Okapi: Generalising Better by Making Statistical Matches Match

Myles Bartlett[1]* **Sara Romiti**[1] **Viktoriia Sharmanska**[1,2] **Novi Quadrianto**[1,3,4]

[1]Predictive Analytics Lab, University of Sussex [2]Imperial College London
[3]BCAM Severo Ochoa Strategic Lab on Trustworthy Machine Learning
[4]Monash University, Indonesia

Abstract

We propose *Okapi*, a simple, efficient, and general method for robust semi-supervised learning based on online statistical matching. Our method uses a nearest-neighbours-based matching procedure to generate cross-domain views for a consistency loss, while eliminating statistical outliers. In order to perform the online matching in a runtime- and memory-efficient way, we draw upon the self-supervised literature and combine a memory bank with a slow-moving momentum encoder. The consistency loss is applied within the feature space, rather than on the predictive distribution, making the method agnostic to both the *modality* and the *task* in question. We experiment on the WILDS 2.0 datasets [63], which significantly expands the range of modalities, applications, and shifts available for studying and benchmarking real-world unsupervised adaptation. Contrary to [63], we show that it is in fact possible to leverage additional unlabelled data to improve upon empirical risk minimisation (ERM) results with the right method. Our method outperforms the baseline methods in terms of out-of-distribution (OOD) generalisation on the iWildCam (a multi-class classification task) and PovertyMap (a regression task) image datasets as well as the CivilComments (a binary classification task) text dataset. Furthermore, from a qualitative perspective, we show the matches obtained from the learned encoder are strongly semantically related. Code for our paper is publicly available at `https://github.com/wearepal/okapi/`.

1 Introduction

Machine learning models have been deployed for safety-critical applications such as disease diagnosis [73] and self-driving cars [77], and in socially important contexts such as the allocation of healthcare, education, and credit (e.g. [23, 35]). Many machine learning algorithms, however, rely on supervision from a large amount of labelled data, and are typically trained to exploit complex relationships and distant correlations present in the training dataset. This strategy has proven to be effective in the setting when we have training (source) and test (target) data that are i.i.d.

In reality, machine learning models are often deployed on target data whose distribution is different from the source distribution they were trained on. For example, in the task of classifying animal species in a camera trap image, one aims to learn a model that can generalise to new camera trap locations despite variations in illumination, background, and label frequencies, given training examples from a limited set of camera trap locations. Exploiting correlations that only hold in these limited locations but not in the new locations can hurt out-of-distribution (OOD) generalisation. While we only have a small subset of camera traps that have their images labelled, we have a large

[*] Corresponding author: `m.bartlett@sussex.ac.uk`.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

amount of unlabelled data from the other camera traps that capture diverse operating conditions. In general, unlabelled data is much more readily available than labelled data and can often be obtained from distributions beyond the source distribution. Taking advantage of these unlabelled data during training is a key element to build robust models that have good OOD performance without sacrificing in-distribution (ID) performance.

Our work is a direct response to the empirical conclusion of [63] for the WILDS 2.0 dataset that existing semi-supervised methods – leveraging the unlabelled data provided in this extended version of the WILDS benchmark [41] – fail to provide consistent benefit over the combination of judicious data-augmentation and standard empirical risk minimisation (ERM). We show that with the right method, however, it is in fact possible to make effective use of large volumes of unlabelled data as supplement to a smaller set of labelled data, from a limited set of domains, to achieve strong generalisation to data from domains outside the training distribution. To develop this method, we turn to a statistical matching (SM) framework [57, 59, 60], a model-based approach for providing joint information on variables and indicators collected through multiple sources. SM has been widely utilised to assess the effect of interventions in numerous fields, such as education, medical and community policies (e.g. [10, 17]). In SM, intervened units are paired with control units and those units without a sufficiently-good match according to a given statistical criteria are excluded when estimating the treatment effect. In the running example of animal-species classification, intervened units may correspond to the limited set of camera trap locations that are fully-annotated, while control units refer to the many more camera trap locations that are only partially annotated. Pairing is beneficial for capturing diverse operating conditions, yet the ability to drop unpaired units is crucial for mitigating the risk of statistically-poor matches corrupting the training signal.

In designing an online method for statistically matching samples from different domains (camera-trap locations) and using this to define a consistency loss, we arrive at our proposed semi-supervised method, *Okapi*. This consistency loss is predicated on the simple idea of pulling together similar samples from different domains within the latent space of the encoder, and using this to bootstrap said encoder such that the distributions become progressively more aligned over the course of training. Since matching samples using the full dataset at each step of training is computationally infeasible, we instead approximate it using a combination of momentum-encoding and a memory-bank that has been well-proven in self-supervised learning [33, 42]. Compared with other consistency-based methods such as FixMatch [67], Okapi has the advantage of being agnostic to both the task and the modality, in addition to being distributionally robust. Contrary, to Sagawa et al. [63], we show that the supplementary unlabelled data and domain information can be leveraged by Okapi to improve upon standard ERM on datasets from the WILDS 2.0 benchmark.

2 Preliminaries

2.1 Problem setting

In the standard supervised setting, one is given a dataset, $\mathcal{D}_l \triangleq \{x_i, y_i\}_{i=1}^{N_l}$, and trains a model, parameterised by θ, to well-approximate the empirical distribution as $p_\theta(y|x)$. Labelled data is limited by the cost of annotation yet one often has access to a far larger corpus of unlabelled data, $\mathcal{D}_u \triangleq \{x_i\}_{i=1}^{N_u}$, which can be used to supplement $\mathcal{D}_l$. Semi-supervised learning is motivated by the idea that this additional data can often be used to improve the ID and/or OOD performance of $p_\theta(y|x)$. We can view unsupervised domain adaptation (UDA) as a special case of semi-supervised learning, where there is assumed to be some distribution shift (adverse to a naïvely-trained predictor) between $\mathcal{D}_l$ and $\mathcal{D}_u$. Here, $\mathcal{D}_u$ comes from the domain on which $p_\theta(y|x)$ is to be evaluated, such that we have $\mathcal{D}_u \triangleq \mathcal{D}_{\text{OOD}}$, where $\mathcal{D}_{\text{OOD}}$ denotes the target domain, that is OOD w.r.t. D_l. In the most general sense, a *domain*, or *environment* [4, 19] describes some partitioning of the data according to its source or some secondary characteristic, such as time of day, weather, location, lighting, or the model of the device used to collect said data; one would hope that a predictor trained under one set of conditions (e.g. day) would perform with minimal degradation under another set of conditions (e.g. night) when those conditions are irrelevant to the task at hand.

