RealPatch: A Statistical Matching Framework for Model Patching with Real Samples

Sara Romiti[1], Christopher Inskip[1], Viktoriia Sharmanska[1,4], and Novi Quadrianto[1,2,3]

[1] Predictive Analytics Lab (PAL), University of Sussex, United Kingdom
[2] BCAM Severo Ochoa Strategic Lab on Trustworthy Machine Learning, Spain
[3] Monash University, Indonesia
[4] Imperial College London
{s.romiti, c.inskip, sharmanska.v, n.quadrianto}@sussex.ac.uk

Abstract. Machine learning classifiers are typically trained to minimise the average error across a dataset. Unfortunately, in practice, this process often exploits spurious correlations caused by subgroup imbalance within the training data, resulting in high average performance but highly variable performance across subgroups. Recent work to address this problem proposes *model patching* with CAMEL. This previous approach uses generative adversarial networks to perform intra-class inter-subgroup data augmentations, requiring (a) the training of a number of computationally expensive models and (b) sufficient quality of model's synthetic outputs for the given domain. In this work, we propose RealPatch, a framework for simpler, faster, and more data-efficient data augmentation based on statistical matching. Our framework performs model patching by augmenting a dataset with real samples, mitigating the need to train generative models for the target task. We demonstrate the effectiveness of RealPatch on three benchmark datasets, CelebA, Waterbirds and a subset of iWildCam, showing improvements in worst-case subgroup performance and in subgroup performance gap in binary classification. Furthermore, we conduct experiments with the imSitu dataset with 211 classes, a setting where generative model-based patching such as CAMEL is impractical. We show that RealPatch can successfully eliminate dataset leakage while reducing model leakage and maintaining high utility. The code for RealPatch can be found at `https://github.com/wearepal/RealPatch`.

Keywords: Classification, subgroup imbalance, model patching, statistical matching, dataset leakage.

1 Introduction

Machine learning models have fast become powerful yet ferocious pattern matching tools, able to exploit complex relationships and distant correlations present in a dataset. While often improving the average accuracy across a dataset, making use of spurious correlations (i.e. relationships that appear causal but in reality are not) for decision making is often undesirable, hurting generalization in the

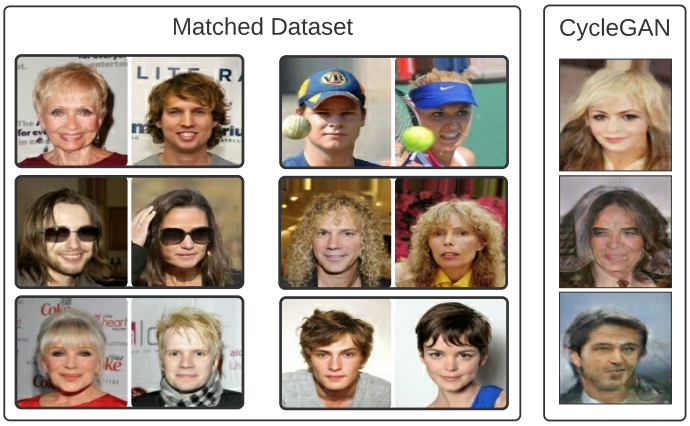

Fig. 1. Examples of images and their counterfactuals on the attribute male/female, retrieved using RealPatch (left); both original and matched images are real samples from the CelebA dataset. RealPatch preserves characteristics across matched pairs such as pose, facial expression, and accessories. We also show CycleGAN synthetic counterfactual results (right) on the same attribute.

case that spurious correlations do not hold in the test distribution, and resulting in models that are biased towards certain subgroups or populations.

Recent works in machine learning have studied the close link between invariance to spurious correlations and causation [28,16,1,24,34]. Causal analysis allows us to ask counterfactual questions in the context of machine learning predictions by relying on attribute-labelled data and imagining "what would happen if" some of these attributes were different. For example, "would the prediction of a smile change had this person's cheeks been rosy"? While simple to answer with tabular data, generating counterfactuals for image data is non-trivial.

Recent advances in generative adversarial networks (GAN) aim to build a realistic generative model of images that affords controlled manipulation of specific attributes, in effect generating image counterfactuals. Leveraging research progress in GANs, several works now use counterfactuals for (a) detecting unintended algorithmic bias, e.g. checking whether a classifier's "smile" prediction flips when traversing different attributes such as "heavy makeup" [11], and (b) reducing the gap in subgroup performance, e.g. ensuring a "blonde hair" classifier performs equally well on male and female subgroups [33,13]. The first relies on an *invert then edit* methodology, in which images are first inverted into the latent space of a pre-trained GAN model for generating counterfactuals, while the latter uses an *image-to-image translation* methodology. One of the most recent approaches, CAMEL [13], focuses on the latter usage of counterfactuals to *patch* the classifier's dependence on subgroup-specific features.

