q(i)$. The objective is to maximize the cosine similarity between positive pairs while minimizing it between negative pairs. Subsequently, the prototypical semantic alignment loss is computed as the average of the supervised contrastive loss over all valid training samples:

$$\mathcal{L}_{psa} = \frac{\sum_{i=1}^{|S^t|} \mathcal{L}_{sc}(x_i^t) + \sum_{i=1}^{|S^a|} \mathbb{1}_{(\eta_i \geq \sigma_{align})} \mathcal{L}_{sc}(x_i^a)}{|S^t| + \sum_{i=1}^{|S^a|} \mathbb{1}_{(\eta_i \geq \sigma_{align})}}. \tag{7}$$

Our proposed prototypical semantic alignment loss strengthens the matching of cross-domain samples. Considering that the visual appearances of cervical images are highly similar across cases, it can also assist in quick concentration on the most important information for classification.

4.3 Cross-Domain Knowledge Transfer

The aforementioned feature alignments assist our model to generate domain-invariant features with a global horizon. Next, we perform cross-domain knowledge transfer based on supervised classification. We adopt a threshold σ_{clf} for cross-domain knowledge transfer with respect to the consistency score computed in Eqn. 5. Let $x^t \in S^t$ and $x^a \in \tilde{S}^a = \{x_i^a | \eta_i \geq \sigma_{clf}\}$ denote the training samples in the target mini-batch and in the filtered auxiliary mini-batch based on σ_{clf}, respectively, then the cross-entropy for classification is computed as

$$\begin{aligned} \mathcal{L}_{clf} &= \mathcal{L}_{intra\_clf} + \gamma \mathcal{L}_{inter\_clf} \\ &= -\frac{1}{|S^t|} \sum_{x_i^t \in S^t} y_i^t \log(\hat{y}_i^t) - \frac{\gamma}{|\tilde{S}^a|} \sum_{x_i^a \in \tilde{S}^a} \eta_i y_i^a \log(\hat{y}_i^a), \end{aligned} \tag{8}$$

where $\hat{y}_i$ and y_i denote the prediction and the ground-truth label of image x_i, respectively. We further weight the auxiliary samples by $\gamma \cdot \eta_i$ to enforce stronger supervision from samples with a larger cross-domain consistency. It serves as the mainstay of our framework, pushing it to constantly focus on extracting informative features for classification.

4.4 Overall Objectives

We optimize our model by jointly considering the classification loss, $\mathcal{L}_{clf}$, and the feature alignment losses, $\mathcal{L}_{eda}$ and $\mathcal{L}_{psa}$. The overall loss function of our proposed prototypical cross-domain knowledge transfer framework is formulated as

$$\mathcal{L} = \mathcal{L}_{clf} + \alpha \mathcal{L}_{eda} + \beta \mathcal{L}_{psa}, \tag{9}$$

where α, β are coefficients controlling the balance between the classification loss and the feature alignment loss functions.

5 EXPERIMENTS

5.1 Dataset

Totally 17,002 cervical images are used in our experiments, which were collected from three separate medical studies: Natural History Study of HPV and Cervical Neoplasia (NHS) [26], ASCUS-LSIL Triage Study (ALTS) [22], and Biopsy Study (Biopsy) [56]. We filter the records that are labeled with ground-truth CIN grades (CIN 0,1,2,3,4) within 1 year of the screening date and formulate it as a binary classification problem to detect abnormal cases, following previous work [65]. The accessibility of these datasets is based on request and constrained agreement. When compared to the state-of-the-art, we performed two sets of experiments by utilizing NHS and Biopsy as the target dataset, respectively, while the NHS dataset is utilized in ablation studies. Please refer to the supplementary material for a detailed description of these datasets.

5.2 Implementation Details

The original resolution of cervical images is generally 2,400×1,600. Following previous work [3], we adopt a cropping scheme to select the region of interest as a preprocessing step. We adopt the ResNet-50 [25] as our backbone and initialize it with the ImageNet self-supervised model Dino [7]. For training stability, we first train

Table 2: Performance comparison between our method and domain adaptation methods on the Biopsy dataset.

Methods	Aux data	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC-AUC
CCSA [40]	✓	56.56±1.16	66.27±7.30	26.67±14.14	36.42±13.30	0.5398±0.0471
PAC [39]	✓	58.78±1.92	62.59±2.72	36.40±7.25	46.02±4.28	0.5511±0.0225
BrAD [23]	✓	59.02±0.06	57.99±1.83	62.22±11.70	59.54±4.66	0.5566±0.0022
JCL [42]	✓	59.84±1.08	61.93±1.73	48.33±11.79	53.79±6.82	0.5801±0.0115
DSN [5]	✓	62.29±1.64	68.03±9.20	47.77±9.62	55.11±4.45	0.6208±0.0078
Ours-mkmmd	✓	63.93±0.92	**72.22±3.29**	43.33±6.42	54.17±1.77	**0.6269±0.0153**
Ours-adv	✓	**64.75±1.16**	63.48±0.22	**66.66±4.72**	**65.00±2.36**	0.6253±0.0023

our model without the prototypical semantic alignment loss for 5 epochs as a warm-up, then continue training by empirically setting the balancing coefficients α, β, γ in the objective function to $0.1, 0.01, 0.1$, respectively. We conduct an ablation study to evaluate the impact of the thresholds σ_{align} and σ_{clf} for cross-domain knowledge transfer, based on which we set $\sigma_{align} = 0.4$ and $\sigma_{clf} = 0.9$ in the rest of the experiments. More details of implementation can be found in the supplementary material.