Assuming the data follows the conditional generative distribution $x \sim p(x|s)$, where s is the domain label, one would ideally use $\mathcal{D}_{\text{OOD}}$ to learn invariance to the marginal distribution, $p(s)$, and thereby achieve the equivalence $p_\theta(y|x) = p_\theta(y|x, s)$. In practice, one typically does not have access to $\mathcal{D}_{\text{OOD}}$ but does have access to training data sourced from a mixture of domains which can be leveraged to learn a more general invariance that extends to those domains outside the training distribution [4]. Such a learning paradigm is referred to as domain generalisation (DG). While some DG works consider the more extreme case of s being unobserved [19], we follow the more conventional setup [4, 43, 63] in which the domain(s) associated with each sample (labelled and unlabelled) is indicated by the discrete label (set of labels) s. We denote the set of possible domains for the in-distribution labelled and unlabeled data, as $\mathcal{S}_l$ and $\mathcal{S}_u$, respectively, and their union as $\mathcal{S} \triangleq S_l \cup S_u$. Following the setup established in [63], $\mathcal{D}_u$ is assumed to be unlabelled only w.r.t the targets and not w.r.t the domain labels and thus that both $\mathcal{D}_l$ and $\mathcal{D}_u$ can be augmented with the latter to give the re-definitions $\mathcal{D}_l \triangleq \{x_i, y_i, s_i\}_{i=1}^{N_l}$ and $\mathcal{D}_u \triangleq \{x_i, s_i\}_{i=1}^{N_u}$.

2.2 Statistical matching

Statistical matching is a sampling strategy which aims to balance the distribution of the observed covariates in the *treated* and *control* groups. In general terms, observed covariates x are measured characteristics of the samples; in our work we refer to the encodings generated by a deep neural network as covariates instead of the original characteristics. The treated and control groups are two partitions of the data; specifically, the treated group is the set of samples having a specific value of a variable of interest (here, the domain indicator, s) and the control group is its complement.

In this work we utilise Nearest Neighbour (NN) matching, a distance-based matching method that pairs sample i of the treated group with the closest sample j belonging to the control group. A distance measure is used to define how close two samples, i and j, are, with *propensity score distance* (PSD) and *Euclidean distance* being two widely-used distances that we employ here – indirectly (as a means of filtering) and directly, respectively.

The propensity score distance is defined as the difference between propensity scores, e_i and e_j, of samples i and j, i.e. $\mathrm{PSD}(e_i, e_j) \triangleq |e_i - e_j|$. In causal inference, the propensity score refers to the probability of sample i belonging to the treated group, given its covariates x_i [58]; in practice, this conditional probability is rarely known a priori and thus requires estimation, typically via logistic regression [68]. We generalise the notion of a propensity score to categorical domains simply by modelling the conditional probability for each domain, with e_i instead a $|\mathcal{S}|$-dimensional probability vector. The Euclidean-distance approach, in contrast, computes the distance between the covariates, x_i and x_j, of a given pair of samples. Despite PSD being the more prevalent of the two distances, it is ill-suited to cases in where pairs are close in value w.r.t. all covariates and in such cases Euclidean distance should be preferred [40]. Nevertheless, propensity scores remain a relevant component of NN-based matching for defining *calipers* that can reduce the likelihood of false-positive matches.

We make use of two types of caliper, *fixed* and *standard deviation*. The fixed caliper [20], t_f, defines a region of common support between the estimated propensity score distribution of the two groups; only those samples within the feasible region are admissible for matching. For binary problems, the feasible region is symmetric such that we have $\{i \,|\, e_i \in (1 - t_f, t_f)\})$ whereas in the more general, categorical case the constraint is one-sided, i.e. $\{i \,|\, \|e_i\|_\infty < t_f)\}$. This selection rule helps by removing samples with extreme propensity scores. The standard deviation-based caliper (std-caliper), on the other hand, [59] defines the maximum discrepancy permitted between paired two samples. The discrepancy is usually expressed in terms of estimated PSD as $|e_i - e_j| < \sigma \cdot t_\sigma$, where σ denotes the mean of the group-wise standard deviations of the propensity scores and t_σ controls the percentage bias-reduction of the covariates. In the categorical case, we can simply substitute the absolute value for the infinity norm: $\|e_i - e_j\|_\infty < \sigma \cdot t_\sigma$. In the following section, we describe how one can leverage this matching framework to define a consistency loss encouraging inter-domain robustness.

3 Method

Here, we introduce *Okapi*, a simple, efficient, and general (in the sense that it is applicable to any task *or* modality) method for robust semi-supervised learning based on online statistical matching. Our method belongs to the broad family of consistency-based methods, characterised by methods such as FixMatch [67], where the idea is to enforce similarity between a model's outputs for two views of an unlabelled image. These semi-supervised approaches based on minimising the discrepancy between two views of a given sample are closely related with self-supervised methods based on instance discrimination [15] and self-distillation [6, 14, 30]. Many of the methods within this family, however, are limited in applicability due to their dependence on modality-specific transformations and only recently has research into self-supervision sought to redress this problem with modality-agnostic alternatives such as MixUp [72], masking [6], and nearest-neighbours [24, 42, 71]. Approaches such as FixMatch, AlphaMatch [28] and CSSL [46] that enforce consistency between the *predictive* distributions suffer further from not being directly generalisable to tasks other than classification. Okapi addresses both of these aforementioned issues through 1) the use of a statistical matching procedure – that we call CaliperNN and detail in Sec. 3.2 – to generate multiple views for a given sample; 2) enforcing consistency between encodings rather than between predictive distributions.

We show that models trained to maximise the similarity between the encoding of a given sample and those of its CaliperNN-generated match are significantly more robust to real-world distribution shifts than the baseline methods, while having the advantage of being both computationally efficient and agnostic to the modality and task in question. Qualitatively speaking, we see that matches produced with the final model are related in semantically-meaningful ways. Furthermore, since the only constraint is that samples be from different domains, the method is applicable whether information about the domain is coarse or fine-grained.

In the following subsections, we begin by giving a general formulation of our proposed semi-supervised loss employing a generic cross-domain k-NN algorithm. We then explain how we can replace this algorithm with CaliperNN in order to mitigate the risk of poorly-matched samples, and how the loss may be computed in an online fashion to give our complete algorithm.