GAN-based counterfactual results are encouraging, however, we should note that GAN models have a number of common issues such as mode collapse, failure

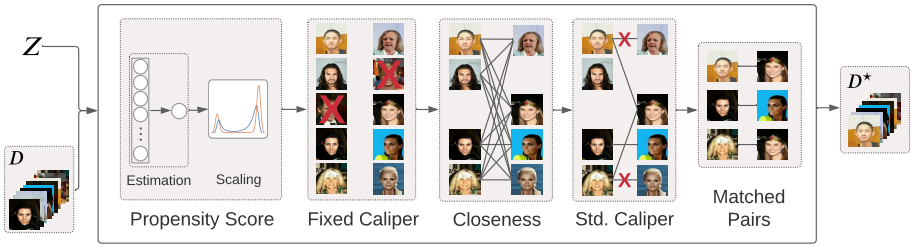

Fig. 2. RealPatch: statistical matching pipeline. Given the dataset D and the spurious attribute Z, the output is a matched dataset $D^\star$. To produce $D^\star$ we: 1) estimate the propensity score, and adjust it with temperature scaling; 2) restrict D using the fixed caliper to remove *extreme* samples; 3) compute the pair-wise closeness for each sample; 4) use the std-caliper to restrict the possible pairs according a maximum propensity score distance ; 5) for each sample, select the closest sample in the opposite group.

to converge, and poor generated results in a setting with a large number of class labels and limited samples per class. We provide an alternative counterfactual-based model patching method that is simpler, faster, and more data-efficient. We focus on a statistical matching technique (see for example [29]) such that for every image, we find an image with similar observable features yet having an opposite attribute value than the observed one; our counterfactual results are shown in Figure 1. A statistical matching framework has been widely utilised to assess causality relationships in numerous fields, such as education [25], medical [4],[6],[32], and community policies [3], [27] to name some. In this work, we explore statistical matching in the context of computer vision, and show its application for model patching with real samples.

Our paper provides the following contributions:

1. We propose an image-based counterfactual approach for model patching called *RealPatch* that uses real images instead of GAN generated images;
2. We provide an empirical evaluation of different statistical matching strategies for vision datasets. Our results can be used as a *guideline* for future statistical matching applications, for example showing the importance of using calipers;
3. We show applications of RealPatch for improving the worst-case performance across subgroups and reducing the subgroup performance gap in a 2-class classification setting. We observe that spurious correlation leads to shortcut learning, and show how RealPatch mitigates this by utilising a balanced dataset to regularise the training;
4. We show applications of RealPatch for reducing dataset leakage and model leakage in a multi 211-class classification setting.

Related Work. Data augmentation strategies using operations such as translation, rotation, flipping, cutout [12], mixup [39], and cutmix [38] are widely used for increasing the aggregate performance of machine learning models in computer vision applications. To improve performance in a targeted fashion, image transformation techniques that learn to produce semantic changes to an

image are used to generate samples for underrepresented subgroups. Sharmanska et al. [33] used a StarGAN model [5] to augment the dataset with respect to a subgroup-specific feature, and subsequently optimized a standard Empirical Risk Minimization (ERM) training objective. Whereas, CAMEL, a framework by Goel et al. [13], used a CycleGAN image transformation approach [42] and minimized a Sub-Group Distributionally Robust Optimization (SGDRO) objective function.

GDRO method [31] aims to minimize the worst-case loss over groups in the training data. CAMEL models [13] minimize the class-conditional worst-case loss over groups. Another approach to reduce the effects of spurious correlation is optimizing a notion of invariance. Invariance serves as a proxy for causality, as features representing "causes" of class labels rather than "effects" will generalize well under intervention. Invariant Risk Minimization (IRM) [1] tries to find a data representation which discards the spurious correlations by enforcing that the classifier acting on that representation is simultaneously optimal in each subgroup. However, more analysis and better algorithms are needed to realize the promise of this framework in practice [24,13].

Model patching focuses on robustness with respect to the unexpected failure of standard classifiers on subgroups of a class. Subgroups can correspond to environments/domains such as water or land, and can also refer to demographic attributes such as females or males [9]. Our work is therefore also related to many works addressing dataset biases in computer vision, particularly, in which the notion of bias relates to demographic attributes (e.g. [35,36,18,17]). Wang et al. [35] showed that even when datasets are balanced e.g. each class label co-occurs equally with each gender, learned models amplify the association between labels and gender, as much as if data had not been balanced. We refine their conclusions about balanced dataset and show that balancing with a statistical matching framework can successfully eliminate dataset leakage while reducing model leakage and maintaining high utility.

2 Our RealPatch Framework

We propose *RealPatch*, a framework that first resamples a dataset such that the *spurious groups* are balanced and equally informative and then utilise such dataset to regularise a classification objective. RealPatch only uses real samples from the original dataset when constructing the augmented dataset, making it faster to perform, and simpler to apply to new tasks, compared to approaches such as CAMEL [13] which require models to generate synthetic samples. Unlike standard data augmentation, our augmentation is in the context of statistical matching; it is a model-based approach for providing joint statistical information based on variables collected through two or more data sources. If we have two sources, e.g. male and female, matching augments the source domain of male images with female images, and the domain of female images with male images. In this section we outline the two stages of RealPatch. In Stage 1, a statistical matching procedure is used to construct a *matched dataset*, a collec-

tion of comparable pairs of images that have opposite values of the spurious attribute. In Stage 2, we learn a model to predict the target label by including the representations of instances in the matched dataset.

Setup. Given a dataset of N samples $D = \{1, \ldots, N\}$ with target label Y and spurious label Z, the dataset is divided into two *spurious groups* D_T and D_C based on the value of Z. These partitions define the so-called *treatment* ($Z = 1$) and *control* ($Z = 0$) groups of size N_T and N_C, respectively. Additionally, we call *target groups* the two partitions created by Y and *subgroups* the four sets caused by both Y and Z. We use X to denote feature representations of the input images extracted from a pre-trained model such as ResNet [15], or Big Transfer BiT [20]. In our framework these encoded representations X are the *observed covariates* that are used to compute the distances M between images and identify the matched pairs. Following the work of causal inference, image representations in X are assigned a propensity score, a measure of how likely an image s belongs to the treatment group, $e_s = \hat{P}(Z_s = 1 | X_s)$. Propensity scores are used during Stage 1 to help prevent the inclusion of instances that would lead to poor matches.