5.3 Comparison to the State-of-the-Art

We compare our proposed framework with nine state-of-the-art methods for both cervical dysplasia visual inspection and domain adaptation. Five commonly used classification measurements including top-1 accuracy, precision, recall, F1-score, and area under the ROC curve (ROC-AUC) are adopted as evaluation metrics. For a fair comparison, we train each model three times to reduce randomness, and report the average results together with the standard deviation of the three independent runs in Tables 1 and 2.

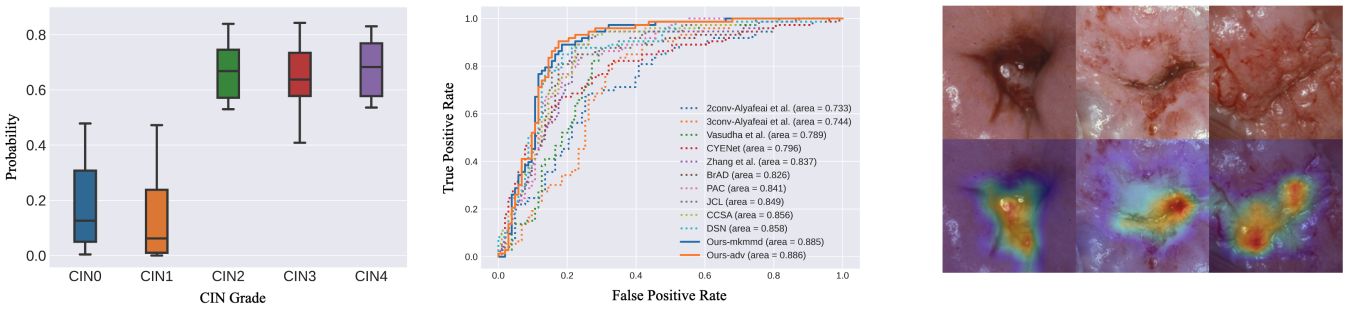
Table 1 illustrates the performance comparison on the NHS dataset. Compared to previous cervical dysplasia visual inspection methods, our proposed framework with either divergence or adversarial alignment surpasses them by an overall large margin. Among the two candidates, the adversarial one performs better, which outperforms the second-best solution [65] by an average improvement of 4.73% in top-1 accuracy, 7.04% in precision, 1.37% in recall, 4.60% in F1 score, and 0.047 in ROC-AUC. A larger gap can be observed in various metrics against the rest of the methods [3, 9, 58], where an improvement of more than 10% in top-1 accuracy, 8% in precision, 6% in recall and 0.1 in ROC-AUC are generally obtained. The experimental results verify our motivation of looking for new auxiliary data sources for medical image analysis. Knowledge can be learned and transferred between medical images that were collected under different trials effectively, whereas the key challenges caused by domain shift and criterion mismatch across medical trials have to be properly solved.

Next, we compare our proposed framework to domain adaptation methods DSN [5], JCL [42], CCSA [40], PAC [39], BrAD [23], where auxiliary data are also utilized for training. Originally designed for unsupervised domain adaptation (UDA) or domain generalization, these methods can be applied to our setting by considering our target domain as the target domain in UDA and our auxiliary domain as the source domain in UDA. This approach allows us to utilize both the target and auxiliary labels by adding a target cross-entropy loss on top of their corresponding objective function. As can be seen, despite the fact that DSN performs the best out of all the existing methods on both datasets, its improvements over cervical dysplasia visual inspection methods are somewhat limited, particularly in terms of top-1 accuracy, recall, and F1 score. We can also see that CCSA outperforms BrAD, PAC, and JCL on the NHS dataset. However, its performance degrades significantly on the Biopsy dataset (refer to Table 2). Such domain adaptation methods mainly focus on solving the domain shift challenge, while ignoring the issue caused by the label inconsistencies that potentially exist in different domains. Performing cross-domain knowledge transfer by fetching auxiliary information without selection may introduce undesirable noise to the target domain. Comparatively, we utilize the auxiliary information by first estimating its transferability and then optimizing the classification model jointly with feature alignment in the shared semantic space. From the results, we can see that our proposed solution with the adversarial module is more robust and less vulnerable to label inconsistencies across domains. It achieves the best performance in almost all five metrics and outperforms the domain adaptation methods by at least 4.17% in terms of top-1 accuracy.

