





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

Sara Romiti¹, Christopher Inskip¹, Viktoriia Sharmanska¹, 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

{s.romiti, c.inskip, sharmanska.v, n.quadrianto}@sussex.ac.uk

A Setup

This section details our experimental setup for reproducibility, including dataset information and training details for the various baselines and proposed RealPatch framework. The code is made available at <https://github.com/wearepal/RealPatch>.

A.1 Dataset

Following the setup used by Goel et al. [1], Table A.1 summarises sizes of each subgroup in both CelebA and Waterbirds. For each dataset, subgroup sizes are kept consistent across the three runs. The same information is provided for iWildCam-small, where 26 and 255 are the IDs of the two camera trap locations considered.

A.2 Baseline Training Details

For CelebA and Waterbirds all four baselines use a fine-tuned ResNet50 architecture, pre-trained on ImageNet. For ERM, GDRO and CAMEL we follow the setup used in [1]. For each baseline, the hyperparameters selected are summarised in Table A.2. For iWildCam-small we use features extracted with a pre-trained BiT model to train both ERM and SGDRO; for ERM we use a logistic regression model with regularisation $C=1$, L2-penalty, tolerance of $1e^{-12}$ and sample weight inversely proportional to its subgroup frequency. For SGDRO we perform model selection using the robust accuracy on the validation set. We consider the following hyperparameters sweep for this baseline. For the Waterbirds dataset, adjustment coefficient is in a range of $\{2, 3, 5, 7\}$, weight decay is in a range of $\{0.005, 0.01, 0.05\}$ and batch size is in a range of $\{64, 128, 256\}$. For the CelebA dataset, adjustment coefficient is in a range of $\{2, 3, 5\}$, weight decay is in a range of $\{0.005, 0.01\}$, and batch size is fixed to 64. For the iWildCam-small dataset, adjustment coefficient is in a range of $\{1, 2\}$, weight decay is fixed to 0.01 and

Table A.1: Number of train/validation/test set images in each dataset.

Dataset	Split	Subgroup Size			
		Non-Blonde Female	Non-Blonde Male	Blonde Female	Blonde Male
CelebA	train	4054	66874	22880	1387
	validation	8535	8276	2874	182
	test	9767	7535	2480	180
Waterbirds		Landbird Land	Landbird Water	Waterbird Land	Waterbird Water
	train	3498	184	56	1057
	validation	467	466	133	133
	test	2255	2255	642	642
iWildCam-small		Meleagris Ocellata ID 26	Crax Rubra ID 255	Meleagris Ocellata ID 26	Crax Rubra ID 255
	train	35	940	980	50
	validation	80	80	80	400
	test	85	80	90	449

batch size is in a range of $\{64, 128\}$. For all datasets, we trained SGDR0 for 100 epochs. The selected hyperparameters for each of the three runs are summarised in Table A.3.

Table A.2: The hyperparameters used for ERM, GDRO, and CAMEL baselines for CelebA and Waterbirds, following [1].

Dataset	Method	Hyperparameters					
		Epochs	Learning Rate	Weight Decay	Batch Size	GDRO Adjustment	λ
CelebA	ERM	50	0.00005	0.05	16	-	-
	GDRO	50	0.0001	0.05	16	3	-
	CAMEL	50	0.00005	0.05	16	3	5
Waterbirds	ERM	500	0.001	0.001	16	-	-
	GDRO	500	0.00001	0.05	24	1	-
	CAMEL	500	0.0001	0.001	16	2	100

A.3 RealPatch Training Details

To give each image a chance of being included in the final matched dataset D^* , we match in both directions, i.e. we consider both values of the spurious attribute to represent the treatment and control group in turn. The size of D^* can therefore be in the range $[0, 2N]$; 0 in the extreme case where no image is paired and $2N$ in the case that all images are. For example, in CelebA we first use our pipeline (Figure 2 in Section 2.1) to match *male* to *female* samples, we then apply it to match *female* to *male* samples (using the same configuration and hyperparameters).

Table A.3: The hyperparameters used for the SGDRO baseline for each of the three runs.

Dataset	Run	Hyperparameters		
		Weight Decay	GDRO Adjustment	Batch Size
CelebA	1	0.005	5	64
	2	0.005	5	64
	3	0.005	5	64
Waterbirds	1	0.01	7	64
	2	0.05	5	64
	3	0.005	2	256
iWildCam-small	1	0.01	2	128
	2	0.01	2	64
	3	0.01	1	64

Reweighting strategy. In our logistic regression models for predicting the propensity score we explore the use of no reweighting, as well as a *spurious-reweighting* strategy. For each sample s , its weight w_s is defined as:

$$w_s = \frac{N}{2 \cdot N_{z_s}},$$

where N_{z_s} is the size of the spurious group ($Z = z_s$).