2.1 Stage 1: Statistical Matching

Matching is a sampling method to reduce model dependency and enforce covariate balance in observational studies across a treatment and control group. In this work, we study the nearest-neighbour (NN) matching algorithm, which for each treatment sample selects the closest control sample. Figure 2 depicts our proposed matching pipeline. The pipeline has the following main building blocks: 1) *propensity score estimation*; 2) *closeness measure*; and 3) *calipers* as a threshold mechanism. Before using the matched dataset in Stage 2, the *matching quality* is measured by assessing the achieved balance of the covariates.

Propensity Score Estimation. In causal inference, a propensity score e_s is the probability of a sample s being in the *treatment* group D_T, given its observed covariates X_s. This conditional probability is usually unknown, therefore it has to be estimated. This is typically done using a logistic regression on the observed X to predict the binary variable Z [8]. Logistic regression allows us to reweight samples when optimising the loss function. We explore the use of *spurious reweighting*, where samples are weighted inversely proportional to the frequency of their spurious label Z; more details are provided in Appendix A.

The shape of the conditional distribution has the potential to impact finding a suitable threshold. In this work we explore the use of *temperature scaling* as a post-processing step to adjust the propensity score distribution before its use for matching. Temperature scaling has become a common approach for recalibrating models [14], but to the best of our knowledge has not been utilised in the context of statistical matching for causal inference. In binary classification cases such as ours, for each sample s the logits z_s are divided by a (learned or fixed) parameter t before applying the sigmoid function:

$$z_s = \log\left(\frac{e_s}{1 - e_s}\right), \; q_s = \frac{1}{1 + e^{-z_s/t}}.$$

With $t=1$ we obtain the original probabilities. When $t<1$ the rescaled probabilities have a sharper distribution reaching a point mass at $t=0$. When $t>1$ the rescaled probabilities are smoother, reaching a uniform distribution as $t\to\infty$. As we show in our ablation study, we found rescaling to be beneficial for improving the achieved covariate balance (Table 3).

Closeness Measure. There are multiple metrics that can be used to measure the distance $M_{i,j}$ between samples $i \in D_T$ and $j \in D_C$, the most commonly used are *Euclidean* and *propensity score* distances. The *Euclidean distance* is defined as $M_{ij} = (X_i - X_j)^\top (X_i - X_j)$ and the *propensity score distance* as the distance between propensity scores $M_{ij} = |e_i - e_j|$. Both *Euclidean* and *propensity score* distances have the advantage of being able to control how many samples are included via a threshold. While propensity score is the most commonly used matching method, Euclidean distance matching should be preferred [19] as the goal is to produce exact balance of the observed covariates rather than balance them on average.

Calipers. Nearest-neighbour matching is forced to find a match for every treatment sample and is therefore at risk of finding poor matched pairs. Caliper matching is a method designed to prevent matching samples with limited covariate overlap. In this work we explore the usage of two different types of caliper, namely *fixed caliper* and *standard deviation (std) based caliper*, both applied to the estimated propensity score. *Fixed caliper* [10] is a selection rule that discards samples that have an estimated propensity score outside of a specific range; i.e. the dataset is restricted to $\{s, \forall\, s \in D \mid e_s \in [c, 1-c]\}$. This allows the exclusion of examples with *extreme* propensity scores; a rule-of-thumb used in previous studies [10] considers the interval defined by $c=0.1$, i.e. $[0.1, 0.9]$. *Standard deviation (std) based caliper* [7] is used to enforce a predetermined maximum discrepancy for each matching pair in terms of propensity score distance. The distance M_{ij} is kept unaltered if $|e_i - e_j| \leq \sigma \cdot \alpha$, and is set to ∞ otherwise. The variable σ is the standard deviation of the estimated propensity score distribution and α is a parameter controlling the percentage of *bias reduction* of the covariates. Cochran and Rubin [7] showed the smaller the α value the more the bias is reduced, the actual percentage of bias reduction depends on the initial standard deviation σ. Commonly used α values are $\{0.2, 0.4, 0.6\}$ [7].

In our application we 1) restrict potential matches based on fixed caliper and 2) follow a hybrid approach selecting the closest sample using Euclidean distance matching while defining a maximum propensity score distance between samples. The final outcome of Stage 1 is a *matched dataset* $D^\star$.

Matching Quality. Matching quality can be assesed through measuring the balance of the covariates across the treatment and control groups. Two commonly used evaluation measures are *standardised mean differences (SMD)* and *variance ratio (VR)* [30]. In the case that high imbalance is identified, Stage 1 should be iterated until an adequate level of balanced is achieved; we provide a guideline of adequacy for each metric below. *Standardised Mean Differences* computes the difference in covariate means between each group, divided by the standard deviation of each covariate. For a single covariate $\mathbf{a}$ from X we have:

$$\text{SMD} = \frac{\bar{\mathbf{a}}_T - \bar{\mathbf{a}}_C}{\sigma} \text{ , where } \sigma = \sqrt{\frac{s_T^2 + s_C^2}{2}}$$

Here, $\bar{\mathbf{a}}_T$ ($\bar{\mathbf{a}}_C$) and s_T^2 (s_C^2) are respectively the sample mean and variance of covariate $\mathbf{a}$ in group D_T (D_C). Intuitively, smaller *SMD* values are better and as a rule of thumb an *SMD* value below 0.1 expresses an adequate balance, a value between 0.1 and 0.2 is considered not balanced but acceptable, and above 0.2 shows a severe imbalance of the covariate [26]. *Variance Ratio* is defined as the ratio of covariate variances between the two groups, with an ideal value close to 1. While in some studies [40] a variance in the interval $(0, 2)$ is defined acceptable, we follow Rubin [30] and use the stricter interval $(4/5, 5/4)$ to indicate the desired proximity to 1. To obtain a single measure for all covariates X, we categorise *SMD* into ≤ 0.1, $(0.1, 0.2)$, and ≥ 0.2, and *VR* into $\leq 4/5$, $(4/5, 5/4)$, and $\geq 5/4$ and assess the distribution of covariates. We show an assessment of matching quality for one run on each dataset in Section 3.1, comparing the covariate balance before and after matching as well the effect of using temperature scaling.