Table 2 reports the performance comparison on the Biopsy dataset. This dataset is highly challenging due to the lower quality and smaller scale compared to the NHS dataset, where only 393 valid records can be utilized for model training. Similarly, our two variants outperform the existing solution in all five metrics, where the adversarial-based early alignment is still better. Compared with domain adaptation methods, it obtains the best result in four out of the five metrics. It outperforms the second best method (*i.e.*, DSN) by 2.46% in top-1 accuracy, 18.89% in recall, and 9.89% in F1 score. Compared with Table 1, we can see that the performance of domain adaptation methods is less stable on different datasets. One reason might be their heavy dependence on the intra-supervision. Most of them were developed based on the assumption that the given labels are accurately annotated by humans, which is actually not always guaranteed in real-world applications. Our method, on the other hand, utilizes both intra-supervision and inter-supervision simultaneously, thus leading to a more practical and robust solution compared to the previous methods.

5.4 Ablation Studies

Model Architecture. We first set our loss function to $\mathcal{L} = \mathcal{L}_{clf} + \alpha\mathcal{L}_{eda}$ with $\sigma_{clf} = 0$ and $\eta_i = 1$ to remove the impact of the *PSA* module, and evaluate our adversarial-based domain-level alignment in Table 3. We compare it with three counterparts given both target



(a) (b) (c)

Figure 4: (a) Predicted probability statistics for each CIN grade. (b) ROC curve comparison among methods from both cervical dysplasia visual inspection and domain adaptation. (c) Visualization of model attention based on GradCAM.

Table 3: Different architectures training with both domains without filtering.

Architecture	E_p	$\mathcal{L}_{eda}$	Acc (%)	Acc Dec. (%)
$\mathcal{L}_{clf} + \alpha\mathcal{L}_{eda}$	✓	✓	82.95	-
- w/o $\mathcal{L}_{eda}$	✓		82.39	-0.56
- w/o E_p		✓	82.39	-0.56
ResNet-50			81.25	-1.70

Table 4: Ablation study for training strategy.

Training Strategy			Acc (%)	Acc Dec.(%)
$\mathcal{L}_{eda}$	$\mathcal{L}_{psa}$	$\mathcal{L}_{inter\_clf}$		
✓	✓	✓	86.55	-
	✓	✓	83.52	-3.03
✓		✓	84.09	-2.46
✓	✓		84.09	-2.46

data and auxiliary data, including ours without $\mathcal{L}_{eda}$, ours without the domain-private encoders, and single-branch structure without both components. We can see that the top-1 accuracy degrades by 0.56% if we remove the adversarial loss $\mathcal{L}_{eda}$ from our method or replace the domain-private encoders with a shared encoder. A further degradation of 1.70% can be observed if both components are removed. Thus, the experimental results verify the effectiveness of our *EDA* module for cervical dysplasia visual inspection.

Training Strategy. Our training strategy consists of multiple loss functions as shown in Eqn. 9. Here we examine their effectiveness by removing each of them from $\mathcal{L}$ and report the results in Table 4. Since the supervision from the target data is necessary for our task, we conduct this ablation study only on $\mathcal{L}_{eda}$, $\mathcal{L}_{psa}$, and $\mathcal{L}_{inter\_clf}$. We observe that, compared to our proposed adversarial-based objective function, removing either one of the individual losses leads to performance degradation ranging from 2.46% to 3.03% in top-1 accuracy. Among them, $\mathcal{L}_{inter\_clf}$ and $\mathcal{L}_{psa}$ both serve as the bridges for integrating auxiliary knowledge, but from different aspects, thus leading to similar performance decrement. The results indicate that all our proposed losses are indispensable components of our method, which work collaboratively to complete the task.

Table 5: Performance comparison when training with different number of target samples.

Percentage	10%	20%	50%	100%
ResNet-50	69.89	76.14	81.82	82.84
Ours-mkmmd	75.57	80.68	82.39	85.80
Ours-adv	**76.70**	**81.82**	**84.66**	**86.55**

5.5 Discussion

Statistical Prediction Distribution. Our binary setting originally comes from five categories (CIN 0,1,2,3,4) — CIN0 and CIN1 are regarded as normal, while the rest of them are regarded as abnormal. In Figure 4(a), we visualize the statistical prediction distribution of our model for each category, where the Y-axis represents the predicted probability of belonging to the abnormal case. We can observe a distinct margin between the first two levels and the last three. The generally non-overlapped phenomenon between the upper bound of normal classes and the lower bound of abnormal classes reveals a clear decision boundary from our model.

ROC Curve. Receiver operating characteristics (ROC) is a probability curve for classification problems at various threshold settings. In Figure 4(b), we present the ROC curves from all methods in Table 1 for comparison. The closer the curve to the left-top (*i.e.*, the larger the area under the curve (AUC)), the better capability the method has. Our method, shown in the orange line, surpasses all other methods with 0.886 in AUC.

Model Attention Visualization. Based on GradCAM [20, 54], we visualize the last convolutional attention map from our framework as shown in Figure 4(c). The brighter color in the second row represents the higher-focus area. We can see that our model focuses more on the areas with obvious pathological features around the cervix, providing a more reasonable prediction for those patients.

