ALFA – Leveraging All Levels of Feature Abstraction for Enhancing the Generalization of Histopathology Image Classification Across Unseen Hospitals

Milad Sikaroudi[1], Maryam Hosseini[1], Shahryar Rahnamayan[2], H.R. Tizhoosh[1,3]

[1]Kimia Lab, University of Waterloo, ON, Canada
[2] Department of Engineering, Brock University, ON, Canada
[3]Rhazes Lab, Mayo Clinic, Rochester, MN, USA

Abstract

We propose an exhaustive methodology that leverages all levels of feature abstraction, targeting an enhancement in the generalizability of image classification to unobserved hospitals. Our approach incorporates augmentation-based self-supervision with common distribution shifts in histopathology scenarios serving as the pretext task. This enables us to derive invariant features from training images without relying on training labels, thereby covering different abstraction levels. Moving onto the subsequent abstraction level, we employ a domain alignment module to facilitate further extraction of invariant features across varying training hospitals. To represent the highly specific features of participating hospitals, an encoder is trained to classify hospital labels, independent of their diagnostic labels. The features from each of these encoders are subsequently disentangled to minimize redundancy and segregate the features. This representation, which spans a broad spectrum of semantic information, enables the development of a model demonstrating increased robustness to unseen images from disparate distributions. Experimental results from the PACS dataset (a domain generalization benchmark), a synthetic dataset created by applying histopathology-specific jitters to the MHIST dataset (defining different domains with varied distribution shifts), and a Renal Cell Carcinoma dataset derived from four image repositories from TCGA, collectively indicate that our proposed model is adept at managing varying levels of image granularity. Thus, it shows improved generalizability when faced with new, out-of-distribution hospital images. The code is available at: https://github.com/miladsikaroudi/ALFA.

1. Motivation

In Domain Generalization (DG), domain alignment techniques are popular [32, 17, 7], as they aim to minimize dif-
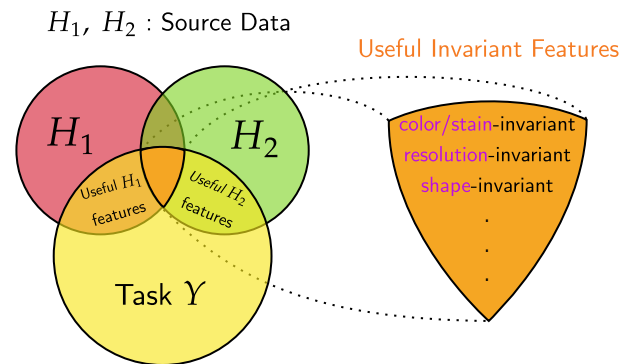

Figure 1. The Venn diagram delineates the feature space of the source hospitals (H_1 and H_2). The yellow area demarcates the label space employed in the classification task. The shared space between feature and label spaces underscores the features that are conducive to executing tasks within the label space.

ferences among source domains to learn domain-invariant features that can withstand unforeseen shifts in the target domain [30].

Ignoring domain-specific information in favor of domain-invariant features may not always lead to the best generalization performance, as noted by Mancini et al. [19] and Shankar et al. [27]. Bui et al. [3] proposed the meta-Domain Specific Domain Invariant (mDSDI) method and provided a mathematical proof for that.

In the field of histopathology, various domains are often represented by different hospitals [24, 28]. In such settings, biases or distribution shifts can emerge due to the process of sample collection, slide preparation, or interpretation [9, 10]. These shifts can potentially create a mismatch between training and testing data. Recognizing these potential biases is paramount for ensuring precise and dependable outcomes in computational pathology [18, 4].
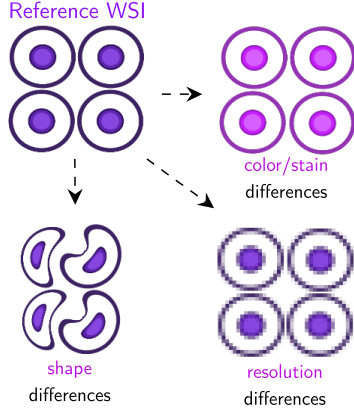
Figure 2. Different distribution shifts happening in digital pathology setups.

Figure 1 illustrates that, apart from the invariant features, there exists a set of unique features specific to each hospital. These features are of significant importance as they can establish a mapping from the feature space to the label space or be useful for performing the task Y.

Although acknowledging these biases may lessen distribution shifts, their total elimination, especially in computational pathology, remains unattainable [2]. Several factors, such as the absence of standardization across pathology labs, the need for diverse inclusion, variable slide preparation guidelines, and inter-observer variability [2], result in *visual* and *histomorphological* differences in histopathology slides. Notably, different races or sexes might display distinct histomorphologic patterns [22], thereby contributing to distribution shifts between training and testing data, as demonstrated in Fig. 2.

As a result, valuable invariant features such as (1) color/stain, (2) resolution, and (3) shape, among others [13], can be identified in digital pathology setups as shown in Fig. 1. The distribution shifts across various hospitals or domains are typically unknown beforehand. However, self-supervision that mimics these differences (through augmentation) can help the model learn representations invariant to these variations.

Self-supervision nudges the model to derive features useful for predicting the data itself, independent of any prevalent hard labels [21, 6]. Consequently, this can lead to the capture of more fundamental perceptual information like edges, corners, and textures in digital images [5]. The amalgamation of these self-supervised representations with other invariant and specific ones allows for a spectrum of representations, covering all levels of *feature abstraction*.