Hyperparameters for Reducing Subgroup Performance Gap. We include the hyperparameter sweep and provide the best hyperparameters found for each dataset and run. To select the hyperparameters for Stage 1 of RealPatch we perform a grid search summarised in Table A.4, selecting the configuration with the best covariates balance in terms of *SMD* and *VR*. Although we need to perform hyperparameters search, we notice the optimal values (Table A.5) are quite stable across different seeds; in practice, the grid search for Stage 1 can be restricted. As per the hyperparameters of Stage 2, we perform model selection utilising the robust accuracy on the validation set. We consider the following hyperparameters sweep. For the Waterbirds dataset, adjustment coefficient is in a range of $\{2, 3, 5, 7\}$, weight decay is in a range of $\{0.005, 0.01, 0.05\}$, regularisation strength λ is in a range of $\{0, 1, 5, 10\}$ and batch size is in a range of $\{64, 128, 256\}$. For the CelebA dataset, adjustment coefficient is in a range of $\{2, 3, 5\}$, weight decay is in a range of $\{0.005, 0.01\}$, λ is in a range of $\{0, 1, 5\}$, and batch size fixed to 64. For the iWildCam-small dataset, adjustment coefficient is in a range of $\{1, 2\}$, weight decay is fixed to 0.01, λ is in a range of $\{0, 1, 2, 7, 10, 12, 15\}$, and batch size is in a range of $\{64, 128\}$. Table A.5 reports the values of the best hyperparameters found.

Hyperparameters for Reducing Dataset and Model Leakage. For the imSitu dataset we perform a grid search over hyperparameters, using *spurious reweighting* in the propensity score estimation model, temperature $t = [0.6, 1]$ with step 0.1, a

Table A.4: Hyperparameter grid search used in Stage 1 of RealPatch for reducing subgroup performance gap.

Hyperparameter	Sweep
PS-reweighting	no reweighting spurious reweighting
PS-temperature (t)	$[0.6, 1.3]$ with step 0.05
Fixed caliper (c)	0.1 0.05 0 (None)
Std-based caliper (α)	0.2 0.4 0.6 ∞ (None)

Table A.5: The hyperparameters values selected for RealPatch on CelebA, Waterbirds and iWildCam-small across three runs.

CelebA dataset							
Run	PS-reweighting	t	c	α	Weight Decay	GDRO Adj.	Reg. λ Batch Size
1	no reweighting	0.7	0.1	0.6	0.01	5	5 64
2	no reweighting	0.7	0.1	0.6	0.005	5	1 64
3	no reweighting	0.7	0.1	0.6	0.005	5	1 64
Waterbirds dataset							
Run	PS-reweighting	t	c	α	Weight Decay	GDRO Adj.	Reg. λ Batch Size
1	no reweighting	0.9	0.1	∞	0.05	3	1 128
2	no reweighting	0.9	0.1	∞	0.05	3	1 128
3	no reweighting	0.7	0.1	∞	0.005	2	1 256
iWildCam-small dataset							
Run	PS-reweighting	t	c	α	Weight Decay	GDRO Adj.	Reg. λ Batch Size
1	spurious-reweighting	1	0.05	∞	0.01	2	5 128
2	spurious-reweighting	1.3	0.1	∞	0.01	1	12 128
3	spurious-reweighting	1	0.05	∞	0.001	1	10 64

fixed caliper with $c = \{0, 0.1\}$, and an std-based caliper with $\alpha = 0.2$. For model selection, we use the covariate balanced achieved on the training set in terms of SMD and VR . The selected hyperparameters are *spurious reweighting*, $t = 0.6$, $c = 0$, and $\alpha = 0.2$.

B Results

In Appendix B.1 we show additional results for our RealPatch framework. In Appendix B.2 we report the results obtained using different setups for the CAMEL baseline.