2.2 Stage 2: Target Prediction

This stage is concerned with predicting a discrete target label Y from covariates X. Inspired by Goel et al. [13] our training process involves the minimization of a loss $\mathcal{L}$ that combines a SGDRO objective function $\mathcal{L}_{SGDRO}$ and a self-consistency regularisation term $\mathcal{L}_{SC}$:

$$\mathcal{L} = \mathcal{L}_{SGDRO} + \lambda \mathcal{L}_{SC}, \tag{1}$$

where λ is a hyperparameters controlling the regularisation strength. The SGDRO loss is inspired by GDRO [31], with the difference of considering a non-flat structure between the *target* and *spurious* labels; the hierarchy between target and spurious labels is included by considering the *spurious groups* difference within each *target group*. The SGDRO component of our loss is computed on the entire dataset D.

Similarly to [13], our $\mathcal{L}_{SC}$ encourages predictions $f_\theta(\cdot)$ of a matched pair (x_T, x_C) in $D^\star$ to be consistent with each other and is defined as:

$$\mathcal{L}_{SC}(x_T, x_C, \theta) = \frac{1}{2}\left[KL(f_\theta(x_T)||\tilde{m}) + KL(f_\theta(x_C)||\tilde{m})\right], \tag{2}$$

where $\tilde{m}$ is the average output distribution of the matched pair. While the SGDRO objective accounts for the worst-case subgroup performance, the form of the regularisation term induces model's predictions to be subgroup invariant [13].

3 Experiments

We conduct two sets of experiments to assess the ability of RealPatch to 1) improve the worst-case subgroup performance and reduce the subgroup performance gap in a binary classification setting, and 2) reduce dataset and model

leakage w.r.t. a spurious attribute in a 211-class classification setting. We describe them in turn.

3.1 Reducing Subgroup Performance Gap

In this section we study the effect of our RealPatch for increasing the worst-case performance across subgroups and reducing the gap in subgroup performance. We evaluate RealPatch against a variety baselines on three datasets, and perform an ablation analysis on configurations of RealPatch. We compare approaches using **Robust Accuracy**: the lowest accuracy across the four subgroups, **Robust Gap**: the maximum accuracy distance between the subgroups, as well as **Aggregate Accuracy**: a standard measure of accuracy. Our goal is to improve the robust accuracy and gap while retaining the aggregate accuracy performance as much as possible. That is because performance degradation on a subgroup(s) might occur if this improves the worst performing subgroup (e.g. [23]).

Datasets. We use three publicly available datasets, CelebA[5] [21], Waterbirds[6] [31] and iWildCam-small[7] [2]. **CelebA** has 200K images of celebrity faces that come with annotations of 40 attributes. We follow the setup in [13], and consider hair colour $Y \in \{\text{blonde, non-blonde}\}$ as target label, and gender $Z \in \{\text{male, female}\}$ as spurious attribute. In this setup, the subgroup ($Y = \text{non-blonde}, Z = \text{female}$) is under-sampled in the training set (from 71,629 to 4,054) as per [13] amplifying a spurious correlation between the target and the demographic attribute. We keep all other subgroups as well as the validation and test sets unchanged. The images are aligned and resized to 128x128. For stability we repeat our experiments three times using different randomly under-sampled subgroups ($Y = \text{non-blonde}, Z = \text{female}$). **Waterbirds** has 11,788 examples of birds living on land or in water. We follow [31] and predict $Y \in \{\text{waterbird, landbird}\}$, and use the background attribute $Z \in \{\text{water, land}\}$ as spurious feature. The spurious correlation between target and background is present in the dataset as waterbirds appear more frequently in a water scene, whereas landbirds on land. In order to perform three runs we randomly define the train/validation/test splits while enforcing the original subgroup sizes as per [13]. **iWildCam-small** is a subset of iWildCam dataset [2], whose task is to classify animal species in camera trap images. Here, we consider two species (meleagris ocellata and crax rubra) within two camera trap locations. The dataset contains $3,349$ images, specifically $2,005$ (train), 640 (val) and 704 (test). These splits have a spurious correlation between animal species and locations.This experiment emphasizes the applicability of RealPatch in a small dataset setting.

Baselines. Here we describe the four baseline methods used for comparison. **Empirical Risk Minimization (ERM)** is a standard stochastic gradient descent model trained to minimize the overall classification loss. **Group Distributionally Robust Optimisation (GDRO)** is a stochastic algorithm proposed by [31] with the aim of optimising the worst-case performance across the

[5] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

[6] https://github.com/kohpangwei/group_DRO

[7] https://github.com/visipedia/iwildcam_comp/tree/master/2020

Table 1. A comparison between RealPatch and four baselines on two benchmark datasets. The results shown are the average (standard deviation) performances over three runs. RealPatch is able to construct a model that is robust across subgroups with high robust accuracy and small robust gap.