This chapter proposes a new method that extracts disentangled feature abstractions at all levels to enhance the model's ability to generalize to new data from different hospitals/domains. Accordingly, the main contributions are:

- The proposed approach, called ALFA, is an extension of the mDSDI [3] technique for DG. ALFA disentangles the components of Self-Supervised Learning (SSL), domain-invariant, and domain-specific representations to reduce redundancy and improve generalization to unseen target data. By exploiting all levels of feature abstraction, ALFA strives to fully utilize the available information in the dataset.

- The mDSDI [3] approach utilizes adversarial training to extract domain-invariant features, but it can be unstable due to a non-differentiable step (gradient reversal). Therefore, a loss function called "*soft class-domain alignment*" is proposed to minimize the average divergence between two domain probability distributions and a target probability distribution representing a soft class label for each class. This loss function provides better stability during optimization and more distinct latent space for the representation.

- To evaluate the effectiveness of the proposed improvement, we conduct experiments on two public datasets: the PACS [16] benchmark for DG and a Renal Cell Carcinoma (RCC) subtyping task extracted Whole Slide Images (WSI) of The Cancer Genome Atlas (TCGA) [14] data portal.

2. Methods

The key concepts underlying the invented approach have been discussed up to this point. In this section, the specific details of how ALFA is implemented will be delved into. A visual representation and a high-level overview of ALFA framework are provided in Fig. 3. It should be noted that the symbols "$\oplus$" and "$\circ$" have been employed as concatenation and union operators, respectively.

Several components are included in the invented integrated network: (1) an SSL representation $z_\alpha^I = \alpha(I; \theta_\alpha)$, with α being the SSL encoder that is parameterized by θ_α; (2) a domain-invariant representation $z_\beta^I = \beta(I; \theta_\beta)$, parameterized by θ_β, which serves as the domain invariant feature extractor; (3) domain-specific representation $z_\gamma^I = \gamma(I; \theta_\gamma)$, where γ stands for the domain-specific feature extractor parameterized by θ_γ; (4) a domain aligner, parameterized by θ_{Δ_β}, denoted as $\Delta_\beta(z_\beta^I; \theta_{\Delta_\beta}) : z_\beta^I \rightarrow \overline{1 : N_c}$; (5) a domain classifier, parameterized by θ_{Δ_γ}, represented as $\Delta_\gamma(z_\gamma^I; \theta_{\Delta_\gamma}) : z_\gamma^I \rightarrow \overline{1 : N_h}$, where N_c and N_h refer to the number of classes, and the number of participating hospitals/domains in the training, respectively; (6) a regular classifier, parameterized by θ_c, i.e., $\Delta_c(z_\alpha^I \oplus z_\beta^I \oplus z_\gamma^I; \theta_c) : z_\alpha^I \oplus z_\beta^I \oplus z_\gamma^I \rightarrow \overline{1 : N_c}$. The hospital and images sample spaces are represented by $\mathcal{H}$ and $\mathcal{I}$ respectively. Images,

or I, are denoted by their target labels, or y, and hospital labels, or h, as (I, y, h).

2.1. Phase I: Extracting different levels of feature abstraction

Phase I consists of multiple steps: Histopathology-tailored self-supervised learning, determining general features across hospitals, finding hospital-specific features, feature pair disentanglement, establishing the classification loss, and finally inference and training.

2.1.1 Construction of Histopathology-tailored Self-supervision Representations: Capturing the "Synthetic-Invariant Features"

At the core of "Phase I" is the training of an SSL feature extractor, represented by $\alpha(I; \theta_\alpha)$. This component is designed to discern and learn generalizable features, which will become instrumental for accomplishing downstream tasks. The learning process employs a technique known as a triplet loss, a methodology similar to the one detailed in the study by Wang et al. [33].

The first step in this learning process involves the creation of pseudo-classes. These are generated by applying a set of transformations, denoted as $\mathcal{T}$, to a single image. This leads to the production of an augmented image, symbolized as $I_t = t(I) | t \sim \mathcal{T}$.

An image originating from a different pseudo-class, denoted as I_d, is also utilized in the training process. Together, these images serve as training inputs for the network, which employs the following triplet loss formula:

$$\mathcal{L}_{\text{SSL}} = \max(||z_\alpha^I - z_\alpha^{I_t}||_2 - ||z_\alpha^I - z_\alpha^{I_d}||_2 + M, 0), \quad (1)$$

In this equation, $||.||_2$ signifies the Euclidean distance and M represents the margin, which is set heuristically at 1.5. This formula guides the network's learning process by maximizing the intra-class similarity and minimizing the inter-class similarity. The distribution shift is not known a priori but the most prevalent shift according to Fig. 2 are color/stain or shape or resolution differences. Accordingly, transformations set $\mathcal{T}$ includes HED jitter (a histopathology tailored jitter that works on HED channel instead of RGB channels) with a jitter parameter of $\theta = 0.05$, as prescribed by Tellez et al., 2018 [31]. Additionally, random affine transformations are applied, including rotation (varying between -10 and 10 degrees), translation (ranging from 0 to 0.1), and shear (from -1 to 1) in both x and y directions. For the resolution differences, we used resizing images to a lower resolution, then resizing them back to their original size, effectively achieving pixelation. These transformations contribute to the generation of diverse pseudo-classes, providing a wide range of inputs to strengthen the learning process. In Eq. 1, the margin, represented by M,

is set to 1.5. The mining strategy employed is semi-hard mining, for which the margin is specified as 0.7.