3.1 Enforcing consistency between cross-domain pairs

We view our predictor as being composed of an encoder (or *backbone*) network, $f_\theta : \mathcal{X} \to \mathbb{R}^d$, generating intermediary outputs (features) $z \triangleq f_\theta(x)$, and a prediction head, g_ϕ, such that the prediction for sample x is given by $\hat{y} \triangleq g_\phi \circ f_\theta(x)$. We similarly consider the aggregate loss $\mathcal{L}$ as having a two-part decomposition given by

$$\mathcal{L} \triangleq \mathcal{L}_{\text{sup}} + \lambda\, \mathcal{L}_{\text{unsup}}, \tag{1}$$

where $\mathcal{L}_{\text{sup}}$ is the supervised component measuring the discrepancy (as computed, for example, by the MSE loss) between $\hat{y}$ and the ground-truth label y, $\mathcal{L}_{\text{unsup}}$ is the unsupervised component based on some kind of pretext task, such as cross-view consistency, and λ is a positive pre-factor determining the trade-off between the two components. For our method, we do not assume any particular form for $\mathcal{L}_{\text{sup}}$ and focus solely on $\mathcal{L}_{\text{unsup}}$.

Given a pair of datasets $\mathcal{D}_l$ and $\mathcal{D}_u$, sourced from the labelled domain $\mathcal{S}_l$, and unlabelled domain $\mathcal{S}_u$ respectively, along with their union $\mathcal{D} \triangleq \mathcal{D}_l \cup \mathcal{D}_u$ our goal is to train a predictor that is robust (invariant) to changes in domain, including those unseen during training. To do this, we propose to regularise $z \triangleq f_\theta(x)$ to be smooth (consistent) within local, cross-domain neighbourhoods. At a high-level, for any given *query* sample x_q sourced from domain s_q, we compute $\mathcal{L}_{\text{unsup}}$ as the mean distance between its encoding $z_q \triangleq f_\theta(x_q)$ and that of its k-nearest neighbours, $V_k(z_q)$ with the constraint that $\{s_q\} \cap \mathbf{s}_n = \emptyset$, where $\mathbf{s}_n$ is the set of domain-labels associated with $V_k(z_q)$. The general form of this loss for a given sample can then be written as

$$V_k(z_q) \triangleq \text{NN}(z_q, \{f_\theta(x) \mid (x, s) \in \mathcal{D}, s \neq s_q\}, k), \tag{2}$$

$$\mathcal{L}_{\text{unsup}} \triangleq \frac{1}{k} \sum_{z_n \in V_k(z_q)} d(z_q, z_n) \tag{3}$$

where $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is some distance function. Here, we follow [30] and define d to be the squared Euclidean distance between normalised encodings for our experiments. Allowing the NN algorithm to select pairs in an unconstrained manner, given the pool of queries and keys, however, can lead to poorly-matched pairs that are detrimental to the optimisation process. To address this, we replace the standard NN algorithm with a propensity-score-based variant, inspired by the statistical matching framework [58].

3.2 Cross-domain matching

For the matching component of our algorithm, we propose to use a variant of k-NN which, in addition to incorporating the above cross-domain constraint, filters the queries and keys that represent probable

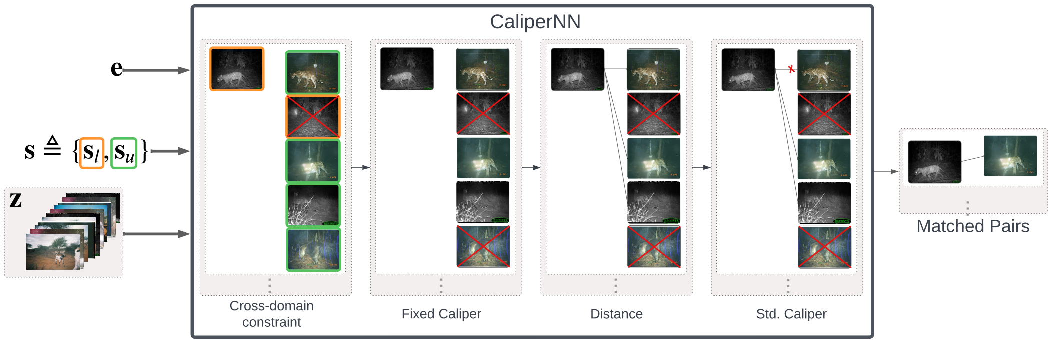

Figure 1: Illustration of our proposed statistical matching algorithm, CaliperNN. Given the anchor image encoding **z**, the corresponding domain label s (we consider the binary case of labelled vs. unlabelled for simplicity), and propensity score e, CaliperNN outputs the closest samples subject to their being from different domains, following filtering by the fixed and std. calipers.

outliers, according to their learned propensity scores. The initial stage of filtering employs a fixed caliper, where samples with propensity scores surpassing a fixed confidence threshold are removed; this is followed by a second stage of filtering wherein any two samples (from different domains) can only be matched if the Euclidean distance between their respective propensity scores is below a pre-defined threshold (std-caliper). See Fig 3.2 for a pictorial representation of these steps and Appendix G for reference pseudocode.

The propensity score, e, for a given sample x is estimated as $p(s|z)$ using a linear classifier f_θ, $h_\psi : \mathbb{R}^d \to \triangle^{|\mathcal{S}|}$ where $\triangle^{|\mathcal{S}|}$ is the probability simplex over possible domain labels, $\mathcal{S}$, induced by the softmax function. h_ψ^d is trained via maximum (weighted) likelihood to predict the domain label of a given sample for all samples within the aggregate dataset $\mathcal{D}$, or (typically) a subset of it, encoded by f_θ. Since we apply both calipers to the learned propensity score, the shape of this distribution can have a significant effect on the outcome of matching. Accordingly, we apply temperature-scaling, with scalar $\tau \in \mathbb{R}_\star^+$, to sharpen or flatten the learned propensity-score distribution. We denote the set of associated parameters ($\{\ t_f, t_\sigma, \tau\}$, as the threshold for the fixed-caliper, the threshold for the std-caliper, and the temperature, respectively) as ξ and discuss in Appendix D how one can determine suitable values for these in practice.