Percentage of target data. We further compare our framework with the baseline model using different percentages of target data and report the results in Table 5. In each column, we randomly select a certain percentage of target samples for training to study the impact of the target data volume. The results show that our adversarial-based approach outperforms the baseline model by a significant margin, with improvements of 6.81%, 5.68%, 2.84%, and 3.71% achieved when the percentage is set to 10%, 20%, 50%, and

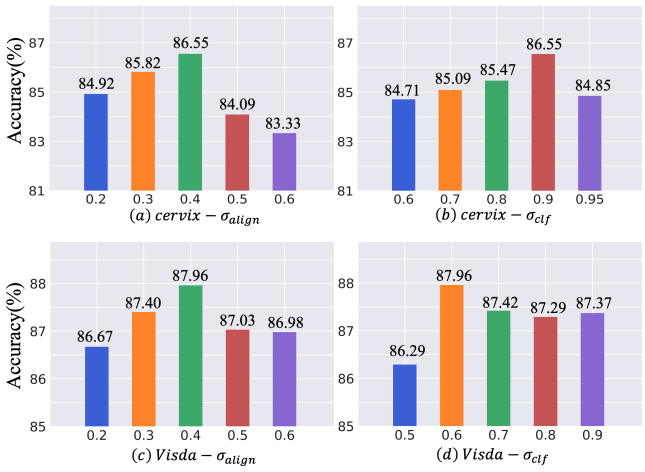


Figure 5: Performance comparison using different thresholds on the NHS and Visda dataset. (a): Different σ_{align} with $\sigma_{clf} = 0.9$ on NHS dataset. (b): Different σ_{clf} with $\sigma_{align} = 0.4$ on NHS dataset. (c): Different σ_{align} with $\sigma_{clf} = 0.6$ on Visda dataset. (d): Different σ_{clf} with $\sigma_{align} = 0.4$ on Visda dataset.

100%, respectively. These findings demonstrate that our framework is able to learn complementary and transferable information from the auxiliary data, which is particularly beneficial when the amount of labeled data in the target domain is limited.

Thresholds. We investigate the impact of two key hyper-parameters, σ_{align} and σ_{clf}, on the NHS dataset and compare the top-1 accuracy of one candidate by holding the other fixed as the best value we found during the experiment. As shown in Figure 5(a), the best classification result was obtained with $\sigma_{align} = 0.4$. Decreasing σ_{align} will result in supervision with noisy labels, while increasing it will result in less information learned and transferred from the auxiliary domain. For σ_{clf}, which defines the transferability threshold for $\mathcal{L}_{inter\_clf}$, Figure 5(b) shows that the best classification result was obtained with $\sigma_{clf} = 0.9$. A similar pattern can be observed that either increasing or decreasing σ_{clf} will lead to performance degradation. However, compared with σ_{clf}, we observe that decreasing σ_{align} has less impact than increasing it. This indicates that contrastive feature alignment is more tolerable with out-of-distribution semantics than direct inter-domain supervision. In supervised contrastive learning, multiple positive pairs are constructed, both in-domain and cross-domain, making it a more robust solution to the potential inconsistency between cross-domain classification boundaries.

5.6 Results on Visda-2017 Dataset

To verify the generalization capability of our model, we conduct additional experiments on Visda-2017 [47]. Different from the cervix dataset, it is a large and general image dataset consisting of synthetic images and real images across 12 classes. The potential noise inside the synthetic images due to the artificial generation process is a large obstacle towards good performance. Therefore, both domain shift and label uncertainty challenges are presented in this dataset. In these experiments, we regard the real images as the target domain and the synthetic images as the auxiliary domain to evaluate our method. Compared with existing domain adaptation methods [5, 23, 40, 42], we report the top-1 and top-5 accuracy in Table 6. We can see that our framework is able to surpass other domain adaptation methods in this general dataset, obtaining a 3.96% improvement in top-1 accuracy compared to the second-best solution. The results show that our method not only works well in small-scale medical datasets that focus on specific binary classification problem, but also generalizes well in large-scale general image datasets and multi-class classification problems. We also investigate the impact of the thresholds on Visda-2017. As shown in Figure 5(c) and (d), the best result is obtained with $\sigma_{clf} = 0.6$ and $\sigma_{align} = 0.4$. Different from the results on the cervix dataset, a lower σ_{clf} is preferable on Visda-2017, possibly due to high cross-domain label consistency. To summarize, the results indicate the potential utilization of our proposed method in applications other than the medical domain, which will be explored as part of our future work.

Table 6: Performance comparison between our method and domain adaptation methods on the Visda-2017 dataset.