2.1.2 Hospital-invariant representations: "general features across hospitals"

Aligning class relationships across hospitals promotes more transferable knowledge for model generalization compared to individual hard label prediction [7]. This study aims to impose an overall pattern of retrieved features that represent the intrinsic similarity between the semantic structures of different classes. Soft labels and class labels are used for consistency across domains.

– Soft Confusion Matrix: Let $\bar{z}_c^{(k)}$ denote the mean of class c in the domain k in the embedding space. We use *softmax* activation for representing the probability of belonging to classes as $s_c^{(k)} = softmax(\Delta_\beta(\bar{z}_c^{(k)})/\tau)$, where $\tau > 1$ is the temperature. The group of soft labels, $[s_c^{(k)}]_{c=1}^C$, serves as a form of *soft confusion matrix* associated with a particular domain/hospital. By combining the class labels with this soft confusion matrix, our goals, i.e., domain alignment and retaining the class relationship, can be fulfilled.

– Soft Class Label Injection: In addition to this soft confusion matrix, proposed in [7], we have injected the soft class labels through a discrete probability density function that represents each class as follows

$$p_c := \{i | i = \delta_i^c + (1 - \delta_i^c)(\frac{\zeta}{n_c - 1})\}, \quad (2)$$

where δ_i^c is the Kronecker delta which is defined as $\delta_i^c := \begin{cases} 1 & \text{if } i = c \\ 0 & \text{otherwise} \end{cases}$, and $\zeta \lessapprox 1$ is a constant value indicating a high probability value (e.g., $\zeta = 0.9$), and n_c is the number of classes.

Overall, the aim is to minimize the average divergence [26] over all the C classes between three distributions: two arbitrary hospitals/domains samples drawn from the two training hospitals' images, $(I_{h_1}, ., h_1) \sim \mathcal{H}_1, (I_{h_2}, ., h_2) \sim \mathcal{H}_2$, and the probability distribution of each class, defined in Eq. 2.

The *Soft Class-Domain Alignment* loss which serves as the domain-invariant loss in our design is defined as

$$\mathcal{L}_i((I_{h_1}, ., h_1), (I_{h_2}, ., h_2); \theta_\beta \circ \theta_{\Delta_\beta}) :=$$
$$\frac{1}{C} \sum_{c=1}^C \frac{1}{6} \Big[D_{\text{KL}}(s_c^{(h_1)} \| s_c^{(h_2)}) + D_{\text{KL}}(s_c^{(h_2)} \| s_c^{(h_1)}) +$$
$$D_{\text{KL}}(p_c \| s_c^{(h_2)}) + D_{\text{KL}}(s_c^{(h_2)} \| p_c) + \quad (3)$$
$$D_{\text{KL}}(s_c^{(h_1)} \| p_c) + D_{\text{KL}}(p_c \| s_c^{(h_1)}) \Big],$$

where $D_{\text{KL}}(p \| q) = \sum_r p_r \log \frac{p_r}{q_r}$, and $\theta_\beta \circ \theta_{\Delta_\beta}$ is the union of all the parameters for β feature extractor and Δ_β classifier.
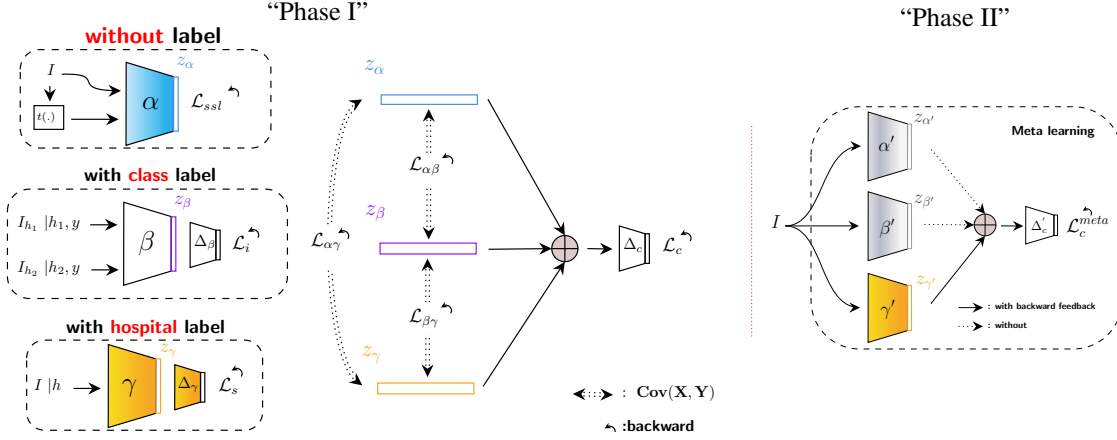
Figure 3. ALFA has two phases: In Phase I, three feature extractors extract different levels of feature abstraction, and disentangled features are concatenated for classification. In Phase II, updated feature extractors' representations are concatenated and fed into the updated classifier to update parameters in a Meta-learning fashion while α' and β' feature extractors remain frozen.