Dataset	Method	Aggregate ↑ Accuracy (%)	Robust ↑ Accuracy (%)	Robust ↓ Gap (%)
CelebA	**ERM**	89.21 (0.32)	55.3 (0.65)	43.48 (0.68)
	GDRO	**90.47** (7.16)	63.43 (18.99)	34.77 (19.65)
	SGDRO	88.92 (0.18)	82.96 (1.39)	7.13 (1.67)
	CAMEL	84.51 (5.59)	81.48 (3.94)	**5.09** (0.44)
	RealPatch (Ours)	89.06 (0.13)	**84.82** (0.85)	5.19 (0.9)
Waterbirds	**ERM**	86.36 (0.39)	66.88 (3.76)	32.57 (3.95)
	GDRO	**88.26** (0.55)	81.03 (1.16)	14.80 (1.15)
	SGDRO	86.85 (1.71)	83.11 (3.65)	6.61 (6.01)
	CAMEL	79.0 (14.24)	76.82 (18.0)	7.35 (5.66)
	RealPatch (Ours)	86.89 (1.34)	**84.44** (2.53)	**4.43** (4.48)

Table 2. Experiments with iWildCam-small[2]. The results shown are the average (standard deviation) performances over three runs. CycleGAN-based CAMEL is not applicable for small training data (2K images).

Method	Aggregate ↑ Accuracy (%)	Robust ↑ Accuracy (%)	Robust ↓ Gap (%)
ERM	**79.97** (1.18)	75.43 (3.01)	19.65 (1.96)
SGDRO	78.55 (2.45)	75.50 (3.58)	14.28 (4.35)
RealPatch (Ours)	79.36 (2.09)	**76.70** (3.19)	**11.36** (4.87)

subgroups. **Sub-Group Distributionally Robust Optimisation (SGDRO)** [13] as described in Section 2.2. **CAMEL** is a two stage approach proposed by [13] that uses the synthetic samples to define a subgroup consistency regulariser for model patching. Conceptually this model is most similar to ours, where we use real samples for model patching. The training details are in Appendix A.

RealPatch Configurations and Hyperparameters. RealPatch can be instantiated in many different configurations. In these experiments hyperparameters of RealPatch include the choice of *calipers*, *temperature*, and *reweighting strategies* in the propensity score estimation model as well as *self-consistency strength* λ, *adjustment coefficients* and *learning rates* in the target prediction model. To select hyperparameters for Stage 1 of RealPatch we perform a grid search, selecting the configuration with the best covariates balance in term of *SMD* and *VR*. An ablation study on such hyperparameters is provided in Section 3.1. As per the hyperparameters of Stage 2, we perform model selection utilising the robust accuracy on the validation set. Further details of hyperparameters used and best configuration selected are summarised in Appendix A.

Results on CelebA. From Table 1, RealPatch is able to significantly improve the worst-case subgroup performance and reduce the subgroup performance gap compared to the other baseline methods such as ERM, GDRO, SGDRO. Our proposed method improves the robust accuracy, robust gaps and aggregate accuracy with respect to the best baseline SGDRO by 1.86%, 1.94% and 0.14% respectively. When compared to CAMEL, RealPatch improves robust accuracy (+3.34%), but slightly worsens the robust gap (+0.1%). Compared with CAMEL, GDRO and SGDRO, RealPatch is very consistent across runs, with a standard deviation of 0.13, 0.85 and 0.9 for aggregate accuracy, robust accuracy and robust gap, in contrast to 5.59, 7.16 and 0.18 (aggregate accuracy), 3.94, 18.99 and 1.39 (robust accuracy) and 0.44, 19.65 and 1.67 (robust gap) for CAMEL, GDRO and SGDRO respectively. On inspection of matched pairs from the dataset $D^\star$, we observe preservation in pose, facial expression (e.g. smiling in most of the examples), hair style (in many cases the colour as well, but not always), and accessories such as hat and glasses. Figures 1 shows samples of retrieved matched pairs, further examples are in Appendix B. Naturally, due to use of real samples used in matching, RealPatch suffers no issues regarding quality of images in the augmented dataset often observed with generative models. Figure 1 shows CycleGAN generated examples used in the consistency regularizer in the CAMEL's loss.

Results on Waterbirds. Our RealPatch model can significantly reduce the gap between subgroup performances and improve the worst-case accuracy compared to all baselines. While GDRO have a better aggregate accuracy up to 1.37%, this model exhibits a higher imbalance over the subgroup performances with a robust gap of 14.80% in comparison to 4.43% of RealPatch and a robust accuracy of 81.03% as opposed to 84.44% of RealPatch. When compared to CAMEL, RealPatch shows improvements across all metrics, with +7.89% aggregate accuracy, +7.62% robust accuracy and −2.92% robust gap. Similar conclusions hold true when comparing RealPatch against the best baseline SGDRO with +1.33% robust accuracy and −2.18% robust gap. The characteristics preserved between matched pairs are less obvious than in CelebA, mainly observing matching across the bird's primary/body colour; examples are shown in Appendix B.

Results on iWildCam-small. We show results comparing RealPatch against ERM and the best baseline SGDRO in Table 2. When compared to the two baseline methods, our RealPatch improves robust accuracy by +1.27% and +1.2% and robust gap by −8.29% and −2.92%. It is worth noticing in this setting CycleGAN-based CAMEL is not applicable due to insufficient data to train CycleGAN. Examples of retrieved matched pairs are in Appendix B.

Please refer to Appendix B for the full results of Table 1 and Table 2 which include the four subgroup performances.