	Top-1 (%)	Top-5 (%)
CCSA [40]	77.64	97.06
JCL [42]	78.12	97.48
BrAD [23]	83.89	98.29
DSN [5]	84.00	97.92
Ours-mkmmd	85.59	98.32
Ours-adv	**87.96**	**98.69**

6 CONCLUSION

Targeted at cervical dysplasia visual inspection, we present a novel prototypical cross-domain knowledge transfer framework to perform robust auxiliary-to-target knowledge transfer. Two key components are introduced in our method, namely the *EDA* module and the *PSA* module. The former addresses the domain shift problem by aligning the intermediate representations, while the latter utilizes a prototype-based strategy to learn useful and reliable semantic information from the auxiliary domain. Experiments on three benchmark cervical image datasets demonstrate the state-of-the-art performance of our proposed approach, with 4.7% improvement in top-1 accuracy and 0.05 in ROC-AUC. Additional result visualizations and ablation studies are presented to validate our framework design, together with the experiments on Visda-2017 dataset to demonstrate the effectiveness of our method in a more general problem setting. In the future, we plan to investigate the potential of our method in not only cross-domain but also cross-modal applications with varying label quality.

ACKNOWLEDGMENTS

This work was supported by Singapore Ministry of Education Academic Research Fund Tier 1 under MOE's official grant number T1 251RES2029, the National Natural Science Foundation of China No. 62272390, and Zhejiang Gongshang University "Digital+" Disciplinary Construction Management Project (Project Number SZJ2022C005).

REFERENCES

[1] 2017. Intel&MobileODT dataset. https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening/data. Accessed 2021-02-10.

[2] 2018. Imbalanced Sampler. https://github.com/ufoym/imbalanced-dataset-sampler. Accessed 2021-02-10.

[3] Zaid Alyafeai and Lahouari Ghouti. 2020. A fully-automated deep learning pipeline for cervical cancer classification. *Expert Systems with Applications* 141 (2020), 112951.

[4] David Berthelot, Rebecca Roelofs, Kihyuk Sohn, Nicholas Carlini, and Alex Kurakin. 2021. Adamatch: A unified approach to semi-supervised learning and domain adaptation. *arXiv preprint arXiv:2106.04732* (2021).

[5] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. *Advances in neural information processing systems* 29 (2016).

[6] Jinming Cao, Oren Katzir, Peng Jiang, Dani Lischinski, Danny Cohen-Or, Changhe Tu, and Yangyan Li. 2018. Dida: Disentangled synthesis for domain adaptation. *arXiv preprint arXiv:1805.08019* (2018).

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9650–9660.

[8] Jinyeong Chae, Ying Zhang, Roger Zimmermann, Dongho Kim, and Jihie Kim. 2022. An Attention-Based Deep Learning Model with Interpretable Patch-Weight Sharing for Diagnosing Cervical Dysplasia. In *Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 3*. Springer, 634–642.

[9] Venkatesan Chandran, MG Sumithra, Alagar Karthick, Tony George, M Deivakani, Balan Elakkiya, Umashankar Subramaniam, and S Manoharan. 2021. Diagnosis of cervical cancer based on ensemble deep learning network using colposcopy images. *BioMed Research International* 2021 (2021).

[10] Sung K Chang, Yvette N Mirabal, Edward Neely Atkinson, Dennis D Cox, Anais Malpica, Michelle Follen, and Rebecca R Richards-Kortum. 2005. Combined reflectance and fluorescence spectroscopy for in vivo detection of cervical precancer. *Journal of biomedical optics* 10, 2 (2005), 024031.

[11] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. 2019. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 627–636.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[13] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems* 33 (2020), 22243–22255.

[14] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[16] Timothy DeSantis, Nahida Chakhtoura, Leo Twiggs, Daron Ferris, Manocher Lashgari, Lisa Flowers, Mark Faupel, Shabbir Bambot, Steven Raab, and Edward Wilkinson. 2007. Spectroscopic imaging as a triage test for cervical disease: a prospective multicenter clinical trial. *Journal of lower genital tract disease* 11, 1 (2007), 18–24.

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

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

[19] Muhammad Ghifary, W Bastiaan Kleijn, and Mengjie Zhang. 2014. Domain adaptive neural networks for object recognition. In *PRICAI 2014: Trends in Artificial Intelligence: 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings 13*. Springer, 898–904.

[20] Jacob Gildenblat and contributors. 2021. PyTorch library for CAM methods. https://github.com/jacobgil/pytorch-grad-cam.

[21] Arnaud Gotlieb, Marine Louarn, Mari Nygard, Tomas Ruiz-Lopez, Sagar Sen, and Roberta Gori. 2017. Constraint-based verification of a mobile app game designed for nudging people to attend cancer screening. In *Twenty-Ninth IAAI Conference*.

[22] The ASCUS-LSIL Triage Study ALTS Group. 2003. A randomized trial on the management of low-grade squamous intraepithelial lesion cytology interpretations. *American journal of obstetrics and gynecology* 188, 6 (2003), 1393–1400.

[23] Sivan Harary, Eli Schwartz, Assaf Arbelle, Peter Staar, Shady Abu-Hussein, Elad Amrani, Roei Herzig, Amit Alfassy, Raja Giryes, Hilde Kuehne, et al. 2022. Unsupervised Domain Generalization by Learning a Bridge Across Domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5280–5290.