The inclusion of our histopathology-tailored SSL features as a level of feature abstraction could raise the question of how distinct SSL features are from the hospital-invariant features in our design. The subsequent theorem and lemma aim to provide clarity on this matter:

Theorem 1. *Given transformations T_1 for α (synthetic distribution shifts assuming in the self-supervision pretext task) and T_2 for differences of hospitals (due to the real distribution shift across domains/ hospitals), and an optimization objective to minimize the covariance between z_α and z_β obtained respectively by T_1 (explicit data augmentation) and T_2 (implicit changes due to sources of distribution shifts), are distinct and uncorrelated in the feature space and both contribute unique information for mapping from feature space to label space.*

Lemma 1. *The following assertions hold for the features $z_\alpha = \alpha(T_1(I))$ and $z_\beta = \beta(T_2(I))$ for a given image I:*

1. *z_α and z_β are uncorrelated:*

   If the optimization objective successfully minimizes the covariance, the covariance between z_α and z_β, $Cov(z_\alpha, z_\beta)$, will be close to zero. This indicates that z_α and z_β are uncorrelated, i.e., changes in z_α do not predict changes in z_β and vice versa.

2. *z_α and z_β contribute unique information to the mapping:*

   Let us consider a mapping function M that maps the feature space to the label space. For z_α and z_β, the mapping function M can be written as $M(z_\alpha, z_\beta)$. If z_α and z_β are uncorrelated, removing one from the mapping will reduce the information provided by M. That is, $M(z_\alpha, \emptyset) \neq M(z_\alpha, z_\beta)$ and $M(\emptyset, z_\beta) \neq M(z_\alpha, z_\beta)$.

Therefore, given transformations T_1 for α and T_2 for domain shifts, and an optimization objective to minimize the covariance between the resulting augmentation-based self-supervised features and invariant features are distinct and uncorrelated in the feature space and both contribute unique information for mapping from feature space to label space.

2.1.3 Hospital-specific representations: "least general or specific features"

To extract the most specific features, similar to [3], $\gamma(I; \theta_\gamma)$ is used for feature extraction followed by the Δ_γ domain classifier that is trained in a supervised manner using cross-entropy loss to predict the domain/hospital label:

$$\mathcal{L}_s := -\mathbb{E}_{(I,.,h)\sim\mathcal{I}} h \log \Delta_\gamma(z_\gamma^I; \theta_{\Delta_\gamma}). \qquad (4)$$

2.1.4 Disentanglement loss between pairs of extracted features

To prevent redundancy and ensure diversity in our feature extractors, we need to disentangle their resulting representations from each other. This can be achieved by zeroing the covariance matrix between pairs of random vectors, such as z_a and z_b. A zero-covariance matrix indicates that the variables are independent and have no correlation or effect on each other. To enforce this disentanglement, we define pairwise covariance loss functions between each pair of α, β, and γ feature extractors' representations as

$$\mathcal{L}_{\alpha\beta} := -\mathbb{E}_{I\sim\mathcal{I}}\big[|| \operatorname{Cov}(z_\alpha^I, z_\beta^I)||_2\big], \qquad (5)$$

$$\mathcal{L}_{\alpha\gamma} := -\mathbb{E}_{I\sim\mathcal{I}}\big[|| \operatorname{Cov}(z_\alpha^I, z_\gamma^I)||_2\big], \qquad (6)$$

$$\mathcal{L}_{\beta\gamma} := -\mathbb{E}_{I\sim\mathcal{I}}\big[|| \operatorname{Cov}(z_\beta^I, z_\gamma^I)||_2\big]. \qquad (7)$$
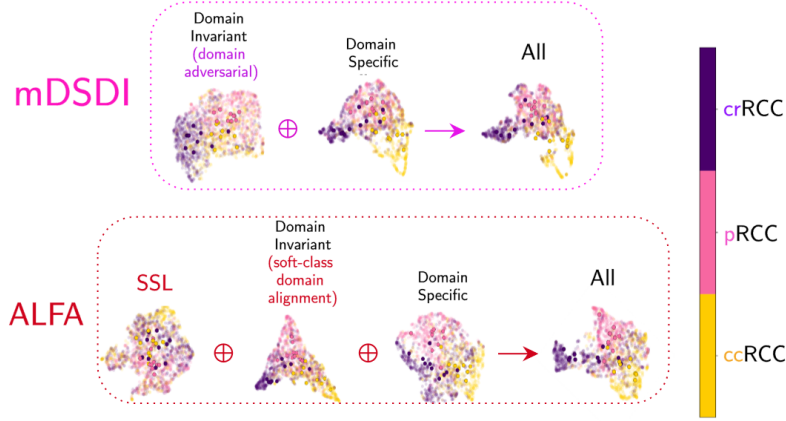
Figure 4. 2D feature embeddings for the feature extractors in mDSDI [3] versus in ALFA: (target hospital: 'NCI'). 'All' is the concatenation of domain-specific and domain-invariant representations for the mDSDI [3] (up), and SSL, domain-invariant, and domain-specific representations for ALFA (bottom). Opaque-shaded scatters are WSIs representations obtained by averaging on patches' representations (transparent-shaded).

2.1.5 Classification loss using aggregation of extracted features

The goal of the classifier $\Delta_c(z_\alpha^I \oplus z_\beta^I \oplus z_\gamma^I; \theta_c) : z_\alpha^I \oplus z_\beta^I \oplus z_\gamma^I \to \overline{1 : N_c}$ is to classify the images according to their hard class labels using a concatenation of all the extracted features:

$$\mathcal{L}_c := -\mathbb{E}_{(I,y,.)\sim\mathcal{I}} \left[y \log \Delta_c(z_\alpha^I \oplus z_\beta^I \oplus z_\gamma^I; \theta_c) \right], \quad (8)$$

where y is the target class label for image I.