Ablation Analysis. We perform ablations on three of the main components of our statistical matching stage, analysing the effect of 1) temperature scaling, 2) fixed caliper and 3) std-based caliper towards matching quality. Since the distribution of propensity scores is altered by temperature scaling, and the

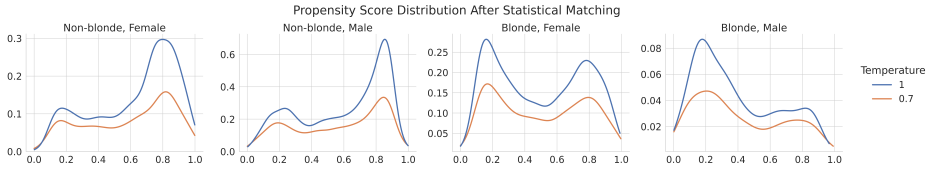

Fig. 3. Estimated propensity score distribution on CelebaA dataset after matching, shown for each of the four subgroups. We compare the original distribution (blue, $t = 1$) with its scaled version using the selected temperature (orange, $t = 0.7$). Post-matching, the propensity score is approximately bimodal, showing that our procedure is balancing the propensity distribution across subgroups. Decreasing t makes the two modes have more similar values, resulting in a matched dataset with better covariate balance in terms of *SMD* and *VR* (Table 3).

Table 3. Comparison of the covariate balance in 1) the original dataset D, 2) the matched dataset $D^\star$ 3) the matched dataset $D^\star$ with no temperature scaling 4) $D^\star$ with no fixed caliper and 5) $D^\star$ with no std-based caliper. The results are reported for a single run per dataset. Our matching procedure can successfully improve the covariate balance in both benchmark datasets, with fixed caliper significantly boosting its quality.

Dataset		**SMD**			**VR**		
		$\leq 0.1 \uparrow$	$(0.1, 0.2) \downarrow$	$\geq 0.2 \downarrow$	$\leq 4/5 \downarrow$	$(4/5, 5/4) \uparrow$	$\geq 5/4 \downarrow$
CelebA	D	348	344	1356	309	859	880
	$D^\star$ (best)	**1977**	71	0	0	**2038**	10
	$D^\star$ ($t=1$)	1957	91	0	2	2032	14
	$D^\star$ ($c=0$)	1522	482	44	13	1797	238
	$D^\star$ ($\alpha=\infty$)	1909	138	1	11	2028	9
Waterbirds	D	376	346	1326	992	723	333
	$D^\star$ (best)	**1436**	512	100	18	**1533**	497
	$D^\star$ ($t=1$)	1409	526	113	13	1482	553
	$D^\star$ ($c=0$)	852	596	600	50	1104	894
	$D^\star$ ($\alpha=\infty$)	1436	512	100	18	1533	497

propensity score is used by both types of calipers to exclude possible matches, these components are fairly coupled. We compare results obtained using different matching configurations, starting with the best configuration optimised in term of covariates balance achieved and (a) removing temperature scaling (setting $t = 1$), (b) removing the fixed caliper (setting $c = 0$) and (c) removing the std-based caliper (setting $\alpha = \infty$). In Table 3 we report the covariate balance before (D) and after ($D^\star$) matching for a single run of CelebA and Waterbirds, under all the three settings. It should be noted that the selected best configuration for all three runs of Waterbirds do not include the usage of std-based caliper, there is therefor no difference between $D^\star$(best) and $D^\star(\alpha = \infty)$ in Table 3. A similar analysis for iWildCam-small is in Appendix B. We evaluate the matching quality through Standardised Mean Difference (*SMD*) and Variance Ratio

(*VR*) as described in Section 2.1. All datasets benefit from matching, resulting in a better covariate balance than the original dataset. In CelebA we are able to produce an adequate balance ($SMD \leq 0.1$ and $4/5 \leq VR \leq 5/4$) for most of the covariates, 1977 out of 2048 for *SMD* and 2038 out of 2048 for *VR*. For the Waterbirds dataset, we achieve a slightly less prominent balance, nevertheless, improvements with respect to the original training dataset are achieved. Across all datasets the strongest effect is obtained by removing the influence of the **fixed caliper**. Since the propensity score characterises how likely an image is to belong to a subgroup, preserving all images at the extremes before calculating possible pairs is seen to be highly detrimental. The impact of **temperature scaling** and **std. caliper** is weaker overall, or absent for Waterbids under the setting $\alpha = \infty$, though still worth investigating for the specific application.

While the use of calipers is relatively well known in causal inference, temperature scaling is not commonly explored. We inspect the effect of its usage on the propensity score distribution. For a single run of CelebA, in Figure 3 we show the estimated propensity score distribution for each of the four subgroups for $D^\star$ (the after matching dataset). Post-matching, the propensity score is approximately bimodal, showing that our procedure is balancing the propensity distribution across the subgroups. We show the distribution of $D^\star$ generated with $t=1$ (no temperature) and $D^\star$ generated with $t=0.7$ (selected temperature). Decreasing t leads to the two modes having more similar values, resulting in matched dataset with better covariate balance in terms of *SMD* and *VR* (Table 3). We observe a similar effect on the Waterbirds dataset, shown in Appendix B.

3.2 Reducing Dataset and Model Leakage

In this section we study the effect of our RealPatch on dataset and model leakage.

Leakage. We use dataset leakage and model leakage [35] to measure dataset bias. Dataset leakage measures how much information the true labels leak about gender, and corresponds to the accuracy of predicting gender from the ground truth annotations. In model leakage, the model is being trained on the dataset, and we measure how much the predicted labels leak about gender.

imSitu dataset. We use the imSitu dataset [37] of situation recognition, where we have images of 211 activities being performed by an agent (person). We follow the setting of prior works [41,35] and study the activity bias with respect to a binarised gender of the agent. The dataset contains 24,301 images (training), 7,730 images (validation), 7,669 images (test).