For convenience we define the set of all encodings, given by f_θ, as $\mathbf{z} \triangleq \{f_\theta(x) \,|\, x \in \mathcal{D}\}$, the set of all associated propensity scores as $\mathbf{e} \triangleq \{h_\psi(x) \,|\, z \in \mathbf{z}\}$, and the set of associated domain labels as $\mathbf{s}$. In the offline case, the matches for $\mathcal{D}$ are then computed as

$$\mathrm{MatchedSamples} \triangleq \{(z, \mathrm{CaliperNN}_\xi(z, \mathbf{z}, \mathbf{e}, s, k)) \,|\, z \in \mathbf{z}, s \in \mathbf{s}\}, \tag{4}$$

with CaliperNN returning the set of k-nearest neighbours according to d, subject to the aforementioned cross-domain and caliper-based constraints. We allow for the fact that there may be no valid matches for some samples due to these constraints; in such cases we have $\emptyset$ as the second element of their tuples, indicating that $\mathcal{L}_{\mathrm{unsup}}$ should be set to 0.

3.3 Scaling up with Online Learning

Re-encoding the dataset following each update of the feature-extractor, in order to recompute $\mathrm{MatchedSamples}$, is prohibitively expensive, with cost scaling linearly with $N \triangleq N_l + N_u$. Moreover, CaliperNN requires explicit computation of the pairwise distance matrices, which can be prohibitive memory-wise for large values of N. We address these problems using a fixed-size memory bank, $\mathcal{M}_z^{N_\mathcal{M}}$ storing only the last $N_\mathcal{M}$ (where $N_\mathcal{M} \ll N$) encodings from a slow-moving momentum encoder [30, 33], $f_{\theta'}$, which we refer to as the *target* encoder, in line with [30], and accordingly refer to f_θ as the *online* encoder. Unlike [30], however, we make use of neither a projector nor a predictor head (in the case of the target encoder) in order to compute the inputs to the consistency loss and simply use the output of the backbone as is – this is possible in our setting due to $\mathcal{L}_{\mathrm{sup}}$ preventing representational collapse. More specifically, the target encoder's parameters,

θ', are computed as a moving average of the online encoder's, θ, with decay rate $\zeta \in (0, 1)$, per the recurrence relation

$$\theta'_t = \zeta\theta'_{t-1} + (1-\zeta)\theta_t, \tag{5}$$

As the associated domain labels are also needed both for matching and to compute the loss for the propensity scorer, we also store the labels associated with $\mathcal{M}_z^{N_\mathcal{M}}$ in a companion memory bank $\mathcal{M}_s^{N_\mathcal{M}}$. We initialise $\mathcal{M}_z^{N_\mathcal{M}}$ and $\mathcal{M}_s^{N_\mathcal{M}}$ to $\emptyset$, resulting in fewer than $N_\mathcal{M}$ samples being used during the initial stages of training when the memory banks are yet to be populated.

Each iteration of training, we sample a batch of size B from $\mathcal{D}$ consisting of inputs $\mathbf{x}$ and $\mathbf{s}$. During the matching phase, the inputs are passed through the *target* encoder to obtain $\mathbf{z}'_q \triangleq \{f_{\theta'}(x) | x \in \mathbf{x}\}$, serving as the queries for CaliperNN. We also experiment with a simpler variant where the *online* encoder is instead used for this query-generation step, such that we instead have $\mathbf{z}'_q \triangleq \{f_\theta(x) | x \in \mathbf{x}\}$, and find this can work equally well if ζ is sufficiently high. The keys are then formed by combining the current queries with the past queries contained in the memory bank: $\mathbf{z}_k \triangleq \mathbf{z}'_q \cup \mathcal{M}_z^{N_\mathcal{M}}$. The domain labels associated with $\mathbf{z}_k$ are likewise formed by concatenating the domain labels in the current batch with those stored in $\mathcal{M}_s^{N_\mathcal{M}}$: $\mathbf{s}_k \triangleq \mathbf{s}_q \cup \mathcal{M}_s^{N_\mathcal{M}}$. Once the matches for the current samples have been computed, the oldest B samples in $M_z^{N_\mathcal{M}}$ and $M_s^{N_\mathcal{M}}$ are overwritten with $\mathbf{z}_k$ and $\mathbf{s}_k$, respectively. The consistency loss is then enforced between each query $\mathbf{z}_q \triangleq \{f_\theta(x) | x \in \mathbf{x}\}$, according to the differentiable *online* encoder, and each of its matches, $V_k(z'_q) \triangleq \text{CaliperNN}_\xi(z_q, \mathbf{z}_k, h_\psi(\mathbf{z}_k), \mathbf{s}_k)$ providing that $V_k(z'_q) \neq \emptyset$ (that is, under the condition that the estimated propensity score for z'_q does not violate the caliper(s) and there are at least k valid matches whose estimated propensity scores also do not), with the loss simply 0 otherwise. Since $f_{\theta'}$ is frozen, $\mathbf{z}_k$ carries an implicit stop-gradient and gradients are computed only w.r.t. θ. These steps are illustrated pictorially in Fig 2 and as pseudocode in Appendix G.

Similarly, rather than solving for the optimal parameters, $\psi^\star$ for the propensity scorer given the current values of $\mathbf{z}_k$, which is infeasible for the large values of $N_\mathcal{M}$ needed to well-approximate the full dataset, we resort to a biased estimate of $\psi^\star$. Namely, we train h_ψ in an online fashion to minimise the per-batch loss

$$\mathcal{L}_{\text{ps}} = \frac{1}{|\mathbf{z}_k|} \sum_{z \in \mathbf{z}_k, s \in \mathbf{s}_k} w_{\mathbf{s}_k}(s) \mathcal{H}(h_\psi(z), s), \tag{6}$$

where $\mathcal{H}$ is the standard cross-entropy loss between the predictive distribution and the (degenerate) ground-truth distribution, given by the one-hot encoded domain labels, and $w_{\mathbf{s}_k} : \mathcal{S} \to \mathbb{R}_*^+$ is a function assigning to each s an importance weight [66] based on the inverse of its frequency in $\mathbf{s}_k$ to counteract label imbalance. In the special case in which the $\mathcal{D}_l$ and $\mathcal{D}_u$ are known to have disjoint support over S (that is, $\mathcal{S}_l \cap \mathcal{S}_u = \emptyset$), we can substitute their domain labels with 1 and 0, respectively (such that we have $\mathcal{D}_l \triangleq \{x_i, y_i, 1\}_{i=1}^{N_l}$ and $\mathcal{D}_u \triangleq \{x_i, 0\}_{i=1}^{N_u}$), thus reducing the propensity scorer and CaliperNN to their binary forms. Knowing whether this condition is satisfied a priori (and thus whether the use of domain labels can be forgone completely from our pipeline) is not unrealistic: one may, for example know that two sets of satellite imagery cover two different parts of the world (e.g. Africa and Asia) yet not know the exact coordinates underlying their respective coverage.