2.1.6 Inference and training in Phase I –

Using all loss functions, the feature extractors and modules are updated via

$$\mathcal{L}_{\text{total}} := a_1\mathcal{L}_{\text{SSL}} + a_2\mathcal{L}_i + a_3\mathcal{L}_s + \quad (9)$$
$$a_4\mathcal{L}_{\alpha\beta} + a_5\mathcal{L}_{\alpha\gamma} + a_6\mathcal{L}_{\beta\gamma} + a_7\mathcal{L}_c,$$

where a_i coefficients are selected as balanced parameters between loss functions and all were set to 1.0 in our experiments as the loss values were in the same range. Through backward the total loss, i.e., $\mathcal{L}_{\text{total}}$, the updated encoders, i.e., $\alpha'(I;\theta_{\alpha'})$, $\beta'(I;\theta_{\beta'})$, $\gamma'(I;\theta_{\gamma'})$, and classifiers $\Delta'_\beta(I;\theta_{\Delta'_\beta})$, $\Delta'_\gamma(I;\theta_{\Delta'_\gamma})$ and $\Delta'_c(I;\theta_{\Delta'_c})$ are obtained.

2.2. Phase II: Meta-learning for generalization improvement

To adapt the domain-specific representation z_γ to the target domain using information from source domains, we use the same meta-learning framework as [3]. The α' and β' feature extractors remain frozen while the γ' feature extractor and Δ'_c classifier are updated. We aim to update

$\omega = \theta_{\gamma'} \circ \theta_{\Delta'_c}$ through meta-learning by dividing each hospital data $\mathcal{H}_k$ into disjoint meta-train $\mathcal{H}_k^{tr}$ and meta-test $\mathcal{H}_k^{te}$ sets and the objective is to

$$\min_\omega \quad \mathcal{L}_c^{\text{meta}} := f(\omega - \nabla f(\omega, \mathcal{H}_k^{tr}), \mathcal{H}_k^{te}), \quad (10)$$

where

$$f(\omega = \theta_{\gamma'} \circ \theta_{\Delta'_c}, \mathcal{H}_k) = \quad (11)$$
$$- \mathbb{E}_{(I_k,y_k,k)\sim\mathcal{H}_k} \Big[$$
$$y_k \log \Delta'_c(z_{\alpha'}^{I_k} \oplus z_{\beta'}^{I_k} \oplus \gamma'(I_k,\theta_{\gamma'}); \theta_{\Delta'_c})$$
$$\Big],$$

where y_k and k are the target class label and hospital label, respectively, for image I_k.

3. Experiments and Results

The study evaluates the effectiveness of the proposed method, ALFA, against mDSDI [3], HA [29], and ERM through a leave-one-*domain/hospital*-out evaluation using data from multiple hospitals/domains. The evaluation includes reporting "accuracy" for the target (hold-out) domain/hospital, as well as "AUROC" and "recall" metrics for RCC subtyping, which is important for cancer diagnosis tasks.

3.1. Datasets

In this study, we evaluated our approach using three unique benchmarks: PACS, a Domain Generalization benchmark; synthetic-MHIST, a synthetic histopathology dataset mimicking staining variances, derived from the MHIST dataset by adding HED jitter; and a task involving Renal Cell Carcinoma (RCC) subtyping (Fig. 6).
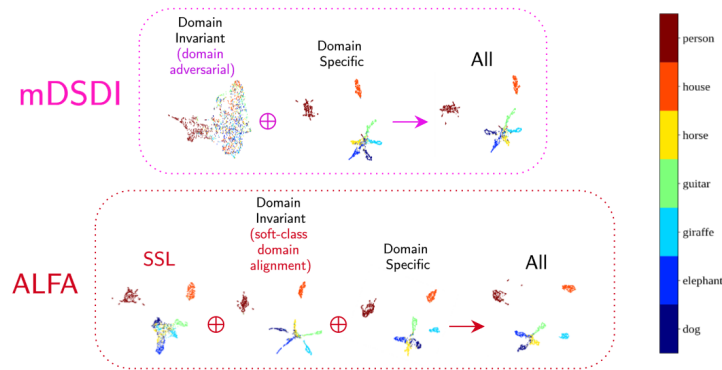
Figure 5. 2D feature embedding for the feature extractors in mDSDI [3] (upper row) versus in ALFA (ours) (bottom row), target domain: 'Photo' on PACS.

– **PACS:** P(hoto), A(rt), C(artoon), S(ketch) is a benchmark for DG on natural images. It has been created by intersecting the classes in Caltech256 (Photo), Sketchy (Photo,Sketch) [23], TU-Berlin (Sketch) [8] and Google Images (Art painting, Cartoon, Photo). This benchmark includes four domains (Photo, Sketch, Cartoon, Painting), and 7 common categories 'dog', 'elephant', 'giraffe'. 'guitar', 'horse', 'house', and 'person' with a total 9991 images.

– **Sythetic-MHIST**: MHIST [34] is a histopathology dataset containing two different class labels, i.e. Hyperplastic Polyp (HP), and Sessile Serrated Adenoma (SSA) with a total of 3152 images. It involves the distinction between the clinically-important binary categories of HPs and SSAs, which is a challenging problem with considerable pathologists' inter-variability. In this study, we synthetically created four different domains by applying HED jitter borrowed from [31] with different degrees to this tiny dataset. HED jitter is a histopathology-tailored color jitter that applies random distortions in the HED color space.