Matching Results. We performed matching on the training data, and include all matched pairs as a rebalanced dataset to analyse the leakage. On this dataset, all samples have been matched, and the dataset size after matching has been doubled. This is expected, given the dataset has 211 classes with $44 - 182$ samples per class, which is significantly less than in CelebA. Doubling the size of the dataset does not mean we include every sample twice. Instead this should be seen as rebalancing/resampling the dataset based on how many times each sample has been matched. For matching we use the features extracted with a pretrained ResNet101 model. The selected hyperparameters are *spurious reweight-*

Table 4. Matching-based rebalancing in imSitu achieves the best leakage-accuracy trade-off. It shows nearly no dataset leakage, leading to a reduction in model leakage while maintaining overall accuracy. This is in contrast to the co-occurrence-based rebalancing based on gender-label statistics (e.g. $\alpha=1$ [35]), where a reduction in dataset leakage does not lead to reduction in model leakage in a meaningful way, and the overall accuracy drops.

Data	**Dataset ↓ leakage** λ_D	**Model ↓ leakage** λ_M	mAP ↑	F1 ↑
original training data	68.35 (0.16)	76.79 (0.17)	41.12	39.91
balancing with $\alpha = 3$ [35]	68.11 (0.55)	75.79 (0.49)	39.20	37.64
balancing with $\alpha = 2$ [35]	68.15 (0.32)	75.46 (0.32)	37.53	36.41
balancing with $\alpha = 1$ [35]	53.99 (0.69)	74.83 (0.34)	34.63	33.94
RealPatch (ours)	55.13 (0.76)	68.76 (0.69)	38.74	38.13

ing in propensity score estimation, a temperature of $t = 0.6$, $c = 0$ in the fixed caliper, and $\alpha = 0.2$ in the std-based caliper. This configuration was selected based on the best covariates balanced achieved on the training set: we can reach an adequate balance with an *SMD* value below 0.1 and *VR* close to 1 for most of the 1024 covariates used, 992 and 1010 respectively compared to 327 and 510 of the original dataset. A table with all the covariates balance is in Appendix B.

Leakage Results. We follow the same architectures and training procedures as [35] to measure the dataset and model leakage. We compare our results with the rebalancing strategies based on gender-label co-occurances proposed in [35]. We report our findings in Table 4. The results clearly show that dataset rebalancing via matching helps to achieve the best trade-off between debiasing (the dataset and the model leakage), and performance (F1 and mAP scores). We achieve significant reduction in dataset leakage (nearly no leakage 55.13 versus original 68.35) and model leakage (68.76 versus 76.79), while maintaining the accuracy of the model with mAP and F1 scores comparable to those achieved with the original training data. This is in contrast to rebalancing based on co-occurrences of gender and activity labels [35]. In the case of a rebalanced dataset with $\alpha=1$ that achieves nearly no dataset leakage (53.99), the model trained on this dataset leaks similarly to the model trained on the original data (74.83 versus 76.79), and has a significant drop in the overall performance. This suggests that statistical matching helps to reduce dataset leakage in a meaningful way as the model trained on the rebalanced dataset can reduce leakage as well.

4 Limitations and Intended Use

While *SMD* and *VR* are valuable metrics to indicate the quality of the matched dataset, there is no rule-of-thumb for interpreting whether the covariates have been *sufficiently* balanced. Supplementing *SMD* and *VR* with manual inspection of matched pairs and evaluating on a downstream task is still required.

Additionally, RealPatch currently only handles binary spurious attributes, requiring additional work (such as [22]) to handle matching over multiple treatments. It is worth noticing that also the baselines considered, GDRO, SGDRO and CAMEL, have only been tested on a binary spurious attribute. We intend to explore the usage of RealPatch for non-binary spurious attributes in the future work. A natural extension would be to use a One-vs-Rest approach for matching: for each sample find the closest sample having a different value of the spurious attribute.

5 Conclusions

We present RealPatch, a two-stage framework for model patching by utilising a dataset with real samples using statistical matching. We demonstrate the effectiveness of RealPatch on three benchmark datasets, CelebA, Waterbirds and iWildCam-small. We show that RealPatch's Stage 1 is successfully balancing a dataset with respect to a spurious attribute and we effectively improve subgroup performances by including such matched dataset in the training objective of Stage 2. We also highlight the applicability of RealPatch in a small dataset setting experimenting with the so-called iWildCam-small. Compared to CAMEL, a related approach that requires the training of multiple CycleGAN models, we see competitive reductions in the subgroup performance gap without depending on the ability to generate synthetic images. We also show the effectiveness of RealPatch for reducing dataset leakage and model leakage in a 211-class setting, where relying on generative model-based patching such as CAMEL is impractical. RealPatch can successfully eliminate dataset leakage while reducing model leakage and maintaining high utility. Our findings show the importance of selecting calipers to achieve a satisfactory covariates balance and serve as a guideline for future work on statistical matching on visual data. We encourage the use of RealPatch as a competitive baseline for strategic rebalancing and model patching, especially in the case where developing models for image generation is prohibitive or impractical.