4 Related Work

Domain Generalisation The goal of domain generalisation (DG) is to produce models that are robust to a wide range of distribution shifts (including those outside the training distribution), given a training set consisting of samples sourced from multiple domains. Despite the various techniques (many well theoretically-motivated) designed to improve the generalisation of deep neural networks current methods continue to fall short in the face of natural distribution shifts [31, 41]. Indeed, ERM has repeatedly shown to be a strong baseline – frequently outperforming dedicated methods that leverage domain information or additional unlabelled data – for DG [31, 63], despite the theoretical problems associated with using it when the training and test sets are misaligned. Until now, only pre-training on larger, more diverse datasets (with harder examples), has consistently proven to improve OOD generalisation, yet allowing pre-trained models to fit the ID data too closely can undo any such benefit conferred by the pre-training [3, 39, 69, 74]. Similar to Okapi, MatchDG [51] draws upon causal matching to tackle DG. Despite the surface-level similarity, there are a number of

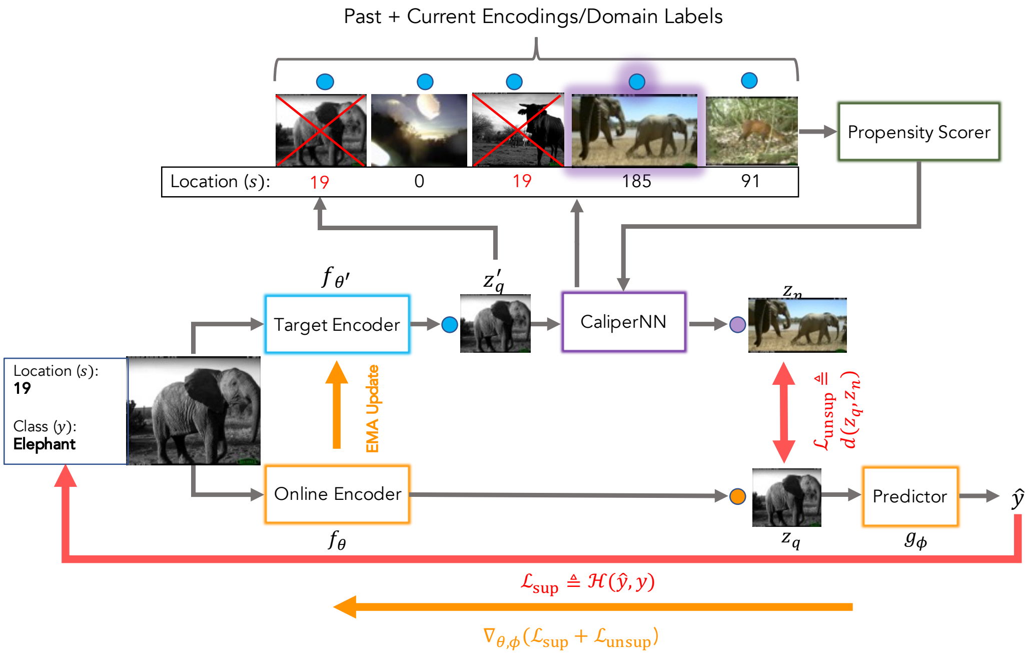
Past + Current Encodings/Domain Labels
Location (s):
19
0
19
185
91
Propensity Scorer
$f_{\theta'}$
z'_q
Target Encoder
CaliperNN
z_n
Location (s):
19
Class (y):
Elephant
EMA Update
$\mathcal{L}_{\text{unsup}} \triangleq d(z_q, z_n)$
Online Encoder
Predictor
$\hat{y}$
f_θ
z_q
g_ϕ
$\mathcal{L}_{\text{sup}} \triangleq \mathcal{H}(\hat{y}, y)$
$\nabla_{\theta,\phi}(\mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{unsup}})$

Figure 2: Overview of Okapi's online-learning pipeline based using the iWildCam dataset for the sake of illustration. For simplicity, we limit k to 1 so that the output of matching is a single vector rather than a set of vectors; for the same reason we illustrate the process for only a single sample taken from the labelled data set $\mathcal{D}_l$, annotated with both domain (s; in this case, *camera location*) and class (y) information. Inspired by recent advances in self-supervised learning, we maintain a copy (the target encoder) of the online encoder, f_θ, whose parameters, θ', are an exponential moving average (EMA) of θ. This EMA update is performed at the beginning of each training set at a rate governed by the decay coefficient, ζ. For a given sample, we first compute its embedding using the target encoder to produce the query vector, z'_q, and by the online encoder to produce z_q, which will serve as the 'anchor' in the consistency loss. This query vector is then used – alongside the output of the propensity scorer – by CaliperNN to compute its cross-domain nearest neighbour, z_n, where the keys are taken to be the current and past (stored in the Memory Bank) $N_\mathcal{M}$ encodings of the data. The cross-domain constraint, prohibiting matching of samples belonging to the same domain, is denoted through a red coloring of the location identifiers, the nearest sample obeying this constraint and the constraints of the calipers with purple highlighting. The consistency loss is the distance between z_q and z_n, defined by function some distance function d. Finally, the supervised loss, $\mathcal{L}_{\text{sup}}$ (here instantiated as the standard cross-entropy loss, $\mathcal{H}$), is computed using the output of the predictor acting on z_q and the ground-truth given by y.

significant differences, principally in the respects that we consider semi-supervised DG (whereas MatchDG requires full-labeling w.r.t. y) and employ an augmented form of k-NN for bias-reduction in the absence of y.

Self-Supervised Learning In self-supervised learning (SelfSL), models are trained to solve pretext tasks constructed from the input data. This learning paradigm has led to significant breakthroughs in unsupervised learning in recent years, with performance now approaching (or even surpassing, along some axes such as adversarial robustness) that of supervised methods for many tasks while requiring significantly less labelled data. Due to its generality, SelfSL has seen use across the complete spectrum of applications and modalities and underlies many of the foundation models [11] that have emerged in NLP [13, 16, 21], Computer Vision [29], and at their intersection [2, 78]. Common pretext tasks include those based on the masked-language-modelling approach – originally popularised by BERT [21] and recently generalised to other modalities [6, 7] – [15, 33], contrastive captioning [56, 78], and instance discrimination and self-distillation [14, 30] which rely on transformations of the data to generate multi-view inputs. Approaches belonging to the latter two categories were originally limited by the fact that the transforms had to be tailored for a particular modality and for

Table 1: A comparison between Okapi and different baselines on two benchmark image datasets. We include both the results of our re-run of the baselines and those of [63]. Both ID and OOD performances are reported. For iWildCam we average over results from 3 different seeds, for PovertyMap we do so over the 5 pre-defined CV folds. Standard deviations are shown in parentheses.