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[26] Rolando Herrero, Allan Hildesheim, Concepcion Bratti, Mark E Sherman, Martha Hutchinson, Jorge Morales, Ileana Balmaceda, Mitchell D Greenberg, Mario Alfaro, Robert D Burk, et al. 2000. Population-based study of human papillomavirus infection and cervical neoplasia in rural Costa Rica. *Journal of the National Cancer Institute* 92, 6 (2000), 464–474.

[27] Liming Hu, David Bell, Sameer Antani, Zhiyun Xue, Kai Yu, Matthew P Horning, Noni Gachuhi, Benjamin Wilson, Mayoore S Jaiswal, Brian Befano, et al. 2019. An observational study of deep learning and automated evaluation of cervical images for cancer screening. *JNCI: Journal of the National Cancer Institute* 111, 9 (2019), 923–932.

[28] Ahmedin Jemal, Freddie Bray, Melissa M Center, Jacques Ferlay, Elizabeth Ward, and David Forman. 2011. Global cancer statistics. *CA: a cancer journal for clinicians* 61, 2 (2011), 69–90.

[29] Xiang Jiang, Qicheng Lao, Stan Matwin, and Mohammad Havaei. 2020. Implicit class-conditioned domain alignment for unsupervised domain adaptation. In *International Conference on Machine Learning*. PMLR, 4816–4827.

[30] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems* 33 (2020), 18661–18673.

[31] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A Plummer, Stan Sclaroff, and Kate Saenko. 2020. Cross-domain self-supervised learning for domain adaptation with few source labels. *arXiv preprint arXiv:2003.08264* (2020).

[32] Chen Li, Dan Xue, Zhijie Hu, Hao Chen, Yudong Yao, Yong Zhang, Mo Li, Qian Wang, and Ning Xu. 2019. A survey for breast histopathology image analysis using classical and deep neural networks. In *International Conference on Information Technologies in Biomedicine*. Springer, 222–233.

[33] Johann Li, Guangming Zhu, Cong Hua, Mingtao Feng, Ping Li, Xiaoyuan Lu, Juan Song, Peiyi Shen, Xu Xu, Lin Mei, et al. 2021. A Systematic Collection of Medical Image Datasets for Deep Learning. *arXiv preprint arXiv:2106.12864* (2021).

[34] Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. *Advances in neural information processing systems* 29 (2016).

[35] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. 2021. Deep Dual Consecutive Network for Human Pose Estimation. In *CVPR*. 525–534. https://doi.org/10.1109/CVPR46437.2021.00059

[36] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. 2019. Towards Natural and Accurate Future Motion Prediction of Humans and Animals. In *CVPR*. 10004–10012. https://doi.org/10.1109/CVPR.2019.01024

[37] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*. PMLR, 97–105.

[38] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2017. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*. PMLR, 2208–2217.

[39] Samarth Mishra, Kate Saenko, and Venkatesh Saligrama. 2021. Surprisingly simple semi-supervised domain adaptation with pretraining and consistency. *arXiv preprint arXiv:2101.12727* (2021).

[40] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. 2017. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*. 5715–5725.

[41] Yanglan Ou, Yuan Xue, Ye Yuan, Tao Xu, Vincent Pisztora, Jia Li, and Xiaolei Huang. 2020. Semi-supervised cervical dysplasia classification with learnable graph convolutional network. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1720–1724.

[42] Changhwa Park, Jonghyun Lee, Jaeyoon Yoo, Minhoe Hur, and Sungroh Yoon. 2020. Joint contrastive learning for unsupervised domain adaptation. *arXiv preprint arXiv:2006.10297* (2020).

[43] Ye Rang Park, Young Jae Kim, Woong Ju, Kyehyun Nam, Soonyung Kim, and Kwang Gi Kim. 2021. Comparison of machine and deep learning for the classification of cervical cancer based on cervicography images. *Scientific Reports* 11, 1 (2021), 1–11.

[44] Paul. 2017. Kaggle bounding box labels. https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening/discussion/31565. Accessed 2021-02-10.

[45] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. 2019. Domain agnostic learning with disentangled representations. In *International Conference on Machine Learning*. PMLR, 5102–5112.

[46] Xingchao Peng and Kate Saenko. 2018. Synthetic to real adaptation with generative correlation alignment networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1982–1991.