– **RCC subtyping dataset from TCGA:** The RCC dataset [29, 12], comprises patches of various Renal Cell Carcinoma (RCC) cancer subtypes collected from five different hospitals. Due to the absence of certain cancer subtypes, two of the hospitals' data have been merged. The dataset comprises 4 image repositories: (1) H-MD, (2) MSKCC, (3) IGC, and (4) NCI from TCGA. The dataset contains $\approx 70k$ patches of size 224×224.

3.2. Experimental Setup

The backbone of all feature extractors was the ResNet18 [11], pre-trained on ImageNet [15], with all of its batch normalization layers frozen as per the guidelines given in [25]. All features were embedded to a size of 512. The Adam optimizer was employed with an initial learning rate of 5e-5. A batch size of 32 was established and set the maximum number of iterations to 3000.
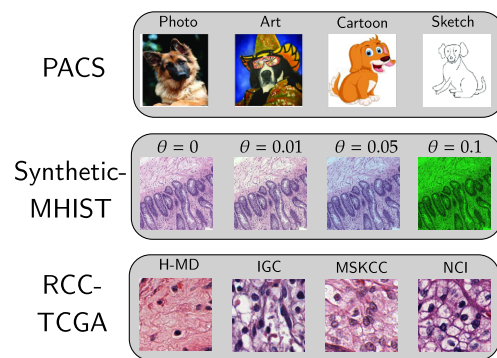

Figure 6. Different datasets and their domains in this study.

3.3. Results

– **Losses convergence:** During the training, it was found that γ feature extractor was dominating over other feature extractors and potentially causing a dampening effect on their contributions to the model's overall performance. Hence, we added layer normalization [1] whenever the representations are concatenated to address this issue. With this modification, all losses converged almost simultaneously according to Fig. 7.

– **Low-dimensional Embedding Visualization:** In accordance with the best practices suggested in the original UMAP paper [20], PCA [35] was applied to obtain the first 50 principal components, followed by UMAP [20] for further dimensionality reduction to 2. According to Figs. 7 and 4, ALFA's domain-invariant approach yields a more powerful discriminatory representation for different RCC subtypes or different categories on PACS, compared to mDSDI [3], for domain-specific representations. In other words, ALFA's domain-invariant encoder learns some features that are also learned by mDSDI's domain-specific encoder. ALFA's SSL representation provides useful representation, which seems even better than mDSDI's domain-invariant features according to these figures.
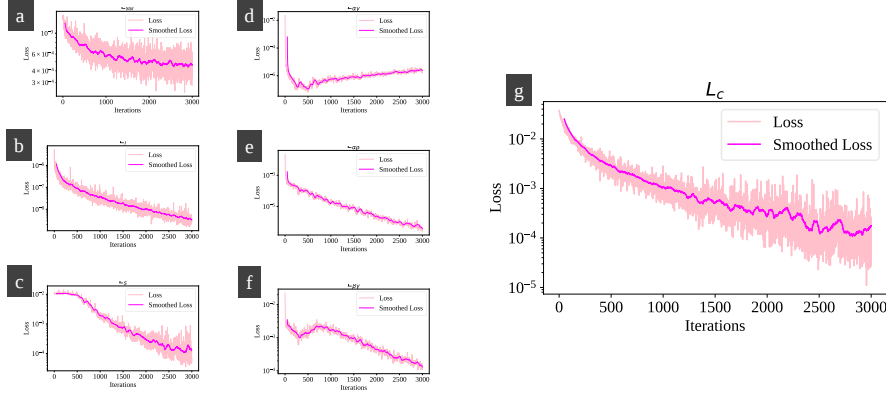
Figure 7. The behavior of ALFA's loss functions during the training when the hold-out set was Art on PACS. (a) $\mathcal{L}_{\text{SSL}}$, (b) $\mathcal{L}_i$, (c) $\mathcal{L}_s$, (d) $\mathcal{L}_{\alpha\gamma}$, (e) $\mathcal{L}_{\alpha\beta}$, (f) $\mathcal{L}_{\beta\gamma}$, (g) $\mathcal{L}_c$

3.3.1 PACS dataset classification task

The accuracy of mDSDI [3] and ALFA applied on PACS have been reported in Table 1. It can be seen in Table 1, except for the 'Sketch' with a high semantic shift in comparison to the rest of the target domains, ALFA outperforms mDSDI [3] with an average accuracy of $83.75\pm6.72\%$ compared to mDSDI's [3] average accuracy of $80.34 \pm 5.60\%$. According to this, ALFA cannot only be effective for generalization to unseen hospitals but it can also be effective for DG tasks for natural images.