Method	**iWildCam**		**PovertyMap**			
	macro F1 ↑		worst U/R corr. ↑		worst U/R MSE ↓	
	ID	OOD	ID	OOD	ID	OOD
ERM [63]	47.0 (1.4)	32.2 (1.2)	0.66 (0.04)	0.49 (0.06)	-	-
FixMatch [63]	46.3 (0.5)	31.0 (1.3)	0.54 (0.10)	0.30 (0.11)	-	-
ERM	48.6 (1.1)	33.3 (0.3)	0.72 (0.03)	0.53 (0.09)	0.23 (0.03)	0.35 (0.12)
FixMatch	51.1 (1.0)	35.2 (0.7)	0.50 (0.13)	0.34 (0.12)	0.59 (0.42)	0.88 (0.61)
Okapi (ours)	50.6 (0.7)	36.1 (0.9)	0.72 (0.02)	0.55 (0.10)	0.22 (0.02)	0.33 (0.10)
Okapi (no calipers)	-	-	0.72 (0.02)	0.54 (0.12)	0.22 (0.02)	0.36 (0.14)

some modalities, such as tabular data, there is no obvious way to define them. A number of recent works have sought to obviate this problem through the use of MixUp [72], masking [6, 34], and k-NN [24, 42, 71], the latter of which is directly relevant to our work. Okapi bears closest resemblance to [42] in combining momentum-encoding with nearest-neighbours lookup to generate the views for a BYOL-style [30] consistency loss. However, a key distinction lies in the use of an augmented form of nearest-neighbours, CaliperNN, which both constrains pairs of samples to being from *different* domains and filters out any queries or keys deemed outliers according to a learned *propensity score*.

Semi-Supervised Learning Semi-supervised learning (SemiSL) encompasses a broad class of algorithms that combine unsupervised learning with supervised learning in order to improve the performance of the latter, especially when labelled data is limited. Many SemiSL methods are based on the self-training paradigm which can trace its roots back decades to the early work in pattern recognition by [65] and continues to be relevant in the modern era due to its generality, both within SemiSL itself and in related fields such as domain adaptation [27], and fledgling field of SelfSL [14] discussed above. Self-training applies to any framework predicated on using a model's own predictions to produce pseudo-labels for the unlabelled data which can either be used as targets for self-distillation [75] or enforcing consistency between predictions that themselves have been perturbed [5, 75] or that have been generated from perturbed/multi-view inputs [67]. FixMatch [67] is one example of a consistency-based method which has proven effective for semi-supervised classification, despite its simplicity, and various works [28, 46] have since built on the its framework prescribing the use of weakly- and strongly-augmented inputs to generate the targets and predictions, respectively. Like these methods, Okapi also makes use of a cross-view consistency loss, however, the alternative views for a given sample are generated not through data-augmentation but through statistical matching [58], with the aim being to achieve invariance to the domain rather than a particular series of perturbations. Another example of particular relevance to our work is [70], which uses a copy of the model with exponentially-averaged weights to generate the targets for the unlabelled data. Okapi also uses such a model to produce the targets for its consistency loss, but is more akin to momentum-encoding [33] in the respect that the loss is imposed on the latent space.

5 Experiments

5.1 Datasets

We evaluate Okapi on three datasets taken from the WILDS 2.0 benchmark [63]. These span a variety of modalities and tasks, allowing us to showcase the generality of our proposed method (Okapi): **iWildCam** (images, multiclass classification), **PovertyMap** (multispectral images, regression), and **CivilComments** (text, binary classification). Details of each dataset can be found in Appendix A.

5.2 Image experiments

Results of our image-data experiments are summarised in Table 1. Due to spacial constraints, we defer the full set of results, including those for the 'offline' (w.r.t. the matching) version of Okapi to Appendix. C. For both datasets in question, we use the same metrics as [63]: macro-F1 for iWildCam and worst-group (with the group defined as urban (U) vs. rural (R)) Pearson correlation for Poverty Map. For completeness, we include mean squared error (MSE) as a secondary metric for the latter dataset. Following [63], we compute the mean and standard deviation (shown in parentheses) over multiple runs for both ID and OOD test sets, with these runs conducted with 3 different random seeds and 5 pre-defined cross-validation folds for iWildCam and PovertyMap, respectively.

We compare Okapi against two baselines, ERM and FixMatch [67], both according to our re-implementation and according to the original implementation given in [63]. We note that since FixMatch, in its original form, is only applicable to classification problems due to its use of confidence-based thresholding, for the PovertyMap dataset, FixMatch represents a simplified variant (following [63]) without such thresholding, that is trained to simply minimise the MSE between *all* regressed values for the weakly- and strongly-augmented images. As described in Appendix D, the main difference between the baselines run included in [63] and our re-runs is in the backbone architecture, with us opting for a ConvNeXt [47] architecture over a ResNet one. For both datasets, and for both baselines we observe significant improvements stemming the change of backbone. Moreover, utilising ConvNeXt seems to be crucial in enabling FixMatch to surpass the ERM baseline in the classification task with 32.2 (ERM) vs 31.0 (FixMatch) and 33.3 (ERM) vs. 35.2 (FixMatch), with ResNet and ConvNeXt architecture respectively.

Okapi, convincingly outperforms the baselines, w.r.t the OOD metric of interest, on both datasets. We observe an improvement of $+0.9$ macro F1, i.e. 36.1 vs 35.2 of Okapi and FixMatch (the best baseline for iWildCam) respectively. For the regression task in PovertyMap, Okapi achieves 0.55 and 0.33 on the OOD test set in terms of Pearson correlation and MSE, respectively, in contrast to the 0.53 and 0.33 of ERM. At the same time, we note that FixMatch fails to generalise well to this task, yielding by far the worst results amongst the evaluated methods.

5.3 Text classification

Method	**Civil Comments** worst-group acc ↑
	OOD
ERM [63]	66.6 (1.6)
ERM (fully-labelled) [63]	69.4 (0.6)
ERM (reproduction)	68.5 (2.2)
Okapi (ours)	69.7 (2.0)

Table 2: Comparison between Okapi and the baselines methods on the Civil Comments dataset. We include both the original results of [63] as well as those of our reproduction of their ERM baseline. Performance is measured in terms of worst-group accuracy and averaged over seeds; standard deviations are shown in parentheses.