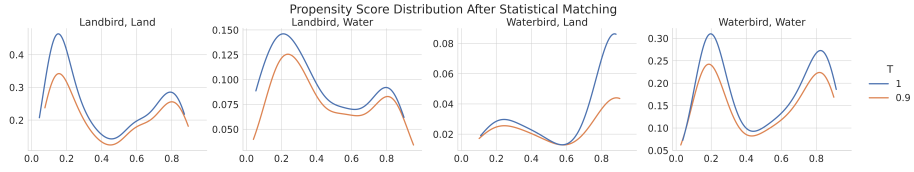


Fig. B.1: Estimated propensity score distributions on the Waterbirds dataset after matching, shown for each of the four subgroups. We compare the original distribution (blue, $t = 1$) with its scaled version using the selected temperature (orange, $t = 0.9$). Post-matching, the propensity score is approximately 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 2 in Section 3.1).

B.1 RealPatch

In this section we include 1) the information to confirm the effect of RealPatch hyperparameters (further to the Ablation Analysis in Section 3.1 of the main paper), 2) additional examples of RealPatch and CycleGAN counterfactuals for both CelebA and Waterbirds datasets, 3) subgroup results for each dataset, 4) examples of matched pairs and achieved matching quality for iWildCam-small, and 5) examples of matched pairs and achieved matching quality for imSitu dataset.

Effect of Temperature on Propensity Score. For a single run of Waterbirds, in Figure B.1 we show the estimated propensity score distribution for each of the four subgroups for the dataset obtained after matching D^* . Similarly to Figure 3 in Section 3.1 we compare the distributions obtained when imposing no temperature scaling ($t = 1$) and when selecting the temperature hyperparameter (here, $t = 0.9$). The figure shows consistent results with what was already observed in CelebA: decreasing t leads to the two modes having more similar values, resulting in matched dataset with better propensity score balance and covariate balance in terms of *SMD* and *VR* (Section 3.1, Table 3).

Additional Counterfactual Examples. In this section we show additional samples of retrieved matched pairs as well as random synthetic examples generated using CycleGAN. For the CelebA dataset, in Figure B.2 we include our results (a) when matching females-to-males and (b) males-to-females. Similarly, for the Waterbirds dataset we include in Figure B.3 the matched pairs (a) land-to-water and (b) water-to-land. In both datasets we notice that CycleGAN often adds artifacts and is frequently unable to recognise birds in the Waterbirds dataset (often removing them when translating from land to water; see Figure B.3a).

Subgroup results. Table B.6 and Table B.7 are an extension of Table 1 and Table 2 to include the accuracy of all the four subgroups. It is worth mentioning that the



(a) Examples of female images and their female counterfactuals.

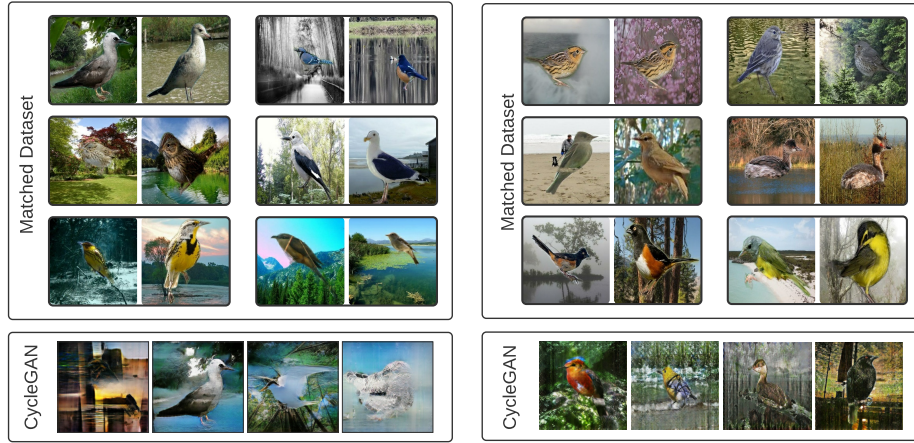
(b) Examples of male images and their female counterfactuals.

Fig. B.2: Examples of pairs retrieved using Stage 1 of RealPatch (top); both original and matched images are real samples from the CelebA dataset. We also show CycleGAN synthetic counterfactual results (bottom) on the same attribute-translation task.

worst-case accuracy can be observed in different subgroups across the three runs; therefore the average robust accuracy does not necessarily correspond to the average accuracy of one of the four subgroups. Although we observe degradation on a subgroup(s) to improve the worst-case in all methods including baselines, our RealPatch makes strikingly better trade-off of the aggregate and robust accuracies than all the baselines.