Table 1. Results on PACS dataset

| Target | Source | Accuracy (%) | | |
		ERM	mDSDI [3]	ALFA(ours)
Photo	{A,C,S}	91.98	90.06	96.15
Art	{P,C,S}	76.85	76.27	83.10
Cartoon	{P,A,S}	74.87	76.20	78.71
Sketch	{P,A,C}	76.76	78.85	78.41
Average		80.11 ± 6.76	80.34 ± 5.60	84.09 ± 7.06

3.3.2 Synthetic-MHIST classification task

Considering the results on the synthetic-MHIST dataset in Table 2, we can see that ALFA (ours) outperforms mDSDI [3] in all cases except for when the target domain is $\theta = 0.05$. On average, ALFA achieves higher accuracies, with an average accuracy of $84.17 \pm 3.17\%$ compared to mDSDI's average accuracy of $82.65 \pm 5.02\%$. For the case when $\theta = 0.5$ is the target domain, which represents the most challenging target domain with the highest degree of corruption and accordingly highest distribution shift in comparison to the rest of the source domains, the ALFA has a significant improvement over mDSDI and ERM. This can be attributed to the fact that incorporating self-supervision representations, as ALFA does, can lead to improved performance in generalization to unseen domains compared to using only specific and domain-invariant features, as mDSDI does.

Table 2. Results on Synthetic-MHIST dataset

Target	Source	ERM	mDSDI	ours
$\theta = 0$	$\theta = \{0.01, 0.05, 0.5\}$	84.65	85.16	86.39
$\theta = 0.01$	$\theta = \{0, 0.05, 0.5\}$	84.75	86.18	86.28
$\theta = 0.05$	$\theta = \{0, 0.01, 0.5\}$	85.26	85.47	85.47
$\theta = 0.5$	$\theta = \{0, 0.01, 0.05\}$	75.64	73.80	78.6
Average		82.57 ± 3.93	82.65 ± 5.02	84.17 ± 3.17

3.3.3 RCC subtyping classification task

The accuracy of mDSDI [3], HA [29] approach, and ALFA applied on RCC subtyping task has been reported in Table 3.

In the context of H-MD, ALFA (ours) exceeds the performance of both ERM and mDSDI, achieving an accuracy of 65.52% as opposed to 72.49% and 51.72% respectively. However, it falls short when compared to the HA [29] method, which reaches 75.29% accuracy. This outcome may indicate that, considering H-MD encompasses two distinct data sources, methods that focus on extracting invariant features are more effective. In this regard, HA [29], which specifically targets hospital-invariant features, outperforms both ALFA and mDSDI.

For the IGC dataset, both ALFA and mDSDI attain the same accuracy of 86.21%, outperforming ERM and HA [29], which scored 75.86% and 70.42%, respectively. Given that both ALFA and mDSDI leverage a combination of invariant and hospital-specific features, it's reasonable to speculate that the inclusion of specific features has enhanced their overall performance in this context. In essence, the unique features sourced from the different hospitals appear to have provided valuable information that contributed to improved performance on the IGC dataset.

At the NCI, ALFA again leads with an accuracy of 86.36%, outperforming ERM (81.82%), mDSDI [3] (72.73%), and HA [29] (83.38%).

In the case of MSKCC, ALFA's performance is notably lower with an accuracy of 84.69%, when compared to both ERM (86.73%) and HA [29] (88.19%). However, ALFA

Table 3. Results on RCC subtyping task

		Accuracy (%)				AUROC (%)				Recall (%)			
Target	Source	ERM	mDSDI [3]	HA [29]	ALFA	ERM	mDSDI [3]	HA [29]	ALFA	ERM	mDSDI [3]	HA [29]	ALFA
IGC	{NCI, MSKCC, H-MD}	75.86	86.20	70.42	86.21	93.23	95.78	88.36	95.33	57.14	82.88	62.38	85.39
NCI	{IGC, MSKCC, H-MD}	81.82	72.73	83.38	86.36	96.49	94.46	97.32	97.83	83.08	71.46	85.48	86.41
MSKCC	{IGC, NCI, H-MD}	86.73	85.71	88.19	84.69	95.91	95.89	96.47	95.99	82.99	87.05	85.32	87.99
H-MD	{IGC, NCI, MSKCC}	72.49	51.72	75.29	65.52	85.38	88.37	90.16	90.48	72.96	51.85	78.42	66.67
Average		79.22 ±5.36	74.09 ±13.72	79.32 ±6.49	80.69 ±8.61	92.75 ±4.34	93.62 ±3.02	93.08 ±3.34	94.90 ±2.66	74.07 ±10.38	73.31 ±13.36	77.90 ±8.44	81.62 ±8.50

does surpass mDSDI, which has an accuracy of 85.71%. This result might suggest that invariant features are more conducive to effective generalization in this context than specific features. Therefore, both mDSDI and ALFA, which rely on a combination of site-specific and site-invariant features, may not perform as well as HA [29] or even ERM.

According to this table, ALFA outperforms mDSDI [3], HA [29], and ERM with average accuracy, AUROC, and recall of $80.69 \pm 8.61\%$, $94.90 \pm 2.66\%$, and $81.62 \pm 8.50\%$, respectively. ALFA performed similarly to mDSDI and ERM in terms of AUROC and accuracy, but slightly better overall. However, ALFA significantly outperformed other methods in terms of recall metric, especially for the "NCI", and "IGC" target hospitals.

3.3.4 Ablation study

In the conducted ablation study, each element of ALFA demonstrated its efficacy in improving overall generalization, as illustrated in Table 4. We assessed feature extractors, labeled as α, β, and γ, in various active or inactive states. Performance was evaluated across four subsets from PACS and RCC, with an average accuracy score calculated. When only α was active, the average accuracy was lowest, at 48.63 ± 14.98 for PACS and 46.50 ± 12.20 for RCC. There was a substantial increase in average accuracy to around 78.88 ± 6.99 for PACS and similar for RCC, when either β or γ was activated while the others were inactive. The highest average accuracies, 84.09 ± 7.06 for PACS and 80.70 ± 8.99 for RCC, were achieved when multiple feature extractors were activated concurrently, highlighting the effectiveness of the complete ALFA method.