In Table 2 we summarise the numerical results for the CivilComments dataset. Remaining consistent with [63], we evaluate models according to the worst-group accuracy – the minimum of the conditional accuracies obtained by conditioning on each of the 8 dimensions of s – averaged over 5 replicates. Since there is no canonical ID test split available for this dataset, we report only the results only for the OOD split that is, rather than doing so for a custom split to avoid misrepresentation. We compare Okapi against both ERM variants featured in [63] – one trained on only the official labelled data and one trained with annotated unlabelled data (fully-labelled) – as well as our re-implementation of the ERM variant trained on only the labelled data with an identical hyperparameter configuration to the former. In contrast to the image datasets, we do not diverge in our choice of architecture, with all models trained with a pre-trained DistilBERT [64] backbone.

We observe marked improvement in the worst-group accuracy of this baseline compared with that reported therein. We attribute this partly to the high variance of the model-selection procedure (inherited from [63]) based on intermittently-computed validation performance (which does not consistently align with test performance) to determine the final model. This aside, we observe that Okapi outperforms the ERM baseline by a significant margin, to the point of parity with the fully-labelled baseline.

5.4 Ablations and qualitatitive analysis

In order to evaluate the importance of the caliper-based filtering to the performance of Okapi, we perform an ablation experiment on PovertyMap dataset (Okapi (no calipers)) with said filtering disabled (and all else constant), such that instead of CaliperNN we have standard k-NN, albeit with

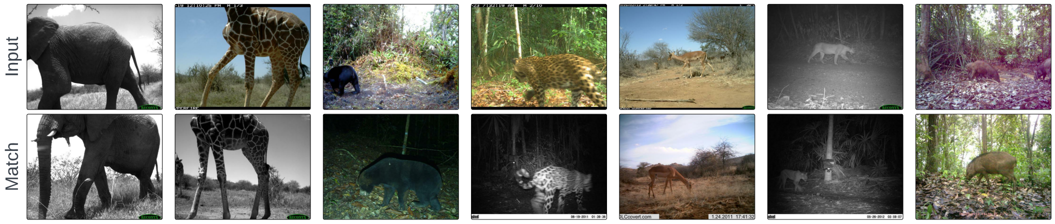

Figure 3: Examples of input (labelled) images and their 1-NN matched (unlabelled) images retrieved using CaliperNN on iWildCam dataset. Here, we match images from the labelled-train set to images from the unlabelled-extra set, taking advantage the fact that their domains are disjoint.

the cross-group constraint still in place (per Eq. 2). We see that performance degrades according to both metrics of interest, and, crucially, that the standard deviation of the runs is significantly higher, in line with our expectation that filtering out poor matches should stabilise optimisation. We provide additional ablation experiments in Appendix F, exploring the relative importance of the two (fixed and std-) calipers, the optimal number of neighbours to use for computing $\mathcal{L}_{\text{unsup}}$, and the feasibility of using the online encoder to generate the queries for CaliperNN.

Finally, in Fig. 3 we show samples of matched pairs retrieved by CaliperNN from the encodings of the learned encoder for the iWildCam dataset. Here, we see that semantic information (encoding the species of animal) is preserved across pairs, while nuisance factors such as illumination, background and contrast vary. Further examples from PovertyMap are shown in Appendix E. In Appendix H, we include matching results for the PACS (photo (P), art painting (A), cartoon (C), and sketch (S)) dataset [45] demonstrating how temperature scaling, in conjunction with the fixed caliper, can be used to control the filtering rate.

6 Conclusion

In this work, we introduced, Okapi, a semi-supervised method for training distributionally-robust models that is intuitive, effective, and is applicable to any modality or task. Okapi is based on the simple idea of supplementing the supervised loss with a cross-domain consistency loss that encourages the outputs of an encoder network to be similar for neighbouring (within the latent space of the encoder itself) samples belonging to different domains, which is made efficient using an online-learning framework. Rather than simply using k-NN with a cross-domain constraint, however, we propose an augmented form based on statistical matching (CaliperNN) that combines propensity scores with calipers to winnow out low-quality matches; we find this to be important for both the end-performance and consistency of Okapi. Our work serves as a response to [63], in that we find that it is in fact possible to effectively incorporate unlabelled data and domain information into a training algorithm in order to improve upon ERM with respect to an OOD test set, assuming an appropriate choice of architecture. Namely, on three datasets from the WILDS 2.0 benchmark, representing two different tasks (classification and regression) and modalities (image and text), we show that Okapi outperforms both the ERM and FixMatch baselines according to the relevant OOD metrics.

Buoyed by these promising results, we intend to apply Okapi to other tasks (e.g. object detection and image segmentation) and other modalities (e.g. audio) to further establish its generality. Furthermore, one limitation of the current incarnation of the method is that the thresholds for the calipers are fixed over the course of training whereas it may be beneficial to set these adaptively with the view to optimise such measures of inter-domain balance as *Variance Ratio* and *Standard Mean Differences* that are commonly used to evaluate the the goodness of statistical matching procedures.

Acknowledgments and Disclosure of Funding

This research was supported by a European Research Council (ERC) Starting Grant for the project "Bayesian Models and Algorithms for Fairness and Transparency", funded under the European Union's Horizon 2020 Framework Programme (grant agreement no. 851538). NQ is also supported by the Basque Government through the BERC 2018-2021 program and by Spanish Ministry of Sciences, Innovation and Universities: BCAM Severo Ochoa accreditation SEV-2017-0718.

References

[1] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[3] A. Andreassen, Y. Bahri, B. Neyshabur, and R. Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021.

[4] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[5] P. Bachman, O. Alsharif, and D. Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014.

[6] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022.

[7] H. Bao, L. Dong, and F. Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[8] S. Beery, E. Cole, and A. Gjoka. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020.

[9] Q. Berthet, M. Blondel, O. Teboul, M. Cuturi, J.-P. Vert, and F. Bach. Learning with differentiable pertubed optimizers. *Advances in neural information processing systems*, 33:9508–9519, 2020.

[10] A. Biglan, D. Ary, and A. C. Wagenaar. The value of interrupted time-series experiments for community intervention research. *Prevention Science*, 1(1):31–49, 2000.

[11] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[12] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

[13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[14] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[16] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[17] P. Christian, L. E. Murray-Kolb, S. K. Khatry, J. Katz, B. A. Schaefer, P. M. Cole, S. C. LeClerq, and J. M. Tielsch. Prenatal micronutrient supplementation and intellectual and motor function in early school-aged children in nepal. *Jama*, 304(24):2716–2723, 2010.