Additional Results for iWildCam-small. In Figure B.4 we show samples of retrieved matched pairs, here we can see how matching is able to preserve the bird species as well as the background colours. Similarly to Table 3, in Table B.8 we compare the effect of the main components of our statistical matching stage for iWildCam-small dataset, analysing the effect of temperature scaling and calipers. The selected best configuration for all three runs do not include the usage of std-based caliper, therefore we do not study the effect of removing such component (i.e. setting $\alpha = \infty$). The results are consistent with what observed for CelebA and Waterbirds: the strongest effect is obtained by removing the influence of the fixed caliper, while the impact of temperature scaling is weaker overall.

Additional Results for imSitu. In Figures B.5 we show examples of matched pairs retrieved using RealPatch on the imSitu dataset. Here, we observe that the activity is generally preserved, though not necessary reflecting an identical *situation* label in the dataset; for example, we have matched images of agents



(a) Examples of birds on land and their counterfactuals in water. (b) Examples of birds on water and their counterfactuals in land.

Fig. B.3: Examples of pairs retrieved using using Stage 1 of RealPatch (top); both original and matched images are real samples from the Waterbirds dataset. We also show CycleGAN synthetic counterfactual results (bottom) on the same attribute-translation task. CycleGAN often adds artifacts and is frequently unable to recognise birds in the Waterbirds dataset (often removing them when translating from land to water; see left-column B.3a).

Table B.6: A comparison between RealPatch and four baselines on two benchmark datasets which includes the subgroup results. This table is an extension of Table 1 in the main paper. The results shown are the average (standard deviation) performances over three runs.

Dataset	Method	Aggregate \uparrow Acc. (%)	Robust \uparrow Acc. (%)	Robust \downarrow Gap (%)	Subgroup \uparrow Y Acc. (%) Z			
CelebA					Non-Blonde, Female	Non-Blonde, Male	Blonde, Female	Blonde, Male
	ERM	89.21 (0.32)	55.3 (0.65)	43.48 (0.68)	80.2 (0.78)	98.78 (0.11)	98.07 (0.39)	55.3 (0.65)
	GDRO	90.47 (7.16)	63.43 (18.99)	34.77 (19.65)	90.03 (10.21)	92.5 (9.57)	87.66 (11.07)	68.75 (26.15)
	SGDRO	88.92 (0.18)	82.96 (1.39)	7.13 (1.67)	90.09 (0.31)	87.67 (0.38)	88.52 (1.29)	82.96 (1.39)
	CAMEL	84.51 (5.59)	81.48 (3.94)	5.09 (0.44)	85.57 (5.48)	82.51 (5.26)	84.15 (2.50)	81.63 (3.70)
	RealPatch (Our)	89.06 (0.13)	84.82 (0.85)	5.19 (0.9)	90.01 (0.05)	87.78 (0.14)	89.52 (0.63)	84.82 (0.85)
Waterbirds					Landbird, Land	Landbird, Water	Waterbird, Land	Waterbird, Water
	ERM	86.36 (0.39)	66.88 (3.76)	32.57 (3.95)	99.45 (0.22)	76.39 (1.36)	66.88 (3.76)	94.95 (0.5)
	GDRO	88.26 (0.55)	81.03 (1.16)	14.80 (1.15)	95.83 (0.36)	81.03 (1.16)	83.01 (0.7)	92.2 (0.81)
	SGDRO	86.85 (1.71)	83.11 (3.65)	6.61 (6.01)	88.53 (4.08)	85.99 (1.95)	84.63 (4.81)	86.19 (1.56)
	CAMEL	79.0 (14.24)	76.82 (18.0)	7.35 (5.66)	77.17 (17.39)	82.08 (12.23)	84.17 (12.86)	78.85 (19.72)
	RealPatch (Our)	86.89 (1.34)	84.44 (2.53)	4.43 (4.48)	88.03 (3.03)	86.39 (1.1)	85.67 (3.54)	85.93 (0.78)

“pumping” and “cleaning” a car (both related to car maintenance) or agents “curling” and “combing” hair (both related to hair styling). Additionally, in

Table B.7: A comparison between RealPatch and two baselines on iWildCam-small datasets which includes the subgroup results. This table is an extension of Table 2 in the main paper. The results shown are the average (standard deviation) performances over three runs.