[47] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. 2017. VisDA: The Visual Domain Adaptation Challenge. *ArXiv* abs/1710.06924 (2017).
[48] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. 2019. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems* 32 (2019).
[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).
[50] Sumindar Kaur Saini, Vasudha Bansal, Ravinder Kaur, and Mamta Juneja. 2020. ColpoNet for automated cervical cancer screening using colposcopy images. *Machine Vision and Applications* 31 (2020), 1–15.
[51] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. 2019. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8050–8058.
[52] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Asymmetric tri-training for unsupervised domain adaptation. In *International Conference on Machine Learning*. PMLR, 2988–2997.
[53] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3723–3732.
[54] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
[55] Dezhao Song, Edward Kim, Xiaolei Huang, Joseph Patruno, Héctor Muñoz-Avila, Jeff Heflin, L Rodney Long, and Sameer Antani. 2014. Multimodal entity coreference for cervical dysplasia diagnosis. *IEEE transactions on medical imaging* 34, 1 (2014), 229–245.
[56] Mark H Stoler, Brigitte M Ronnett, Nancy E Joste, William C Hunt, Jack Cuzick, and Cosette M Wheeler. 2015. The interpretive variability of cervical biopsies and its relationship to HPV status. *The American journal of surgical pathology* 39, 6 (2015), 729.
[57] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 443–450.
[58] Ajay Mittal Vasudha and Mamta Juneja. 2018. Cervix cancer classification using colposcopy images by deep learning method. *Int J Eng Technol Sci Res* 5 (2018), 426–432.
[59] World Health Organization WHO. 2022. WHO-cervical-cancer. https://www.who.int/health-topics/cervical-cancer. Accessed 2021-07-15.
[60] Tao Xu, Han Zhang, Xiaolei Huang, Shaoting Zhang, and Dimitris N Metaxas. 2016. Multimodal deep learning for cervical dysplasia diagnosis. In *International conference on medical image computing and computer-assisted intervention*. Springer, 115–123.
[61] Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. 2020. Label-driven reconstruction for domain adaptation in semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*. Springer, 480–498.
[62] Yifang Yin, Harsh Shrivastava, Ying Zhang, Zhenguang Liu, Rajiv Ratn Shah, and Roger Zimmermann. 2021. Enhanced audio tagging via multi-to single-modal teacher-student mutual learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 10709–10717.
[63] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811* (2017).
[64] Xu Zhang, Felix Xinnan Yu, Shih-Fu Chang, and Shengjin Wang. 2015. Deep transfer network: Unsupervised domain adaptation. *arXiv preprint arXiv:1503.00591* (2015).
[65] Ying Zhang, Yifang Yin, Zhenguang Liu, and Roger Zimmermann. 2021. A Spatial Regulated Patch-Wise Approach for Cervical Dysplasia Diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 733–740.

A APPENDIX

A.1 Cervix dataset

We utilize a totally of 17,002 cervical images from the Natural History Study of HPV and Cervical Neoplasia (NHS) [26], ASCUS-LSIL Triage Study (ALTS) [22] and Biopsy Study (Biopsy) [56] in this paper. They are three separate clinical studies by the National Cancer Institute (NCI) during previous decades.

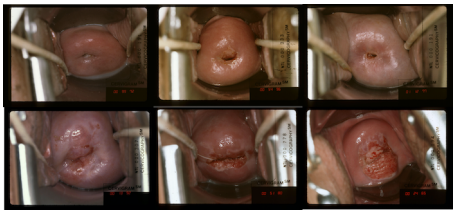


Figure 6: The original cervical images from the NHS dataset. Images in the first row are normal cases while images in the second row are abnormal cases.

- NHS is a longitudinal study in Costa Rica started in June 1993, which focuses on studying the role of human papillomavirus infection in the etiology of high-grade cervical neoplasia and evaluating new cervical cancer screening technologies. During 7 years, 10,000 women were enrolled and two cervigrams were taken at each clinic visit as shown in Fig. 6. It consists of high-resolution cervical images with the shape of around $2400 \times 1600 \times 3$.
- ALTS was designed to evaluate 3 alternative methods (immediate colposcopy, repeat PAP tests and testing for HPV) for managing atypical squamous cells of undetermined significance (ASCUS) and low-grade squamous intraepithelial lesions (LSIL). It is a randomized clinical trial started in November 1996, where women age 18+ with ASCUS (n=3488) or LSIL (n=1572) cytology were enrolled at 4 colposcopy clinics in the United States. Similarly, two cervigrams with shapes similar to the NHS dataset were taken during each visit in this study.
- Biopsy was a cross-sectional study designed to understand cervical disease on the lesion level and to establish criteria for conducting cervical biopsies. Out of 2,270 women referred for colposcopy, 690 eligible women consented to participate in the study.

During these projects, each patient may have participated in multiple screening sessions, where two photographs of the cervix (cervigrams) were taken during each recruitment and clinic visit as shown in Figure 6.

The cervical intraepithelial neoplasia (CIN) level normally serves as the criterion to judge the severity of cervical cancer. In our dataset, cervical images are labeled from CIN0 to CIN4, where histologic CIN2 or worse (CIN2+: CIN2, CIN3, CIN4) indicates the cancer precursor or cancer. To construct an appropriate dataset for the model training, which aims at alerting potential patients for further medical examination, we model this problem as a binary classification problem. Cases with CIN2+ are regarded as abnormal cases, while others are regarded as normal cases. Also, abnormal cases whose screening dates surpass one year are discarded due to the possible noise introduced by these samples. In this way, we have 885 images for the NHS dataset, 15,724 images for the ALTS dataset, and 393 images for the Biopsy dataset. The positive and negative ratios are 354:531 for the NHS dataset, 1961:13763 for the ALTS dataset and 151:242 for the Biopsy dataset. Two target datasets (NHS and Biopsy) in our case are not largely imbalanced, while the auxiliary dataset is. Thus, we apply a balance sampler [2] to handle

the imbalance problem in the auxiliary dataset. For each epoch, we randomly select a balanced subset of the auxiliary samples, which has the same number of images as the target training dataset. A train-test ratio of 4:1 is further employed on these datasets following [65]. Specifically, we split the samples based on the session ID (i.e. patient). During each session, two photographs of a patient was taken. Both pictures from the same session (i.e. same patient) will be assigned to either the training set or the testing set. The accessibility of these datasets is based on request and constrained agreement.