Acknowledgments. 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. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. CoRR **abs/1907.02893** (2019)
2. Beery, S., Cole, E., Gjoka, A.: The iWildCam 2020 competition dataset. CoRR **abs/2004.10340** (2020)
3. Biglan, A., Ary, D., Wagenaar, A.C.: The value of interrupted time-series experiments for community intervention research. Prevention Science **1**(1), 31–49 (2000)
4. Chastain, R.L., et al.: Estimated full scale IQ in an adult heroin addict population. (1985)
5. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
6. Christian, P., Murray-Kolb, L.E., Khatry, S.K., Katz, J., Schaefer, B.A., Cole, P.M., LeClerq, S.C., Tielsch, J.M.: Prenatal micronutrient supplementation and intellectual and motor function in early school-aged children in nepal. JAMA **304**(24), 2716–2723 (2010)
7. Cochran, W.G., Rubin, D.B.: Controlling bias in observational studies: A review. Sankhyā: The Indian Journal of Statistics, Series A pp. 417–446 (1973)
8. Cox, D.R., Snell, E.J.: The Analysis of Binary Data. London: Chapman and Hall (1989)
9. Creager, E., Jacobsen, J., Zemel, R.S.: Environment inference for invariant learning. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 2189–2200. PMLR (2021)
10. Crump, R.K., Hotz, V.J., Imbens, G.W., Mitnik, O.A.: Dealing with limited overlap in estimation of average treatment effects. Biometrika **96**(1), 187–199 (2009)
11. Denton, E., Hutchinson, B., Mitchell, M., Gebru, T., Zaldivar, A.: Image counterfactual sensitivity analysis for detecting unintended bias. arXiv preprint arXiv:1906.06439v3 (2020)
12. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
13. Goel, K., Gu, A., Li, Y., Re, C.: Model patching: Closing the subgroup performance gap with data augmentation. In: International Conference on Learning Representations (2020)
14. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
16. Heinze-Deml, C., Meinshausen, N., Peters, J.: Invariant causal prediction for nonlinear models. Journal of Causal Inference **6(2)** (2018)
17. Kehrenberg, T., Bartlett, M., Sharmanska, V., Quadrianto, N.: Addressing missing sources with adversarial support-matching. arXiv preprint arXiv:2203.13154 (2022)
18. Kehrenberg, T., Bartlett, M., Thomas, O., Quadrianto, N.: Null-sampling for interpretable and fair representations. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVI. Lecture Notes in Computer Science, vol. 12371, pp. 565–580. Springer (2020)

19. King, G., Nielsen, R.: Why propensity scores should not be used for matching. Political Analysis **27**(4), 435–454 (2019)
20. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (BiT): General visual representation learning. In: European conference on computer vision. Springer (2020)
21. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
22. Lopez, M.J., Gutman, R.: Estimation of causal effects with multiple treatments: a review and new ideas. Statistical Science pp. 432–454 (2017)
23. Martinez, N., Bertran, M., Sapiro, G.: Minimax pareto fairness: A multi objective perspective. In: International Conference on Machine Learning. pp. 6755–6764. PMLR (2020)
24. Mitrovic, J., McWilliams, B., Walker, J.C., Buesing, L.H., Blundell, C.: Representation learning via invariant causal mechanisms. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=9p2ekP904Rs
25. Morgan, S.L.: Counterfactuals, causal effect heterogeneity, and the catholic school effect on learning. Sociology of education pp. 341–374 (2001)
26. Normand, S.L.T., Landrum, M.B., Guadagnoli, E., Ayanian, J.Z., Ryan, T.J., Cleary, P.D., McNeil, B.J.: Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. Journal of clinical epidemiology **54**(4), 387–398 (2001)
27. Perry, C.L., Williams, C.L., Veblen-Mortenson, S., Toomey, T.L., Komro, K.A., Anstine, P.S., McGovern, P.G., Finnegan, J.R., Forster, J.L., Wagenaar, A.C., et al.: Project northland: outcomes of a community wide alcohol use prevention program during early adolescence. American Journal of Public Health **86**(7), 956–965 (1996)
28. Peters, J., Bühlmann, P., Meinshausen, N.: Causal inference using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society Series B **78(5)** (2016)
29. Rubin, D.B.: Matching to remove bias in observational studies. Biometrics pp. 159–183 (1973)
30. Rubin, D.B.: Using propensity scores to help design observational studies: application to the tobacco litigation. Health Services and Outcomes Research Methodology **2**(3), 169–188 (2001)
31. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
32. Saunders, A.M., Strittmatter, W.J., Schmechel, D., George-Hyslop, P.S., Pericak-Vance, M.A., Joo, S., Rosi, B., Gusella, J., Crapper-MacLachlan, D., Alberts, M., et al.: Association of apolipoprotein E allele ∈4 with late-onset familial and sporadic alzheimer's disease. Neurology **43**(8), 1467–1467 (1993)
33. Sharmanska, V., Hendricks, L.A., Darrell, T., Quadrianto, N.: Contrastive examples for addressing the tyranny of the majority. arXiv preprint arXiv:2004.06524 (2020)
34. Veitch, V., D'Amour, A., Yadlowsky, S., Eisenstein, J.: Counterfactual invariance to spurious correlations: Why and how to pass stress tests. CoRR **abs/2106.00545** (2021), https://arxiv.org/abs/2106.00545

35. Wang, T., Zhao, J., Yatskar, M., Chang, K.W., Ordonez, V.: Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: ICCV (2019)
36. Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 547–558. Association for Computing Machinery (2020)
37. Yatskar, M., Zettlemoyer, L., Farhadi, A.: Situation recognition: Visual semantic role labeling for image understanding. In: CVPR (2016)
38. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: International Conference on Computer Vision (ICCV) (2019)
39. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018)
40. Zhang, Z., Kim, H.J., Lonjon, G., Zhu, Y., et al.: Balance diagnostics after propensity score matching. Annals of translational medicine **7**(1) (2019)
41. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2941–2951 (2017)
42. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)