Method	Aggregate \uparrow	Robust \uparrow	Robust \downarrow	Subgroup \uparrow Y			
	Acc. (%)	Acc. (%)	Gap (%)	Acc. (%)	Z		
				Meleagris Ocellata, Meleagris Ocellata, Crax Rubra, Crax Rubra,			
				26	255	26	255
ERM	79.97 (1.18)	75.43 (3.01)	19.65 (1.96)	84.31 (7.33)	87.07 (2.95)	92.22 (4.8)	75.43 (3.01)
SGDRO	78.55 (2.45)	75.5 (3.58)	14.28 (4.35)	85.49 (4.93)	87.5 (3.06)	79.25 (2.76)	75.5 (3.58)
RealPatch (Our)	79.36 (2.09)	76.7 (3.19)	11.36 (4.87)	87.06 (4.8)	84.58 (2.95)	80.37 (1.38)	76.76 (3.26)

Table B.8: Comparison of the covariate balance in 1) the original dataset D , 2) the matched dataset D^* 3) the matched dataset D^* with no temperature scaling 4) D^* with no fixed caliper. The results are reported for a single run per dataset. Our matching procedure can successfully improve the covariate balance in iWildCam-small dataset, with fixed caliper significantly boosting its quality.

	SMD			VR		
	$\leq 0.1 \uparrow$	$(0.1, 0.2) \downarrow$	$\geq 0.2 \downarrow$	$\leq 4/5 \downarrow$	$(4/5, 5/4) \uparrow$	$\geq 5/4 \downarrow$
D	413	354	1281	612	471	965
D^* (best)	1125	656	267	161	1005	882
D^* ($t=1$)	753	615	680	191	695	1162
D^* ($c=0$)	1037	641	370	331	930	787



Fig. B.4: Matched samples on a subset of iWildCam dataset using the spurious attribute camera trap location.

Table B.9 we show the comparison of the achieved covariates balance imSitu: RealPatch is able to produce a matched dataset with the majority of covariates perfectly balanced in term of *SMD* and *VR*.



(a) Examples of male images and their female counterfactuals (b) Examples of female images and their male counterfactuals

Fig. B.5: Examples of pairs retrieved using using Stage 1 of RealPatch; both original and matched images are real samples from the imSitu dataset. Note that activities are generally preserved across pairs despite not conditioning on the target class during matching.

Table B.9: A comparison of the covariates balance in imSitu, before matching (D) and after matching (D^*). Our procedure is able to produce a dataset with the majority of covariates perfectly balanced (992 and 1010 out of 1024) in term of SMD and VR .

	SMD				VR	
	≤ 0.1	$\uparrow (0.1, 0.2)$	$\downarrow \geq 0.2$	$\downarrow \leq 4/5$	$\downarrow (4/5, 5/4)$	$\uparrow \geq 5/4$
D	327	271	426	272	510	242
D^*	992	32	0	4	1010	10

B.2 CAMEL

In Table B.10 we report three results for the CAMEL model: (a) the metrics obtained after training the model for full 50 epochs for CelebA (and 500 epochs for Waterbirds) as per [1]; (b) the results from the epoch where the model achieved the best robust metric on the validation set; in accordance with RealPatch, we report the average (standard deviation) across three repeats over different data splits for both (a) and (b) results; (c) the results from Table 2 in [1] are also included since the authors have a different setup, namely they keep the default train/validation/test split while *changing the random seed to initialise the model*. We include both settings (a) and (b) since the exact procedure in [1] is somewhat unclear; we use authors' implementation of CAMEL to produce them. The

Table B.10: Three different results for the CAMEL model: 1) metrics obtained at the last epoch after training the model; 2) results from the epoch where the model achieved the best robust gap on the validation set; 3) the results included in Table 2 in [1].

Dataset	Method	Aggregate \uparrow Acc. (%)	Robust \uparrow Acc. (%)	Robust \downarrow Gap (%)
CelebA	CAMEL (re-run epoch 50)	96.6 (0.51)	57.96 (3.55)	40.12 (4.18)
	CAMEL (re-run SGDRO)	84.51 (5.59)	81.48 (3.94)	5.09 (0.44)
	CAMEL [1], Table 2	92.90 (0.35)	83.90 (1.31)	-
Waterbirds	CAMEL (re-run epoch 500)	89.63 (7.84)	68.12 (6.93)	29.59 (3.91)
	CAMEL (re-run SGDRO)	79.0 (14.24)	76.82 (18.0)	7.35 (5.66)
	CAMEL [1], Table 2	90.89 (0.87)	89.12 (0.36)	-

results in Section 3.1 Table 1 show the output of the method described in (b) as it appears to be the closest comparison.

References

1. 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)