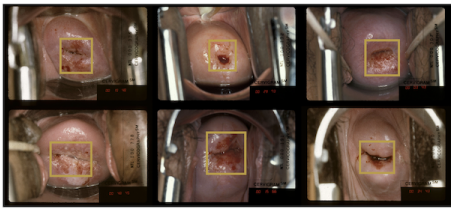


Figure 7: Detection result from our detector with confidence threshold 0.8.

A.2 Implementation Details

Our implementation is based on PyTorch 1.10 and Ubuntu 20.04. During preprocessing, due to the high resolution of cervical images of 2,400×1,600, we train a cervix detector following [3] and use the detector to focus on the important area. We crop out the cervix area from the original cervical images, where medical instruments and background still exist, so as to alleviate the problem of information loss when we resize them into 224×224 for the better utilization of the ImageNet pre-trained model. To achieve that, we utilize the Intel&MobileODT dataset [1, 44] to train a cervix detector, which is adopted to detect the cervix areas (RoI) from our original cervical images. These areas are later cropped out so as to build a fresh dataset with smaller-shape cervical images (cropped cervical images). The confidence threshold of detection is set to 0.8 so as to discard invalid or unreliable cropping results that may interfere with model training. The detection results are shown in Figure. 7. After that, we adopt random color jitter, random grayscale, random gaussian blur and random horizontal flip to perform data augmentation for each image. In our experiment, the baseline ResNet-50 model achieves 2.16% improvement in top-1 accuracy by using the above augmentation. One possible reason is that the scale of our target dataset is small and the visual similarity across training samples is high.

For model training, we adopt the ResNet-50 [25] model as our backbone in this paper, with the first three stages as the domain-private encoders and the fourth stage as the shared encoder. We append a projection head and a classification head on top of the shared encoder (*i.e.*, the last stage of ResNet-50) to perform contrastive learning and classification, respectively. Both the projection head and the classification head are implemented as two fully-connected layers with ReLU activation. The input to the *PSA* module is the middle activated output from the classification head. The encoders are initialized with the ImageNet self-supervised model Dino [7] and the domain-private encoders are trained in an end-to-end fashion without pre-training on domain data separately.

Table 7: Ablation studies for loss efficients α, β, γ.

α	β	γ	Top-1	Top-5
0.1	**0.01**	**0.1**	**87.96**	**98.69**
0.05	-	-	86.43	98.63
0.2	-	-	87.75	98.80
0.3	-	-	87.77	98.85
0.4	-	-	87.25	98.68
0.5	-	-	86.63	98.69
-	0.001	-	86.89	98.65
-	0.05	-	87.32	98.77
-	0.1	-	86.87	98.56
-	0.2	-	86.96	98.44
-	0.5	-	87.00	98.52
-	-	0.01	86.99	98.70
-	-	0.05	87.19	98.16
-	-	0.2	87.08	98.57
-	-	0.3	86.74	98.60
-	-	0.4	86.82	98.62

We train our model using the Adam optimizer with weight decay set to 10^{-3}. We adopt a mini-batch size of $|S^t| = |S^a| = 128$ and an initial learning rate of 10^{-4}. As the auxiliary domain can be much larger than the target domain, we find it to be beneficial by sampling balancedly from the two domains with a ratio of $1:1$. For training stability, we first train our model without the prototypical semantic alignment loss for 5 epochs as a warm-up, then continue training by empirically set the balancing coefficients α, β, γ in the objective function to $0.1, 0.01, 0.1$, respectively. We conduct an ablation study to evaluate the impact of the thresholds σ_{align} and σ_{clf} for cross-domain knowledge transfer, based on which we set $\sigma_{align} = 0.4$ and $\sigma_{clf} = 0.9$ in the rest of the experiments.

A.3 Ablation Studies for Loss Coefficients

Here we conduct three ablation studies for α, β, γ for $\mathcal{L}_{ada}$, $\mathcal{L}_{psa}$, $\mathcal{L}_{inter\_clf}$, respectively, on the Visda-2017 dataset due to its more stable results from a large testset. As shown in Table 7, we compare the top-1 and top-5 accuracy of one candidate by holding the other two fixed as the best value we found during the experiment, *i.e.*, $\alpha = 0.1$, $\beta = 0.01$ and $\gamma = 0.1$. We can see that even though our method achieved better top-1 and top-5 accuracies regardless of the coefficients setting compared with the existing solutions as shown in our main text, a proper one still serves as an important factor for further improvement.