[18] E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pages 1436–1445. PMLR, 2019.

[19] E. Creager, J.-H. Jacobsen, and R. Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.

[20] R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[22] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems*, 31, 2018.

[23] J. A. Dunnmon, D. Yi, C. P. Langlotz, C. Ré, D. L. Rubin, and M. P. Lungren. Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology*, 290(2):537–544, 2019.

[24] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021.

[25] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[26] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[27] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[28] C. Gong, D. Wang, and Q. Liu. Alphamatch: Improving consistency for semi-supervised learning with alpha-divergence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13683–13692, 2021.

[29] P. Goyal, Q. Duval, I. Seessel, M. Caron, M. Singh, I. Misra, L. Sagun, A. Joulin, and P. Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*, 2022.

[30] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33: 21271–21284, 2020.

[31] I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.

[32] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[33] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[34] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

[35] M. Hurley and J. Adebayo. Credit scoring in the era of big data. *Yale Journal of Law and Technology*, 18:148–216, 2017.

[36] B. Y. Idrissi, M. Arjovsky, M. Pezeshki, and D. Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022.

[37] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

[38] T. Kehrenberg, M. Bartlett, O. Thomas, and N. Quadrianto. Null-sampling for interpretable and fair representations. In *European Conference on Computer Vision*, pages 565–580. Springer, 2020.

[39] D. Kim, K. Wang, S. Sclaroff, and K. Saenko. A broad study of pre-training for domain generalization and adaptation. *arXiv preprint arXiv:2203.11819*, 2022.

[40] G. King and R. Nielsen. Why propensity scores should not be used for matching. *Political Analysis*, 27(4):435–454, 2019.

[41] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[42] S. A. Koohpayegani, A. Tejankar, and H. Pirsiavash. Mean shift for self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10326–10335, 2021.

[43] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[44] P. Lahoti, K. Gummadi, and G. Weikum. Operationalizing individual fairness with pairwise fair representations. *Proceedings of the VLDB Endowment*, 13(4):506–518, 2019.

[45] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[46] J. Lienen and E. Hüllermeier. Credal self-supervised learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14370–14382. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/7866c91c59f8bffc92a79a7cd09f9af9-Paper.pdf`.

[47] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.

[48] I. Loshchilov and F. Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=Skq89Scxx`.

[49] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[50] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[51] D. Mahajan, S. Tople, and A. Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021.

[52] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013.

[53] L. Oneto, M. Donini, G. Luise, C. Ciliberto, A. Maurer, and M. Pontil. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. *Advances in Neural Information Processing Systems*, 33:15360–15370, 2020.

[54] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[55] N. Quadrianto, V. Sharmanska, and O. Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8227–8236, 2019.

[56] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[57] S. Romiti, C. Inskip, V. Sharmanska, and N. Quadrianto. Realpatch: A statistical matching framework for model patching with real samples. *CoRR*, abs/2208.02192, 2022.

[58] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[59] P. R. Rosenbaum and D. B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1): 33–38, 1985.

[60] D. B. Rubin. Matching to remove bias in observational studies. *Biometrics*, pages 159–183, 1973.

[61] D. B. Rubin. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3):169–188, 2001.

[62] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[63] S. Sagawa, P. W. Koh, T. Lee, I. Gao, S. M. Xie, K. Shen, A. Kumar, W. Hu, M. Yasunaga, H. Marklund, S. Beery, E. David, I. Stavness, W. Guo, J. Leskovec, K. Saenko, T. Hashimoto, S. Levine, C. Finn, and P. Liang. Extending the WILDS benchmark for unsupervised adaptation. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=z7p2V6KROOV`.

[64] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[65] H. Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.

[66] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[67] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020.

[68] E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Stat Sci*, 25 (1):1–21, 2010. ISSN 0883-4237.

[69] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[70] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[71] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, and L. V. Gool. Revisiting contrastive methods for unsupervised learning of visual representations. *Advances in Neural Information Processing Systems*, 34, 2021.

[72] V. Verma, T. Luong, K. Kawaguchi, H. Pham, and Q. Le. Towards domain-agnostic contrastive learning. In *International Conference on Machine Learning*, pages 10530–10541. PMLR, 2021.

[73] D. S. Watson, J. Krutzinna, I. N. Bruce, C. E. Griffiths, I. B. McInnes, M. R. Barnes, and L. Floridi. Clinical applications of machine learning algorithms: beyond the black box. *Bmj*, 364, 2019.

[74] O. Wiles, S. Gowal, F. Stimberg, S.-A. Rebuffi, I. Ktena, K. D. Dvijotham, and A. T. Cemgil. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=Dl4LetuLdyK`.

[75] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[76] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):1–11, 2020.

[77] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[78] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Yes; we show through our experiments detailed in Sec. 5 that our proposed method is able to outperform our baseline methods, and the original implementations of them given in [63], according to three recent OOD benchmark datasets.
   (b) Did you describe the limitations of your work?
   Yes; we describe the foremost limitation of our work in Sec. 6. Namely, we point out that using fixed thresholds for the caliper the entire duration of training is likely suboptimal and these hyperparameters should instead be set adaptively.
   (c) Did you discuss any potential negative societal impacts of your work?
   No, we do not anticipate there being any negative societal impacts specific to our work. On the contrary, we propose a method for training more distributionally-robust models with the view to develop safer, less biased, machine-learning systems.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them?
   Yes, we have read the review guidelines and have made sure that our paper complies with them.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results?
   N/A
   (b) Did you include complete proofs of all theoretical results?
   N/A
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?
   Yes; the code – containing with the requisite instructions and configuration files needed to run all of the experiments detailed in Sec 5 – is publicly available and linked to in the Abstract.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
   Yes; all implementation details, including those related to optimisation and hyperparameter-selection, are given in Appendix D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?
   Yes, we ran all methods with multiple random seeds/folds; we report the standard deviation over the runs for each method-dataset combination in parentheses in Table 1 and Table 2.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)?
   No, however do provide estimates of the carbon footprint for a single run of our method and of the ERM and FixMatch baselines for the iWildCam dataset.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators?
   Yes, we use datasets from the WILDS 2.0 benchmark developed by [63] for evaluating our models and cite the associated paper in both Sec. 1 and Sec. 5.
   (b) Did you mention the license of the assets?
   No.
   (c) Did you include any new assets either in the supplemental material or as a URL?
   Yes, we include a URL link to the code for the paper at the end of the Abstract.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating?
   N/A

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?
N/A

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable?
N/A

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable?
N/A

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation?
N/A