4. Conclusions

In this study, we introduced an inventive methodology named ALFA (All Levels of Features Abstraction), which exploits multiple feature extractors to disassociate features at disparate levels of abstraction. The efficiency of ALFA is substantiated across a range of benchmarks, such as the commonly utilized DG benchmark PACS, a synthetic-MHIST dataset, and Renal Cell Carcinoma subtype identification utilizing the TCGA database. ALFA successfully

Table 4. Ablation on PACS and RCC: Active feature extractor(s) is/are blue, Deactivated one(s): gray

PACS	Photo	Art	Cartoon	Sketch	Average
α, β, γ	69.70	53.17	44.28	27.38	48.63 ± 14.98
α, β, γ	91.13	73.33	75.68	75.38	78.88 ± 6.99
α, β, γ	94.43	77.34	72.05	71.46	78.82 ± 9.11
α, β, γ	90.11	75.48	76.23	75.99	79.45 ± 6.03
α, β, γ	94.73	78.61	73.93	72.02	79.82 ± 8.75
α, β, γ	95.38	82.03	78.37	78.67	83.61 ± 6.80
ALFA	96.15	83.10	78.71	78.41	84.09 ± 7.06

RCC	IGC	NCI	MSKCC	H-MD	Average
α, β, γ	62.34	48.23	40.12	35.30	46.50 ± 12.20
α, β, γ	77.43	62.34	64.28	64.08	67.03 ± 6.28
α, β, γ	80.34	65.78	61.38	61.10	67.15 ± 8.11
α, β, γ	76.56	63.84	64.21	64.11	67.18 ± 5.99
α, β, γ	80.23	66.41	62.43	60.98	67.51 ± 7.85
α, β, γ	80.77	72.43	73.11	64.42	72.68 ± 5.67
ALFA	86.21	86.36	84.69	65.52	80.70 ± 8.99

surpassed the existing state-of-the-art mDSDI approach in performance on PACS, underscoring its superior capabilities. Moreover, in comparison to mDSDI and ERM, ALFA showcased exceptional capability to generalize to previously unseen hospitals. In our approach, we employed a straightforward augmentation-based self-supervision technique that was tailored for histopathology, enabling the acquisition of invariant features. Concurrently, we introduced a novel domain alignment loss function, termed "soft class-domain alignment" loss. This technique skillfully extracts domain-invariant features, aligning data from two disparate sources with soft class labels serving as the reference distribution, thereby enhancing our design significantly. As advancements continue in the realm of self-supervision, the incorporation of more effective SSL features could potentially augment our design even further.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Andrew H Beck, Ankur R Sangoi, Samuel Leung, Robert J Marinelli, Torsten O Nielsen, Marc J Van De Vijver, Robert B West, Matt Van De Rijn, and Daphne Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science translational medicine*, 3(108):108ra113–108ra113, 2011.

[3] Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems*, 34:21189–21201, 2021.

[4] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

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

[7] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, pages 6450–6461, 2019.

[8] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012.

[9] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[10] Metin N Gurcan, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147–171, 2009.

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

[12] Milad Hosseini, S Maryamand Sikaroudi, Morteza Babaie, and HR Tizhoosh. Proportionally fair hospital collaborations in federated learning of histopathology images. *IEEE Transactions on Medical Imaging*, 2023.

[13] Frederick M Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I Olopade, Jakob N Kather, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature communications*, 12(1):4423, 2021.

[14] National Cancer Institute. The cancer genome atlas. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga, n.d. Accessed: [insert date accessed].

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[16] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

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

[18] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical image analysis*, 33:170–175, 2016.

[19] Massimiliano Mancini, Samuel Rota Bulo, Barbara Caputo, and Elisa Ricci. Best sources forward: domain generalization through source-specific nets. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 1353–1357. IEEE, 2018.

[20] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[21] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[22] Prashanth Rawla. Epidemiology of prostate cancer. *World journal of oncology*, 10(2):63, 2019.

[23] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016.

[24] Marin Scalbert, Maria Vakalopoulou, and Florent Couzinié-Devy. Test-time image-to-image translation ensembling improves out-of-distribution generalization in histopathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 120–129. Springer, 2022.

[25] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 68–83. Springer, 2020.

[26] Andrea Sgarro. Informational divergence and the dissimilarity of probability distributions. *Calcolo*, 18(3):293–302, 1981.

[27] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

[28] Milad Sikaroudi, Maryam Hosseini, Ricardo Gonzalez, Shahryar Rahnamayan, and HR Tizhoosh. Generalization of vision pre-trained models for histopathology. *Scientific reports*, 13(1):6065, 2023.

[29] Milad Sikaroudi, Shahryar Rahnamayan, and Hamid R Tizhoosh. Hospital-agnostic image representation learning in digital pathology. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3055–3058. IEEE, 2022.

[30] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[31] David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, et al. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9):2126–2136, 2018.

[32] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[33] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015.

[34] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Mustafa Nasir-Moin, Naofumi Tomita, et al. Learn like a pathologist: curriculum learning by annotator agreement for histopathology image classification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2473–2483, 2021.

[35] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.